A Foundation Model for Intensive Care: Unlocking Generalization across Tasks and Domains at Scale

Manuel Burger[1,+], Daphné Chopard[1,2,*], Malte Londschien[3,4,*], Fedor Sergeev[1,5,*],
Hugo Yèche[1], Rita Kuznetsova[1], Martin Faltys[6], Eike Gerdes[7],
Polina Leshetkina[8], Peter Bühlmann[3], Gunnar Rätsch[1,4,9,10,+]

[1] Department of Computer Science, ETH Zurich, Switzerland; [2] Department of Intensive Care and Neonatology and Children's Research Center, University Children's Hospital Zurich, University of Zurich, Switzerland; [3] Seminar for Statistics, ETH Zurich, Switzerland; [4] AI Center, ETH Zurich, Switzerland; [5] Helmholtz AI, Germany; [6] Department of Intensive Care Medicine, University Hospital and University of Bern, Switzerland; [7] University Hospital Zurich, Switzerland; [8] Department of Health Science and Medicine, University of Luzern, Switzerland; [9] Swiss Institute for Bioinformatics, Zurich, Switzerland; [10] Medical Informatics Unit, University Hospital Zurich, Switzerland

[+] Joint-corresponding authors: `manuel.burger@inf.ethz.ch` and `gunnar.raetsch@inf.ethz.ch`
[*] These authors contributed equally and share second authorship.

Abstract

Intensive care departments generate vast multivariate time series data capturing the dynamic physiological states of critically ill patients. Despite advances in AI-driven clinical decision support, existing models remain limited. They are tailored to specific conditions or single institutions and require extensive adaptation for new settings. To make such generalization feasible, we introduce ICareFM, a novel foundation model for intensive care, trained on a harmonized dataset of unprecedented scale. The dataset contains 650,000 patient stays, accumulating more than 4,000 patient years of data, and over one billion measurements from hospitals in the US, several European countries, and China. ICareFM employs a novel self-supervised time-to-event objective that extracts robust patient representations from noisy, irregular, multivariate time series. As a result, ICareFM can generalize to new tasks and beyond its training distribution, a property we demonstrate through evaluations in a range of out-of-distribution scenarios, including transfer to unseen hospitals and zero-shot inference on previously unobserved tasks. ICareFM consistently outperforms conventional machine learning models and recent foundation model baselines, demonstrating strong generalization, improved data efficiency, and the ability to generate interpretable forecasts. These results establish ICareFM as a scalable and adaptable foundation model for critical care time series, enabling zero-shot clinical prediction and working towards the development of digital patient twins for precision medicine.

1 INTRODUCTION

Intensive care units (ICUs) and emergency departments (EDs) generate high-frequency, heterogeneous, irregularly sampled multivariate time series data reflecting critically ill patients' dynamic physiological states [1–3]. Continuous monitoring of organ function parameters during an ICU stay enables early recognition of physiological deterioration and timely intervention. However, detecting subtle patterns across multiple physiological variables over time is challenging for physicians, particularly under the pressure of a high-stakes environment [4]. This complexity contributes to cognitive overload and increases the risk of delayed or suboptimal decision-making.

Artificial intelligence (AI) promises to transform these complex data streams into timely, clinically actionable insights [5]. Researchers have developed a range of methods from clinical scoring functions to advanced machine learning algorithms [6–9]. Yet these approaches share critical limitations: they are designed for a specific clinical condition affecting a particular organ system and are tailored to a distinct study cohort, often a single hospital. On large data cohorts, such approaches require time-intensive redevelopment to achieve reliable performance.

Inspired by recent advances in large-scale AI models [10;11], we address these limitations by introducing the first foundation model specifically designed for heterogeneous, irregularly sampled multivariate time series from ICUs and EDs. We present, to our knowledge, the first intensive care foundation model, called **ICareFM**. The model's capabilities critically depend on our foundational effort to create the largest and most comprehensive, harmonized critical care time series dataset to date. This dataset is an essential requirement for developing a foundation model and we present the dataset as one of the major contributions of this work, alongside the model itself.

This large and diverse dataset enables our foundation model to learn generalizable representations across the entire intensive care domain. Due to regulatory aspects most academically accessible datasets from critical care domains (e.g., ICUs or EDs) are shared by a single hospital in a unique data format due to specificities in the local data management infrastructure and policies. To build a comprehensive dataset for a foundation model, we build on the `ricu` framework [12] to harmonize a heterogeneous collection of source datasets, by extending its set of extracted vital signs and laboratory test results, adding harmonized treatment concepts—including exact dosage and administration rate extraction of core intensive care drugs—and incorporating additional hospital datasets to increase demographic diversity. The final collective dataset considered for the development of the foundation model contains up to 650,000 patient stays and up to one billion individual measurements can be extracted. To the best of our knowledge, this is the first work to assemble a critical care dataset of this scale, combining sources from the US, Europe, and, for the first time, China.

Different to past efforts in developing data-driven decision support systems, be it clinical scores [9;13;14] or trained machine learning algorithms [5;15], a foundation model is not tailored to a specific task. Guided by the intuitions in clinical scoring functions [13;14] and recent foundation models in the health care domain [16;17] we propose a novel self-supervised time-to-event training objective tailored to learn useful patterns and patient representations from the noisy and confounded intensive care data streams. An essential attribute of foundation models is their ability to improve downstream task performance with increasing training data scale. Reflecting to this property, our pretrained foundation model demonstrate excellent adaptability to a diverse set of domain-relevant downstream applications, surpassing the performance of conventional time series forecasting foundation model training strategies [18;19] or prior time-to-event foundation models [16].

Traditional machine learning models developed for intensive care have been trained on datasets from a single hospital with a potential external validation on another second hospital's dataset [5;8;20]. The underlying assumptions here have often been, that the model would, if at all, only be used in the same hospital where it has been trained. The data distributions observed in different hospitals vary strongly due to different patient populations and treatment policies [8], which raised the question whether it is even useful to assess generalizability through external validation procedures [21]. A foundation model's purpose is the wide application on different downstream tasks and data sources (hospitals) and thus by developing a foundation model we challenge the belief "all models are local". External validation results represent a core pillar of this work. We show that pretrained foundation models can have relevant performance implications when used with or even without adaptation to a previously unseen hospital's patient distribution. We propose and rigorously apply a data efficiency focused out-of-distribution evaluation scheme, where the value of externally trained machine learning models is expressed as the amount of locally required data to surpass the foundation model. Different to task-specific discriminative performance metrics, our study hence represents an immediately actionable guide to choose the best modeling approach for deployment given a specific target application constrained by local computational resources and data availability.

A key driver of the success of large language models (LLM) [22;23] is the capability to be adapted to a desired downstream task in a zero-shot fashion without retraining or tuning any of the models parameters but rather through in-context learning and prompting of the model even without any labeled examples. Modern LLMs are well-known for these capabilities, but researchers have successfully demonstrated similar behaviors for computer vision models[24], multi-modal text-image models[25;26], time-series forecasting[19], and even multi-modal early event prediction in intensive care[27]. We present a novel time series foundation model that provides zero-shot early event risk estimates by leveraging generalized time-to-event prediction capabilities to probabilistically approximate diagnostic clinical event definitions and project biomarker evolutions, using only external hospital data. This is a crucial step towards data-driven digital patient twins[28] for personalized risk assessment and counterfactual treatment strategy evaluation.

In contrast, models trained on coded data[16;29;30] are able to scale to larger patient populations, but do not learn from physiological patterns directly. They are driven by data not available in real time, operate on longer time scales, and are further limited by fixed coding systems and integration challenges.

This work comprises the publication of an extensive dataset harmonization effort to create the biggest critical care time series dataset to date. The dataset enabled us to train the first foundation model for intensive care medicine. The model generalizes across relevant downstream applications in hospitals outside of its training distribution, shows exceptional utility especially for smaller hospitals, and can be prompted to provide zero-shot prognostic risk estimates for unseen target tasks in unseen hospitals, alongside interpretable anticipated future value distributions of core biomarkers.

2 RESULTS

2.1 The largest critical care time series dataset

By harmonizing over 11 different source datasets (**Fig. 1**) across a comprehensive and enlarged set of clinical concepts we create the biggest harmonized critical care dataset and make the resources to reproduce it available to the medical and research community. To the best of our knowledge, this is the first work to bring together critical care datasets from the US, Europe, and, for the first time, China. Harmonizing datasets across different continents can improve generalization and is crucial to the fairness and inclusiveness of ML research on critical care data. The additional datasets increase the number of processed and harmonized stays from less than 400,000[12] close to 800,000. After further processing and selection, the final dataset comprises around 650,000 patient stays, with most omissions of stays due to k-anonymization (Details in Supplementary Section A.6). In the following we will exclusively only refer to and report results on the subset of patients considered and evaluated during this study. The final harmonized dataset considered in the study contains over 4,000 patient years worth of data. The expansion both in terms of absolute number of patient stays and considered variables increases the total number of extractable individual data points from approximately 400 million in prior work to one billion.

The diversity of our dataset enables research for small but specific cohorts of patients. For example, PICdb[33] is a small pediatric dataset. The median age is around two years, while it is mostly over 60 on other datasets (see Supplementary Table 1 and Fig. 1). By providing an accessible way to pretrain on large amounts of harmonized data, we create an opportunity for smaller-scale targeted studies to benefit from the existing larger-scale research on modeling for critical care time series.

In addition to increasing the dataset size, we also increase the number of considered variables. Crucially, we enable comparison of treatment policies between hospitals by introducing a comprehensive set of treatment concepts (see Supplementary Section A.3). This set comprises 44 variables, ranging from specific core intensive care medications (e.g., continuous rates of norepinephrine) to broader classes of treatments (e.g., indicator of antibiotics administration). The variable set balances granularity with missingness, ensuring that as much signal as possible is extracted from each source while maintaining high presence of variables across different datasets.

By integrating our harmonization efforts and the development of new concepts with existing tools[12] and additionally providing a light-weight version of our entire processing pipeline we make the largest harmonized critical care time series dataset available to the research community and increase the accessibility of an otherwise difficult to handle complex collection of source data. This dataset will not only support future foundation models, but foster research in robust machine learning algorithms, domain-adaptation methods, as well as fairness and bias investigations.

2.2 ICareFM: a foundation model for intensive care medicine

We leverage our large-scale harmonized dataset to train the first foundation model tailored for intensive care medicine. We propose a novel self-supervised time-to-event training objective (Fig. 1) to train ICareFM. The

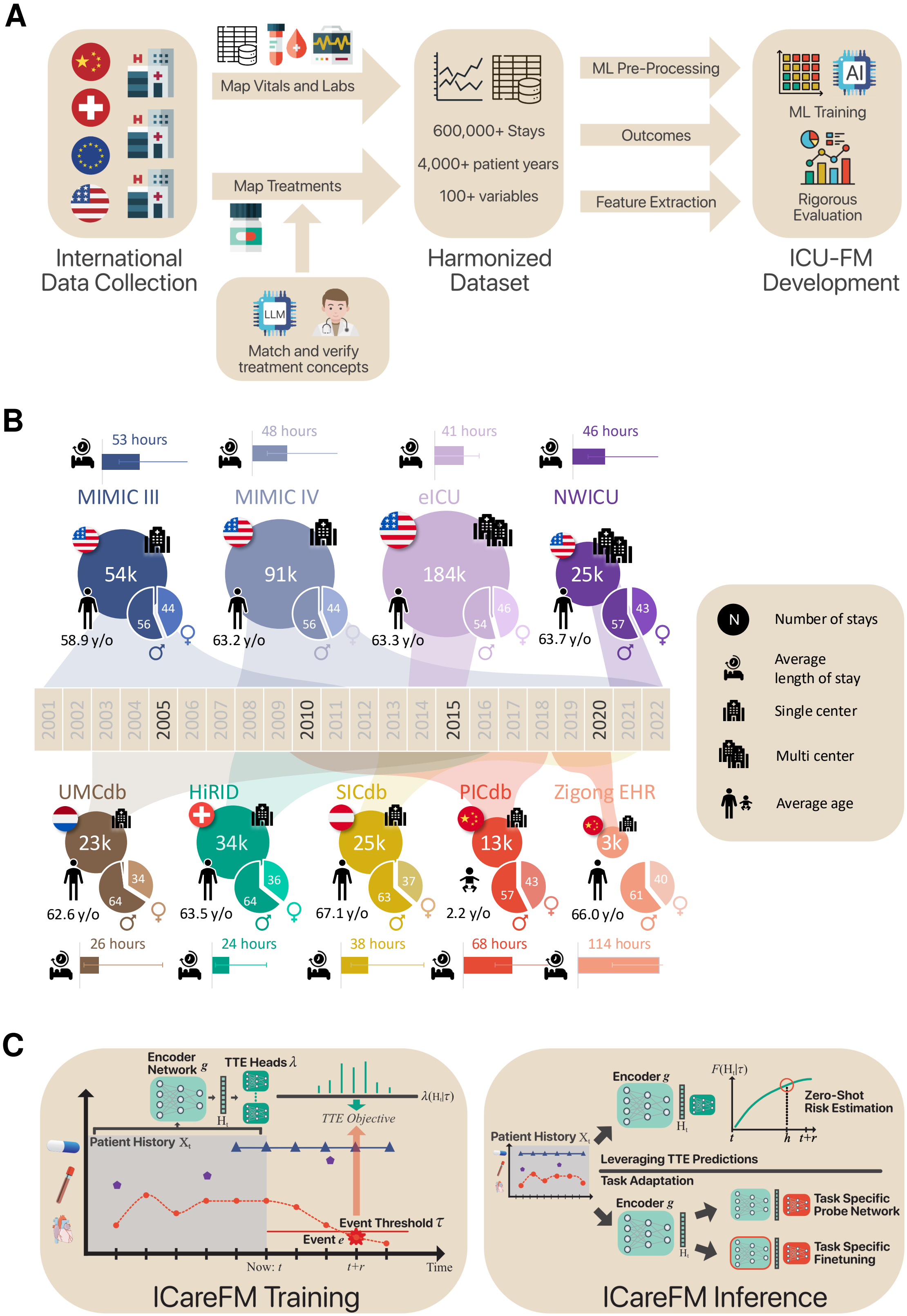
A
Map Vitals and Labs
Map Treatments
LLM
Match and verify treatment concepts
International Data Collection
600,000+ Stays
4,000+ patient years
100+ variables
Harmonized Dataset
ML Pre-Processing
Outcomes
Feature Extraction
AI
ML Training
Rigorous Evaluation
ICU-FM Development
B
53 hours
MIMIC III
54k
44
56
58.9 y/o
48 hours
MIMIC IV
91k
44
56
63.2 y/o
41 hours
eICU
184k
46
54
63.3 y/o
46 hours
NWICU
25k
43
57
63.7 y/o
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
UMCdb
23k
34
64
62.6 y/o
26 hours
HiRID
34k
36
64
63.5 y/o
24 hours
SICdb
25k
37
63
67.1 y/o
38 hours
PICdb
13k
43
57
2.2 y/o
68 hours
Zigong EHR
3k
40
61
66.0 y/o
114 hours
N
Number of stays
Average length of stay
Single center
Multi center
Average age
C
Encoder Network g
TTE Heads λ
λ(H_t|τ)
TTE Objective
H_t
Patient History X_t
Event Threshold τ
Event e
Now: t
t+r
Time
ICareFM Training
Encoder g
F(H_t|τ)
Zero-Shot Risk Estimation
t
h
t+r
Patient History X_t
Leveraging TTE Predictions
Task Adaptation
Encoder g
Task Specific Probe Network
Task Specific Finetuning
ICareFM Inference

Figure 1: Data Pipeline and Model Design A) Data collection and processing overview. We process a large heterogeneous collection of source datasets into a harmonized dataset covering a comprehensive set of intensive care variables. **B)** Summary statistics of core source datasets. More details in Supplementary Table 1 and Supplementary Table 2. **C)** ICareFM Training and Inference modes. *Left:* ICareFM is trained with a self-supervised time-to-event survival objective. *Right:* The pretrained ICareFM model can be used without any further adaptation in zero-shot mode. Alternatively, the model can be adapted in various ways. Figures are created using resources from Flaticon[31] and Bioicons[32].

large survival model is trained to regress the time until key observations cross into critical regions, where thresholds are dynamically varied during training to ensure the model is learning patterns across the full range of the distribution.

The proposed objective is more robust to noisy data, is closely aligned with downstream online prediction tasks facilitating zero-shot predictions, while incentivizing a generalized understanding of how patient trajectories evolve into the future across relevant observations.

We assess the quality of different pretraining strategies by fitting logistic regressions (linear probes[34]) on top of the fixed learned patient state representations for a selection of early event prediction tasks.

In **Fig. 2A** ICareFM's dynamic event thresholding outperforms established time-series forecasting pretraining approaches[18;19;35] and prior time-to-event foundation model objectives[16;17] for structured electronic health records defining events purely based on the presence of certain targets (indicator events). We further investigate intermediate objectives where we consider a fixed threshold τ for each target variable to define events and a mixed event definition, where events on vital parameters are defined with a fixed threshold τ and events on sparsely observed laboratory tests are defined based on presence alone.

2.3 ICareFM can perform arbitrary zero-shot early event predictions

While task-specific early warning systems can alert clinicians effectively[5], they lack flexibility to adapt to changing treatment policies, patient comorbidities, or practical constraints like shift schedules.

ICareFM can perform arbitrary risk estimations given a certain clinical event definition based on typically observed parameters in intensive care. The selection of variables used to detect an event, the associated critical thresholds, and the prediction horizon can all be varied at inference time. **Fig. 3B,C** visualize such a zero-shot risk approximation for circulatory and liver failure respectively. Considering circulatory failure (Fig. 3B), risk estimates for both mean arterial pressure and blood serum lactate to reach critical thresholds in the upcoming 8 hours are convolved to obtain a circulatory failure estimate. For liver failure we show forecasted MELD scores (Fig. 3C).

Most relevant here is, that these approximations can be flexibly adapted at inference time. As such, while for zero-shot kidney failure prediction we can consider a creatinine threshold at 3.4 mg/dl, this can be dynamically adapted to reflect patient specific baseline estimates (following KDIGO AKI guidelines[36]) of past blood sample results. The clinician's expertise and access to additional patient record information and comorbidities can guide the decision on which variables to build the risk estimate on and at which levels. This provides an incredibly interactive way to obtain personalized risk estimates for both the patient's health state (driven by data) and the clinician's needs in terms of decision support given clinical training, hospital policies, and competing risks.

Furthermore, certain conditions can be measured and diagnosed in different ways. For example to diagnose respiratory failure a common clinical parameter is the P/F ratio (partial pressure of oxygen in blood (PaO2) over fraction of inspired oxygen (FiO2)). However, under varying circumstances clinicians might be interested in different severities of respiratory failure defined at different thresholds on the P/F ratio[8] or approximate the P/F ratio using the ratio of pulse oxymetry oxygen saturation (SpO2) over FiO2 to avoid expensive blood sample testing[37]. While standard machine learning models would have to be trained separately for each severity and type of definition, ICareFM can be dynamically adjusted to forecast events of varying severity and considering an appropriate definition given the available patient data and patient situation.

2.4 ICareFM can transfer its predictive capabilities to unseen target hospitals

We evaluate ICareFM's generalization capabilities by performing multiple pretraining rounds, every time completely excluding a different target hospital from its pretraining dataset. We then assess ICareFM's performance on the hold-out target in a transfer setting (domain generalization) without using any target data and in a domain adaptation setting by gradually adapting the model to the target hospital data and task.

We consider four deployment regimes to structure this evaluation:

1. *Zero-Shot (Domain and Task).* The model is used in a fully zero-shot setting, without any adaptation to the target task or domain (hospital or dataset). It relies solely on the capabilities learned during self-supervised pretraining and can be flexibly adjusted to perform various types of predictions at inference time.

2. *FM Domain Generalization.* For deployment in dedicated early warning applications, the model is fine-tuned on large amounts of harmonized external data for the specific task of interest, and then applied directly in an unseen target hospital. This regime enables robust performance without any access to local patient data and is particularly relevant for scalable deployment.

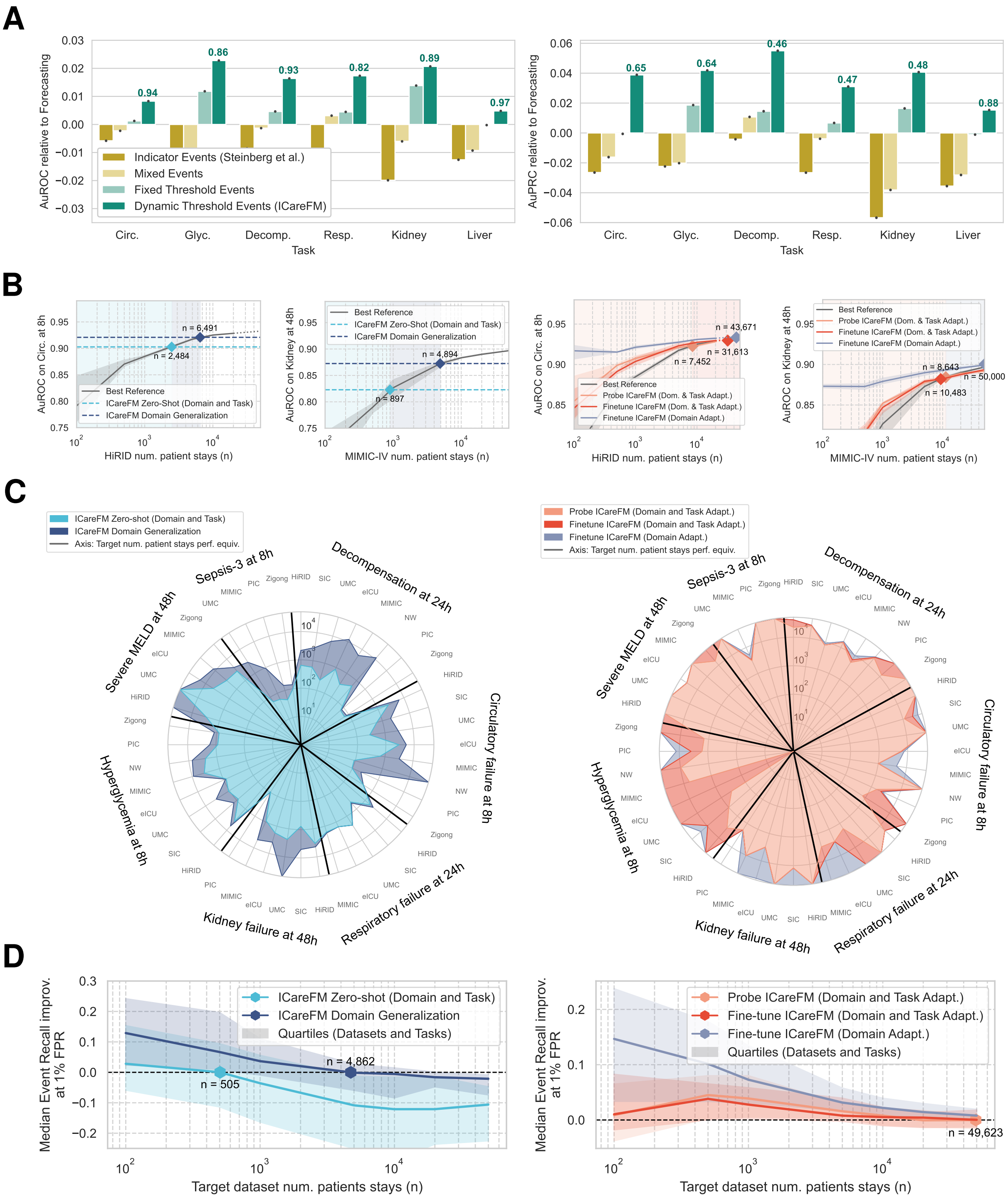

Figure 2: Core Model Performance A) Linear probe performance for different time-to-event objectives. Reported metrics are computed relative to performance achieved using established forecasting regression objectives. **B)** Examples for domain transfer and adaptation scaling studies with increasing amounts of local data. Blue colored plots compare models trained using only external hospital data with state-of-the-art models trained using only increasing amounts of local data. Red shades show results using increasing amounts of local data to adapt the externally pretrained ICareFM. **C)** Summary overview of domain transfer and adaptation studies, each point on the radar chart represents the number of local patients at the intersection of ICareFM's performance curve and the reference model's performance curve. **D)** Relative performance gains of different ICareFM modes expressed in discrete event recall at low false positive rate as a function of local training data used for adaptation or training state-of-the-art local reference models.

3. *Domain and Task Adaptation.* For long-term deployment in a larger hospital with permissive patient data usage policies the model can be adapted using available local data to achieve optimal local performance. The third regime investigates model adaptation starting from the core ICareFM network, trained solely with its self-supervised learning objective on external data. In this setting, only local labeled data are used—no external labeled data are involved. It is particularly well suited to support clinical research focused on deploying advanced AI warning systems for studies involving newly developed biomarkers or treatment strategies.
4. *Domain Adaptation.* The fourth regime is intended for long-term deployment, the pretrained foundation model is adapted to the task using labeled data in the external datasets and then finally tuned to state-of-the-art performance using the local data.

In **Fig. 2 B-D** we visualize the achieved performance equivalent in numbers of patients needed to train a task-specific machine learning model on the target with similar performance to the foundation model. Fig. 2 B) shows two examples for a circulatory failure prediction on the HiRID[1] dataset and kidney failure prediction on the MIMIC-IV[38] dataset. Blue colored plots evaluate the two zero-shot regimes and red/purple shaded plots evaluate the domain adaptation regimes using additional local data. We see that on the HiRID dataset a complete task-agnostic model can achieve the same performance as a locally trained model using around 2,500 patients, whereas models additionally adapted for the specific task using externally labeled data perform very competitively and require several thousand local patients to match. In the *Domain Adaptation* regime we note that models building from the foundation model generally perform state-of-the-art.

The radar chart in Fig. 2C summarizes the entire study results as represented in Fig. 2B for each dataset and task using area under the receiver operater curve (AuROC) as the metric to evaluate scaling plots as in Fig. 2B. We note that *Zero-Shot (Domain and Task)* performance is extremely competitive in low prevalence settings such as severe liver failure (by severe MELD[39] score) or very specific early event predictions relying on few measurements to diagnose such as the benchmarked hyperglycemia event. If local data is available for adaptation the foundation models latent space alone serves as an excellent basis to train high performing probe networks on top without adapting the actual weights of the foundation model. Given large amounts of local data, switching to fine-tuning often yields state-of-the-art performance. We also note that the foundation model essentially never underperforms locally trained models and is always within margin or noticeably better.

In Fig. 2C we analyze the relative performance gains of the different foundation model modes over best performing locally trained models now specifically for discrete event detection[7] at low false positive rates. *Zero-Shot (Domain and Task)* median performance equivalent reduces from a median of around 900 patients to 500 in such a high-precision setting. However, the models leveraging externally labeled data show inferior performance only after training reference models with 5,000 and more patients and also then performing almost on par. As soon as local data is available for adaptation the foundation models robust latent space is key to achieve SOTA performance with considerable gains if less than 10,000 local patients are available for training and adaptation.

Naturally, the model is limited by the patterns observed in its training dataset. This becomes apparent in the cases of PICdb[33], which is the only pediatric source dataset or Zigong[40], which is an infectious disease cohort on a dataset with low measurement density Supplementary Table 1. Further we observe challenges in detecting and transferring patterns predictive for sepsis.

2.5 ICareFM performance and robustness scales with increased training data

The fundamental effort to create a harmonized large-scale dataset facilitates one of the first scaling studies performed in this domain to understand how increased training data can unlock more robust machine learning models, which can be trusted to maintain their performance while deployed in previously unseen hospitals. For each target hospital, we perform the ICareFM pretraining with an increasing amount of patients from all the other hospitals in the collective dataset. For each sample size, we take four different sub-samples. We then evaluate the out-of-distribution (unseen hospital) performance across all downstream tasks and compute the performance equivalent in number of target patients for both *FM Domain Generalization* (using external labeled patients after pretraining to task adapt the model) and pure *Zero-Shot (Domain and Task)* using the model's generalized survival prediction heads. **Fig. 3A** indicates that by scaling our pretraining dataset, we can respectively achieve enhanced out-of-distribution performance across datasets and tasks. We derive a scaling law by fitting a power function. For pure zero-shot predictions we observe a square root relationship, implying that if locally available data is doubled, similar performance gains could have been achieved by collecting 4 times more

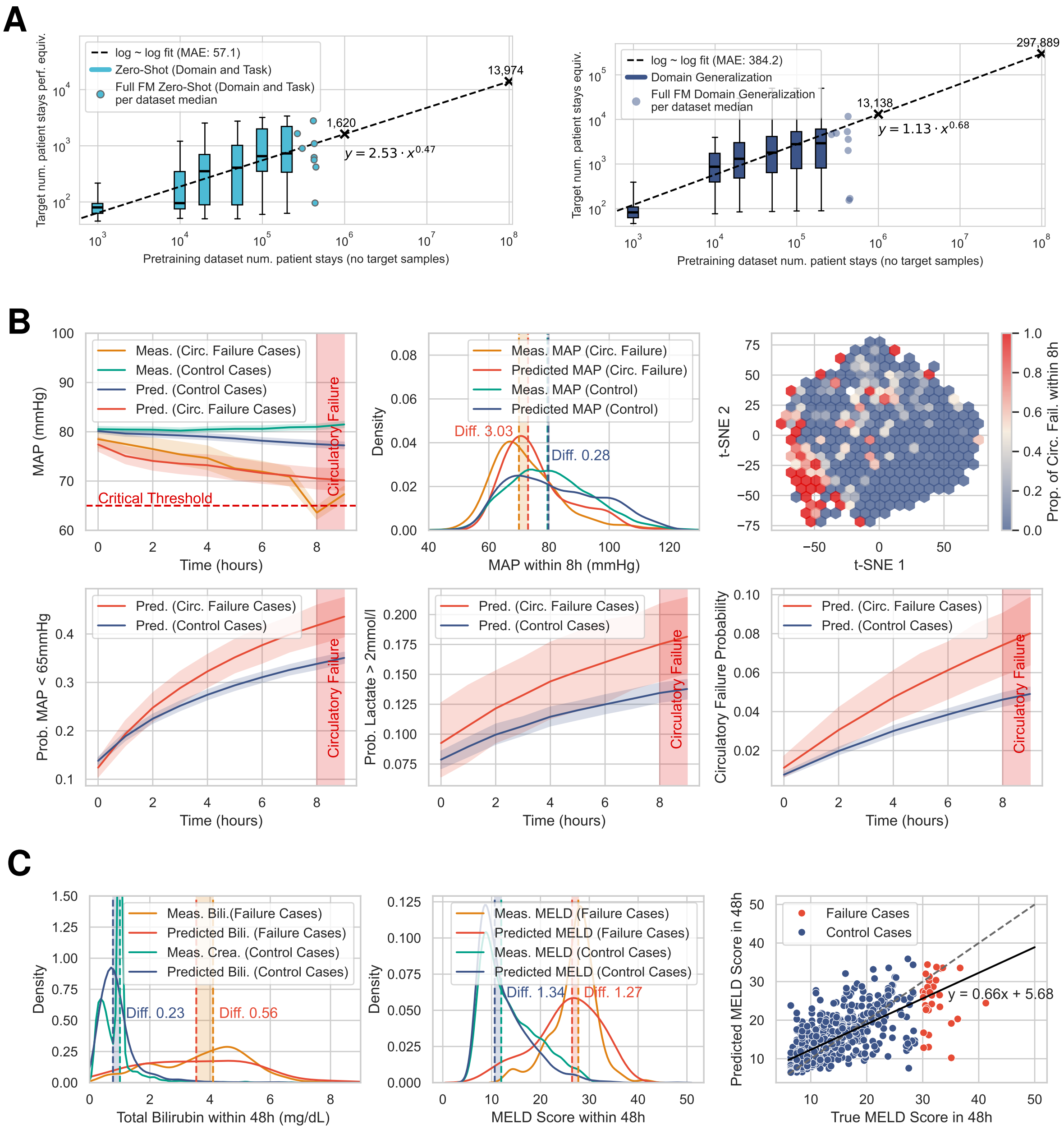

Figure 3: Extended Model Results All results do not leverage target hospital data for training: **A)** Scaling studies expressing zero-shot performances in achieved target hospital patient equivalent value as a function of number of patients used for pretraining. The panels **B)** and **C)** visualize zero-shot prediction procedures and results comparing a subsample of patients with failure events and control patients without failure events. **B)** results on the HiRID[1] dataset, from left to right and top to bottom: forecasted MAP values over time, forecasted MAP distributions within 8 hours, t-SNE[41] projections of the models latent space colored by early event predicton label, cumulative failure probability for MAP < 65 mmHg, cumulative failure probability for lactate > 2mmol/l, and convolved cumulative failure probability for zero-shot circulatory failure. **C)** results on the MIMIC-IV[38] from left to right: forecasted bilirubin distribution within 48 hours, forecasted MELD score (function of bilirubin, INR, and creatinine) within 48 hours, and calibration plot of ground truth future MELD scores and forecasted MELD scores.

external data. If additionally, external labels are used to precondition the model for the target task in the unseen domain, the efficiency improves a bit and increasing the pretraining dataset by a factor 4, would yield similar gains than collecting approximately 2.5 times more local data. These results encourage the further collection and harmonization of critical care data to build increasingly larger models, which cover wider demographics, working towards physiological patient twins, which can reliably be deployed at scale without the need to adapt to every single hospital.

Our results suggest, that a global dataset of 100 million patient stays (magnitudes available to large health care data management system providers) could yield domain generalization performance competitive to models using close to 300,000 local patient stays, a number that few hospitals in the world alone could collect locally. A foundation model trained at that scale would also show excellent performance without adapting to any specific task in pure zero-shot fashion, being competitive with local models trained with over 10,000 patients, implying robust predictions from a highly flexible model and relevant for a large fraction of smaller to medium sized hospitals world-wide.

2.6 ICareFM provides interpretable patient state prognosis

When ICareFM is used to perform predictions in either zero-shot mode, or even with a task-specific probe on top, it can additionally provide an estimate of the expected value distribution within an upcoming time horizon. Fig. 3B,C show respective predicted value distributions for mean arterial pressure, bilirubin, and even the complete MELD score. We compare a set of patients experiencing a certain type of organ failure in the upcoming horizon against a set of control patients and annotate differences in medians of the measured and predicted distributions. ICareFM's forecasted value distribution closely approximates the measured distribution of values.

ICareFM learned to distinguish likely value distributions of high and low risk patient groups. These predictions allow the clinician to interrogate the models belief on how certain clinical parameters might evolve in the future and expose the complex internal state of the model as interpretable concepts[42]. This is a highly relevant step towards a future where foundation models can serve as digital patient twins and physicians can interactively simulate treatment strategies and observe their statistically estimated impact on the patient's trajectory.

3 Discussion

We have developed ICareFM, a novel time series foundation model for intensive care medicine, built upon the largest open and harmonized critical care time series dataset. The dataset contains around 650,000 patient stays, accumulating over 4,000 patient years worth of time series data. Our study highlights the capabilities that can be unlocked by such a large harmonized collection of health care data aggregated from different hospitals spread across the globe. We hope to inspire the research community to continue creating ever-growing resources to drive the next generation of foundation models transforming intensive care medicine.

We showed that our novel pretraining strategy outperforms established self-supervised learning objectives for time series data and enables a multitude of capabilities. The large time-to-event foundation model can provide zero-shot early event prediction risk estimates purely from its generalized survival objective pretraining. The study shows the merit of developing domain-specific pretraining objectives beyond simply transferring successful objectives from the computer vision or natural language processing domain, while at the same time not losing any generality w.r.t. models capabilities after pretraining.

ICareFM demonstrates multiple capabilities without task-specific training. It can provide zero-shot early event prediction risk estimates, forecast distributions of future values, and shows excellent adaptability. The models' predictions can be customized to the specific needs of the treating physician, the patient state, and comorbidities. This is a crucial step towards the development of digital patient twins and truly personalized intensive care medicine. We envision future AI models to advance the presented capabilities further and enable counterfactual testing of different treatment strategies next to an interactive digital representation of the patient state and physiology.

In this study, we rigorously evaluated a single foundation model training procedure by performing multiple rounds of external evaluation and domain adaptation studies across different downstream use cases. We emphasize the importance to evaluate AI models intended for wide-scale deployment rigorously for robustness against distribution shifts across hospitals and countries globally. Our work shall serve as a marker for future work to move beyond single hospital development and single hospital external evaluation procedures. Our harmonized dataset provides the foundations for the community to perform extensive and robust evaluation procedures across many datasets. We further set a corner stone by presenting a single foundation model performing a multitude of downstream applications often without the need to further adapt it for specific tasks. This should inspire future work to look beyond the development of task specific classifiers and develop foundation models with a comprehensive patient state understanding.

The study is limited by the heterogeneity observed in the harmonized dataset, which is mostly determined through data accessibility. Hence, we do observe for example worse transfer results to the only pediatric hospital in the collected data. The model has little opportunity to learn patterns specific for newborns and children when the single pediatric dataset is held out for testing. Furthermore, a large portion of publicly accessible data in this domain is stored at an one hour temporal resolution and, as such, the training and evaluation was performed at a harmonized one hour temporal resolution. However, some datasets and studies expose higher frequency recordings[1;2;43] and show benefits in more timely predictions by considering higher resolution data where available[5]. We believe there is potential in exploring foundation models for the domain of intensive care operating at higher and multiple resolutions to provide even more accurate and timely predictions. The model was extensively tested for its robustness under distribution shifts observed by transferring to hospitals outside of its training distribution. However, we have not evaluated the model w.r.t. fairness and biases on a patient level. Biases in the model are a complex function of its training procedure and most importantly its training dataset. While this study focused on scalability and population-level performance advancements and capabilities, it is of utmost importance to further investigate the bias and fairness implications emerging by training on such a heterogeneous dataset and how changes in the training set affect the biases in the model.

ICareFM establishes a new paradigm for intensive care prediction models through three transformative capabilities. First, it achieves remarkable data efficiency in cross-hospital generalization, matching locally-trained model performance with none to little institution-specific data. Second, its zero-shot inference capability empowers clinicians to query arbitrary clinical events without model retraining, functioning as an adaptive inference engine that responds to real-time clinical decision needs. Third, ICareFM provides a foundational infrastructure for future intensive care algorithms, enabling researchers to build specialized applications upon its robust learned representations rather than developing models from scratch. These capabilities position ICareFM not merely as another predictive model, but as the cornerstone of a new ecosystem of AI-driven intensive care tools that will democratize access to AI driven tools for low resource settings, accelerate innovation, and improve patient outcomes across diverse healthcare settings.

4 METHODS

4.1 Related Work

Foundation Models for Healthcare In the clinical domain, existing foundation models have primarily focused on structured Electronic Health Records (EHR), such as billing codes[44] and medical reports[23;45], rather than on in-hospital critical care time series data. Some studies[46;47] have examined the adaptability of public EHR models for clinical prediction tasks, while others[48] have assessed their performance on both in-distribution and out-of-distribution data. Models trained on coded data[16;29;30] can scale to large patient populations but are limited in their ability to capture physiological patterns. These limitations arise because coded data are not available in real time, represent delayed representations of patient states, operate at coarse temporal resolutions, and are constrained by fixed coding schemas and integration challenges.

Time Series Foundation Models Following advances in natural language processing (NLP), large-scale multi-purpose pretrained models - referred to as "foundation models" - have been developed across various data types and applications. In the context of time series, a large body of work has focused exclusively on forecasting non-medical data[19;49–51], typically considering only univariate time series.

Benchmarks on ICU time series In the literature, two benchmarking strategies are commonly observed: single-center and multi-center.[52] introduced the first standardized and reproducible single-center benchmarks using MIMIC-III[53]. Following this seminal work, subsequent studies defined new tasks or explored alternative datasets, including[54] on eICU[55],[56] on HiRID[1], and[57], which proposed an alternative approach on MIMIC-III. The abundance of work around single-center data has motivated efforts to aggregate them into multi-center studies, such as[15] and[58].

Self-Supervised Learning for Time Series There's a growing body of work presenting unsupervised representation learning approaches applied to EHR data[59–62], employing various strategies such as autoencoding, masked autoencoding, and forecasting-based objectives.

Building on the success of contrastive learning in computer vision benchmarks[63;64], researchers have extended this methodology to medical time series[65] and biosignals[66;67] more broadly. Some attempts have also been made toward unsupervised domain adaptation for medical time series, although these efforts have so far been limited to experiments involving only two datasets[68].

Survival Analysis Survival analysis rapidly shifted away from proportional linear hazard models, as originally was proposed[69]. Subsequent studies proposed modeling the hazard function directly with neural networks[70;71], and later work[72] extended this by discarding the proportional hazards assumption. Most recently, survival analysis has emerged as a viable self-supervised pretraining objective for large deep neural networks. Multiple promising results[16;17] indicate broader utility beyond modeling task and study specific survival distributions, but to learn comprehensive patient state representations, which naturally captures the event driven nature of modern health care systems.

4.2 Dataset creation

To advance intensive care modeling with deep learning, we created the largest harmonized critical care time series dataset to date. To scale the number of high-quality harmonized patient stays with relatively dense data observations from intensive care units, but also emergency departments and other closely monitored critical care settings, we include the following published datasets from the USA (MIMIC-III[53], MIMIC-IV[38], eICU[55], and NWICU[73]), Europe (AmsterdamUMCdb[43] as UMCdb, SICdb[2], and HiRID[1]), and China (PICdb[33] and Zigong EHR[40]). Additionally, we incorporate an emergency department (ED) dataset MIMIC-IV-ED[3]. We further harmonize a structured electronic health records dataset EHRSHOT[74], consisting of coded data but offering a relevant amount of vital measurements, lab test results, and treatment information.

4.2.1 Data Harmonization

The datasets we consider are recorded using different, non-standardized formats. The `ricu`[12] package serves as a basis for dataset harmonization. It defines data source agnostic concepts as an abstraction for encoding clinical concepts. These include static information about the patient (e.g., height), observations (e.g., heart rate), and treatments (e.g., administration of antibiotics). By mapping the concepts to source variables in each dataset, the package facilitates exporting a unified view across them. We expand `ricu` to include new observation concepts and incorporate new datasets, namely SICdb, NWICU, PICdb, Zigong EHR, MIMIC-IV-ED, and EHRSHOT. We define new concepts using clinical expert opinion informed by what variables were reported as most important for various tasks and models in the literature (Supplementary Section A.2.3).

Crucially, we introduce a principled way to harmonize a wide range of treatment variables. Next to continuous amounts of supplemental oxygen, a selection of key ventilator measurements, and amounts of liquid filtered from the patient by means of continous renal replacement therapy (CRRT), we extract exact administration rates for core intensive care drugs, such as vasopressors (e.g., Epinephrine), blood thinners (e.g., Heparin), sedatives (e.g., Propofol), and loop diuretics. Further administered drugs are aggregated into higher level treatment concepts such as administration of any type of antibiotics, opioid pain management, non-opioid pain management, etc. In total we collect around 40 harmonized treatment concept variables covering many critical intensive care treatment procedures.

To comprehensively harmonize 16,000 raw treatment descriptors across the entire dataset collection we implemented a processing pipeline to assign raw data entries to our curated treatment concepts. We leverage the power of LLMs[75] to create an initial automated assignment of each raw source descriptor to a target treatment concept. The assignment is then verified by data scientists and further finalized by two trained physicians, ensuring clinical accuracy of the variable mapping. This process results in significantly increased number of concepts compared to previous works[15;58] to 141 (see Supplementary Section A.3).

4.2.2 Processing pipeline

We use anonymized data (Supplementary Section A.6) with permissive exclusion criteria (Supplementary Section A.2.1) to include as many patients as possible. The time series are extracted as a uniform grid at resolutions of 60 minutes using temporally causal median aggregation strategies if multiple observations are present (some datasets can support higher resolution extraction, but we opt for a harmonized view at a single resolution). Further, similar to Yèche et al.[56], we remove outliers, impute missing values, scale variables, extract features for tree-based models, and define task labels.

Finally, we export the processed data into two formats consumable by modern deep learning and classical machine learning algorithms. First, a dense fully imputed time-grid of all considered variables and second a large set of statistical features to drive classical machine learning method performance. Further details on data harmonization and processing are described in Supplementary Section A.

4.2.3 Task Annotations

Our study focuses on clinically relevant real-time prediction tasks where patient outcomes in the ICU can be influenced by timely intervention. These include: circulatory failure[5], severe respiratory failure[8], and severe kidney failure[20]. Further, we consider a liver related objective building on the MELD score[39] and a simple glucose related hyperglycemia event[76]. Additionally, we consider all-cause mortality with the prediction of decompensation[52] and early event prediction of sepsis-3 annotations[15]. All of these are modeled as binary early event prediction tasks[77] (i.e. forecasting and prognosing future patient states modeled as a binary classification) with a clinically relevant prediction horizon. On the emergency department data[3] we perform early event prediction with a 2 hour horizon for hypotension (mean arterial blood pressure below 65 mmHg) and hypoxia (SpO2 blood oxygen saturation below 90%) as previously suggested by Chen et al.[11]. Further details on the exact definitions applied in Supplementary Section A.5.

4.3 Foundation model development

A strong pretraining objective for a time series foundation model encourages the model to learn how past patient observations affect future trajectory changes. This procedure should ultimately drive relevant early event prediction downstream applications. Most recent published time series foundation models focus on exact trajectory forecasting[19;50]. However, these models are limited to regularly sampled univariate time series forecasting and thus unsuitable for the irregularly sampled, multivariate time series observed in intensive care data. Lyu et al.[35] have explored a similar type of pretraining on a single intensive care dataset.

Exact trajectory forecasts can be challenging to train on noisy and heavily confounded healthcare data (clinician's interventions act as a hidden confounder for the future trajectory). To this end Steinberg et al.[16;17] have proposed time-to-event objectives to pretrain foundation models for structured electronic health records (EHR), such as billing codes, and image data paired with structured EHR. The high-level idea is to consider a set of relevant target diagnostic, laboratory test, and treatment codes as targets and learn the distribution of time until a certain event occurs. As proposed for structured EHR this focuses on the presence of a certain billing code in the future and ignores potentially associated values (e.g. the actual result of a lab test).

4.3.1 A novel time-to-event objective

The objective as proposed by Steinberg et al.[16] could be applied to intensive care data by defining events based on the presence of laboratory test results as the frequency of performed tests is indicative for suspected conditions. By doing so, the model would learn to distill the clinicians' diagnostic expertise. However, while this approach might work well for structured EHR with clear diagnostic codes and typically longer term disease progression modeling goals, it has several drawbacks on the data observed in intensive care. First, regularly sampled variables such as any type of vital do not contribute to the learning process as their measurement pattern is trivially learned, potentially with the exception of measurements performed by equipment only attached to the patient under certain circumstances such as a ventilator. Second, measurement patterns are not only indicative of suspected conditions, but it's common practice in intensive care to perform certain measurements routinely, which is trivially learned by the algorithm irrespective of the patient state and the learned policy will be specific to the distribution of training hospitals and generalize less to unseen target hospitals. Finally, the actual value of the test as an outcome is completely ignored in this approach.

Guided by the structure of established clinical scores[14;78] we propose a novel time-to-event self-supervised pretraining objective. We select a core set $\mathcal{S}_{target}$ of observational variables (from vitals and laboratory tests) as our target set of variables to learn from. This set of variables is chosen to be well represented in most hospitals in the training pool and in alignment with relevant potential downstream applications. Note that we explicitly exclude treatments from this set to avoid learning hospital-specific treatment policies. Given patient data $\mathbf{X} \in \mathbb{R}^{T,K}$ (for a patient with length of stay T and total of K observed variables). For each of these target variable we sample a threshold τ and define an event for a variable $k \in \mathcal{S}_{target}$ at time t as $e_{t,k} = \mathbb{1}_{|\mathbf{X}_{t,k}|>\tau}$ where due to prior centering of the data this defines events as moving away from the mean of the data into the tails of the distribution (abnormal observed values). Let the annotated event time be T_e. The model is trained to learn the time-to-event distribution $\mathbb{P}(T_e = t + r \,|\, k,\, X_t,\, \tau)$ for a variable $k \in \mathcal{S}$ conditioned on the patient's history up to time t, some offset in time into the future r, and a specific threshold τ. In practice we sample τ at random during training and learn a time-to-event distribution over the entire value range of each variable $k \in \mathcal{S}$.

Given a patient history $\mathbf{X}_t$ at time t the model function g learns a latent representation $\mathbf{H}_t = g(\mathbf{X}_t)$. To learn the time-to-event distribution we learn a hazard function[77;79] for $\lambda_k : (\mathbf{H_t}, \tau) \mapsto \mathbb{P}^{T_{surv}}$ for each target variable, which maps the latent patient state to a probability of an event occurring at each time step from the current time point t up to $t + T_{surv}$, which is the maximum truncated survival distribution horizon considered.

Further details on model architecture, objective function, and hyperparameter ablations are in Supplementary Section B.

4.3.2 Inference capabilities

Zero-shot prognosis At inference, the trained hazard prediction heads can be used to obtain a probabilistic estimate on whether the variable with index k will go above (or below) a desired threshold τ at a fixed distance into the future. However, typically at inference, we are interested in the cumulative risk up to a certain horizon h into the future. Following the approach by Yèche et al.[77], this can be obtained by performing inference on the cumulative failure function $F_k(h|\mathbf{H_{i,t}},\, \tau)$ defined as follows:

$$F_k(h \,|\, \mathbf{H_t},\, \tau) = 1 - \prod_{r=1}^{h} (1 - \lambda_k(r \,|\, \mathbf{H_t},\, \tau)) \tag{1}$$

For example to perform a zero-shot prediction of the estimated cumulative circulatory failure risk within the next 8 hours as defined by Hyland et al.[5] (mean arterial pressure (MAP) < 65 mmHg and blood lactate (Lact.) > 2mmol/l) we can use the following derivation:

$$F_{\text{Circ. Failure}}(\text{8hr} \,|\, \mathbf{H_t}) \approx F_{\text{MAP}}(\text{8hr} \,|\, \mathbf{H_t},\, < 65\,\text{mmHg}) * F_{\text{Lact.}}(\text{8hr} \,|\, \mathbf{H_t},\, > 2\,\text{mmol/l}) \tag{2}$$

Hence, the pretrained foundation model can be used to perform zero-shot early event predictions right after pretraining solely by providing the clinical event definition of the target that is aimed to be predicted several hours in advance. This capability allows clinicians to customize model predictions on-the-fly to the specific patient's comorbidities and treatment regime needs without any adaptation of the model. Note that this approach assumes statistical independence among variables beyond conditioning on the same patient state. In practice, our results show that this performs well enough and is highly scalable due to limited additional overhead with a growing target variable set.

Zero-shot forecasting Because the model has been trained across the whole value range of each target variable, at inference, the threshold τ can be arbitrarily varied. Let Y_k the be random variable of values observed for a specific target k in the upcoming horizon and let $\mathbb{P}(Y = y \,|\, \mathbf{H}_i, h)$ be the probability distribution of Y_k taking value y within the horizon h. We can thus approximate the expected value of a specific variable k within the upcoming horizon h hours into the future by discretizing the value space into a set of bins $(\tau_l, \tau_u) \in \mathcal{B}$ with a lower and upper threshold boundary:

$$\mathbb{E}[Y_k \,|\, \mathbf{H}_i, h] \approx \sum_{(\tau_l, \tau_u) \in \mathcal{B}} \big(F_k(h \,|\, \mathbf{H_t},\, \tau_l) - F_k(h \,|\, \mathbf{H_t},\, \tau_u)\big) * \frac{(\tau_u - \tau_l)}{2} \tag{3}$$

This capability can be leveraged in two ways. First, by approximating expected value distributions within an upcoming horizon h we can approximate clinical scores such as Apache[78] or MELD[39] several hours into the future, giving yet another angle to perform zero-shot predictions without any form of task adaptation on the model. Second, this can be used to interrogate the model's internal belief, where the patient's future trajectory is headed, and provide a degree of interpretability. Note that given the exact training objective (Supplementary Section B) this is really an approximation, but our results suggest that it is sufficiently accurate to serve as additional output to interpret the model's internal beliefs on how a patient's trajectory might evolve when performing zero-shot inference.

4.3.3 ICareFM inference and deployment modes

We evaluate ICareFM across four distinct deployment scenarios that reflect different clinical implementation contexts, each requiring varying levels of local data availability and computational resources. All evaluations are performed with the foundation model pretrained exclusively on external data, completely excluding the target hospital from pretraining.

Zero-Shot (Domain and Task). In this most restrictive scenario, ICareFM is deployed without any task-specific training or local data adaptation. The model leverages solely the capabilities acquired during self-supervised pretraining to perform inference. For clinical event prediction, we utilize the model's learned hazard prediction heads λ_k to directly estimate event probabilities by combining predicted probabilities for reaching critical value regions inspired by clinical scores. However, different to a clinical scoring function, which combines currently observed measurements, ICareFM prognostically applies a diagnostic event definition in the future. This mode enables flexible, on-the-fly prediction adjustments at inference time without requiring any adaptation procedures, making it suitable for immediate deployment in resource-constrained settings or for exploratory clinical applications, such as pilot studies which do not require patient consent for training.

Domain Generalization. This regime represents the most scalable deployment scenario for dedicated early warning systems. Here, we leverage the large amounts of harmonized external data to tune ICareFM for specific clinical tasks while assessing direct domain zero-shot transfer performance on new target hospitals. We remove the original hazard prediction heads λ_k and replace them with task-specific linear prediction heads and either only the prediction head or the entire network is tuned. The resulting task-adapted model can then be deployed across any target hospital without requiring local data, enabling robust global scalability while maintaining task-specific performance. However, different to the prior more general mode, this model is then bound to predict events according to the label definition used during task adaptation.

Domain and Task Adaptation. This mode is designed for clinical research settings where local labeled data is available but there are no external labels available. This could be the case if the model is being tested to diagnose and estimate hospital or study specific biomarkers or disease annotations not available in the harmonized data collection. Starting from the core ICareFM network trained solely with the self-supervised objective on external data, we extract latent patient state representations $\mathbf{H_t}$ at each time step and train lightweight prediction heads (linear classifiers or multi-layer perceptrons with two layers and a non-linear activation) using only local labeled data. This approach is particularly valuable for accelerating clinical research involving newly developed biomarkers or treatments, where task-specific external labeled data may not be available. We refer to this as *Probe FM* in Fig. 2 C). If large amounts of local data are available the entire ICareFM backbone can be tuned (*Finetune FM*) to detect task-specific patterns to obtain state-of-the-art performance in most observed target datasets and benchmarked donwstream tasks.

Domain Adaptation. This regime represents the optimal long-term deployment strategy for hospitals with permissive data usage policies and sufficient computational resources. The core pretrained foundation model network is first adapted to specific tasks using labeled data from external datasets, then fine-tuned using local data to achieve state-of-the-art performance. We investigate both full fine-tuning (updating all model parameters) and parameter-efficient fine-tuning techniques such as LoRA[80]. During fine-tuning, we replace the original hazard prediction heads with randomly initialized task-specific heads and optimize the entire model end-to-end using local data.

For all adaptation modes involving local data, we systematically vary the amount of available local training data to assess the local data efficiency of each approach. This evaluation protocol enables us to quantify the equivalent number of local patients required to match foundation model performance when training task-specific models from scratch, providing practical guidance for clinical deployment decisions depending on data and label availability, hospital size and resources, and local patient data usage policies.

4.3.4 Evaluation procedures

Intensive care data shows strong distribution shifts across hospitals[8], which makes it challenging to outperform a locally trained model using only external data. However, foundation models are typically pretrained using external data and hopefully show strong performance when used out-of-the-box or with minor adaptation to reduce the effective local model development and training time. Typically, when the unseen target hospital has large amounts of patient data available for developing its local model, state-of-the-art machine learning algorithms can achieve strong predictive performance. However, for smaller hospitals or hospitals in countries with stricter regulations on patient data usage for AI model training, leveraging external data to achieve optimal performance can be key.

For each hospital and downstream prediction task, we train a strong set of reference models using state-of-the-art gradient boosting trees[81] using a comprehensive set of statistical features over multiple historical horizons into the past (based on work by[82] and further expanded to improve performance Supplementary Section A.4.3) and advanced deep neural network sequence architectures[6;56]. For each of the two types of algorithms, we train models for an increasing amount of patients from the target hospital and select the best method across a collection of hyperparameters on a validation set, which is part of the budget for the target hospital. The optimal model and its hyperparameter configuration are selected on the validation set performance averaged across three different training runs on the target dataset. Each run uses a different random initialization, randomness for optimization, and a different subsample from the target. Final performance is always reported on a complete held-out test set.

We then perform zero-shot evaluations and adaptations of the foundation model (trained on external data) using probe and fine-tuning approaches with an increasing budget of target patients for adaptation. For each of the four foundation model usage modes (Section 4.3.3), we compute the intersection of the performance curves of the foundation model and the prior computed best reference performances of models leveraging solely data from the target hospital and trained using best practices for state-of-the-art machine learning model development. This yields a performance assessment expressed as *target dataset #patient performance equivalence*. Given an unseen target hospital and downstream task, if the available number of patients eligible for adaptation is lower than the reported number, then it is advisable to use the foundation model in its suggested mode at that sample size. If larger amounts of local training data are available, it might be advisable to develop a local task-specific solution if pure statistically observed performance is the main concern. However, development time, flexibility during deployment, and model robustness might still favor the use of the foundation model in one of its proposed deployment modes.

Acknowledgments

We gratefully acknowledge the patients and their families whose willingness to share health-care data makes this work possible. We also thank the clinicians, data engineers, and investigators who collected, curated, de-identified, and released the intensive-care datasets on which this study relies. Their sustained commitment to open, high-quality data has created an extraordinary research community that continues to advance critical-care science. Specifically, this work uses and acknowledges the MIMIC-III, MIMIC-IV, and MIMIC-IV-ED, eICU Collaborative Research Database 2.0, AmsterdamUMCdb, HIRID (High time-resolution ICU dataset), Northwestern ICU (NWICU) database, Salzburg Intensive Care database (SICdb), Pediatric Intensive Care database (PICdb), Critical care database comprising patients with infection at Zigong Fourth People's Hospital (Zigong EHR), EHRSHOT datasets.

This work was supported as part of the Swiss AI Initiative (`swiss-ai.org`) by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a02 on Alps.

Computational data analysis was performed at Leonhard Med (`https://sis.id.ethz.ch/services/sensitiveresearchdata/`) secure trusted research environment at ETH Zurich. SNF Funding (213236) to Andre Kahles contributed to compute resources.

The project was supported by grant #2022-278 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology) and ETH Core funding (to Gunnar Rätsch).

Malte Londschien was supported by the ETH Foundations of Data Science and the ETH AI Center. Fedor Sergeev was supported by grant #902 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology). Daphné Chopard received funding from grant #2021-911 of the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology).

We would like to thank Zeynep Babür, Jasmina Bogojeska from ZHAW School of Engineering for their contribution to extract further harmonized data points from the MIMIC-IV dataset.

We would like to thank the Swiss AI community members and associated students participating in our weekly meetings to push the state-of-the-art in health care time series modeling further for the discussions and their feedback, most notably: Alexander Morgenroth, Fabian Baldenweg, Simon Boutry, and Xiaochen Zheng. Fedor Sergeev thanks Vincent Fortuin for his helpful suggestions to improve the manuscript.

Ethics

This project received an ethics waiver from the cantonal ethics committee of the canton of Zurich, Switzerland (`BASEC-Nr.: Req-2024-00528`).

Data and Code Availability

Harmonized large-scale dataset If legally and organisationally feasible, the harmonized and processed multi-center dataset will be published and shared on the PhysioNet platform[83], significantly simplifying access to a large collection of critical care time-series data and accelerating research in this domain.

Source datasets All the source datasets are available via PhysioNet[83] upon completion of CITI's "Data or Specimens Only Research" course (HiRID and SICdb require additional project specific provider approval). Further datasets are accessible directly via the providers (UMCdb[43] and EHRSHOT[74]).

Data Harmonization Tooling To harmonize the different source datasets, we built upon the `ricu`[12] tool. Upon publication we will release our fork of the package containing code and configuration to harmonize the source datasets and full multi-center dataset.

Data processing Code and other required resources to reproduce the full data processing pipeline to obtain the harmonized multi-center dataset will be added or linked in the projects root Github repository `https://github.com/ratschlab/icarefm`.

Model development and experiments Code and other required resources to reproduce model training and all presented evaluation experiments will be added or linked in the projects root Github repository `https://github.com/ratschlab/icarefm`.

References

[1] Faltys, M. *et al.* HiRID, a high time-resolution icu dataset (version 1.1.1). *PhysioNet* (2021). URL `https://doi.org/10.13026/nkwc-js72`.

[2] Rodemund, N., Kokoefer, A., Wernly, B. & Cozowicz, C. Salzburg intensive care database (sicdb), a freely accessible intensive care database. *PhysioNet https://doi. org/10.13026/ezs8-6v88* (2023).

[3] Johnson, A. *et al.* Mimic-iv-ed (2023). URL `https://doi.org/10.13026/jzz5-vs76`.

[4] Ehrenfeld, J. M. & Cannesson, M. *Monitoring technologies in acute care environments : a comprehensive guide to patient monitoring technology* (Springer, New York, 2014).

[5] Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine* **26**, 364–373 (2020).

[6] Tomašev, N. *et al.* Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nature Protocols* **16**, 2765–2787 (2021).

[7] Yèche, H., Pace, A., Ratsch, G. & Kuznetsova, R. Temporal label smoothing for early event prediction. In Krause, A. *et al.* (eds.) *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 of *Proceedings of Machine Learning Research*, 39913–39938 (PMLR, 2023). URL `https://proceedings.mlr.press/v202/yeche23a.html`.

[8] Hüser, M. *et al.* A comprehensive ml-based respiratory monitoring system for physiological monitoring & resource planning in the icu. *medRxiv* (2024). URL `https://www.medrxiv.org/content/early/2024/01/23/2024.01.23.24301516`. `https://www.medrxiv.org/content/early/2024/01/23/2024.01.23.24301516.full.pdf`.

[9] Plečko, D. *et al.* A framework and analytical exploration for a data-driven update of the sequential organ failure assessment (sofa) score in sepsis. *Critical Care and Resuscitation* **27**, 100105 (2025). URL `https://www.sciencedirect.com/science/article/pii/S1441277225000092`.

[10] Chen, Z. *et al.* Meditron-70b: Scaling medical pretraining for large language models, 2023. *URL https://arxiv. org/abs/2311.16079* (2023).

[11] Chen, E. *et al.* Multimodal clinical benchmark for emergency care (mc-bec): A comprehensive benchmark for evaluating foundation models in emergency medicine. *Advances in Neural Information Processing Systems* **36** (2024).

[12] Bennett, N., Plečko, D., Ukor, I.-F., Meinshausen, N. & Bühlmann, P. ricu: R's interface to intensive care data. *GigaScience* **12**, giad041 (2023).

[13] Smith, G. B. *et al.* The national early warning score 2 (news2). *Clinical Medicine* **19**, 260 (2019). URL `https://www.sciencedirect.com/science/article/pii/S1470211824011862`.

[14] Moreno, R. *et al.* The sequential organ failure assessment (sofa) score: has the time come for an update? *Critical Care* **27**, 15 (2023). URL `https://doi.org/10.1186/s13054-022-04290-9`.

[15] Moor, M. *et al.* Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *EClinicalMedicine* **62** (2023).

[16] Steinberg, E., Fries, J., Xu, Y. & Shah, N. Motor: A time-to-event foundation model for structured medical records (2023). URL `https://arxiv.org/abs/2301.03150`. `2301.03150`.

[17] Huo, Z. *et al.* Time-to-event pretraining for 3d medical imaging (2024). URL `https://arxiv.org/abs/2411.09361`. `2411.09361`.

[18] Zhou, H. *et al.* Informer: Beyond efficient transformer for long sequence time-series forecasting (2021). URL `https://arxiv.org/abs/2012.07436`. `2012.07436`.

[19] Ansari, A. F. *et al.* Chronos: Learning the language of time series. *Transactions on Machine Learning Research* (2024). URL `https://openreview.net/forum?id=gerNCVqqtR`.

[20] Lyu, X. *et al.* An empirical study on kdigo-defined acute kidney injury prediction in the intensive care unit. *Bioinformatics* **40**, i247–i256 (2024).

[21] Youssef, A. *et al.* External validation of ai models in health should be replaced with recurring local validation. *Nature Medicine* **29**, 2686–2687 (2023). URL `https://doi.org/10.1038/s41591-023-02540-z`.

[22] Brown, T. B. *et al.* Language models are few-shot learners (2020). URL `https://arxiv.org/abs/2005.14165`. `2005.14165`.

[23] Chen, Z. *et al.* Meditron-70b: Scaling medical pretraining for large language models (2023). `2311.16079`.

[24] Kirillov, A. *et al.* Segment anything. *arXiv:2304.02643* (2023).

[25] Radford, A. *et al.* Learning transferable visual models from natural language supervision. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 8748–8763 (PMLR, 2021). URL `https://proceedings.mlr.press/v139/radford21a.html`.

[26] Zhang, S. *et al.* A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI* **2**, AIoa2400640 (2025). URL `https://ai.nejm.org/doi/full/10.1056/AIoa2400640`. `https://ai.nejm.org/doi/pdf/10.1056/AIoa2400640`.

[27] Baldenweg, F., Burger, M., Rätsch, G. & Kuznetsova, R. Multi-modal contrastive learning for online clinical time-series applications (2024). URL `https://arxiv.org/abs/2403.18316`. `2403.18316`.

[28] Katsoulakis, E. *et al.* Digital twins for health: a scoping review. *npj Digital Medicine* **7**, 77 (2024). URL `https://doi.org/10.1038/s41746-024-01073-0`.

[29] Shmatko, A. *et al.* Learning the natural history of human disease with generative transformers. *medRxiv* 2024–06 (2024).

[30] Callaway, E. Medical ai trained on whopping 57 million health records. *Nature* (2025).

[31] HAJICON *et al.* Flaticon. `https://www.flaticon.com` (2025). Accessed: May 26, 2025.

[32] Tisch, M., Dürr, S. & Servier. Bioicons. `https://bioicons.com` (2025). Accessed: May 26, 2025.

[33] Li, H., Zeng, X. & Yu, G. Paediatric intensive care database (2019).

[34] Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes (2018). URL `https://arxiv.org/abs/1610.01644`. `1610.01644`.

[35] Lyu, X., Hueser, M., Hyland, S. L., Zerveas, G. & Raetsch, G. Improving clinical predictions through unsupervised time series representation learning (2018). URL `https://arxiv.org/abs/1812.00490`. `1812.00490`.

[36] Kellum, J. & et al., L. N. Kdigo clinical practice guideline for acute kidney injury. section 2: Aki definition. *Crit. Care Med.* **Suppl. 2**, 19–36 (2012).

[37] Bilan, N., Dastranji, A. & Ghalehgolab Behbahani, A. Comparison of the spo2/fio2 ratio and the pao2/fio2 ratio in patients with acute lung injury or acute respiratory distress syndrome. *J. Cardiovasc. Thorac. Res.* **7**, 28–31 (2015).

[38] Johnson, A. *et al.* MIMIC-IV" (version 3.1). PhysioNet (2024) (2024). URL `https://physionet.org/content/mimiciv/3.1/`.

[39] Malinchoc, M. *et al.* A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology* **31** (2000). URL `https://journals.lww.com/hep/fulltext/2000/04000/a_model_to_predict_poor_survival_in_patients.8.aspx`.

[40] Xu, P., Chen, L. & Zhang, Z. Critical care database comprising patients with infection at zigong fourth people's hospital (version 1.1) (2022).

[41] van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008). URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

[42] Koh, P. W. *et al.* Concept bottleneck models. In III, H. D. & Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 5338–5348 (PMLR, 2020). URL `https://proceedings.mlr.press/v119/koh20a.html`.

[43] Thoral, P. J. *et al.* Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: the amsterdam university medical centers database (amsterdamumcdb) example. *Critical care medicine* **49**, e563–e577 (2021).

[44] Wornow, M., Thapa, R., Steinberg, E., Fries, J. A. & Shah, N. H. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models (2023). URL `https://arxiv.org/abs/2307.02028`. `2307.02028`.

[45] Saab, K. *et al.* Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416* (2024).

[46] Guo, L. L. *et al.* A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine* **7**, 171 (2024).

[47] Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine* **6**, 135 (2023).

[48] Guo, L. L. *et al.* Ehr foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports* **13**, 3767 (2023).

[49] Das, A., Kong, W., Sen, R. & Zhou, Y. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688* (2023).

[50] Goswami, M. *et al.* Moment: A family of open time-series foundation models. In *International Conference on Machine Learning* (2024).

[51] Zhou, T., Niu, P., Sun, L., Jin, R. *et al.* One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems* **36**, 43322–43355 (2023).

[52] Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Scientific data* **6**, 1–18 (2019).

[53] Johnson, A., Pollard, T. & Roger, M. MIMIC-III Clinical Database (2016). URL `https://physionet.org/content/mimiciii/1.4/`.

[54] Sheikhalishahi, S., Balaraman, V. & Osmani, V. Benchmarking machine learning models on multi-centre eicu critical care dataset. *Plos one* **15**, e0235424 (2020).

[55] Pollard, T. J. *et al.* The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data* **5**, 1–13 (2018).

[56] Yèche, H. *et al.* Hirid-icu-benchmark — a comprehensive machine learning benchmark on high-resolution icu data. In Vanschoren, J. & Yeung, S. (eds.) *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1 (2021). URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/5878a7ab84fb43402106c575658472fa-Paper-round1.pdf`.

[57] Wang, S. *et al.* Mimic-extract: a data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, ACM CHIL '20 (ACM, 2020). URL `http://dx.doi.org/10.1145/3368555.3384469`.

[58] Van De Water, R. *et al.* Yet another icu benchmark: A flexible multi-center framework for clinical ml. *arXiv preprint arXiv:2306.05109* (2023).

[59] Lyu, X., Hueser, M., Hyland, S. L., Zerveas, G. & Raetsch, G. Improving clinical predictions through unsupervised time series representation learning. *arXiv preprint arXiv:1812.00490* (2018).

[60] Tipirneni, S. & Reddy, C. K. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **16**, 1–17 (2022).

[61] Labach, A. *et al.* Effective self-supervised transformers for sparse time series data (2023). URL `https://openreview.net/pdf?id=HUCgU5EQluN`.

[62] Im, S., Oh, J. & Choi, E. Labtop: A unified model for lab test outcome prediction on electronic health records. *arXiv preprint arXiv:2502.14259* (2025).

[63] Misra, I. & Maaten, L. v. d. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6707–6717 (2020).

[64] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738 (2020).

[65] Yèche, H., Dresdner, G., Locatello, F., Hüser, M. & Rätsch, G. Neighborhood contrastive learning applied to online patient monitoring. In *International conference on machine learning*, 11964–11974 (PMLR, 2021).

[66] Wang, Y., Han, Y., Wang, H. & Zhang, X. Contrast everything: A hierarchical contrastive framework for medical time-series. *Advances in Neural Information Processing Systems* **36**, 55694–55717 (2023).

[67] Kiyasseh, D., Zhu, T. & Clifton, D. A. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, 5606–5615 (PMLR, 2021).

[68] Ozyurt, Y., Feuerriegel, S. & Zhang, C. Contrastive learning for unsupervised domain adaptation of time series. *arXiv preprint arXiv:2206.06243* (2022).

[69] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972).

[70] Yousefi, S. *et al.* Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports* **7**, 1–11 (2017).

[71] Katzman, J. L. *et al.* Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology* **18**, 24 (2018).

[72] Gensheimer, M. & Narasimhan, B. A scalable discrete-time survival model for neural networks. peerj 2019; 7: e6257.

[73] Moukheiber, D. *et al.* Northwestern icu (nwicu) database (version 0.1.0). *PhysioNet* (2024). URL `https://doi.org/10.13026/s84w-1829`.

[74] Wornow, M., Thapa, R., Steinberg, E., Fries, J. & Shah, N. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models (2023). `2307.02028`.

[75] Team, G. *et al.* Gemma 2: Improving open language models at a practical size (2024). URL `https://arxiv.org/abs/2408.00118`. `2408.00118`.

[76] Mehdizavareh, H., Khan, A. & Cichosz, S. L. Enhancing glucose level prediction of icu patients through hierarchical modeling of irregular time-series (2025). URL `https://arxiv.org/abs/2411.01418`. `2411.01418`.

[77] Yèche, H., Burger, M., Veshchezerova, D. & Rätsch, G. Dynamic survival analysis for early event prediction (2024). URL `https://arxiv.org/abs/2403.12818`. `2403.12818`.

[78] Knaus, W. A., Zimmerman, J. E., Wagner, D. P., Draper, E. A. & Lawrence, D. E. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit. Care Med.* **9**, 591–597 (1981).

[79] Maystre, L. & Russo, D. Temporally-consistent survival analysis. In Koyejo, S. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 35, 10671–10683 (Curran Associates, Inc., 2022). URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/455e1e30edf721bd7fa334fffabdcad8-Paper-Conference.pdf`.

[80] Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models (2021). URL `https://arxiv.org/abs/2106.09685`. `2106.09685`.

[81] Ke, G. *et al.* LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30**, 3146–3154 (2017).

[82] Soenksen, L. R. *et al.* Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine* **5**, 149 (2022). URL `https://doi.org/10.1038/s41746-022-00689-4`.

[83] Goldberger, A. L. *et al.* Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* **101**, e215–e220 (2000).

[84] Grechko, A. V. *et al.* Ricd: Russian intensive care dataset. *General Reanimatology* **20**, 22–31 (2024).

[85] Oliver, M., Allyn, J., Carencotte, R., Allou, N. & Ferdynus, C. Introducing the blendedicu dataset, the first harmonized, international intensive care dataset. *Journal of Biomedical Informatics* **146**, 104502 (2023). URL `https://www.sciencedirect.com/science/article/pii/S153204642300223X`.

[86] Ellis, R. K. Determination of po2 from saturation. *Journal of Applied Physiology* **67**, 902–902 (1989). URL `https://doi.org/10.1152/jappl.1989.67.2.902`. PMID: 2793692.

[87] Wiesner, R. *et al.* Model for end-stage liver disease (meld) and allocation of donor livers. *Gastroenterology* **124**, 91–96 (2003). URL `https://doi.org/10.1053/gast.2003.50016`.

[88] Singer, M. *et al.* The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* **315**, 801–810 (2016). URL `https://doi.org/10.1001/jama.2016.0287`. `https://jamanetwork.com/journals/jama/articlepdf/2492881/jsc160002.pdf`.

[89] Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus) (2023). URL `https://arxiv.org/abs/1606.08415`. `1606.08415`.

[90] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

[91] Li, Y., Si, S., Li, G., Hsieh, C.-J. & Bengio, S. Learnable fourier features for multi-dimensional spatial positional encoding (2021). URL `https://arxiv.org/abs/2106.02795`. `2106.02795`.

[92] Schuppli, S. *et al.* Evolving hpc services to enable ml workloads on hpe cray ex (2025). URL `https://arxiv.org/abs/2507.01880`. `2507.01880`.

[93] Fusco, L. *et al.* Understanding data movement in tightly coupled heterogeneous systems: A case study with the grace hopper superchip (2024). URL `https://arxiv.org/abs/2408.11556`. `2408.11556`.

Supplementary Material for “A Foundation Model for Intensive Care: Unlocking Generalization across Tasks and Domains at Scale”

Manuel Burger[1,+], Daphné Chopard[1,2,*], Malte Londschien[3,4,*], Fedor Sergeev[1,5,*],
Hugo Yèche[1], Rita Kuznetsova[1], Martin Faltys[6], Eike Gerdes[7],
Polina Leshetkina[8], Peter Bühlmann[3], Gunnar Rätsch[1,4,9,10,+]

[1] Department of Computer Science, ETH Zurich, Switzerland; [2] Department of Intensive Care and Neonatology and Children’s Research Center, University Children’s Hospital Zurich, University of Zurich, Switzerland; [3] Seminar for Statistics, ETH Zurich, Switzerland; [4] AI Center, ETH Zurich, Switzerland; [5] Helmholtz AI, Germany; [6] Department of Intensive Care Medicine, University Hospital and University of Bern, Switzerland; [7] University Hospital Zurich, Switzerland; [8] Department of Health Science and Medicine, University of Luzern, Switzerland; [9] Swiss Institute for Bioinformatics, Zurich, Switzerland; [10] Medical Informatics Unit, University Hospital Zurich, Switzerland

[+] Joint-corresponding authors: `manuel.burger@inf.ethz.ch` and `gunnar.raetsch@inf.ethz.ch`
[*] These authors contributed equally and share second authorship.

A Data

A.1 Data Collection

The overview of datasets is presented in Supplementary Table 1 and Supplementary Table 2. All these datasets are available freely via PhysioNet[83] or directly from the providers (UMCdb[43] and EHRSHOT[74]). We do not include RICD[84] because it is not freely accessible (a separate contract and payment are required).

A.2 Data Harmonization

We include the following published datasets from the USA (MIMIC-III (v1.4)[53], MIMIC-IV (v3.1)[38], eICU (v2.0)[55], and NWICU (v0.1.0)[73]), Europe (AmsterdamUMCdb (v1.0)[43] as UMCdb, SICdb (v1.0.6)[2], and HiRID (v1.1.1)[1]), and China (PICdb (v1.1.0)[33] and Zigong EHR (v1.1)[40]). Additionally, we incorporate an emergency department (ED) dataset MIMIC-IV-ED (v2.2)[3]. We further harmonize a structured electronic health records dataset EHRSHOT (v2.1)[74], consisting of coded data but offering a relevant amount of vital measurements, lab test results, and treatment information.

MIMIC-III The “Medical Information Mart for Intensive Care” is a large single-center dataset from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. It has relatively high data density, accurate reporting of treatment rate administration, and ventilation information. Most notably the dataset contains data from two critical care data management systems (Philips CareVue Clinical Information System and iMDsoft Metavision ICU). There are hence already within a single dataset harmonization efforts required to comprehensively model the entire dataset patient population. We extract individual intensive care stays from this dataset.

MIMIC-IV The MIMIC-IV dataset represents a comprehensive update to MIMIC-III. It contains only data from the period after introducing the iMDsoft Metavision ICU data management system and in version v3.1 contains patients up to the year 2022 (including the COVID-19 period). There is a non-trivial overlap of patients between the MIMIC-III and MIMIC-IV dataset. This overlap could be mostly removed by only considering patient data stored in the CareVue system, when using data from MIMIC-III, if this is desirable. We extract individual intensive care stays from this dataset, while the dataset could support extracting entire hospital admissions potentially including multiple subsequent ICU stays.

eICU The “eICU Collaborative Research Database” is the largest dataset in the collection and contains patient data from 207 hospitals across the United States. The provided data density is comparatively lower, but the data is comprehensive and facilitates accurate treatment rate extractions, large set of laboratory test results, and ventilations settings amongst other data. We extract individual intensive care stays from this dataset, while the dataset could support extracting entire hospital admissions potentially including multiple subsequent ICU stays.

NW-ICU The "Northwestern ICU (NWICU) database" is a multi-center dataset from a recent period (2020–2022) spanning the COVID-19 pandemic period. The dataset has lower data densities compared to, e.g., the MIMIC-III/IV datasets and we found it more challenging to retrieve accurate treatment information from the dataset. Surprisingly in its current version (v0.1.0) we were not able to retrieve comprehensive ventilation information from the dataset, which would be highly relevant to accurately study COVID-19 patient populations. We extract individual intensive care stays from this dataset, while the dataset could support extracting entire hospital admissions potentially including multiple subsequent ICU stays.

HiRID The "HiRID, a high time-resolution ICU dataset" is a high-quality intensive care dataset. The data contains highly dense vital measurements and laboratory test results, accurate reporting of treatment administration and ventilation settings. We extract individual intensive care stays from this dataset. The dataset does not support associating multiple subsequent stays with the same patient.

UMCdb The "AmsterdamUMCdb" is a high-quality intensive care dataset. The data contains highly dense vital measurements and laboratory test results, accurate reporting of treatment administration and ventilation settings. We extract individual intensive care stays from this dataset. The dataset does not support associating multiple subsequent stays with the same patient.

SICdb The "Salzburg Intensive Care database (SICdb)" is a high-quality intensive care dataset. The data contains highly dense vital measurements and laboratory test results, accurate reporting of treatment administration and ventilation settings. The dataset shares the highest frequency for vitals across all datasets. Vitals can be extracted to minute-level updates. We extract individual intensive care stays from this dataset. The dataset does not support associating multiple subsequent stays with the same patient.

PICdb The "Pediatric Intensive Care database" is a chinese pediatric care dataset and hence, except for a few children in the MIMIC-III dataset, the only dataset in the harmonized collection containing children or newborns. The measurement density on this dataset is relatively low. Treatments are mostly extracted only as indicators due to lack of administration rate information. We extract individual intensive care stays from this dataset, while the dataset could support extracting entire hospital admissions potentially including multiple subsequent ICU stays.

Zigong EHR The "Critical care database comprising patients with infection at Zigong Fourth People's Hospital" is a small dataset focused on infectious diseases and hence a patient population with more severe outcomes compared to the other datasets in the collection. We extract individual intensive care stays from the dataset, while the dataset could support extracting entire hospital admissions potentially including multiple subsequent ICU stays. Data recording density is also much lower, which makes this dataset more challenging both to accurately annotate and for the machine learning models to perform well. Treatments are mostly extracted only as indicators due to lack of administration rate information.

EHRSHOT This dataset contains data from Stanford medicine. It has not been built with the primary objective to share high frequency data from intensive care such as continuous vital measurements. It contains mostly coded data points recorded for billing and insurance purposes. However, after closer inspection we found it contains a considerable amount of coded vital, lab, and treatment measurements paired with the actually observed numeric values. We harmonize the data source and develop a heuristic (Supplementary Section A.7) to extract more general hospital stays, which might include periods of intensive care stays, but do not need to. The extracted stays are hence much sparser in available information and represent a comparatively healthier patient population.

MIMIC-IV ED This dataset contains emergency department admissions at the Beth Israel Deaconess Medical Center from 2011–2019. The ED cohort is very heterogeneous, with varying disease severities. We extract sparse vital information from the datasets `triage` and `vitalsigns` tables. There is no continuous treatment data available, but we extract treatment indicators based on the information in the `pyxis` table. We additionally merge in missing demographic information from the core MIMIC-IV dataset, where possible.

A.2.1 Inclusion criteria

We consider patient stays that after extraction have

- a valid admission and discharge time,
- a valid length of stay (LoS) that is longer than 4 fours,
- a maximum gap between measurements smaller than 48 hours,
- and more than 4 measurements.

Compared to Van De Water et al.[58], we broaden the inclusion criteria by reducing the LoS requirement from 6 to 4 hours and increasing the allowed maximum gap between measurements from 12 to 48 hours.

By including as many patients as possible, we aim to create a more general version of the dataset, that can be further trimmed down for specific studies. Additionally, a wide range of stays can improve the generalizability of predictive models. Furthermore, as this work focuses on the development of a foundation model it is important to keep exclusion criteria minimal and include as many patients as possible irrespective of their clinical trajectory to large dataset covering a wide demographic.

A.2.2 Intensive Care data types

Intensive care data relevant for timely predictions and harmonized in our dataset can be grouped into the following data types:

- Demographic patient information, which is data considered to remain constant during the time period of an intensive care stay such as gender, age, height, etc. Some of these data points can be categorical such as ethnicity (not available in all datasets).
- Numeric physiological vital data measured by various equipment such as heart rate, oxygen saturation or data measured by ventilators such as plateau pressure, peak pressure, etc. This data is observed continously and also stored with high frequency and density in most of the harmonized datasets.
- Categorical data recorded by equipment or entered manually by nursing staff such as glascoma score or RASS scores. Human derived scores are available more sparsely and availability also varies across hospitals depending on nursing staff policies.
- Numeric laboratory test results from blood or other body fluid samples. These are available sparsely and the data density varies across hospitals due to different patient population needs, hospital policies on regular monitoring activities, and financial trade-offs.
- Treatments:
  - Life support equipment to support or take over functions of failing organs: Ventilation settings or CRRT (continuous renal replacement therapy) settings
  - Drugs given via infusions, we extract both accurate treatment administration rates and treatment indicators (pure presence of the treatment while the infusion is active)
  - Drugs given via other pathways as bolus form (injections, tablets, etc.). If dosage information is available we harmonize these drug administrations together with the infusions by computing an instantaneous application rate, otherwise a simple administration indicator is extracted.

A.2.3 Concepts for treatment variables

To create a set of treatment concepts, we select variables that were identified as most predictive for various labels in published literature (circulatory failure[5], respiratory failure[8], kidney failure[20], sepsis[15]). Then, based on the input from consulting ICU clinicians, some high priority concepts are made more granular (e.g., vasopressors) while the low priority ones are left out (e.g., laxatives).

In about half of the datasets, treatment variables are stored as plain text, amounting to over 16,000 unique entries. Manual mapping of that amount of variables would be slow and error-prone. Instead, we first use LLMs to pre-label the variables according to the provided concept set. This allows grouping of similar treatments, making manual verification much quicker. The manual verification is then performed in three rounds: first by data scientists, and then independently by two physicians, ensuring the mapping is correct.

Previous works Oliver et al.[85] have proposed a processing pipeline for harmonizing treatments by including drug exposure information as indicators. We expand on this by (1) considering not only indicators, but also administration rates for core medications used in critical care settings, and (2) grouping individual drugs into abstract treatment concepts, thereby increasing the overlap across datasets in concepts while maintaining relevance for downstream applications.

More generally, previous works[15] have suggested that including medication variables harms the accuracy of predictive models when transferring models due to changes in treatment policies[8], but little research has been done on the reasons behind this effect and what can be done to mitigate it. The information about administered medications is an insight into the actions of the clinicians, and could drastically improve the model accuracy and transfer, if the model can learn proper causal treatment effects. By including these variables in the harmonization pipeline, we prepared the ground for this future research on learning generalizable treatment effect estimations on critical care time series.

A.3 Concept reference table

The full concept (or variable) reference table is shown in Supplementary Section A.3 (Table to be added in future version). It includes 141 variables: 6 static demographic, 80 observations, and 55 treatment variables.

A.4 Data Processing

A.4.1 Scaling

Data is scaled depending on its type:

- continuous observations are standardized (i.e. centered and scaled to unit variance) and if deemed suitable a log transformation is applied before standardization,
- categorical observations are one-hot encoded and each variable has a dedicated class to encode missing information,
- continuous treatments are quantile-transformed and mapped to the $[0, 1]$ range such that a 0 represents *no medication given*,
- treatment indicators are binary encoded using $\{0, 1\}$ where 0 represents *no medication given*.

A.4.2 Imputation

Gridded time-step data as inputs for model training are forward-filled indefinitely for all observation variables. The remaining missing values are then imputed with 0 for continuous variables, which corresponds to a population mean imputation after considering standard scaling before the imputation stage. The remaining categorical entries are imputed with a value corresponding to the dedicated class that encodes missing information for each categorical variable.

Any treatment variable is excluded from forward-filling operations and missing data points are strictly filled using 0, which given the previously introduced scaling and encoding scheme always corresponds to no treatment being applied.

A.4.3 Feature extraction

We build on the feature set proposed by Soenksen et al.[82] to process the MIMIC-IV[38] dataset. To improve performance we then further expand this set of features and select specific features for each variable type. For each time step, each feature is computed over three history sizes of 8, 24, and 72 hours:

- For continuous observations and continuous treatment variables, we compute:
  - mean on raw and imputed data,
  - standard deviation on raw data,
  - slope of a linear fit on the raw data and imputed data,

  - mean absolute change over imputed data,
  - fraction of non-missing data points,
  - quantiles: 0% (Min.), 10%, 50%, 90%, 100% (Max.).
- For categorical variables, we compute the mode, number of missing points, and a binary indicator of whether there are any missing points at all.
- For treatment indicators we compute the number of points with treatment and a binary indicator whether any treatment was applied.

A.5 Task Annotations

We define sample labels by annotating the time series following the clinical definitions used by prior work for each specific task. As such, these labels constitute proxy labels derived and computed from data given a clinical definition to diagnose a patient state for a certain condition. Most importantly, we annotate a positive and a negative case only if there's enough evidence in favor of either. As such, a label is only computed if the source data, conditioned on a task-specific imputation scheme proposed by prior work, provides all required inputs to compute the score or state annotation of the clinical task definition at a certain time step. Based on these cases, we define early event prediction labels (e.g., respiratory failure) that are then used for online classification of the future state of the patient.

A.5.1 Event Annotations

Events are diagnostically annotated based on available data using diagnostic criteria and scoring functions. Given a patient time-series resampled to a 1 hour grid we initialize the event annotation vector with unknown states.

Circulatory Failure We follow Hyland et al.[5] relying on mean arterial pressure (at 65 mmHg) and blood serum lactate (at 2 mmol/l) measurements to annotate circulatory failure events. We apply the lactate imputation scheme as proposed by Hyland et al.[5] to create a denser availability of the patient's possible lactate measurement. The imputation scheme assumes less frequent measurements for stable patients and as such indefinitely forward and backward fills those measurements. Critical measurements above the mentioned threshold are only forward and backward filled up to 6 hours. If two measurements are closer than 6 hours the values in between are linearly interpolated. Mean arterial pressure is assumed to be typically recorded with high frequency on almost all considered datasets and as such no further imputation scheme is applied same as Hyland et al.[5]. A positive event is annoted if observed mean arterial pressure is below 65 mmHg and imputed lactate is above 2 mmol/l. A negative event is annotated if observed mean arterial pressure is above 65 mmHg and imputed lactate is below 2 mmol/l. Time-steps with any missing information for mean arterial pressure or lactate after imputation or if neither of the two boolean expressions match the data are annotated with an unknown current patient state.

Respiratory Failure A patient's respiratory state can be assessed by calculating the ratio of partial pressure of oxygen in blood (PaO2) over the fraction of inspired oxygen (FiO2, 21% in ambient air or higher during ventilation)[8], short the P/F ratio (also known as Horovitz index). Hüser et al.[8] proposed an intricate imputation scheme with a parametric imputation algorithm to obtain a dense approximation of the patients PaO2 measurements. Alternatively Yèche et al.[56] rely on a simple rule-based algorithm proposed by Ellis et al.[86] to estimate PaO2 from densely available pulse oxymetry oxygen saturation (SpO2) measurements. Due to the heterogenous nature of a large-scale harmonized multi-source datasets, we decided to use minimal assumptions and not use any form of rule-based or parametric imputation algorithm for PaO2. We backward and forward fill both FiO2 and PaO2 measurements each by at most 1 hour (essentially 1 time-step on the hourly grid). We aim to predict severe respiratory failure which is defined as a P/F ratio below 100 mmHg[8]. If after imputation both PaO2 and FiO2 information are available we annotate a positive event for P/F below 100 mmHg and a negative event for P/F over 100 mmHg, time points with missing data are annotated as unknown event state. Note that we do not make any assumption on the patient's FiO2 measurements outside the 1 hour imputation window around ground truth data. We do not impute missing data points with a 21%, which would correspond to ambient air assuming no ventilation. While the harmonization efforts aim to extract ventilation information, we could not reliably extract dense ventilation information from each dataset and as such do not impose any assumptions here, which could falsely impute ambient air inspired oxygen concentration while the patient is actually on a ventilator but with unreliable data recording and export. This decision causes the label to only annotate events for ventilated patients where FiO2 measurements are recorded by the used ventilation equipment. We do believe the task maintains high clinical relevance while relying largely only on ground truth data for annotations.

Kidney Failure We annotate most severe form (stage 3) of acute kidney injury as defined by the *Kidney Disease: Improving Global Outcomes* (KDIGO) guidelines[36]. Based on work by Lyu et al.[20] we linearly interpolate missing creatinine measurements between ground truth measurements if they are closer than 48 hours. Borders are backward and forward filled with a maximum horizon of 48 hours.[20] backward filled missing urine information indefinitely. Due to the heterogeneity in the harmonized multi-center dataset we decided to backward fill urine rates only up to 48 hours at most. After imputation the KDIGO AKI Stage 3 definitions are applied to annotate events. To summarize a patient is in stage 3 AKI if on continuous renal replacement therapy (CRRT), creatinine values have increased drastically (either absolutely measured or w.r.t. patient's determined baseline values), or the patient has little to no urine output over prolonged period of time. If any of these conditions can be confirmed from the imputed data, the time step is annotated with a positive event. If both urine and creatinine state can be determined from the imputed data and are assessed to be stable, the event annotation is negative. All other time steps remain with an unknown event annotation.

Liver Failure To derive an early event prediction task related to liver failure we leverage the model for end-stage liver disease (MELD)[39]. The score is a linear combination over the logarithms of creatinine, total bilirubin, and the international normalized ratio (INR) all as measured from blood samples. We choose a threshold of 30 for the score to detect rare and severe forms of liver disease. According to[87] this threshold corresponds to a 50% expected mortality on their study cohort. We minimally impute the three blood sample values by backward and forward filling for at most 1 hour (implying at most 1 step of imputation in either temporal direction) and then compute the score where all three measurements are available. A positive event is annotated for time steps with available score above 30 and a negative event is annotate for time steps with available scores below 30.

Hyperglycemia We incorporate an additional simple early event prediction task, which requires minimal assumptions compared to other tasks relying on imputation schemes and heuristics. Based on work by Mehdizavareh et al.[76] we propose to perform the early event prediction of elevated blood glucose measurements. We do not apply any form of imputation. A positive hyperglycemic event is annotated where a ground truth glucose measurement is available and above 180 mg/dL and a negative event is annotated for measurements below said threshold. The event status remains unknown for time-points without ground truth data.

Sepsis We follow Moor et al.[15] to annotate sepsis onsets based on the international consensus definition sepsis-3 definition[88]. However, different to[15] we only annotate suspected infection on datasets with available body fluid sampling and bacterial culture growth data. We do not apply their antibiotic administration heuristic to annotate suspected infections. In summary, the Sepsis-3 definition requires two components to be observed. The first is a suspected infection typically confirmed by bacterial growth in body fluid samples. The second is a rapid increase in SOFA[14] score. The sepsis onset is annotated where the SOFA score increase has been observed inside a window around the suspected infection.

Decompensation Following Harutyunyan e al.[52] we model severe decompensation as the early event prediction of death. A positive event is death annotated at the end of a deceased patients intensive care stay. A negative event is annotated at the end of a patient's stay if not mortality information is recorded, hence, we impute missing information by assuming discharged alive if no information is available to annotate otherwise.

A.5.2 Early Event Prediction Labels

Early event prediction (EEP) label for a given time step is computed as follows: (1) a detection (positive EEP label) is marked if any time-point in the future within the horizon is annotated as the patient is in a failure state (positive event). (2) a negative EEP label (a stable patient without any upcoming failure state) is annotated only if there is no failure state annotation and there is at least one confirmed stable state (negative event) within the horizon. (3) if there is no data confirmed evidence for either the patient being in failure or being stable within the horizon, no EEP label is assigned and no training and evaluation is performed for that specific time step. We use task-specific and clinically relevant prediction horizons from existing literature (8 hours for circulatory failure[5], hyperglycemic events, and sepsis, 24 hours for decompensation[52] and respiratory failure[8], and 48 hours for kidney failure[20] and liver failure).

More specifically given event annotations at time t $e_t \in \{nan, 0, 1\}$ (where nan corresponds to an unknown diagnostic event state) we compute the early event prediction label y_t for a task specific early event prediction horizon h as follows (Eq. (4)):

$$y_t = \begin{cases} 1 & \exists r \in \{t, \ldots, t+h\} : e_r = 1 \land \neg(e_t = 1) \\ 0 & \exists r \in \{t, \ldots, t+h\} : e_r = 0 \land \forall r \in \{t, \ldots, t+h\} : e_r \neq 1 \land \neg(e_t = 1) \\ nan & \text{otherwise} \end{cases} \tag{4}$$

Sepsis Prior work on detection of sepsis[15] used positive labels both before but also after the annotated sepsis onset. We consider the strict early event prediction of the annotated sepsis onset. We consider only the first annotated sepsis onset for any given patient and apply the prior function in Eq. (4). This yields a positive early event prediction label up to 8 hours before the annotated onset and negative early event prediction labels more than 8 hours before the onset. Any patient without an annotated onset in a source dataset with available body fluid sampling data is considered a negative case and all time steps are annotated using a negative (stable) early event prediction label. Note that this is a prior towards a stable patient in the absence of data, hence the choice to only model the task on datasets with available body fluid sampling data.

A.6 K-anonymization

To align with ethical requirements to obtain approval for this study, we perform k-anonymization w.r.t. to age, gender, height, and weight on the following datasets in the harmonized collection as their published version is only considered de-identified: MIMIC-III[53], MIMIC-IV[38], MIMIC-IV ED[3], eICU[55], Zigong EHR[40], PICdb[33], EHRSHOT[74], and NWICU[73].

Available age, weight, and height are binned to bins of size 5. The procedure then ensures there are at least k patients with the same properties in the dataset. Any patient with a value combination (including missingness of a certain attribute) not present at least k times is removed from the dataset.

A.7 EHRSHOT Data Extraction

To extract valid patient stays from the EHRSHOT dataset, a multi-step heuristic was applied to clean and consolidate visit records from the original long-format table provided by the authors. First, rows corresponding to visit information (i.e., those with omop_table equal to "visit_occurrence" or "visit_detail") were extracted, yielding over one million candidate stays. Stays with zero duration (i.e., where the start and end times were identical) were removed. For each unique combination of patient and visit identifier (visit_id), an overarching stay was defined using the earliest start and latest end times among all visits with the same identifier. If only one visit was found for a given overarching stay, it was retained unchanged; otherwise, a new consolidated stay was created to span the full time window, with a merged list of codes from all constituent entries. Then, for each patient, visits that were fully contained within another were removed, but only after their codes were merged into the encompassing stay. Events associated with these removed visits were not discarded: they were re-mapped to the appropriate merged stay based on timestamp overlap. Finally, temporally adjacent or overlapping visits were further merged into composite stays, with unified time windows and concatenated codes. This process yielded a final set of 118'898 non-overlapping, chronologically ordered stays across 5'900 patients. These validated stays were then used to map all relevant events, including special handling of death records, producing a clean, gap-free timeline of patient visits suitable for downstream clinical modeling and analysis.

B Model

B.1 Architecture

We consider preprocessed irregularly-sampled, multi-variate time-series data, which has been imputed and transformed to a regular hourly time-grid (for details see Supplementary Section A). Given a dataset of patients we obtain a tensor $\mathbf{X} \in \mathbb{R}^{N \times T \times C}$ where we have N patients, T time-steps, and V channels of information after preprocessing the raw data. Furthermore, let $\mathbf{M} \in \mathbb{B}^{N \times T \times C}$ be a binary presence mask[6] indicating whether a measurement in $\mathbf{X}$ is imputed or ground truth measured data.

The proposed neural network architecture can be split into three distinct components: a time-step encoder, a sequence encoder, and time-to-event prediction heads. First, a time-step encoder module $f : \mathbb{R}^m \mapsto \mathbb{R}^d$, which has shared learned parameters across each time-step to extract time-independent variable interactions. It projects each time-step vector of dimensionality C into a d dimensional vector. Second, a parametrized sequence model $g : \mathbb{R}^{T \times d} \mapsto \mathbb{R}^{T \times d}$ processing the sequence of encoded time-steps and extracting temporal patterns in a causal manner, while preserving the model dimension d.

$$\mathbf{H} = g(f(\mathbf{X})) \in \mathbb{R}^{N \times T \times d} \tag{5}$$

The time-step encoder f is implemented as a two layer multi-layer perceptron (MLP) with SiLU activations[89]. Following work by Tomasev et al.[6] the time-step encoder is additionally regularized with an L1 penalty on its weights during training. The sequence encoder g is a vanilla Transformer architecture as introduced by Vaswani et al.[90] but we use the more recent SiLU activations. The feedforward hidden dimension is kept at a constant scaling factor 4 from the model dimension (as proposed by Vaswani et al.). We keep the latent dimension of each attention head at a constant 64 and as such the number of attention heads scale linearly with the model dimension. We use learnable Fourier Positional Encodings[91].

Finally, a collection of K prediction heads for time-to-event (TTE) estimation $\lambda : \mathbb{R}^d \mapsto [0,1]^{T_{surv}}$ at a prediction horizon of T_{surv} discrete steps. Each hazard prediction head λ_k is implemented as a two layer MLP with SiLU activations and a sigmoid to map to a probability estimate.

$$\forall_{k=1}^{K} \mathbf{Y}_k = \lambda_k(\mathbf{H}) \text{ where } \mathbf{Y} \in \mathbb{R}^{N \times K \times T \times T_{surv}} \tag{6}$$

Typically if we retrieve embeddings from the model we consider the output $\mathbf{H}$ of g containing a vector representation of every time-step of a given patient.

B.2 Training Infrastructure

The ICareFM foundation model was trained on the Alps Research Infrastructure[92] at the CSCS (Swiss National Supercomputing Centre). A training run of ICareFM uses a single node of Nvidia's Quad GH200, which tightly couples four Grace Hopper Superchip (GH200) GPUs (and an Arm processors) on a single compute node[93]. Each GPU has 96GB of HBM3 memory for a total of 384 GB of GPU memory paired with a total of 480GB LPDDR5 ECC memory attached to the Arm processors. Each Arm Neoverse V2 CPU has 72 processing cores for a total of 288 cores on a single compute node.

	MIMIC-III	MIMIC-IV	eICU	NWICU	HiRID	UMCdb	SICdb	PICdb	Zigong EHR
Demographic information									
Country	USA	USA	USA	USA	Switzerland	Netherlands	Austria	China	China
Multi-Center	No	No	Yes	Yes	No	No	No	No	No
Years	2001 - 2012	2008 - 2022	2014 - 2015	2020 - 2022	2008 - 2016	2003 - 2016	2013 - 2021	2010 - 2018	2019 - 2020
Num. Stays	53,632	91,039	183,581	24,943	33,547	22,883	24,510	13,268	2,525
Gender (% Female)	43.6	44.0	45.9	43.1	35.7	34.0	36.6	42.6	39.2
Gender (% Male)	56.4	56.0	54.1	56.9	64.3	63.9	63.4	57.4	60.8
Age	58.9 (24.6)	63.2 (16.6)	63.3 (17.0)	63.7 (16.2)	63.5 (15.2)	62.6 (16.5)	67.1 (14.6)	2.2 (3.9)	66.0 (16.5)
LoS (Median hours (IQR))	53.0 (29.0, 114.0)	48.0 (27.0, 94.0)	41.0 (22.0, 75.0)	46.0 (25.0, 92.0)	24.0 (20.0, 52.0)	26.0 (21.0, 90.0)	38.0 (24.0, 78.0)	68.0 (23.0, 214.0)	114.0 (65.0, 259.0)
Patient stay prevalence (num. stays (%))									
Mortality	4,123 (7.7)	8,244 (9.1)	9,529 (5.2)	973 (3.9)	2,001 (6.0)	2,242 (9.8)	1,334 (5.4)	888 (6.7)	533 (21.1)
Circ. Failure	7,720 (14.4)	18,644 (20.5)	14,356 (7.8)	1,919 (7.7)	10,392 (31.0)	5,657 (24.7)	7,665 (31.3)	2,357 (17.8)	833 (33.0)
Sepsis-3	17,143 (32.0)	31,485 (34.6)	-	-	-	5,357 (23.4)	-	1,872 (14.1)	944 (37.4)
Hyperglyc.	15,350 (28.6)	27,974 (30.7)	64,257 (35.0)	9,018 (36.2)	15,226 (45.4)	12,833 (56.1)	7,746 (31.6)	5,185 (39.1)	1,365 (54.1)
Resp. Failure	5,594 (10.4)	13,597 (14.9)	11,340 (6.2)	-	3,970 (11.8)	4,794 (21.0)	4,616 (18.8)	-	-
Kidney Failure	4,308 (8.0)	10,771 (11.8)	19,050 (10.4)	-	1,375 (4.1)	1,827 (8.0)	1,641 (6.7)	133 (1.0)	-
Liver Failure	2,249 (4.2)	4,652 (5.1)	3,224 (1.8)	-	472 (1.4)	154 (0.7)	-	-	99 (3.9)
Time-step label prevalence (positive %, (negative %, unknown %))									
Decomp. 24h	1.9 (98.1, 0.0)	2.6 (97.4, 0.0)	1.7 (98.3, 0.0)	1.1 (98.9, 0.0)	2.4 (97.6, 0.0)	2.8 (97.2, 0.0)	1.8 (98.2, 0.0)	1.3 (98.7, 0.0)	3.4 (96.6, 0.0)
Circ. Fail. 8h	1.4 (35.3, 63.2)	2.5 (46.6, 50.9)	0.8 (21.5, 77.7)	0.8 (9.8, 89.4)	3.7 (63.3, 33.0)	2.0 (43.5, 54.5)	4.2 (68.0, 27.8)	1.4 (1.9, 96.7)	3.5 (0.5, 96.0)
Sepsis-3 8h	1.4 (54.5, 44.1)	1.6 (54.0, 44.4)	-	-	-	0.6 (49.6, 49.8)	-	0.5 (84.8, 14.6)	1.4 (87.2, 11.4)
Hyperglyc. 8h	4.9 (34.9, 60.2)	7.8 (39.9, 52.3)	13.6 (49.6, 36.8)	13.5 (43.9, 42.7)	14.9 (55.8, 29.4)	12.1 (71.4, 16.5)	8.7 (78.3, 13.0)	4.3 (32.2, 63.6)	6.9 (15.6, 77.5)
Resp. Fail. 24h	2.9 (20.7, 76.4)	5.1 (23.1, 71.8)	2.2 (17.5, 80.3)	-	5.0 (36.4, 58.6)	7.5 (47.5, 45.0)	3.8 (27.7, 68.5)	-	-
Kidney Fail. 48h	2.4 (58.9, 38.7)	3.9 (78.1, 18.0)	4.6 (58.8, 36.6)	-	2.5 (77.9, 19.6)	2.6 (82.7, 14.7)	2.6 (77.0, 20.4)	0.4 (11.0, 88.6)	-
Liver Fail. 48h	3.1 (13.2, 83.6)	4.3 (22.9, 72.8)	1.4 (8.7, 89.9)	-	1.5 (18.0, 80.5)	0.4 (16.6, 83.0)	-	-	1.2 (24.6, 74.1)
Data density (Median % (IQR %))									
Vital Density	26.9 (22.6, 30.9)	27.2 (23.9, 31.0)	23.3 (20.3, 28.1)	16.2 (0.0, 20.1)	34.0 (27.9, 43.2)	35.2 (26.9, 43.4)	26.0 (22.3, 29.6)	0.7 (0.3, 1.8)	0.7 (0.4, 0.9)
Lab Density	2.7 (1.9, 3.7)	3.1 (2.0, 4.6)	2.0 (1.4, 2.9)	2.0 (1.4, 2.8)	5.5 (3.0, 7.6)	8.8 (5.7, 11.5)	8.9 (7.2, 10.5)	5.0 (2.2, 8.1)	2.2 (1.5, 3.0)
Vital Density adj.	26.9 (22.6, 30.9)	28.2 (24.8, 32.1)	23.3 (20.3, 28.1)	52.2 (0.0, 64.7)	36.5 (30.0, 46.4)	37.8 (28.9, 46.7)	44.4 (38.1, 50.6)	1.4 (0.6, 3.5)	1.5 (1.0, 2.0)
Lab Density adj.	2.7 (2.0, 3.8)	3.3 (2.1, 4.8)	2.2 (1.4, 3.1)	2.4 (1.7, 3.5)	6.4 (3.5, 8.7)	9.0 (5.9, 11.7)	10.1 (8.1, 11.9)	5.4 (2.4, 8.7)	2.8 (1.9, 3.7)

Supplementary Table 1: Dataset statistics for the core training support set showing demographic information, patient prevalences for different downstream tasks, time-step prevalances for specific early event prediction labels of each downstream task, data densities for vitals and laboratory tests (regularly and irregularly measured variables respectively). Adjusted densities (adj.) exclude variables that are never observed for a specific dataset. Statistics for further harmonized data sources are summarized in Supplementary Table 2.

	EHRSHOT	MIMIC-IV ED
Demographic information		
Country	USA	USA
Multi-Center	No	No
Years	1990 - 2023	2011 - 2019
Num. Stays	42,562	149,546
Gender (% Female)	-	53.8
Gender (% Male)	-	46.2
Age	-	61.1 (16.8)
LoS (Median hours (IQR))	49.0 (24.0, 84)	10.0 (7.0, 15.0)
Patient stay prevalence (num. stays (%))		
Mortality	233 (0.5)	48 (0.0)
Circ. Failure	535 (1.2)	-
Sepsis-3	-	-
Hyperglyc.	4,440 (10.4)	-
Resp. Failure	424 (1.0)	-
Kidney Failure	-	-
Liver Failure	-	-
Time-step label prevalence (positive %, (negative %, unknown %))		
Decomp. 24h	0.2 (99.8, 0.0)	0.0 (100.0, 0.0)
Circ. Fail. 8h	0.3 (9.5, 90.2)	-
Sepsis-3 8h	-	-
Hyperglyc. 8h	4.8 (18.1, 77.1)	-
Resp. Fail. 24h	0.5 (2.0, 97.5)	-
Kidney Fail. 48h	-	-
Liver Fail. 48h	-	-
Data density (Median % (IQR %))		
Vital Density	1.9 (1.0, 3.4)	10.3 (8.0, 13.3)
Lab Density	0.6 (0.2, 1.3)	0.0 (0.0, 0.0)
Vital Density adj.	3.8 (1.8, 6.7)	42.9 (33.3, 55.1)
Lab Density adj.	0.7 (0.2, 1.4)	0.0 (0.0, 0.0)

Supplementary Table 2: Dataset statistics for further harmonized datasets (complementing Supplementary Table 1) showing demographic information, patient prevalences across the collection of considered downstream tasks, time-step prevalances for the specific early event prediction labels of each downstream task, and data densities for vitals (regularly measured) and laboratory test (irregularly measured) variables. Adjusted densities do not consider variables, which are never observed for a specific dataset.