

Pretraining on Adult ICU Data Accelerates Data Efficient AI Model Development in Critically Ill Children

Yuxuan Liu

Imperial College London

Jake Ormond

Imperial College London

Emma C. Alexander

Imperial College Healthcare NHS Trust

Alexander Philip Yehuda Brown

University College London

Tuan Chen Aw

Royal Brompton Hospital

Angela Aramburo

Imperial College London

Samiran Ray

Great Ormond Street Hospital for Children

Padmanabhan Ramnarayan

p.ramnarayan@imperial.ac.uk

Imperial College Healthcare NHS Trust

A. Aldo Faisal

Imperial College London

Article

Keywords: Transfer Learning, Adult-to-paediatric Transfer, Paediatric Critical Care, Machine Learning, Clinical Decision Support

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11365696/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Electronic health record data from paediatric intensive care units (PICUs) are scarce and rarely freely accessible. We trained task-specific models on adult ICU data (MIMIC-IV; 71,965 stays) and evaluated them in three UK PICUs (17,765 stays) across six clinical tasks. We compared zero-shot deployment without local data, target-only training from scratch, and fine-tuning of the adult model at local sample sizes ranging from 100 stays to the full cohort, using nested cross-validation. Zero-shot deployment underperformed full-cohort target-only training in all 17 feasible task–site comparisons. At 100 PICU stays, all 17 point estimates favoured fine-tuning and 16 paired 95% confidence intervals excluded zero. At the full cohort, this advantage persisted in 14 of 17 comparisons but was smaller. Adult ICU pretraining is a promising strategy for paediatric AI model development, most valuable where local data are scarcest—but requires local adaptation and recalibration rather than direct deployment.

Introduction

Clinical prediction models are difficult to develop locally in paediatric intensive care. Paediatric intensive care units (PICUs) are concentrated in specialist centres^{1,2} and admit far fewer patients than adult critical care: UK national audit reports recorded 224,748 adult ICU admissions across 263 units in 2018–2019, compared with 18,498 PICU admissions in 2023.^{3,4} Large, accessible adult ICU datasets and shared benchmarks have meanwhile enabled reproducible model development at scale;^{5–7} MIMIC-IV alone provides more than 70,000 adult ICU admissions.⁸ Freely accessible paediatric critical-care datasets are rare and remain much smaller; the single-centre PIC database, located in Zhejiang Province, People's Republic of China, contains approximately 13,500 admissions, and does not include high-frequency bedside monitoring data.⁹ The US-based PICU Data Collaborative and emerging multisite networks offer an important complementary route to larger paediatric datasets,¹⁰ but require substantial multi-institutional coordination and data transfer; individual PICUs may still be unable to assemble sufficiently large cohorts for robust, generalisable modelling.

Transfer learning—initialising a model on a data-rich source domain and adapting it to a data-poor target domain—is an obvious potential solution and has reduced labelled-data requirements in computer vision and natural language processing.^{11,12} Its value for structured electronic health record time-series data is less established. Adult-to-paediatric transfer may be complicated by age-dependent differences in vital signs and laboratory values, for example serum creatinine.^{13,14} The few studies that examined transfer between adult and paediatric populations report mixed results. More importantly, most of them examined only a single outcome or centre, and differed in their tasks, adaptation procedures, comparators and institutional settings.^{15–19} The value of adult pretraining across multiple tasks, independent PICUs and varying target-data volumes therefore remains unclear.

In this study, we evaluated transfer learning from MIMIC-IV adult ICU data to three UK PICUs across six clinical tasks (Fig. 1). On features harmonised across all four cohorts, we trained a task-specific gated

recurrent unit (GRU) on the adult data and compared three strategies at each site: *zero-shot deployment* without local training, *target-only training* from scratch on local data, and *fine-tuning* the adult model on local data. Each strategy is evaluated at local training sizes ranging from 100 ICU stays up to the full cohort, so that the value of adult pretraining is tracked as paediatric data accumulate. Overall, we investigated whether adult pretraining, used through ordinary fine-tuning, lowers the amount of paediatric data needed for competitive performance, and whether this varies by task, site and data volume. Diagnostic comparisons then tested whether zero-shot deficits are addressed by paediatric preprocessing statistics, a locally fitted output layer, or full network adaptation.

Results

Study cohorts and clinical tasks

MIMIC-IV contributed 71,965 adult ICU stays (median age 65 years, 55.8% male). The three PICU cohorts contributed 9,400 (Unit A), 3,594 (Unit B) and 4,771 (Unit C) stays, totalling 17,765 paediatric stays with median ages of 0.7, 2.1 and 0.7 years; physiology and case mix differed substantially from the adult cohort (Table 1).

Table 1

Cohort characteristics. Stay-level statistics for the adult source cohort (MIMIC-IV) and the three paediatric intensive care units; continuous variables as median [IQR]; categorical as count or percentage. Ventilation rate is the percentage of stays with at least one hour of mechanical ventilation (Unit A “—”: the harmonised dataset carries no derived ventilation indicator). Physiological values (heart rate, respiratory rate, MAP, creatinine) are each stay’s mean over its observed timesteps, summarised across stays as median [IQR].

Characteristic	MIMIC-IV	Unit A	Unit B	Unit C
Setting	Adult ICU	Paediatric ICU	Paediatric ICU	Paediatric ICU
Case mix	Medical–surgical	General	General	Cardiorespiratory
Country	USA	UK	UK	UK
Study years	2008–2019	2013–2022	2013–2022	2013–2022
<i>Demographics</i>				
ICU stays, n	71,965	9,400	3,594	4,771
Age, median (IQR), y	65 (53–76)	0.7 (0.1–4.5)	2.1 (0.5–8.1)	0.7 (0.1–4.2)
Male, n (%)	40,157 (55.8)	5,345 (56.9)	2,009 (55.9)	2,639 (55.3)
<i>Clinical course and severity</i>				
ICU mortality, %	7.2	5.9	4.0	3.5
Ventilation rate, %	38.0	—	59.6	63.0
Length of stay, median (IQR), h	47 (27–91)	85 (41–180)	93 (41–169)	75 (34–167)
<i>Physiology (per-stay mean)</i>				
Heart rate, bpm	83 (74–94)	128 (113–141)	122 (107–135)	126 (110–139)
Respiratory rate, /min	19 (17–21)	30 (23–38)	26 (21–33)	28 (23–34)
Mean arterial pressure, mmHg	78 (71–86)	67 (57–77)	64 (56–72)	61 (54–67)
Creatinine, mg/dL	0.93 (0.70–1.40)	0.34 (0.25–0.52)	0.44 (0.39–0.55)	0.37 (0.26–0.55)

Zero-shot performance across tasks and sites

Deployed directly to each PICU without local adaptation, the adult-trained GRU was inferior to full-cohort target-only training across the six tasks (Fig. 2). Zero-shot AKI AUROC ranged from 0.60 to 0.70 against 0.82–0.87 for target-only training; mortality showed a shortfall (0.64–0.75 vs. 0.80–0.87) and similar findings were seen in sepsis at all three sites (0.62–0.65 vs. 0.76–0.81). Zero-shot kidney-function

prediction was likewise inferior (MAE 0.15–0.24 vs. 0.10–0.14 mg/dL). Across all 17 task–site cells, the paired zero-shot deficit relative to full-cohort target-only training excluded zero using outer-fold inference (Supplementary Table S3).

Fine-tuning benefit across paediatric data volumes

Fine-tuning the adult model on local PICU data improved on target-only training at matched sample sizes. We summarise the benefit as the paired contrast FT – Tgt across all 17 primary task–site comparisons (five tasks at three sites each, extubation readiness at two). At $n = 100$ stays, all 17 point estimates favoured fine-tuning, 16 of the paired 95% confidence intervals excluded zero (the exception being AKI at Unit B), and none favoured training from scratch (Table 2). Among the classification tasks, the largest $n = 100$ gains were for mortality (AUROC + 0.14 to + 0.17), with smaller but positive gains for sepsis, AKI and extubation readiness.

Table 2

Paired fine-tuning benefit in the low-data and full-cohort regimes. For each task–site comparison, the table reports the mean paired contrast between fine-tuning (FT) and target-only training (Tgt) across the five outer folds, with a t_4 95% confidence interval. Classification contrasts are AUROC differences (FT – Tgt); regression contrasts are MAE reductions (Tgt – FT), so positive values consistently favour fine-tuning. Length-of-stay MAE is in hours and kidney-function MAE in mg/dL. Bold contrasts have intervals that exclude zero. Zero-shot–Tgt contrasts are reported in Supplementary Table S3; absolute per-method estimates are in Supplementary Table S5.

Task	Site	n = 100 [95% CI]	Full cohort [95% CI]
<i>Classification (AUROC difference)</i>			
Mortality	Unit A	+ 0.139 [+ 0.080, +0.197]	+ 0.010 [+ 0.001, +0.019]
Mortality	Unit B	+ 0.169 [+ 0.116, +0.222]	+ 0.021 [+ 0.006, +0.036]
Mortality	Unit C	+ 0.158 [+ 0.119, +0.197]	+ 0.010 [– 0.009, +0.029]
AKI	Unit A	+ 0.135 [+ 0.051, +0.219]	+ 0.034 [+ 0.017, +0.052]
AKI	Unit B	+ 0.031 [– 0.030, +0.091]	+ 0.014 [+ 0.001, +0.028]
AKI	Unit C	+ 0.046 [+ 0.014, +0.078]	+ 0.011 [– 0.001, +0.024]
Sepsis	Unit A	+ 0.053 [+ 0.004, +0.102]	+ 0.0038 [+ 0.0003, +0.0072]
Sepsis	Unit B	+ 0.046 [+ 0.013, +0.080]	+ 0.002 [– 0.008, +0.012]
Sepsis	Unit C	+ 0.072 [+ 0.036, +0.107]	+ 0.019 [+ 0.012, +0.026]
Extubation readiness	Unit B	+ 0.054 [+ 0.027, +0.082]	+ 0.014 [+ 0.003, +0.025]
Extubation readiness	Unit C	+ 0.029 [+ 0.004, +0.053]	+ 0.023 [+ 0.008, +0.037]
<i>Length of stay (MAE reduction)</i>			
LOS	Unit A	+ 9.21 [+ 7.88, +10.54]	+ 1.79 [+ 1.26, +2.33]
LOS	Unit B	+ 9.11 [+ 8.16, +10.06]	+ 4.69 [+ 3.97, +5.41]
LOS	Unit C	+ 10.51 [+ 8.39, +12.63]	+ 2.52 [+ 2.14, +2.91]
<i>Kidney function (MAE reduction, mg/dL)</i>			
Kidney fn.	Unit A	+ 0.120 [+ 0.100, +0.139]	+ 0.017 [+ 0.012, +0.023]
Kidney fn.	Unit B	+ 0.052 [+ 0.040, +0.064]	+ 0.023 [+ 0.019, +0.028]
Kidney fn.	Unit C	+ 0.057 [+ 0.043, +0.070]	+ 0.021 [+ 0.016, +0.026]

The benefit diminished as local data accumulated (Fig. 2a). Across all 17 task–site comparisons, the paired fine-tuning benefit was smaller at the full cohort than at $n = 100$ (Table 2). For example, mortality AUROC gains narrowed from 0.139–0.169 to 0.010–0.021, while length-of-stay MAE reductions decreased from 9.11–10.51 to 1.79–4.69 hours. Despite these smaller effect sizes, fine-tuning remained

favoured in all 17 full-cohort comparisons, with 14 paired intervals excluding zero. The pooled-PICU analysis showed the same pattern across all six tasks: from $n = 100$ to the full pooled cohort, the mortality gain decreased from 0.141 to 0.004 AUROC and the length-of-stay MAE reduction from 8.62 to 1.50 hours. Evaluated separately at each PICU, full-cohort pooled target-only models matched or exceeded the corresponding site-specific target-only point estimate in 15 of 17 comparisons; both exceptions were extubation readiness (Supplementary Table S4).

As a complementary absolute benchmark, we compared fine-tuning at each data budget with target-only training on all available local data. Fine-tuning matched or exceeded the full-cohort target-only point estimate within 500 stays in 2 of 17 task–site comparisons and within 1,000 stays in 8 of 17 (Supplementary Table S5).

Precision–recall and calibration

On a precision–recall basis, which complements AUROC by reflecting positive predictive value under class imbalance,²⁰ fine-tuning matched or exceeded target-only training in 10 of 11 classification task–site cells, the exception being AKI at Unit C (Fig. 2b; Supplementary Table S7). AKI nonetheless remained a low-prevalence, difficult task, with fine-tuned AUPRCs of 0.06–0.08 against a 0.6–1.0% event rate. In supplementary calibration analyses, uncalibrated zero-shot scores substantially over-predicted absolute risk (mean predicted risk 32.5–65.4% versus event rates of 0.6–18.5%), whereas intercept-only recalibration using local validation predictions aligned the mean risk from both locally trained methods with the observed event rate (Supplementary Table S8).

Diagnostic analyses of adaptation and feature reliance

To characterise where local adaptation acts, we compared source-normalised zero-shot deployment, zero-shot deployment using preprocessing statistics from the other two PICUs, linear probing with the adult encoder frozen, and full fine-tuning (Fig. 3a). External-PICU normalisation produced limited and task-dependent changes: its largest gains were for mortality at Unit B and AKI at Unit A, but it did not generally close the gap to local adaptation. Linear probing uses labelled target data and therefore represents supervised local adaptation rather than zero-shot transfer; it recovered 65–92% of the zero-shot-to-full-fine-tuning AUROC gap across the 11 classification cells. The full-fine-tuning point estimate remained higher in every cell (Supplementary Table S9).

In the leave-one-feature-out permutation-importance analysis, creatinine ranked first for AKI and kidney function, and FiO₂ ranked first for mortality, sepsis, length of stay and extubation readiness (Fig. 3b).

Discussion

In the three PICUs studied, adult ICU pretraining was reusable after local adaptation but was not sufficient on its own. Zero-shot deployment underperformed full-cohort target-only training in all 17 task–site comparisons, with paired intervals excluding zero throughout; its raw classification scores also substantially over-predicted absolute risk. Conversely, at the full local cohort, all 17 point estimates favoured fine-tuning over target-only training, with 14 paired intervals excluding zero. Adult pretraining therefore transferred as an initialisation: it improved on local training but did not remove the need for target-population data.

Previous studies have reported mixed results under different transfer settings. Adult-source transfer remained inferior to paediatric target-only training for ICU mortality even after domain adaptation,¹⁵ and cross-age deployment reduced discrimination and calibration in younger children.¹⁶ In contrast, adult-to-paediatric transfer outperformed direct paediatric training in 200 children with sepsis receiving continuous renal replacement therapy,¹⁷ while adult-only self-supervised pretraining and task-specific training outperformed paediatric logistic regression across 11 outcomes within one health system.¹⁹ In the reverse direction, an age-normalised paediatric sepsis model retained similar performance in adult data.¹⁸ Because these studies used different tasks, comparators, transfer procedures and amounts of target supervision, they do not resolve whether poor zero-shot performance also implies that adult pretraining offers no benefit after local adaptation. By evaluating both within the same framework, we show that these are distinct: the adult predictors were not directly portable, but their weights remained useful initialisations for paediatric training.

The value of that starting point changed as paediatric data accumulated. At $n = 100$ stays, all 17 point estimates favoured fine-tuning over target-only training on the same data; the advantage narrowed as local cohorts grew, although it remained positive at the largest sample sizes studied. Against target-only training on all available local data, fine-tuning reached or exceeded the corresponding point estimate within 500 stays in 2 of 17 comparisons and within 1,000 stays in 8 of 17. The sample size at which fine-tuning matched the full-cohort target-only point estimate therefore varied across tasks and sites. The principal value of adult pretraining was greater sample efficiency during early local model development. Future studies should therefore examine how performance changes as more local data are added and determine, for each task and hospital, how many paediatric stays are needed before an adapted model can be taken forward for prospective testing.

Pooling paediatric development data offers a complementary route, although it involves multi-institutional collaboration and data exchange as well as considerable resources. Full-cohort pooled target-only training matched or exceeded the corresponding site-specific point estimate in 15 of 17 comparisons, while fine-tuning the adult model on the pooled paediatric cohort retained more favourable point estimates across all six tasks, although the gaps were smaller than at $n = 100$. Adult data can therefore provide an initial representation, while multisite paediatric data provide broader evidence about the target population and tasks. The value of combining these routes, particularly when evaluated at an

unseen PICU, remains to be established directly but has the potential to accelerate AI model development in children.

The diagnostic analyses further clarify what was required to make the adult models useful. Replacing adult preprocessing statistics with statistics from the other two PICUs produced limited, task-dependent changes and left the zero-shot deficit largely unresolved. With the adult encoder frozen, fitting a new output layer on paediatric labels recovered 65–92% of the zero-shot-to-full-fine-tuning AUROC gap across the 11 classification cells; full fine-tuning had a higher point estimate in every cell. Regression results were task dependent. We found that linear probing was worse than target-only training for length of stay but better for kidney-function prediction by point estimate at all three sites. Thus, supervised remapping of the frozen representation recovered much of the classification gap, while the adaptation required varied by task. Clinically plausible feature rankings provided a task-relevance check on the adapted models, while leaving the origin of those dependencies unresolved.

There are several limitations relevant to this study. The controlled comparison used 32 shared inputs, excluding paediatric-specific variables and wider schema differences; operational AKI and sepsis definitions also differed between adult and paediatric pipelines. Comparing plain fine-tuning with the same GRU trained from scratch isolates the value of adult initialisation but does not identify an optimal adaptation algorithm or model family. The three target cohorts were drawn from UK PICUs of limited size, and even their pooled sample could not establish whether the marginal benefit of adult pretraining would disappear at greater scale. The site-level pooled comparison had no paired uncertainty, included each evaluated site's own development data and did not test transfer to an unseen PICU. ICU-stay-level splitting may have placed repeat admissions from the same child in different folds, and retrospective performance does not establish useful decision thresholds, workflow improvement or patient benefit.

Within these boundaries, adult pretraining improved sample efficiency during early local model development but did not remove the need for target-population evidence. Future work should determine how adult initialisation and cross-PICU collaboration can be combined and validated at previously unseen hospitals.

Methods

Data sources and harmonisation

The adult source cohort was MIMIC-IV version 2.2,⁸ de-identified ICU data from Beth Israel Deaconess Medical Center (Boston, USA), restricted to patients aged ≥ 18 years with ICU stays of at least 6 hours; 71,965 stays were retained after preprocessing. The paediatric target cohorts were three UK PICUs: Unit A and Unit B, general PICUs, and Unit C, a cardiorespiratory-specialist PICU, each including patients aged 0–18 years admitted for at least 6 hours (Table 1). Extraction and preprocessing followed the YAIB framework⁵ with paediatric-specific modifications for age-stratified laboratory reference ranges (Clinical prediction tasks below).

Of 33 candidate dynamic variables available at the intersection of all four harmonised datasets, four (fibrinogen, lymphocyte percentage, neutrophil percentage, platelet count) were excluded for inconsistent units or measurement scales, leaving 29 dynamic and 3 static features (age, sex, weight): 32 model inputs (Supplementary Table S1). The derived ventilation-episode indicator used to define extubation was separate from these inputs and was unavailable at Unit A.

PICU source tables were harmonised to a common schema, linked at the ICU-stay level and aggregated into hourly intervals. After required-field validation, we excluded stays with fewer than four observed dynamic measurements or more than 12 consecutive hourly rows in which all dynamic variables were missing.

Each ICU stay was represented at hourly resolution. For each variable we added a binary missingness indicator, forward-filled the last observation and zero-filled remaining gaps. Continuous features were standardised; categorical static variables were mode-imputed and label-encoded. All preprocessing statistics were fitted on the training split only. Target-only and fine-tuned models used target-training statistics, whereas zero-shot transfer used the source-cohort statistics unchanged.

Clinical prediction tasks

Six tasks were evaluated: ICU mortality at a 24-hour landmark, remaining length of stay (LOS), onset of acute kidney injury (AKI), sepsis onset, kidney function (the h24–48 median serum creatinine), and extubation readiness. Five follow YAIB benchmark definitions with target-specific adaptations; extubation readiness follows a previously published benchmark definition.^{5,21} Per-cohort prevalences are in Supplementary Table S2.

Given the information available across cohorts, we derived study-specific paediatric labels for AKI and sepsis that differ from the adult definitions: the adult labels use the YAIB KDIGO and Sepsis-3 definitions,^{22,23} whereas the paediatric AKI labels index the creatinine baseline to an age- and sex-specific normative reference rather than a rolling patient minimum, and the paediatric sepsis labels use an organ-dysfunction proxy adapted from the Phoenix criteria²⁴ without a separate suspected-infection criterion. Reported AKI and sepsis prevalences are proportions of positive hourly prediction instances, not stay-level incidence.

Task-specific eligibility criteria were applied within each dataset's base analytic cohort:

1. **Mortality:** ICU mortality among patients alive and remaining in PICU at the 24-hour landmark (died vs. survived during the ICU stay), predicted once per stay from the first 24 h of data. Stays with total length of stay <30 h were excluded. Metric: AUROC.
2. **Length of stay (LOS):** Continuous regression of remaining ICU hours (capped at 168 h, i.e. 7 days) at each timestep, in hours. Metric: mean absolute error (MAE).

3. **Acute kidney injury (AKI):** Hourly binary prediction of imminent creatinine-based KDIGO stage ≥ 1 onset over the first 168 h; as for sepsis, the 12 hours immediately preceding onset are labelled positive. Urine-output criteria were not included because they were not harmonised across cohorts. KDIGO staging is defined relative to a creatinine baseline, which the two dataset pipelines construct differently: the adult (MIMIC-IV) labels use each patient's rolling-minimum creatinine over a trailing 48-hour ($+0.3$ mg/dL absolute criterion) to 7-day ($1.5\times$ relative criterion) window, following the standard ICU KDIGO implementation;^{22,25} the paediatric labels instead use a fixed normative reference creatinine indexed by age and, in adolescence, sex.¹⁴ The 0.6–1.0% prevalences in Supplementary Table S2 are proportions of positive hourly prediction instances, not stay-level cumulative AKI incidence. Metric: AUROC.
4. **Sepsis:** Hourly binary prediction of imminent sepsis onset over the first 168 h. For adults the label is the Sepsis-3 definition (a rise in SOFA ≥ 2 with suspected infection). For children we use an operational organ-dysfunction proxy informed by the Phoenix paediatric sepsis criteria,²⁴ adapted to the variables available across cohorts: points accrue from (i) a respiratory component using pSOFA²⁶ SpO₂/FiO₂ thresholds (292, 264, 220 and 148; evaluated only when SpO₂ < 97%, to avoid the pulse-oximetry ceiling effect), and (ii) a cardiovascular component scoring vasoactive/inotrope use, or an age-thresholded hypotension as defined in PALS²⁷ (systolic blood pressure below 60, 70, $70+2\times\text{age}[\text{yr}]$ and 90 mmHg for ages ≤ 28 d, ≤ 1 y, 1–10 y and >10 y respectively). Sepsis onset is the first hour at which the total score reaches 2; the paediatric proxy is organ-dysfunction-based and does not impose a separate suspected-infection criterion. For both domains the 12 hours immediately preceding onset are labelled positive; timesteps from onset onward are not used, and stays with fewer than 12 h of pre-onset lead-time are excluded. Metric: AUROC.
5. **Kidney function:** A static per-stay regression predicting a single downstream serum-creatinine value (the h24–48 median) from the first 24 h of features. Metric: MAE, reported in clinical units (mg/dL).
6. **Extubation readiness:** Readiness during the first invasive-ventilation episode,²¹ predicted at each hour within the episode and positive within 12 h of a *successful* extubation (no reintubation within 48 h). Episode-level exclusions removed ventilation shorter than 12 h or longer than 28 days, death on the ventilator, and post-extubation death within 48 h. This operational proxy reflects physiological suitability, the decision to attempt extubation and local ventilation practice. Evaluated at Unit B and Unit C only (Unit A lacks a derived ventilation indicator). Metric: AUROC.

Models and transfer conditions

Pretraining was task specific. For each task, a separate gated recurrent unit (GRU)²⁸ sequence encoder with a task-specific linear head was trained on the corresponding adult task.

The head was applied at every timestep for hourly tasks and at the prediction timestep for static tasks. Bayesian hyperparameter search selected a two-layer GRU for AKI and three-layer GRUs for the other tasks, with hidden dimension 32–86, dropout 0.15–0.24 and Adam learning rate 3×10^{-5} to 3×10^{-4} . All

models used Adam (weight decay 10^{-5}), batch size 128, FP32 precision and up to 100 epochs with early stopping. Classification used class-balanced focal loss ($\gamma=2$);²⁹ regression used mean squared error.

Three conditions were compared for every task–hospital pair: *zero-shot*, the source-trained model deployed without adaptation; *target-only*, an otherwise identical GRU trained from scratch on PICU data, the reference for whether transfer helps; and *fine-tuning*, all parameters updated on PICU data from the source initialisation with early stopping on validation loss. Because the aim was to isolate the value of adult initialisation rather than to find the best model per task, we did not benchmark other model families.

Evaluation and statistical analysis

Performance was estimated by outcome-stratified nested cross-validation: an outer 5-fold split defined five disjoint held-out test sets, and an inner 5-fold split within each outer training partition produced the train/validation sets, giving 25 fits over five distinct test folds. Splits were made at the ICU-stay level with a fixed random seed. Point estimates are the mean of the 25 held-out test metrics; to retain the outer test fold as the unit of uncertainty, inner-fit metrics were averaged within each outer fold before computing standard errors across the five folds.

Paired contrasts were computed within each outer fold after averaging its five inner-fit metrics and are reported as mean differences with two-sided t_4 95% confidence intervals; intervals excluding zero were considered statistically supported. Classification contrasts are AUROC differences and regression contrasts are MAE reductions, so positive values favour fine-tuning; area under the precision–recall curve (AUPRC), which is informative under class imbalance, was additionally reported for classification.²⁰

Few-shot sampling and pooled-PICU analysis

Sample efficiency was assessed at target-development sizes $n \in \{100, 200, 500, 1,000, 2,000, 3,000\}$ and at the full cohort, where n counts unique PICU stays sampled without replacement from the outer development partition within each outer fold (outcome-stratified for classification). The same stays and held-out test fold were used for target-only training and fine-tuning, so every task–site– n comparison is paired by construction. The matched-size target-only result is therefore the reference for whether adult initialisation improves on local training alone.

The n stays formed the complete development pool for the five inner train/validation splits; n therefore denotes the number of unique paediatric ICU stays available across the five fits, not the number available to each fit separately. For $n=\text{all}$, all stays in the corresponding outer development partition were used. A fixed sample size was evaluated only where it was available in every outer fold. Both methods used the per-task architecture and optimisation settings described in Models and transfer conditions above.

As a scale-sensitivity analysis, we repeated target-only training and fine-tuning after pooling the three PICU development cohorts. We retained the same nested cross-validation design and estimators, but calculated performance as a micro metric over the combined held-out observations; n denotes the total number of pooled paediatric development stays. The pooled zero-shot reference was calculated by combining the source model's held-out predictions across the available PICUs before computing the metric. To assess whether pooled paediatric training also benefited individual centres, we evaluated each full-cohort pooled target-only model separately on the held-out observations from each PICU and compared its point estimate with that of the corresponding site-specific target-only model. This site-level comparison was descriptive; no paired uncertainty interval was estimated.

Calibration

For locally trained classification models, we fitted an intercept-only logistic recalibration model to each inner model's validation predictions and applied it unchanged to the held-out outer-test predictions. This adjusts the overall risk level without using test labels but leaves ranking and calibration slope unchanged.³⁰ We report event rate, mean predicted risk and Brier score. Zero-shot predictions remained uncalibrated because target-domain recalibration would no longer represent target-free deployment. Calibration metrics use the same mean-of-25 convention as the discrimination metrics.

Representation probing and feature reliance

To identify where adaptation acted, we compared source-normalised zero-shot deployment, zero-shot deployment normalised using the other two PICUs without evaluated-site data, linear probing with the adult encoder frozen and a new target-trained output layer, and full fine-tuning.

Feature reliance was assessed by leave-one-feature-out permutation importance in fine-tuned models across the 25 folds, excluding missingness indicators and pooling rankings over sites and folds.

Both diagnostic analyses used the primary partitions and estimators.

Declarations

Ethics approval and consent

Patient consent for publication: Not applicable.

Ethics approval: Approval for the secondary use of routinely collected clinical data for research without individual patient consent was provided in the UK through the Health Research Authority and Research Ethics Committee (reference 23/HRA/0416), with confirmation of capacity and capability from each participating NHS Trust. The collection of patient information and creation of the MIMIC-IV research

resource were reviewed by the Institutional Review Board at Beth Israel Deaconess Medical Center, which granted a waiver of informed consent and approved the data-sharing initiative.

Data availability

MIMIC-IV is available via PhysioNet (<https://physionet.org>) to credentialed users under a data use agreement. The three UK paediatric ICU datasets (Unit A, Unit B, Unit C) are governed by data use agreements and cannot be shared publicly; they are available to qualified researchers upon reasonable request, subject to appropriate ethical and information-governance approval.

Code availability

The analysis code is available from the corresponding author upon reasonable request.

Acknowledgements

This work was supported by the UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1 and the Rosetrees Trust grant number ID2022\100044. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

Conceptualization: Y.L., A.A.F., P.R.

Data curation: Y.L., J.O., E.C.A., A.P.Y.B.

Formal analysis: Y.L.

Funding acquisition: A.A.F., P.R.

Methodology: Y.L., A.A.F., P.R.

Project administration: Y.L., P.R.

Resources: T.C.A., A.A., S.R., P.R., A.A.F.

Software: Y.L.

Supervision: A.A.F., P.R.

Validation: Y.L., J.O.

Visualization: Y.L., J.O.

Writing – original draft: Y.L.

Writing – review & editing: Y.L., J.O., E.C.A., A.P.Y.B., T.C.A., A.A., S.R., P.R., A.A.F.

Competing interests

The authors declare no competing interests.

References

1. Ramnarayan, P. & Polke, E. The state of paediatric intensive care retrieval in Britain. *Arch. Dis. Child.* 97, 145–149 (2012).

2. Roussak, P. Centralisation of paediatric intensive care and a 24-hour retrieval service. *Br. J. Nurs.* 23, 25–29 (2014).
3. Intensive Care National Audit & Research Centre (ICNARC). CMP Annual Quality Report 2018/19 for adult critical care. ICNARC. <https://account.icnarc.org/Our-Audit/Audits/Cmp/About/2019/12/04/Cmp-Annual-Quality-Report-201819> (2019).
4. Paediatric Intensive Care Audit Network (PICANet). National Paediatric Critical Care Audit State of the Nation Report 2024. Universities of Leeds and Leicester. <https://www.hqip.org.uk/report/picanet-dec24/> (2024).
5. van de Water, R. et al. Yet another ICU benchmark: A flexible multi-center framework for clinical ML in Proceedings of the twelfth international conference on learning representations (ICLR) (2024).
6. Pollard, T. J. et al. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* 5, 180178 (2018).
7. Faltys, M. et al. HiRID, a high time-resolution ICU dataset (version 1.1.1). PhysioNet 10.13026/nkwc-js72 (2021).
8. Johnson, A. E. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* 10, 1 (2023).
9. Zeng, X. et al. PIC, a paediatric-specific intensive care database. *Sci. Data* 7, 14 (2020).
10. Farris, R. W. D. et al. The PICU data collaborative: A novel, multi-institutional, pediatric critical care dataset. *Pediatr. Crit. Care Med.* 26, e941–e951 (2025).
11. Li, X., Grandvalet, Y. & Davoine, F. Explicit inductive bias for transfer learning with convolutional networks in Proceedings of the 35th international conference on machine learning 2825–2834 (PMLR, 2018).
12. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification in Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers) 328–339 (2018).
13. Fleming, S. et al. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: A systematic review of observational studies. *Lancet* 377, 1011–1018 (2011).
14. Uemura, O. et al. Age, gender, and body length effects on reference serum creatinine levels determined by an enzymatic method in Japanese children: A multicenter study. *Clin. Exp. Nephrol.* 15, 694–699 (2011).
15. Zhu, Y. et al. Domain adaptation using convolutional autoencoder and gradient boosting for adverse events prediction in the intensive care unit. *Front. Artif. Intell.* 5, 640926 (2022).
16. Yeo, J.-Y., So, E. S. & Youm, S. The pediatric transfer gap: How adult-ICU-trained machine-learning models degrade on pediatric intensive care patients. *Int. J. Med. Inform.* 220, 106597 (2026).
17. Li, X.-Q., Wang, R.-Q., Wu, L.-Q. & Chen, D.-M. Transfer learning-enabled outcome prediction for guiding CRRT treatment of the pediatric patients with sepsis. *BMC Med. Inform. Decis. Mak.* 24, 266 (2024).

18. Marassi, C., Socia, D., Larie, D., An, G. & Cockrell, R. C. Children are small adults (when properly normalized): Transferrable/generalizable sepsis prediction. *Surg. Open Sci.* 16, 77–81 (2023).
19. Lemmon, J. et al. Self-supervised machine learning using adult inpatient data produces effective models for paediatric clinical prediction tasks. *J. Am. Med. Inform. Assoc.* 30, 2004–2011 (2023).
20. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 10, e0118432 (2015).
21. Liu, Y. et al. Dual-phase machine learning framework for paediatric ventilator liberation: Forecasting extubation readiness and nowcasting extubation outcomes. *BMJ Digit. Health AI* 1, e000138 (2025).
22. Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int. Suppl.* 2, 1–138 (2012).
23. Singer, M. et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* 315, 801–810 (2016).
24. Schlapbach, L. J. et al. International consensus criteria for pediatric sepsis and septic shock. *JAMA* 331, 665–674 (2024).
25. Bennett, N., Plečko, D., Ukor, I.-F., Meinshausen, N. & Bühlmann, P. Ricu: R's interface to intensive care data. *GigaScience* 12, giad041 (2023).
26. Matics, T. J. & Sanchez-Pinto, L. N. Adaptation and validation of a pediatric Sequential Organ Failure Assessment score and evaluation of the Sepsis-3 definitions in critically ill children. *JAMA Pediatr.* 171, e172352 (2017).
27. Kleinman, M. E. et al. Part 14: pediatric advanced life support: 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation* 122, S876–S908 (2010).
28. Cho, K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 1724–1734 (2014).
29. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection in Proceedings of the IEEE international conference on computer vision 2980–2988 (2017).
30. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 17, 230 (2019).

Figures

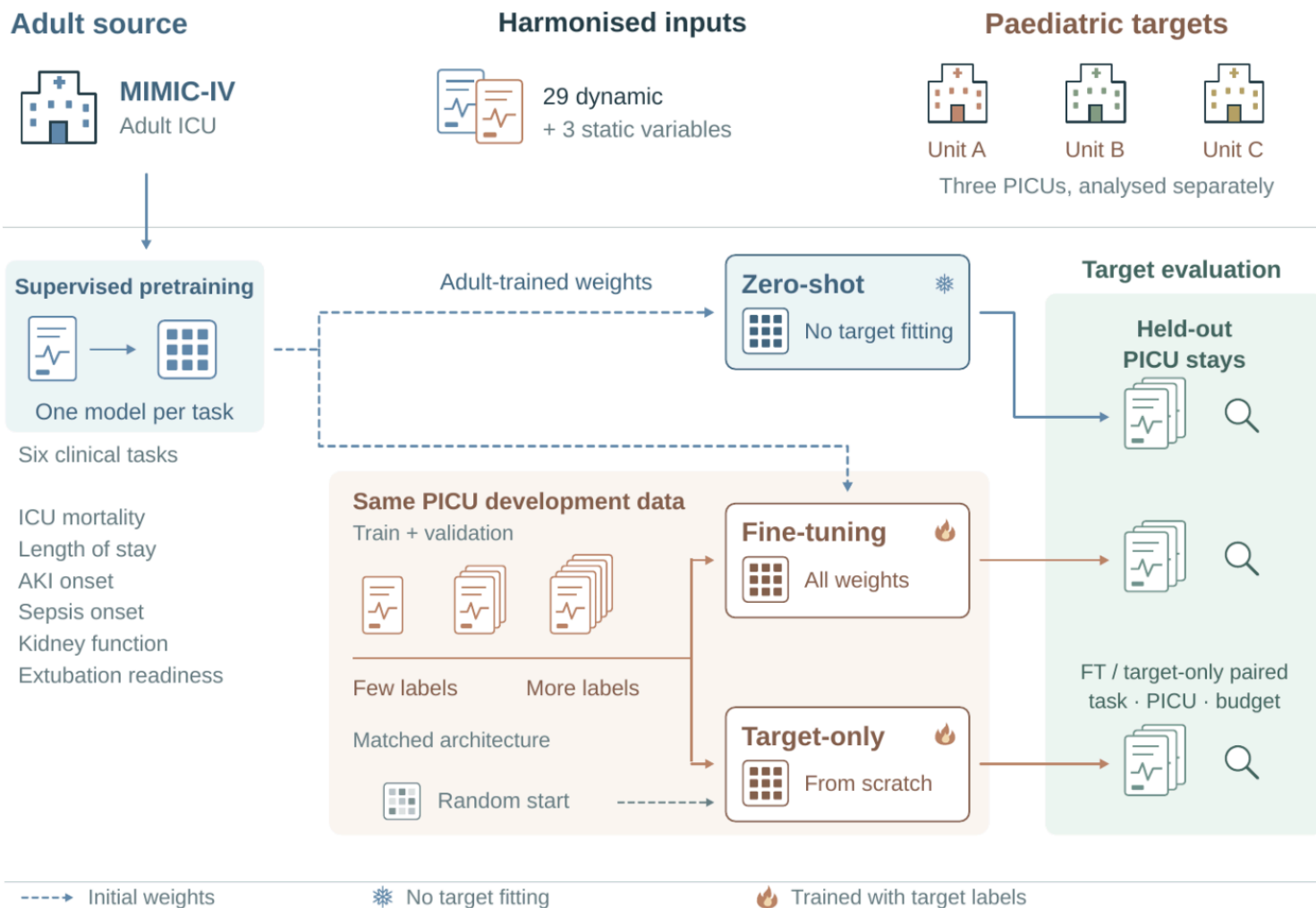
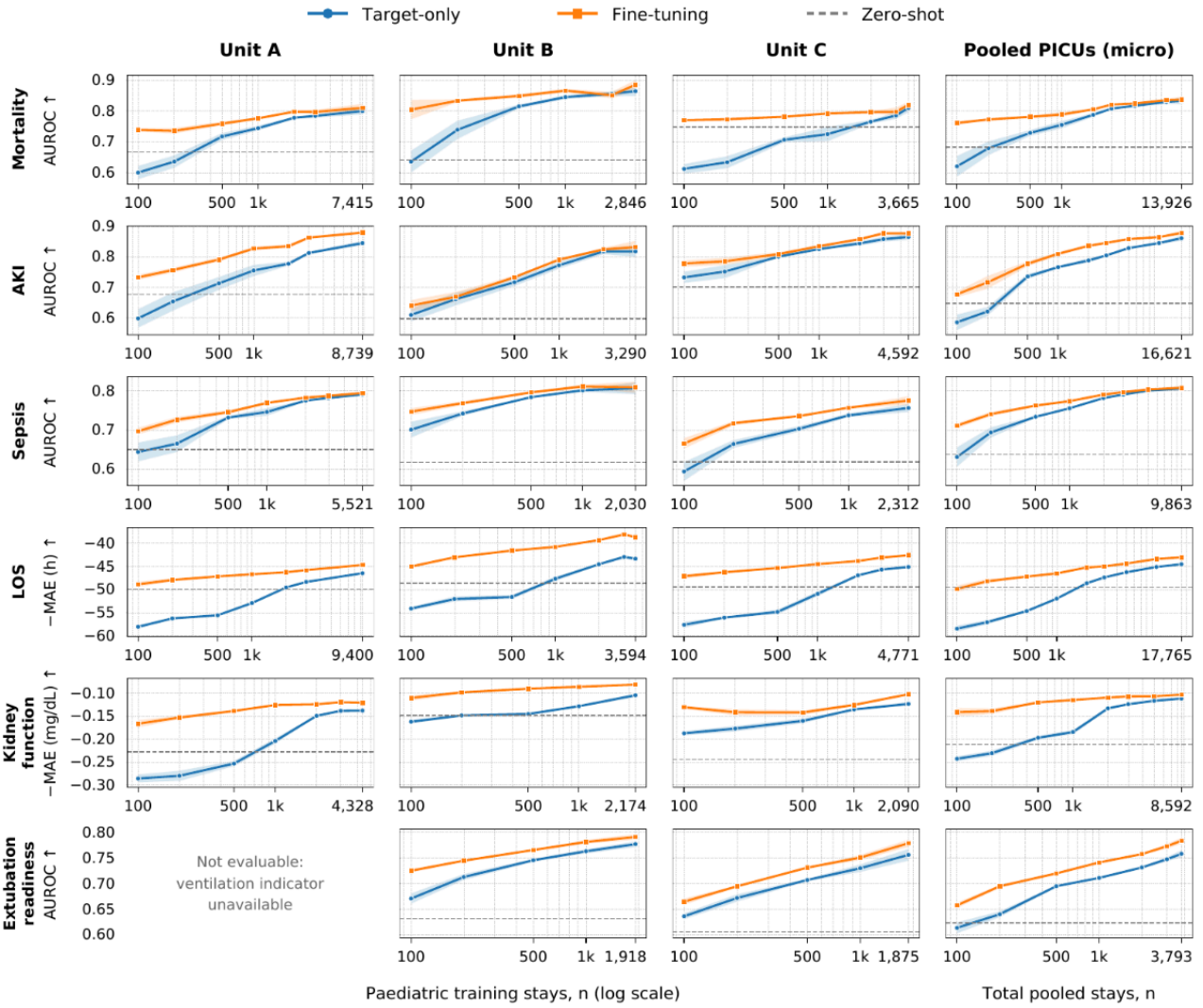


Figure 1

Study design and transfer conditions. One adult source cohort (MIMIC-IV) and three paediatric target cohorts (Units A–C, analysed separately) contributed 32 harmonised input variables: 29 dynamic and 3 static. For each of the six clinical tasks—ICU mortality, length of stay, acute kidney injury onset, sepsis onset, kidney function and extubation readiness—a separate model was trained on the adult cohort. Adult-trained weights were then used in two ways. Under zero-shot deployment the adult model was applied to the PICU without any target fitting. Under fine-tuning the adult weights initialised a model whose parameters were all updated on PICU data. Target-only training started from random initialisation on the same PICU data. Fine-tuning and target-only training used the same PICU development data, the same architecture and the same held-out stays at every combination of task, unit and label budget, so those two conditions are paired throughout; the stacked records indicate label budgets from few to many.

a) Performance across paediatric training sizes



b) Full-cohort AUPRC at individual PICUs

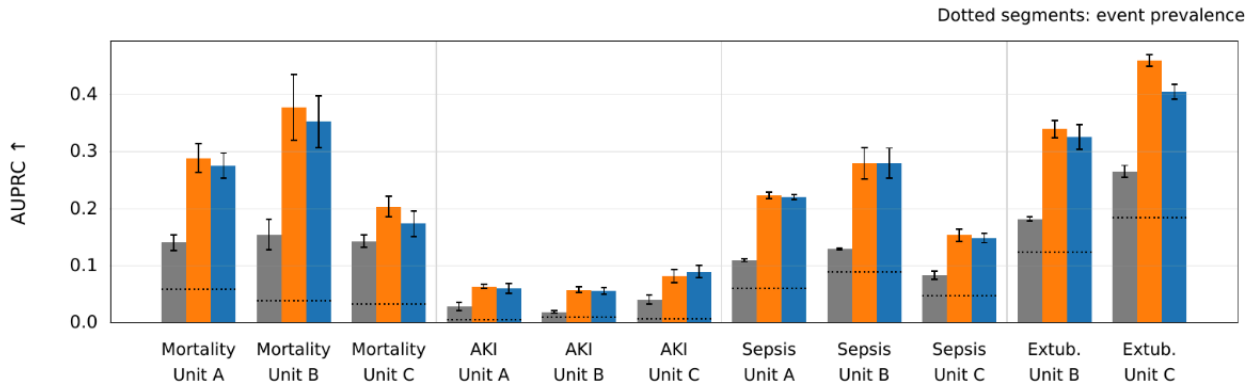


Figure 2

Model performance across paediatric data regimes and at the full paediatric cohort. Orange denotes fine-tuning, blue target-only training and grey zero-shot deployment. Points and bars are means across the 25 nested cross-validation fits; ribbons and error bars show standard errors across the five outer folds. **a)** Held-out performance against the number of paediatric development stays, n , on a logarithmic axis. The grey dashed horizontal line shows zero-shot performance, which does not depend on

paediatric training size. Columns show Units A–C and the pooled-PICU analysis, in which locally trained models used the combined paediatric development data and were evaluated as a micro metric over the combined held-out observations; pooled n is the total across units. AUROC is shown for classification and –MAE for regression, so higher is better; LOS MAE is in hours and kidney-function MAE in mg/dL. The right-most tick is the full task-evaluable cohort and every evaluated sample size is plotted. Extubation readiness was unavailable at Unit A. Target-only and fine-tuning values are in Supplementary Tables S5 and S6. **b)** Full-cohort AUPRC for the 11 classification task–site cells. Dotted segments mark event prevalence, the no-skill reference. Numerical values are in Supplementary Table S7.

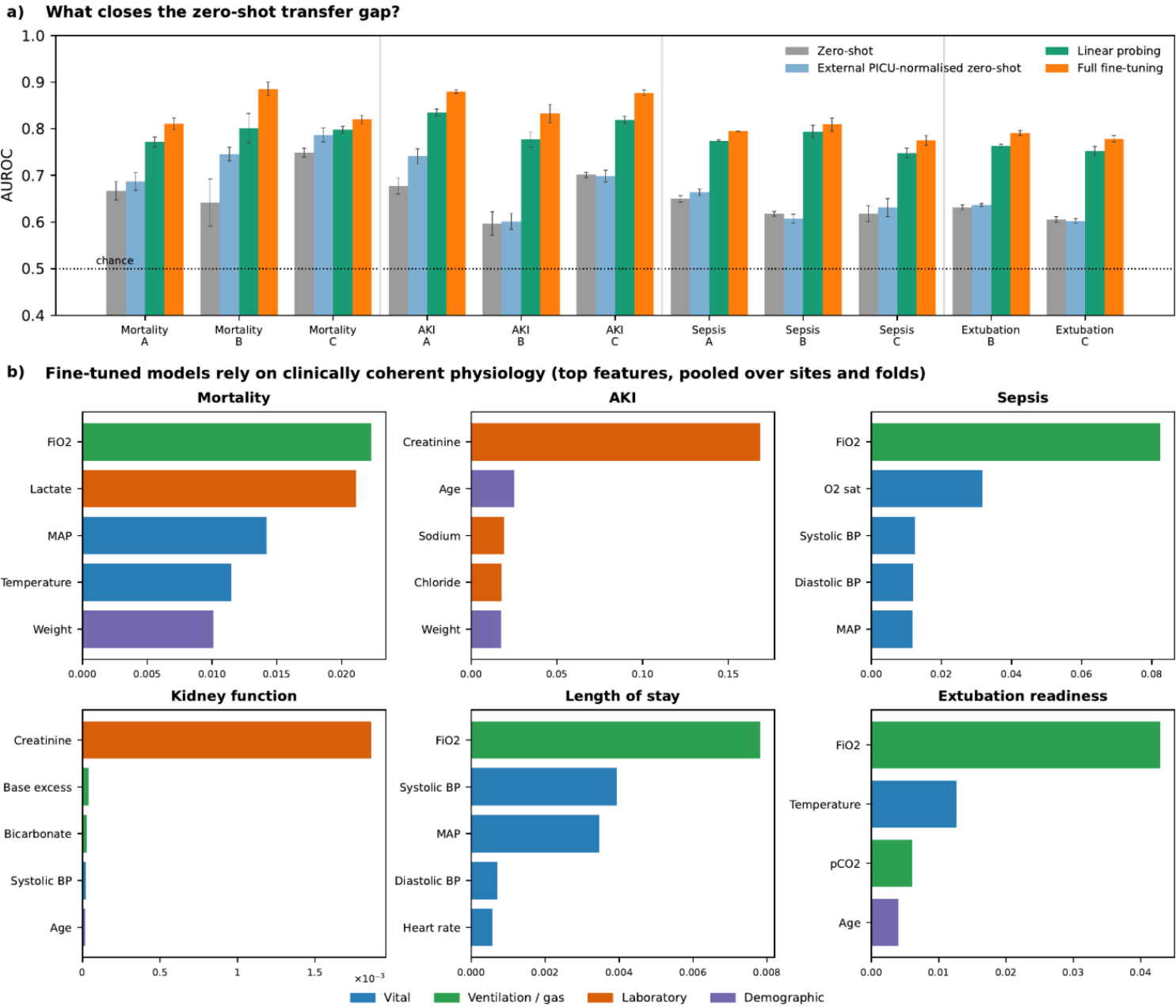


Figure 3

Performance under local adaptation and feature reliance in adapted PICU models. a) Mean held-out AUROC across the 25 nested cross-validation fits for 11 classification task–site cells under four

conditions: source-normalised zero-shot deployment (grey); external-PICU-normalised zero-shot deployment using standardisation statistics from the other two PICUs and no data from the evaluated site (light blue); linear probing with the adult encoder frozen and a new output layer fitted using labelled target data (green); and full fine-tuning of all model parameters using labelled target data (orange). Error bars show standard errors across the five outer folds; the dotted horizontal line marks chance performance. **b)** Top-five leave-one-feature-out permutation-importance features of the fine-tuned models for each task, pooled over the available sites and the 5×5 cross-validation (missing-indicator channels excluded), coloured by feature type. Bar length measures predictive reliance rather than effect direction or causality.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [02SupplementaryInformationNPJ.pdf](#)