Benchmarking Machine Learning Architectures for Antimicrobial Stewardship in Pediatric ICUs

Niklas Raehse[1,2] NIKLAS.RAEHSE@HEST.ETHZ.CH
Luregn J. Schlapbach[1] LUREGN.SCHLAPBACH@KISPI.UZH.CH
Daphné Chopard[1,3] DAPHNE.CHOPARD@KISPI.UZH.CH
[1] *Department of Intensive Care and Neonatology and Children's Research Center, University of Zurich, University Children's Hospital Zurich, Zurich, Switzerland*
[2] *Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland*
[3] *Department of Computer Science, ETH Zurich, Zurich, Switzerland*

Abstract

Antimicrobial stewardship (AMS) is critical for reducing unnecessary antibiotic exposure, particularly in pediatric intensive care units (PICUs), where clinical uncertainty leads to frequent broad-spectrum use, and overuse amplifies antimicrobial resistance which have long-term consequences. Machine learning has been proposed to support AMS by identifying patient-level opportunities for intervention from electronic health record data. However, prior work focuses on adult populations and predominantly relies on static tabular representations, leaving open questions about target design, temporal modeling, and generalizability in pediatric settings. In this work, we present a systematic benchmarking study of AMS intervention prediction in the PICU. Using the publicly available Paediatric Intensive Care database from China and a private PICU cohort from the University Children's Hospital Zurich, Switzerland, we define four clinically relevant proxy targets for reducing antibiotic use: intravenous-to-oral switching, de-escalation, discontinuation, and short-course therapy. We then compare tabular, sequence-based, and graph-based temporal models under a unified evaluation framework. We find that model performance is primarily driven by target prevalence and data characteristics rather than model complexity. Sequence models provide improvements in precision-recall trade-off over tabular approaches at coarse (24-hour) resolution, with limited additional gains when finer temporal structure is incorporated. However, these gains come at the cost of poorer calibration, with simpler tabular models producing more reliable probability estimates. Our results highlight the importance of target selection, temporal representation, and calibration in clinical machine learning, and provide practical guidance for developing reliable decision support systems for AMS in pediatric critical care. The code is publicly available[1].

1. Introduction

Antimicrobial resistance is a growing global threat, and optimizing antibiotic use through antimicrobial stewardship (AMS) is essential to mitigate its impact (World Health Organization, 2015; Barlam et al., 2016). In the pediatric intensive care unit (PICU), antibiotics are among the most frequently administered therapies (Willems et al., 2021) and are often initiated by default due to non-specific signs of systemic inflammation in critically ill

1. https://github.com/pinnacle-kispi/AMS_intervention_prediction

children (Weiss et al., 2020; Evans et al., 2018). This leads to substantial overuse, much of which may be unnecessary or overly broad (Chiotos et al., 2022; Abbas et al., 2016), contributing to toxicity and antimicrobial resistance, with amplified long-term consequences in children (Romandini et al., 2021; Neuman et al., 2018; Sivasankar et al., 2023; Fanelli et al., 2020a). AMS programs aim to identify opportunities to modify therapy, including de-escalation, discontinuation, or switching from intravenous to oral treatment (Pollack and Srinivasan, 2014; Khdour et al., 2018; Deresinski, 2007). However, these processes are resource-intensive and may miss timely intervention opportunities (World Health Organization, 2019), motivating the development of a clinical decision support system (CDSS) to assist clinicians in identifying when antibiotic therapy could be safely adjusted (Giacobbe et al., 2024). Machine learning (ML) methods using electronic health record (EHR) data have been proposed to support AMS by predicting when stewardship interventions are (Giacobbe et al., 2024; Tran-The et al., 2024). These approaches typically rely on retrospective targets derived from historical clinical decisions, which reflect observed practice rather than optimal care, but remain useful for capturing real-world stewardship behavior at scale. Identifying contexts in which therapy modification is likely to be appropriate is particularly valuable, as clinicians may be hesitant to adjust treatment under uncertainty (Chiotos and Tamma, 2020). Despite these advances, AMS intervention prediction in the PICU remains an open challenge (Fanelli et al., 2020b). Compared to adult settings, pediatric data are more heterogeneous (Bruns et al., 2022), and little work has addressed stewardship prediction in this population. Consequently, it remains unclear which intervention targets are most informative and which modeling approaches best capture the temporal and clinical structure of stewardship decisions. Commonly used targets such as intravenous-to-oral switching are underutilized in pediatric settings (Berthe-Aucejo et al., 2012; McMullan et al., 2016), highlighting the need to reassess target definitions. From a modeling perspective, patient data can be represented as either static snapshots or temporal sequences, but the impact of these approaches has not been systematically studied in AMS literature, and model selection is often driven by convention rather than evidence. In this work, we address these gaps by systematically benchmarking ML architectures for AMS intervention prediction in the PICU. We evaluate models across two cohorts, a publicly available PICU dataset (Zeng et al., 2020) and a private dataset, to obtain a broader understanding of findings across pediatric populations. We define four stewardship proxy outcomes clinically relevant to the PICU setting: intravenous-to-oral switching, de-escalation, discontinuation, and short-course therapy. We compare tabular, sequence-based, and graph-based temporal models under a consistent framework, providing insights into how temporal representation and model choice affect predictive performance and offering guidance for the development of a reliable CDSS for AMS in pediatric critical care.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our findings highlight two broader insights for applying machine learning in clinical settings beyond antimicrobial stewardship. First, evaluation metrics can lead to different conclusions about model utility: similar AUROC values can obscure meaningful differences in F1 score and AUPRC, especially for rare outcomes. Metrics should therefore be selected according to the intended clinical use, target prevalence, and relative costs of false-positive and false-negative predictions, and should include prevalence-sensitive metrics that better

reflect a model's ability to identify clinically relevant positive cases. Second, prediction targets developed in well-studied populations may be clinically irrelevant or too rare in underrepresented groups; for example, intravenous-to-oral antibiotic switching is common in adults but arises less often in pediatric care. Clinical machine-learning studies should therefore develop and validate both models and targets specifically for populations whose care processes and outcome distributions differ from those represented in existing research.

2. Related Work

AMS intervention prediction targets. Several studies have proposed predicting AMS interventions from EHR data (Beaudoin et al., 2014; Goodman et al., 2022; Bystritsky et al., 2020; Tran-The et al., 2024), operationalizing stewardship actions such as intravenous-to-oral switching, de-escalation, and discontinuation as prediction targets. However, most work focuses on a single intervention type (Bolton et al., 2024, 2025) or collapses multiple actions into a single binary outcome (Bystritsky et al., 2020; Goodman et al., 2022), thereby potentially obscuring differences between stewardship decisions. Furthermore, existing studies are primarily conducted in on adult populations, and it remains unclear how well these targets transfer to pediatric settings, where intervention patterns and data characteristics differ (Fay and Bryant, 2020; Shah et al., 2023).

Modeling approaches for AMS prediction. Prior work spans a range of modeling approaches. Early studies rely on tabular representations with classical models such as logistic regression or gradient boosting (Beaudoin et al., 2014; Tran-The et al., 2024), prioritizing interpretability and ease of deployment. These approaches represent patients as snapshots by aggregating dynamic features into summary statistics, thereby discarding temporal structure for simplicity (Beaudoin et al., 2014; Goodman et al., 2022; Bystritsky et al., 2020; Tran-The et al., 2024). More recent work has explored deep neural networks with post-hoc explanation (Bolton et al., 2024), but these models operate on similar aggregated inputs and therefore do not explicitly model temporal dynamics. While sequence-based models can capture longitudinal information, they have not been systematically evaluated for AMS prediction. Existing studies remain fragmented across private and public datasets, limiting reproducibility and cross-site comparison, and do not jointly assess the impact of temporal representations and model choice.

Causal outcome estimation for AMS. A related line of work estimates outcomes under alternative treatment decisions as opposed to predicting observed interventions. For example, Bolton et al. (2022) use a recurrent neural network and synthetic controls to model counterfactual antibiotic cessation. While this approach targets decision support more directly, it relies on strong assumptions and unobserved confounding. In contrast, predicting observed interventions provides a simpler and more reproducible supervised learning setting, and prior work has shown that such predictions can improve the prioritization of cases for stewardship review (Tran-The et al., 2024). However, as observed interventions remain proxies for optimal decisions, prospective evaluation is required to determine whether model-guided review improves stewardship efficiency and patient outcomes.

PICU-specific challenges. ML for pediatric AMS remains underexplored (Fanelli et al., 2020b). PICU data often have high missingness likely due to the heterogenous casemix

(Bruns et al., 2022), practical sampling challenges with small patients (Whitehead et al., 2019), as well as smaller cohort sizes which limits the applicability of adult-derived models (Shah et al., 2023). While ML has been applied to related pediatric tasks such as infection detection (Martin et al., 2022; Lee et al., 2022; Fenta et al., 2024; Clarke et al., 2022), antibiotic susceptibility prediction (Oonsivilai et al., 2018), and clinical deterioration (Velez et al., 2025), to our knowledge no prior work has studied AMS intervention prediction models in the PICU.

To address these gaps, we provide a systematic, cross-cohort evaluation of AMS intervention prediction in PICU, using both a public and a private dataset. We jointly examine target formulation, temporal representation, and modeling paradigm, enabling a direct comparison of tabular, sequence, and graph-based models across clinically relevant stewardship tasks.

3. Methods

We formulate AMS intervention prediction as a patient-day classification task using EHR data. At the start of each antibiotic patient-day we want to predict whether an AMS intervention is likely to occur during that day. This section describes the prediction setting, data representation pipeline, model classes, and evaluation protocol (also see Figure 1), whereas cohort-specific preprocessing details are provided in Section 4. The code is made publicly available for reproducibility.[2]

3.1. Prediction setting

Predictions are made once a day at a fixed time, using only information available up to that point. Each prediction is defined for a patient-day with active antibiotic therapy and aims to predict whether an intervention occurs within the following 24 hours. For tabular models, predictions are based on a single patient-day feature vector summarizing the most recent patient state. Temporal models use the same patient-day representation at each time step, but operate on sequences of consecutive patient-days, allowing them to incorporate information from prior days. This unified formulation ensures that all models receive the same per-day input, while differing only in their ability to leverage longitudinal context.

3.2. Feature space

We extract a consistent set of clinical variables across cohorts, including (i) patient and admission context, such as age and admission type, (ii) physiological measurements, namely vital signs and laboratory values, and (iii) antibiotic exposure features (e.g., ATC code, route, concurrent therapies). Full listing of all input features is available in Appendix Table 1 alongside more detailed descriptions.

3.3. Patient-day and temporal representations

We formulate the task at the patient-day level, where each instance corresponds to a 24-hour interval ending at a fixed daily prediction time. This defines a consistent prediction

2. https://github.com/pinnacle-kispi/AMS_intervention_prediction

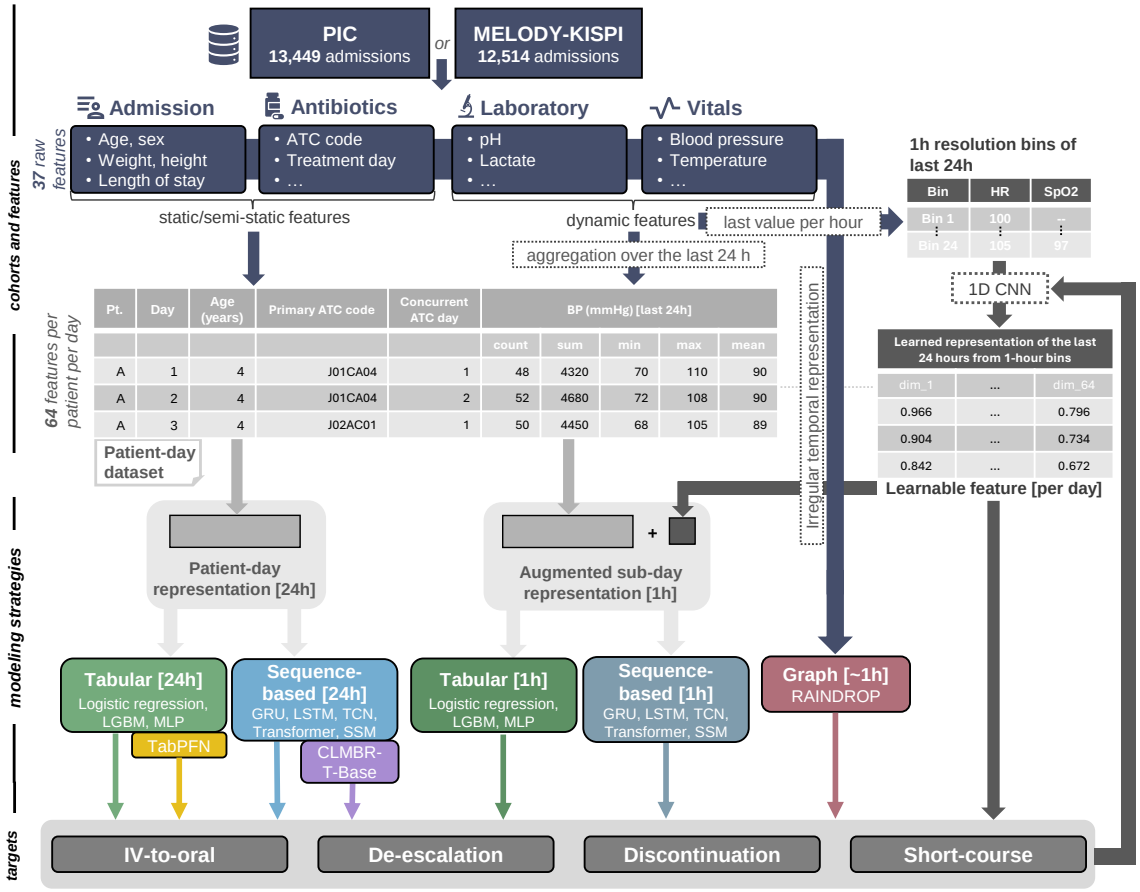

Figure 1: Benchmarking overview. EHR are represented at two temporal resolutions: (i) A X-day lookback tabular representation based on summary statistics of dynamic variables over the preceding day(s), and (ii) hourly bins augmented by a 1-hour representation capturing intra-day dynamics through a learned embedding of hourly bins using a CNN trained jointly with the prediction task. We benchmark three classes of models under a unified framework. All models are evaluated on the four antimicrobial stewardship intervention targets.

unit across all models. We consider multiple temporal resolutions of this patient-day representation to disentangle the effect of temporal granularity and modeling assumptions (see Figure 1). These representations provide different levels of temporal detail over the same prediction unit, enabling a controlled comparison of modeling approaches.

Patient-day representation (24h resolution). Physiological measurements are aggregated into summary statistics over the preceding 24 hours (e.g., minimum, maximum, mean) and combined with static and semi-static features, yielding a single feature vector per patient-day of the recent state. This corresponds to the standard tabular representation.

Augmented sub-day representations (1h resolution). To capture intra-day dynamics, we construct hourly representations within the same 24-hour window. Observations are aggregated into hourly bins by retaining the last recorded value in each bin. Empty bins are then forward-filled using last observation carried forward (LOCF) (Engels and Diehr, 2003). We then concatenate the hourly event grid with the aggregated patient-day representation. A convolutional encoder summarizes this sequence into a compact embedding, which is concatenated with the 24h-resolution feature vector. As a result, the 1h representation augments the 24h representation, allowing models to jointly leverage coarse summary statistics and fine-grained temporal structure. The embedding is thus learned jointly with the prediction task. This representation allows tabular models to incorporate temporal information, and enables direct comparison across model classes.

Irregular temporal representation. We additionally consider a representation that preserves the native irregular sampling of EHR data within each patient-day. This is used by graph-based models, which directly operate on irregular observations and capture temporal and inter-variable relationships without discretization.

3.4. Model classes

We evaluate three classes of models, each corresponding to a different inductive bias about how clinical decisions arise from patient data. Implementations details can be found in Appendix B.

Tabular models. Tabular models operate on independent patient-day feature vectors and assume that relevant information is captured by the current state (i.e., previous patient-day). In line with previous work on AMS intervention prediction (Goodman et al., 2022; Bystritsky et al., 2020; Tran-The et al., 2024; Bolton et al., 2024), we include logistic regression (McCullagh, 2019), LightGBM (Ke et al., 2017), and a multi-layer perceptron (MLP) classifier. We also include the transformer-based tabular foundation model TabPFN (Hollmann et al., 2025) as a strong baseline.

Sequence-based temporal models. Sequence models explicitly capture temporal dependencies across patient trajectories. They process patient data as sequences of observations and incorporate information from previous timesteps (here, patient-days) to inform predictions; either through recurrent state updates or through architectures that aggregate context across the sequence. We evaluate recurrent architectures (GRU, (Cho et al., 2014), LSTM (Hochreiter and Schmidhuber, 1997)), a Temporal Convolutional Network (TCN) (Bai et al., 2018), a Transformer (Vaswani et al., 2017), and a State Space Model (SSM) inspired by Mamba (Gu and Dao, 2024). We additionally evaluate CLMBR-T-Base (Wornow et al., 2023; Guo et al., 2024), a transformer foundation model pretrained on structured adult EHR code sequences.

Graph-based temporal models. Graph-based temporal models provide an alternative to tabular and sequence-based approaches by representing clinical variables and observations as nodes, modeling their dependencies through message passing. Some architectures operate directly on irregular and asynchronous measurements, without fixed-interval discretization. In this work, we include RAINDROP (Zhang et al., 2022), a state-of-the-art

graph neural network designed for irregular multivariate time series classification. While newer classification methods, such as Hi-Patch (Luo et al., 2025) or WaveGNN (Hajisafi et al., 2025) have been proposed, they are not yet established as standard baselines. RAINDROP captures temporal and inter-variable dependencies through neural message passing and self-attention mechanisms.

3.5. Targets

We define four AMS proxy targets derived from antibiotic treatment trajectories (see Figure 2 below and Appendix Section B.3 for details):

- **IV-to-oral**: Transition from intravenous to non-intravenous formulation within a treatment course.
- **De-escalation**: Switch to an antibiotic with a strictly lower antibiotic spectrum index (ASI) (Gerber et al., 2017).
- **Discontinuation**: Termination of antibiotic therapy with at least 72 hours remaining before hospital discharge (Tran-The et al., 2024).
- **Short-course**: Total antibiotic duration longer than 24 hours and shorter than 96 hours, excluding prophylactic cefazolin.

Each target is defined at the patient-day level, and labeled positive only on the day the event occurs. IV-to-oral and de-escalation correspond to established stewardship actions aimed at reducing treatment invasiveness and narrowing the antimicrobial spectrum. Discontinuation and short-course targets capture opportunities to limit overall antibiotic exposure.

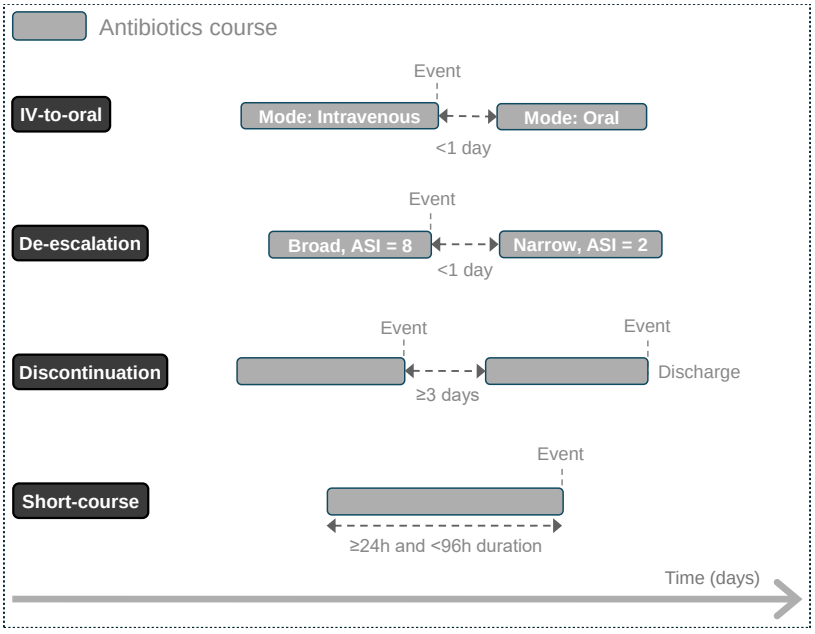

Figure 2: AMS intervention targets.

Target design. We build on prior work that operationalizes stewardship interventions from retrospective prescription data (Tran-The et al., 2024), but adapt these definitions to the PICU setting and extend them in two important ways. First, we introduce *short-course* therapy as an additional proxy outcome. In pediatric critical care, treatments are often initiated per default at admission and then de-escalated early (Chiotos et al., 2022; Abbas et al., 2016). Short courses may therefore reflect situations where antibiotic therapy was initiated under uncertainty and subsequently discontinued when deemed unnecessary, providing a complementary signal for potential over-treatment. Second, we define de-escalation strictly in terms of a reduction in ASI rather than also a reduction in the number of agents. This distinction is important: Reducing the number of antibiotics does not necessarily correspond to narrower antimicrobial coverage (Ilges et al., 2023). By relying on spectrum-based definitions, we better align the target with stewardship intent and avoid conflating treatment simplification with true de-escalation.

Interpretation as intervention opportunities. These targets correspond to observed stewardship interventions and reflect real-world clinical decision-making, rather than adjudicated stewardship intent or optimal treatment decisions. Because they are derived from retrospective clinical practice, they capture when clinicians chose to modify or discontinue therapy-not necessarily when such actions were clinically appropriate. Consequently, models trained on these targets are likely to learn local prescribing and stewardship patterns, including potential biases, and may preferentially identify situations that resemble past intervention decisions. The direction and magnitude of this bias is difficult to quantify and may vary across institutions, clinical teams, and stewardship programs.

Despite this limitation, such proxies remain valuable for a CDSS. Rather than identifying definitive opportunities for antimicrobial optimization, the models are intended to flag patient-days that may warrant stewardship review. This approach provides a scalable and reproducible way to identify opportunities for prospective audit and feedback workflows, enabling more frequent and targeted assessment of antimicrobial therapy.

3.6. Evaluation

Datasets are split per patient-day into training, validation, and test sets (70-10-20), ensuring that all stays of a given patient are assigned to the same split. The test-set is split at the beginning and not used until final evaluation. Models are trained using the same evaluation protocol across all model classes to ensure fair comparison. Performance is evaluated at the patient-day level using standard classification metrics, including Area Under the Receiver Operating Characteristic (AUROC) (Hanley and McNeil, 1982), F1 Score (van Rijsbergen, 1979), and Area Under the Precision Recall Curve (AUPRC) (Davis and Goadrich, 2006).

4. Cohorts

4.1. Cohort Selection

To evaluate AMS intervention prediction across heterogeneous pediatric settings, we use two complementary PICU cohorts: A publicly available dataset and the private *MELODY-KISPI* dataset. This dual-cohort design enables both reproducibility and assessment of cross-site comparison, addressing a key limitation of prior work which typically relies on

a single type of data source. In both cohorts, we extract antibiotic treatment trajectories and construct patient-day observations corresponding to days on which antibiotic therapy is active, following the prediction setting described in Section 3.

PIC. The Chinese Pediatric Intensive Care (PIC) database (Li et al., 2020) is a publicly available dataset containing de-identified EHR data from patients admitted to the PICU of the Children's Hospital of Zhejiang University School of Medicine between 2010 and 2018. It includes demographics, vital signs, laboratory measurements, and medication records, with physiological variables primarily recorded manually. We include all patients with at least one antibiotic prescription during their stay, resulting in 5,671 admissions for 5,515 distinct patients. The cohort is characterized by a young population (median age 0.83 years, IQR 0.16-3.81) and relatively long ICU stays (median 10.16 days, IQR 4.59-19.52). As a public dataset, PIC provides a reproducible benchmark for method comparison.

MELODY-KISPI. The *MELODY-KISPI* cohort comprises de-identified EHR data from PICU admissions at the University Children's Hospital Zurich, Switzerland between 2015 and 2023. The retrospective use of these data was approved by the Cantonal Ethics Committee of Zurich[3]. In contrast to PIC, this dataset reflects a different clinical environment and data collection process, including both sensor-derived and nurse-validated physiological measurements. After applying the same inclusion criteria, the cohort contains 2,789 admissions for 2,059 distinct patients, with a younger and shorter-stay population (median age 0.40 years, IQR 0.04-2.91; median length of stay 5.18 days, IQR 1.77-15.78). This cohort enables evaluation of model robustness and generalization across institutions.

4.2. Cohort-specific characteristics of AMS targets

The distribution of AMS intervention targets differs substantially across cohorts and highlights important pediatric-specific considerations. In particular, intravenous-to-oral switching, commonly used as a proxy for stewardship interventions in adult settings, is extremely rare in both PICU cohorts (0.34% in PIC and 0.98% in the *MELODY-KISPI* cohort, see Table 6 in the Appendix). In contrast, short-course therapy is substantially more frequent, especially in the *MELODY-KISPI* cohort, and often co-occurs with discontinuation, suggesting that it captures a meaningful and prevalent pattern of early treatment cessation. Other combinations of intervention types are rare, indicating that stewardship actions are typically isolated events (see Figure 6 in the Appendix). These differences motivate the inclusion of short-course therapy as an additional proxy outcome tailored to the pediatric setting, complementing established intervention targets and enabling a broader characterization of opportunities for reducing unnecessary antibiotic exposure.

4.3. Data Extraction

We extract a set of 37 clinical variables from both cohorts, covering demographics, admission context, antibiotic prescriptions, vital signs, and laboratory measurements (see Table A in the Appendix for the full list). To enable consistent modeling and fair comparison, both datasets are mapped to a shared feature space covering demographics, clinical measurements, and antibiotic exposure. Antibiotics records are limited to an overlapping set and

3. BASEC 2023-02077

harmonized using ATC (WHO Collaborating Centre for Drug Statistics Methodology, 2024) codes (see Table 2 in the Appendix). We assume that prescription timestamps approximate administered therapy in PIC. For the *MELODY-KISPI* cohort, the available timestamps correspond to actual antibiotic administrations. Laboratory measurements are aligned using LOINC codes (Forrey et al., 1996).

4.4. Feature construction

We apply the unified feature construction pipeline described in Section 3 to both cohorts, ensuring that differences in performance reflect underlying data characteristics and not preprocessing choices. Here, we highlight cohort-specific implementation details. Antibiotic prescriptions are mapped to systemic antibiotics and merged into clinically contiguous courses based on temporal proximity (within 24 hours). This step is particularly important for the *MELODY-KISPI* cohort, where prescriptions were mostly recorded as individual bolus administrations as opposed to continuous treatments. Very short courses (< 24 hours), typically reflecting prophylactic use, are excluded. When multiple antibiotic courses overlap within a patient-day, all active agents are retained, jointly contributing to feature construction and target labeling, instead of collapsing to a single representative treatment. For the *MELODY-KISPI* cohort, antibiotic prescription timestamps correspond to actual administrations, while in PIC they approximate administered therapy. This difference does not affect feature construction but is relevant for interpreting temporal alignment. All remaining preprocessing and feature aggregation steps follow the shared pipeline described in Section 3.

5. Experiments and Results

5.1. Comparison to previous work in adults

For contextualization, we compare the results of our tabular models trained on 24-hour aggregated pediatric data (LightGBM and MLP) with previously reported adult ICU results (full results in Table 7 in Appendix C.2). The feature set used is available to both pediatric datasets (Appendix Section A), comparable to Tran-The et al. (2024) and includes 7/10 of the base vital features used in Bolton et al. (2024). Intervention prevalence differs substantially between adult and pediatric settings. IV-to-oral switching is much rarer in PICU cohorts (0.3-1.0%) than in adult data (3.0%), and discontinuation is also less frequent (1.7-5.7% vs 9%). We observe high AUROC values in pediatric cohorts (e.g., 0.93 for de-escalation), but precision-recall performance is markedly lower for rare targets. For IV-to-oral for example, LightGBM shows low AUPRC and TPR scores compared to Tran-The et al. (2024), reflecting extreme class imbalance. This highlights that AUROC alone can overestimate performance in such settings. For more prevalent targets, strong AUROC does not consistently translate to improved AUPRC, indicating differences in label distribution and clinical practice. Overall, these results show that performance does not directly transfer from adult to pediatric cohorts. Model behavior is strongly influenced by target prevalence and cohort characteristics, motivating the need for PICU-specific evaluation and the inclusion of additional targets such as short-course therapy.

5.2. Effect of sequence modeling at 24-hour temporal resolution

We then investigate whether sequence-based models can better exploit longitudinal structure in patient trajectories than tabular models when both use the same 24-hour aggregated input representation. Specifically, we compare sequence-based models (recurrent, convolutional, and attention-based) with tabular baselines, using the same 24-hour aggregated representation across the *PIC* and *MELODY-KISPI* cohorts. AUROC, F1 score and AUPRC are shown in Figure 3, with full results provided in Appendix C.4. We report the performance of the pretrained models TabPFN and CLMBR-T-Base separately because, unlike the other models, they are not trained from scratch on these cohorts. The F1 score threshold is optimized using the validation split, and averages of three seeds are always reported on the held-out test set in the results below.

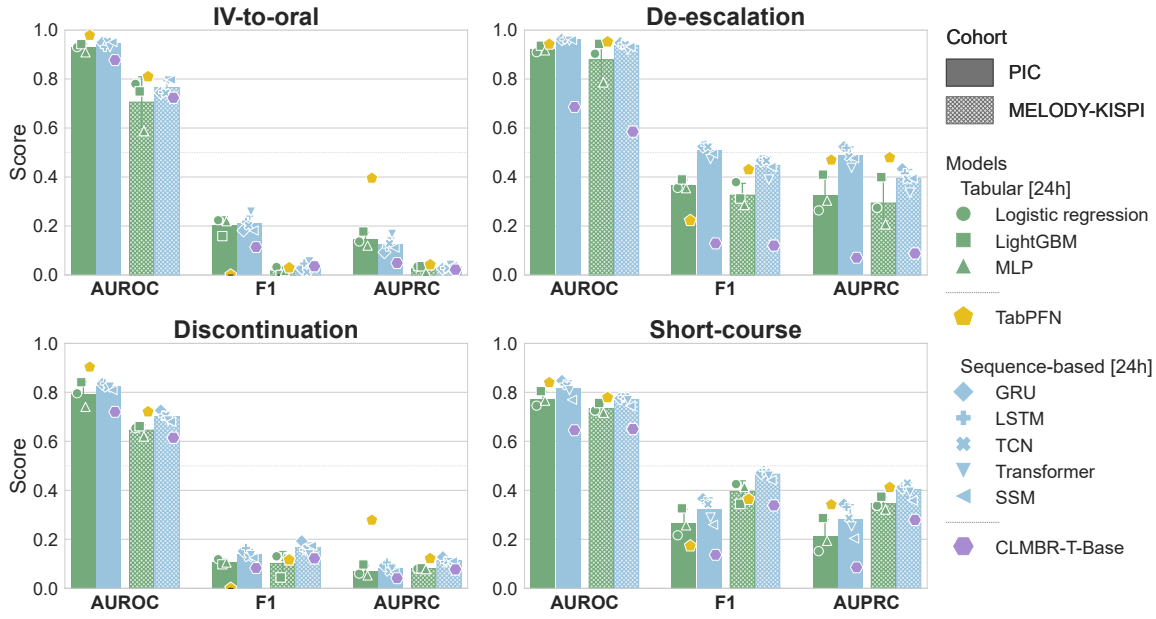

Figure 3: Comparison of tabular and sequence models using 24-hour aggregated representations across the *PIC* and *MELODY-KISPI* cohorts and across four antimicrobial stewardship targets. Bars show the mean performance within each model group. Individual model results are overlaid as points. The pretrained models TabPFN and CLMBR-T-Base are shown separately for reference.

Across tasks and cohorts, sequence models achieve similar AUROC to tabular models (e.g., ∼0.85-0.95 for IV-to-oral and de-escalation), indicating that most predictive signal is already captured by the current patient-day representation. Differences are more apparent in AUPRC and threshold-optimized F1 score. For rare targets (IV-to-oral, discontinuation), sequence models provide modest gains over tabular models, while for more prevalent targets (de-escalation, short-course), the gains are substantially larger. Similar trends are observed across both cohorts. In *PIC*, the average F1 score increased from 0.33 across tabular models to 0.51 across sequence models for de-escalation, and from 0.24 to 0.32 for short-course. Smaller but consistent gains are also observed for *MELODY-KISPI*. F1 score

differences were even larger at the default threshold of 0.5 (Appendix C.4). The results suggest that explicit temporal modeling can improve the prediction of interventions over tabular approaches, and that exploring finer-grained temporal representations may offer additional improvements.

5.3. Effect of finer temporal resolution and irregular time series modeling

We next evaluate whether increasing temporal resolution improves performance. We augment 24-hour features with 1-hour event grids (See Section 3.3) and compare tabular and sequence models, together with the graph-based model (RAINDROP) on the *PIC* cohort (Figure 4). We exclude the pretrained TabPFN and CLMBR-T-Base models from this follow-up experiment because they performed poorly in the preceding 24-hour-resolution comparison. Corresponding results for the *MELODY-KISPI* cohort are available in Figure 7 of the Appendix. Incorporating finer-grained temporal information only slightly im-

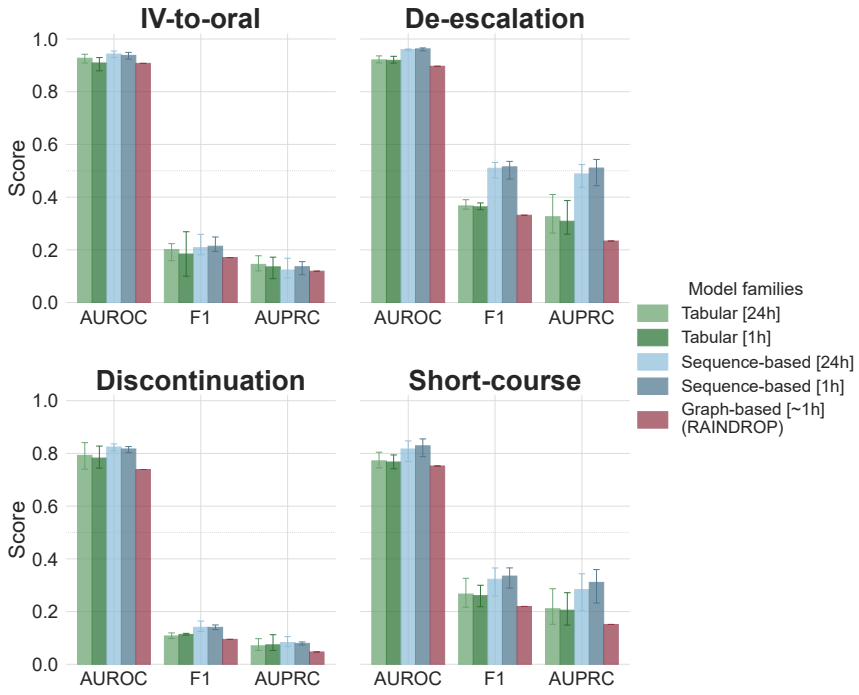

Figure 4: Performance of tabular and sequence-based models at 24-hour resolution and with augmented 1-hour representation, together with the graph-based model RAINDROP on irregular time series on the *PIC* cohort. Bars and error bars show the model class mean and minimum-maximum range, respectively.

proves performance. For de-escalation and short-course, F1 and AUPRC for de-escalation and short-course targets improve ∼0.01-0.03 on average across three seeds in *PIC*, and similar results were observed in *MELODY-KISPI* (Appendix Figure C.5). The CNN-based learnable encoder does add some predictive power over a simple linear projection or direct hourly modeling (Appendix Section C.8, Figures 15 and 16). Tabular models augmented with temporal embeddings show less consistent results and remain below sequence-based ap-

proaches. RAINDROP, also modeling independent patient-days, shows performance comparable to but below that of the tabular models, suggesting that the graph-based embeddings of irregularly sampled data confuse the predictors for these tasks.

Overall, these results indicate that temporal resolution and structure matter, although they provide only limited incremental benefits, and fully leveraging them requires models that can naturally handle sequences as opposed to static snapshots.

5.4. Evaluating reliability across model classes

Trust is a critical factor in the adoption of a CDSS for AMS (Bolton et al., 2025). In this context, well-calibrated predictions are essential, as they determine how reliably predicted probabilities reflect true intervention likelihood. We therefore evaluate calibration across tabular, sequence-based, and graph-based temporal models using different temporal representations. Results for the *PIC* cohort are shown in Figure 5, with corresponding results for *MELODY-KISPI* provided in Appendix Figure 8. Calibration varies substantially

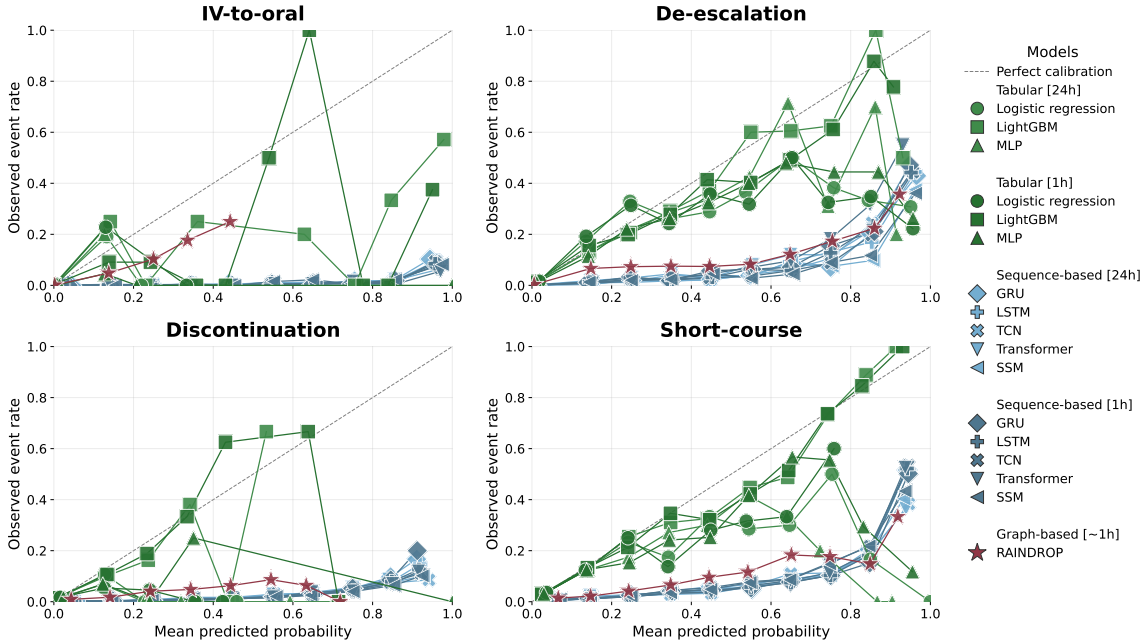

Figure 5: Calibration plots for tabular, sequence-based, and graph-based models across temporal representations for the four AMS intervention predictions tasks in the *PIC* cohort.

across model classes and tasks. Tabular models, particularly LightGBM, show the best calibration, with predicted probabilities closely aligned with observed event rates for more prevalent targets, such as de-escalation and short-course. In contrast, sequence models are over-confident, with predicted probabilities consistently higher than the observed event rates across bins. Complementary experiments highlight that predicted probabilities generally increase more towards the event day for sequence models compared to tabular ones (Figures 11 and 12 in the Appendix). We note that RAINDROP exhibits similar calibration challenges, with weak probability separation and poor calibration. Importantly, increasing temporal

resolution does not substantially improve calibration. Models using 1-hour representations exhibit similar over-confidence as their 24-hour counterparts, indicating that calibration is largely independent of temporal resolution in this setting. For rare targets (IV-to-oral, discontinuation), all models show unstable calibration, with noisy estimates and unreliable behavior at higher probability ranges due to extreme class imbalance. However, for more prevalent targets, like short-course in the *MELODY-KISPI* cohort (Appendix Figure 8), we observe that the sequence models and RAINDROP recover calibration. Overall, these results highlight a trade-off between discrimination and calibration. While more complex temporal models can improve ranking performance, simpler tabular models provide more reliable probability estimates, which is essential for building a trustworthy CDSS.

5.5. Additional Results

We use patient-random splitting for consistency, as *PIC* employs random time offsets that preclude patient-temporal splitting. On *MELODY-KISPI*, we compared both strategies on held-out test sets and found that patient-temporal splitting consistently yields lower performance, suggesting that patient-random splitting provides more optimistic estimates (see Table 15 in the Appendix). Furthermore, we assume patient-days are independent which may inflate the effective sample size due to within-admission correlation. However, nearly identical naïve and admission-level cluster bootstrap confidence intervals (Appendix Table 16) indicate that this correlation has negligible impact on performance uncertainty. Because of conceptual overlap between AMS actions and prior modeling efforts that did not distinguish intervention targets (Bystritsky et al., 2020; Goodman et al., 2022), we also explored multi-task learning in Appendix Section C.6. However, it provided limited benefit, likely due to limited stay-level target overlap (Appendix Figure 6), hierarchical relationships between targets, and differences in label quality. These findings support task-specific modeling, particularly for common or heterogeneous intervention targets.

6. Discussion

In this work, we systematically evaluate multiple modeling paradigms for AMS intervention prediction in the PICU across cohorts, targets, and temporal representations, yielding several key insights.

First, we observed that performance was driven by data characteristics and sequential memory rather than model complexity. Outcome prevalence and temporal structure largely determined both ranking and absolute performance. In particular, IV-to-oral antimicrobial switching was uncommon in our PICU cohorts, consistent with the clinical realities of pediatric critical care (Berthe-Aucejo et al., 2012; McMullan et al., 2016), highlighting the limited transferability of models developed in adult cohorts. One reason for the low prevalence of this target in pediatric data, could be the limited number of oral administrations in PICU. Intravenous administration accounted for 85-98% of antimicrobial courses, likely reflecting sedation, mechanical ventilation, gastrointestinal dysfunction, limited antimicrobial formulations available in both intravenous and oral forms, and the very young age of the cohorts (median age: *PIC* ∼10 months; *MELODY-KISPI* ∼3 months). This was also reflected in treatment initiation, where patients receiving intravenous therapy were younger than those receiving oral therapy in both cohorts: in *PIC*, median age at course

start was 10.8 months (IQR 2.0–48.4) for IV versus 14.6 months (IQR 2.2–60.4) for oral therapy; in the *MELODY-KISPI* cohort, median age was 5.0 months (IQR 0.5–42.3) for IV versus 8.7 months (IQR 0.1–26.8) for oral therapy. Despite its low frequency in the PICUs, IV-to-oral switching remains an important AMS intervention strategy and should be routinely considered during treatment review in critically ill patients (Willems et al., 2021). Oral therapy reduces catheter-related infections and thrombosis, pain and anxiety associated with intravenous access, hospital length of stay, and healthcare costs (Cotter et al., 2021; Keren et al., 2015; Shah et al., 2016; Christensen et al., 2019; Ruebner et al., 2006; Girdwood et al., 2020). Although switching is often limited by clinical instability, impaired gastrointestinal absorption, and restricted oral treatment options, it is feasible in carefully selected children who are clinically improving, able to tolerate enteral therapy, and have an appropriate oral antimicrobial available (McMullan et al., 2016; Willems et al., 2021). The low prevalence of IV-to-oral switching likely contributed to the relatively low F1 scores and AUPRC observed across models despite reasonable AUROC, underlining the challenges of predicting rare clinical events. Nevertheless, the similar data characteristics and model performance observed across both PICU cohorts suggest that development of generalizable pediatric-specific AMS prediction models is feasible.

Second, we noticed that more fine-grained temporal representations provided only modest additional gains beyond 24-hour aggregated features, suggesting that much of the predictive signal was already captured by daily summaries. This supports the use of simpler time-agnostic model development, such as done by Tran-The et al. (2024) and Bolton et al. (2024). Nevertheless, sequence models consistently improved F1 score and AUPRC over tabular approaches, indicating that longitudinal context remains useful for intervention prediction. RAINDROP did not outperform the tabular or sequence-based model classes, providing no clear evidence that graph-based temporal modeling of irregularly sampled events benefits these tasks. Its underperformance may partly reflect our computationally motivated design choice to train the models on independent patient-days, thereby limiting temporal context. However, because RAINDROP also underperformed tabular models with the same patient-day context, limited context alone is unlikely to fully explain the result. Further work is needed to determine whether the difference arises from model complexity, optimization, or data scale. Moreover, the limited performance of CLMBR-T-Base further suggests that representations pretrained on adult data may not transfer directly to pediatric cohorts, consistent with prior evidence of challenges in transferring adult-derived models to children (Sabharwal et al., 2022). Overall, the results highlight the importance of temporal representation alongside model architecture and pretraining population.

Lastly, we observe a trade-off between discrimination and reliability. While sequence-based models improve ranking performance, they are consistently overconfident compared to tabular models. Since trust and reliability are critical for the adoption of an antibiotic CDSS (Laka et al., 2021; Bolton et al., 2025), and finer temporal resolution does not improve calibration, explicit post-hoc calibration may be necessary before deployment.

Overall, these findings suggest that, in the context of PICU AMS intervention prediction, target definition, temporal representation, and calibration are as critical as model choice, and increasing model complexity alone is insufficient for reliable clinical decision support.

Practical deployment considerations. We envision integrating AMS modeling into PICU practice through a dashboard providing daily risk predictions to support prioritization of ongoing antibiotic courses for stewardship review. Since AMS rounds at the University Children's Hospital Zurich are conducted twice weekly, these predictions could prompt ad hoc reviews and support earlier, more systematic stewardship interventions. LightGBM provided the best overall balance of discrimination, calibration, and interpretability. Calibrated GRU models may be preferred when intervention targets are common, identifying patients for review is particularly important, and antibiotic courses are longer (Appendix C.7).

Limitations and future work. Importantly, the prediction targets are derived from retrospective stewardship decisions and therefore represent clinically enacted intervention opportunities. Although these events likely reflect AMS decisions, some may have occurred for other reasons than reducing unnecessary exposure, such as adverse effects or changes in diagnosis. The labels therefore represent proxy intervention events and not confirmed stewardship actions and may also fail to capture missed or earlier opportunities for intervention. A key next step is to assess their clinical validity through review by pediatric infectious disease specialists and PICU clinicians, using the full clinical record and notes to determine the indication, appropriateness, and timing of each observed change. Nonetheless, Tran-The et al. (2024) already achieved better AMS review prioritization using the proxy labels. After validation, a prospective evaluation similar to Bolton et al. (2025) will be required to determine whether model-guided review improves stewardship efficiency, prescribing, and patient outcomes in practice. We further observed that class imbalance substantially affected both discrimination and calibration, limiting the clinical utility of the models. Future work should therefore investigate uncertainty-aware methods, including Bayesian approaches, ensemble Bayesian model averaging, and conformal prediction, to improve calibration and provide reliable confidence estimates for a CDSS (Ruhe et al., 2024; Koumantakis et al., 2026; Kalita et al., 2026). Finally, our framework relies on fixed temporal aggregation windows, which may obscure clinically relevant patterns happening at a finer-grained resolution, while missing data are handled primarily through imputation. Although a recent study on PICU data found that LOCF imputation performed competitively and often outperformed more complex methods (Digitale et al., 2025), it may be less suitable during periods of rapid physiological change (Lachin, 2016), commonly encountered in the PICU. Future work should therefore explore missingness-aware temporal models, such as GRU-D (Che et al., 2018), which explicitly model informative missingness and have shown promise for clinical time-series prediction(Lipton et al., 2016).

7. Acknowledgments

This research project and related results were made possible with the support of the NOMIS foundation. Niklas Raehse and Daphné Chopard are both supported by the grant "Using artificial intelligence to improve early recognition and treatment of critically ill children (AI in pediatric ICU)" from the NOMIS foundation. We would also like to thank Catherine Jutzeler from ETH Zurich for providing the supervisory and institutional infrastructure to enable this project and Philipp Baumann from the University Children's Hospital Zurich, for his support in contextualizing the study and providing valuable clinical insights.

References

Qalab Abbas, Aysha Ul Haq, Ritika Kumar, Syed Asad Ali, Karim Hussain, and Sadia Shakoor. Evaluation of antibiotic use in pediatric intensive care unit of a developing country. *Indian Journal of Critical Care Medicine*, 20(5):291–294, 2016. doi: 10.4103/0972-5229.182197. 2, 8

Bert Arnrich, Edward Choid, Jason A. Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom J. Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. Medical Event Data Standard (MEDS): Facilitating Machine Learning for Health. In *ICLR 2024 Workshop on Learning from Time Series For Health*, March 2024. URL https://github.com/Medical-Event-Data-Standard/meds. 36

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 6, 34, 35

Tamar F Barlam, Sara E Cosgrove, Lilly M Abbo, et al. Implementing an antibiotic stewardship program: guidelines by the infectious diseases society of america and the society for healthcare epidemiology of america. *Clinical Infectious Diseases*, 62(10):e51–e77, 2016. 1

Mathieu Beaudoin, Froduald Kabanza, Vincent Nault, and Louis Valiquette. An antimicrobial prescription surveillance system that learns from experience. *AI Magazine*, 35(1):15–15, 2014. doi: 10.1609/aimag.v35i1.2500. 3

Aurore Berthe-Aucejo, Martine Postaire, Alaa Cheikhelard, Jean Ralph Zahar, and Philippe Bourget. Antibiotic treatment of appendicular peritonitis in children: is the oral route done? *Archives de Pediatrie: Organe Officiel de la Societe Francaise de Pediatrie*, 19(12):1303–1307, 2012. 2, 14

William J. Bolton, Richard Wilson, Mark Gilchrist, et al. Machine learning and synthetic outcome estimation for individualised antimicrobial cessation. *Frontiers in Digital Health*, 4:997219, 2022. doi: 10.3389/fdgth.2022.997219. 3

William J. Bolton, Richard Wilson, Mark Gilchrist, Pantelis Georgiou, Alison Holmes, and Timothy M. Rawson. Personalising intravenous to oral antibiotic switch decision making through fair interpretable machine learning. *Nature Communications*, 15(1):506, 2024. doi: 10.1038/s41467-024-44740-2. 3, 6, 10, 15, 41

William J. Bolton, Richard Wilson, Mark Gilchrist, Pantelis Georgiou, Alison Holmes, and Timothy M. Rawson. The impact of artificial intelligence-driven decision support on uncertain antimicrobial prescribing: A randomised, multimethod study. *The Lancet Digital Health*, 7(11):100912, 2025. doi: 10.1016/j.landig.2025.100912. 3, 13, 15, 16

Nora Bruns, Anna-Lisa Sorg, Ursula Felderhoff-Müser, Christian Dohna-Schwake, and Andreas Stang. Administrative data in pediatric critical care research—Potential, challenges, and future directions. *Frontiers in Pediatrics*, 10, September 2022. ISSN 2296-2360. doi: 10.3389/fped.2022.1014094. 2, 4

Rachel J. Bystritsky, Alex Beltran, Albert T. Young, Andrew Wong, Xiao Hu, and Sarah B. Doernberg. Machine learning for the prediction of antimicrobial stewardship intervention in hospitalized patients receiving broad-spectrum agents. *Infection Control & Hospital Epidemiology*, 41(9):1022–1027, 2020. doi: 10.1017/ice.2020.213. 3, 6, 14

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1):6085, April 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24271-9. 16

Kathleen Chiotos and Pranita D. Tamma. Antibiotics: How can we make it as easy to stop as it is to start? *Clinical Microbiology and Infection*, 26(12):1600–1601, December 2020. ISSN 1198-743X. doi: 10.1016/j.cmi.2020.08.029. 2

Kathleen Chiotos, Jennifer Blumenthal, Juri Boguniewicz, Debra L. Palazzi, Erika L. Stalets, Jessica H. Rubens, Pranita D. Tamma, Stephanie S. Cabler, Jason Newland, Hillary Crandall, Emily Berkman, Robert P. Kavanagh, Hannah R. Stinson, and Jeffrey S. Gerber. Antibiotic indications and appropriateness in the pediatric intensive care unit: A 10-center point prevalence study. *Clinical Infectious Diseases*, 76(3):e1021–e1030, 2022. doi: 10.1093/cid/ciac698. 2, 8

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 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)*, pages 1724–1734, 2014. doi: 10.3115/v1/D14-1179. 6, 34, 35

Eric W Christensen, Alicen Burns Spaulding, William F Pomputius, and Steven P Grapentine. Effects of Hospital Practice Patterns for Antibiotic Administration for Pneumonia on Hospital Lengths of Stay and Costs. *Journal of the Pediatric Infectious Diseases Society*, 8(2):115–121, May 2019. ISSN 2048-7207. doi: 10.1093/jpids/piy003. 15

Sarah LN Clarke, Kevon Parmesar, Moin A Saleem, and Athimalaipet V Ramanan. Future of machine learning in paediatrics. *Archives of Disease in Childhood*, 107(3):223–228, 2022. 4

Jillian M Cotter, Matt Hall, Sonya Tang Girdwood, John R Stephens, Jessica L Markham, James C Gay, and Samir S Shah. Opportunities for Stewardship in the Transition From Intravenous to Enteral Antibiotics in Hospitalized Pediatric Patients. *Journal of Hospital Medicine*, 16(2):70–76, February 2021. ISSN 1553-5592. doi: 10.12788/jhm.3538. 15

Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 233–240, 2006. doi: 10.1145/1143844.1143874. 8

Stan Deresinski. Principles of Antibiotic Therapy in Severe Infections: Optimizing the Therapeutic Approach by Use of Laboratory and Clinical Data. *Clinical Infectious Diseases*, 45(Supplement_3):S177–S183, September 2007. ISSN 1058-4838. doi: 10.1086/519472. 2

Jean Digitale, Deborah Franzon, Mark J. Pletcher, Charles E. McCulloch, and Efstathios D. Gennatas. Methods for Addressing Missingness in Electronic Health Record Data for Clinical Prediction Models: Comparative Evaluation. *JMIR Medical Informatics*, 13(1): e79307, November 2025. doi: 10.2196/79307. 16

Jean Mundahl Engels and Paula Diehr. Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology*, 56(10):968–976, October 2003. ISSN 0895-4356. doi: 10.1016/S0895-4356(03)00170-7. 6, 32

Idris V. R. Evans, Gary S. Phillips, Elizabeth R. Alpern, Derek C. Angus, Marcus E. Friedrich, Niranjan Kissoon, Stanley Lemeshow, Mitchell M. Levy, Margaret M. Parker, Kathleen M. Terry, R. Scott Watson, Scott L. Weiss, Jerry Zimmerman, and Christopher W. Seymour. Association Between the New York Sepsis Care Mandate and In-Hospital Mortality for Pediatric Sepsis. *JAMA*, 320(4):358–367, July 2018. ISSN 0098-7484. doi: 10.1001/jama.2018.9071. 2

Umberto Fanelli, Vincenzo Chiné, Marco Pappalardo, Pierpacifico Gismondi, and Susanna Esposito. Improving the Quality of Hospital Antibiotic Use: Impact on Multidrug-Resistant Bacterial Infections in Children. *Frontiers in Pharmacology*, 11, May 2020a. ISSN 1663-9812. doi: 10.3389/fphar.2020.00745. 2

Umberto Fanelli, Marco Pappalardo, Vincenzo Chinè, Pierpacifico Gismondi, Cosimo Neglia, Alberto Argentiero, Adriana Calderaro, Andrea Prati, and Susanna Esposito. Role of Artificial Intelligence in Fighting Antimicrobial Resistance in Pediatrics. *Antibiotics*, 9(11):767, November 2020b. ISSN 2079-6382. doi: 10.3390/antibiotics9110767. 2, 3

Michael-John Fay and Penelope A Bryant. Antimicrobial stewardship in children: Where to from here? *Journal of Paediatrics and Child Health*, 56(10):1504–1507, 2020. 3

Haile Mekonnen Fenta, Temesgen T Zewotir, Saloshni Naidoo, Rajen N Naidoo, and Henry Mwambi. Factors of acute respiratory infection among under-five children across sub-saharan african countries using machine learning approaches. *Scientific Reports*, 14(1): 15801, 2024. 4

Arden W Forrey, Clement J Mcdonald, Georges DeMoor, Stanley M Huff, Dennis Leavelle, Diane Leland, Tom Fiers, Linda Charles, Brian Griffin, Frank Stalling, et al. Logical observation identifier names and codes (loinc) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical chemistry*, 42 (1):81–90, 1996. 10, 37

Jeffrey S. Gerber, Adam L. Hersh, Matthew P. Kronman, Jason G. Newland, Rachael K. Ross, and Talene A. Metjian. Development and application of an antibiotic spectrum index for benchmarking antibiotic selection patterns across hospitals. *Infection Control & Hospital Epidemiology*, 38(8):993–997, 2017. doi: 10.1017/ice.2017.114. 7, 28, 33

Daniele Roberto Giacobbe, Cristina Marelli, Sabrina Guastavino, Sara Mora, Nicola Rosso, Alessio Signori, Cristina Campi, Mauro Giacomini, and Matteo Bassetti. Explainable

and interpretable machine learning for antimicrobial stewardship: Opportunities and challenges. *Clinical Therapeutics*, 46(6):474–480, 2024. doi: 10.1016/j.clinthera.2024.02.010. 2

Sonya C Tang Girdwood, Maria N Sellas, Joshua D Courter, Brianna Liberio, Michael J Tchou, Lisa E Herrmann, Maya L Dewan, Angela M Statile, and Ndidi I Unaka. Improving the Transition of Intravenous to Enteral Antibiotics in Pediatric Patients with Pneumonia or Skin and Soft Tissue Infections. *Journal of Hospital Medicine*, 15(1):9–15, 2020. ISSN 1553-5606. doi: 10.12788/jhm.3253. 15

Ary L. Goldberger, Luis A. N. Amaral, Jeffrey M. Glass, Leon andHausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23):e215–e220, june 2000. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215. 26

Katherine E. Goodman, Emily L. Heil, Kimberly C. Claeys, Mary Banoub, and Jacqueline T. Bork. Real-world antimicrobial stewardship experience in a large academic medical center: Using statistical and machine learning approaches to identify intervention "hotspots" in an antibiotic audit and feedback program. *Open Forum Infectious Diseases*, 9(7):ofac289, 2022. doi: 10.1093/ofid/ofac289. 3, 6, 14

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, May 2024. 6, 35

Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, and Lillian Sung. A multi-center study on the adaptability of a shared foundation model for electronic health records. *NPJ Digital Medicine*, 7(1):171, 2024. 6, 35

Arash Hajisafi, Maria Despoina Siampou, Bita Azarijoo, Zhen Xiong, and Cyrus Shahabi. WaveGNN: Integrating Graph Neural Networks and Transformers for Decay-Aware Classification of Irregular Clinical Time-Series, December 2025. 7

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747. 8

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. 6, 34, 35

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 01 2025. doi: 10.1038/s41586-024-08328-6. URL https://www.nature.com/articles/s41586-024-08328-6. 6, 34

Daniel T. Ilges, David J. Ritchie, Tamara Krekel, Elizabeth A. Neuner, and Scott T. Micek. Spectrum scores: Toward a better definition of de-escalation. *Infection Control & Hospital*

Epidemiology, 44(6):938–940, June 2023. ISSN 0899-823X, 1559-6834. doi: 10.1017/ice.2022.207. 8

Amit Kalita, Avirup Chattopadhyay, Madhushree Bhattacharjee, and Kaushik Das. A Simulation-Based Validation Framework for Uncertainty-Aware Clinical Triage: Conformal Prediction and Ensemble Learning Applied to a Synthetic ICU Benchmark Cohort, June 2026. ISSN 3067-2007. 16

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 6, 34

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. doi: 10.1109/CVPR.2018.00781. 35

Ron Keren, Samir S. Shah, Rajendu Srivastava, Shawn Rangel, Michael Bendel-Stenzel, Nada Harik, John Hartley, Michelle Lopez, Luis Seguias, Joel Tieder, Matthew Bryan, Wu Gong, Matt Hall, Russell Localio, Xianqun Luan, Rachel deBerardinis, Allison Parker, and for the Pediatric Research in Inpatient Settings Network. Comparative Effectiveness of Intravenous vs Oral Antibiotics for Postdischarge Treatment of Acute Osteomyelitis in Children. *JAMA Pediatrics*, 169(2):120–128, February 2015. ISSN 2168-6203. doi: 10.1001/jamapediatrics.2014.2822. 15

Maher R. Khdour, Hussein O. Hallak, Mamoon A. Aldeyab, Mowaffaq A. Nasif, Aliaa M. Khalili, Ahamad A. Dallashi, Mohammad B. Khofash, and Michael G. Scott. Impact of antimicrobial stewardship programme on hospitalized patients at the intensive care unit: A prospective audit and feedback study. *British Journal of Clinical Pharmacology*, 84(4): 708–715, 2018. ISSN 1365-2125. doi: 10.1111/bcp.13486. 2

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. 32

Emanuele Koumantakis, Konstantina Remoundou, Carmen Fava, Ioanna Roussaki, Alessia Visconti, and Paola Berchialla. Improving machine learning and deep learning models for 30-day ICU readmission prediction using Ensemble Bayesian Model Averaging, May 2026. ISSN 3067-2007. 16

John M. Lachin. Fallacies of Last Observation Carried Forward. *Clinical trials (London, England)*, 13(2):161–168, April 2016. ISSN 1740-7745. doi: 10.1177/1740774515602688. 16

Mah Laka, Adriana Milazzo, and Tracy Merlin. Factors That Impact the Adoption of Clinical Decision Support Systems (CDSS) for Antibiotic Management. *International Journal of Environmental Research and Public Health*, 18(4):1901, January 2021. ISSN 1660-4601. doi: 10.3390/ijerph18041901. 15

Bongjin Lee, Hyun Jung Chung, Hyun Mi Kang, Do Kyun Kim, and Young Ho Kwak. Development and validation of machine learning-driven prediction model for serious bacterial infection among febrile children in emergency departments. *PLoS One*, 17(3):e0265500, 2022. 4

Haomin Li, Xian Zeng, and Gang Yu. Paediatric Intensive Care database. *PhysioNet*, November 2020. doi: 10.13026/32x9-wv38. URL https://doi.org/10.13026/32x9-wv38. Version 1.1.0. 9, 26

Zachary C Lipton, David C Kale, Randall Wetzel, et al. Modeling missing data in clinical time series with rnns. 56(56):253–270, 2016. 16

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Modeling missing data in clinical time series with rnns. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 41494–41519. PMLR, October 2025. 7

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1930–1939, New York, NY, USA, July 2018. Association for Computing Machinery. ISBN 978-1-4503-5552-0. doi: 10.1145/3219819.3220007. 35

Blake Martin, Peter E DeWitt, Halden F Scott, Sarah Parker, and Tellen D Bennett. Machine learning approach to predicting absence of serious bacterial infection at picu admission. *Hospital pediatrics*, 12(6):590–603, 2022. 4

Peter McCullagh. *Generalized linear models*. Routledge, 2019. 6

Brendan J McMullan, David Andresen, Christopher C Blyth, Minyon L Avent, Asha C Bowen, Philip N Britton, Julia E Clark, Celia M Cooper, Nigel Curtis, Emma Goeman, et al. Antibiotic duration and timing of the switch from intravenous to oral route for bacterial infections in children: systematic review and guidelines. *The Lancet Infectious Diseases*, 16(8):e139–e152, 2016. 2, 14, 15

Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. Normalized names for clinical drugs: Rxnorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4):441–448, 2011. 37

Hadar Neuman, Paul Forsythe, Atara Uzan, Orly Avni, and Omry Koren. Antibiotics in early life: dysbiosis and the damage done. *FEMS Microbiology Reviews*, 42(4):489–499, 2018. doi: 10.1093/femsre/fuy018. 2

Mathupanee Oonsivilai, Yin Mo, Nantasit Luangasanatip, Yoel Lubell, Thyl Miliya, Pisey Tan, Lorn Loeuk, Paul Turner, and Ben S. Cooper. Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia. *Wellcome Open Research*, 3:131, October 2018. ISSN 2398-502X. doi: 10.12688/wellcomeopenres.14847.1. 4

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 31, 34

Loria A. Pollack and Arjun Srinivasan. Core Elements of Hospital Antibiotic Stewardship Programs From the Centers for Disease Control and Prevention. *Clinical Infectious Diseases*, 59(suppl_3):S97–S100, October 2014. ISSN 1058-4838. doi: 10.1093/cid/ciu542. 2

Alessandra Romandini, Arianna Pani, Paolo Andrea Schenardi, Giulia Angela Carla Pattarino, Costantino De Giacomo, and Francesco Scaglione. Antibiotic resistance in pediatric infections: Global emerging threats, predicting the near future. *Antibiotics*, 10(4): 393, 2021. doi: 10.3390/antibiotics10040393. 2

Rebecca Ruebner, Ron Keren, Susan Coffin, Jaclyn Chu, David Horn, and Theoklis E. Zaoutis. Complications of central venous catheters used for the treatment of acute hematogenous osteomyelitis. *Pediatrics*, 117(4):1210–1215, April 2006. ISSN 1098-4275. doi: 10.1542/peds.2005-1465. 15

David Ruhe, Giovanni Cinà, Michele Tonutti, Daan de Bruin, and Paul Elbers. Bayesian Modelling in Practice: Using Uncertainty to Improve Trustworthiness in Medical Applications, July 2024. 16

Paul Sabharwal, Jillian H. Hurst, Rohit Tejwani, Kevin T. Hobbs, Jonathan C. Routh, and Benjamin A. Goldstein. Combining adult with pediatric patient data to develop a clinical decision support tool intended for children: Leveraging machine learning to model heterogeneity. *BMC Medical Informatics and Decision Making*, 22(1):84, March 2022. ISSN 1472-6947. doi: 10.1186/s12911-022-01827-4. 15

Neel Shah, Ahmed Arshad, Monty B. Mazer, Christopher L. Carroll, Steven L. Shein, and Kenneth E. Remy. The use of machine learning and artificial intelligence within pediatric critical care. *Pediatric Research*, 93(2):405–412, January 2023. ISSN 1530-0447. doi: 10.1038/s41390-022-02380-6. 3, 4

Samir S. Shah, Rajendu Srivastava, Susan Wu, Jeffrey D. Colvin, Derek J. Williams, Shawn J. Rangel, Waheeda Samady, Suchitra Rao, Christopher Miller, Cynthia Cross, Caitlin Clohessy, Matthew Hall, Russell Localio, Matthew Bryan, Gong Wu, Ron Keren, and for the Pediatric Research in Inpatient Settings Network. Intravenous Versus Oral Antibiotics for Postdischarge Treatment of Complicated Pneumonia. *Pediatrics*, 138(6): e20161692, December 2016. ISSN 0031-4005. doi: 10.1542/peds.2016-1692. 15

Shah Lab. FEMR: Framework for Electronic Medical Records. https://github.com/som-shahlab/femr, 2026. Version 0.2.3. 36

Shivani Sivasankar, Jennifer L Goldman, and Mark A Hoffman. Variation in antibiotic resistance patterns for children and adults treated at 166 non-affiliated US facilities using EHR data. *JAC-Antimicrobial Resistance*, 5(1):dlac128, 2023. doi: 10.1093/jacamr/dlac128. 2

Tam Tran-The, Eunjeong Heo, Sanghee Lim, Yewon Suh, Kyu-Nam Heo, Eunkyung Euni Lee, Ho-Young Lee, Eu Suk Kim, Ju-Yeun Lee, and Se Young Jung. Development of machine learning algorithms for scaling-up antibiotic stewardship. *International Journal of Medical Informatics*, 181:105300, 2024. doi: 10.1016/j.ijmedinf.2023.105300. 2, 3, 6, 7, 8, 10, 15, 16, 41

Cornelis J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 2nd edition, 1979. 8

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 6, 34, 35

Tom Velez, Oluwakemi Badaki-Makun, Danielle Hirsch, Danielle Claire Mercurio, Holly Depinet, Maya Dewan, Rishikesan Kamaleswaran, Jocelyn Grunwell, Maria Triantafyllou, Fehima Abdelrahman, Charles Macias, and Ioannis Koutroulis. Early prediction of antibiotic need and bacteremia risk in non-immunocompromised pediatric emergency patients using machine learning. *Pediatric Research*, pages 1–9, December 2025. ISSN 1530-0447. doi: 10.1038/s41390-025-04656-z. 4

Scott L. Weiss, Mark J. Peters, Waleed Alhazzani, Michael S. D. Agus, Heidi R. Flori, David P. Inwald, Simon Nadel, Luregn J. Schlapbach, Robert C. Tasker, Andrew C. Argent, Joe Brierley, Joseph Carcillo, Enitan D. Carrol, Christopher L. Carroll, Ira M. Cheifetz, Karen Choong, Jeffry J. Cies, Andrea T. Cruz, Daniele De Luca, Akash Deep, Saul N. Faust, Claudio Flauzino De Oliveira, Mark W. Hall, Paul Ishimine, Etienne Javouhey, Koen F. M. Joosten, Poonam Joshi, Oliver Karam, Martin C. J. Kneyber, Joris Lemson, Graeme MacLaren, Nilesh M. Mehta, Morten Hylander Møller, Christopher J. L. Newth, Trung C. Nguyen, Akira Nishisaki, Mark E. Nunnally, Margaret M. Parker, Raina M. Paul, Adrienne G. Randolph, Suchitra Ranjit, Lewis H. Romer, Halden F. Scott, Lyvonne N. Tume, Judy T. Verger, Eric A. Williams, Joshua Wolf, Hector R. Wong, Jerry J. Zimmerman, Niranjan Kissoon, and Pierre Tissieres. Surviving sepsis campaign international guidelines for the management of septic shock and sepsis-associated organ dysfunction in children. *Intensive Care Medicine*, 46(1):10–67, February 2020. ISSN 1432-1238. doi: 10.1007/s00134-019-05878-6. 2

Nedra S. Whitehead, Laurina O. Williams, Sreelatha Meleth, Sara M. Kennedy, Nneka Ubaka-Blackmoore, Sharon M. Geaghan, James H. Nichols, Patrick Carroll, Michael T. McEvoy, Julie Gayken, Dennis J. Ernst, Christine Litwin, Paul Epner, Jennifer Taylor, and Mark L. Graber. Interventions to prevent iatrogenic anemia: A Laboratory Medicine Best Practices systematic review. *Critical Care*, 23:278, August 2019. ISSN 1364-8535. doi: 10.1186/s13054-019-2511-9. 4

WHO Collaborating Centre for Drug Statistics Methodology. Anatomical therapeutic chemical (ATC) classification, 2024. URL https://www.whocc.no/atc_ddd_index/. Accessed: 2026-04-14. 10, 33, 37

Rand R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, December 2011. ISBN 978-0-12-387015-5. 31

Jef Willems, Eline Hermans, Petra Schelstraete, Pieter Depuydt, and Pieter De Cock. Optimizing the Use of Antibiotic Agents in the Pediatric Intensive Care Unit: A Narrative Review. *Paediatric Drugs*, 23(1):39–53, 2021. ISSN 1174-5878. doi: 10.1007/s40272-020-00426-y. 1, 15

World Health Organization. Global action plan on antimicrobial resistance. https://iris.who.int/server/api/core/bitstreams/1a487887-e162-46a0-8aef-802907c66070/content, 2015. 1

World Health Organization. Antimicrobial stewardship programmes in health-care facilities in low- and middle-income countries: A practical toolkit. https://iris.who.int/server/api/core/bitstreams/e2e59ce4-c373-4066-9ff7-a0cce3e6ccdd/content, 2019. 2

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. 2023. 6, 35

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836, 2020. 35

Xian Zeng, Gang Yu, Yang Lu, Linhua Tan, Xiujing Wu, Shanshan Shi, Huilong Duan, Qiang Shu, and Haomin Li. PIC, a paediatric-specific intensive care database. *Scientific Data*, 7(1):14, 2020. doi: 10.1038/s41597-020-0355-4. 2

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations, ICLR*, 2022. 6, 39

Appendix A. Data

A.1. Cohorts

A.1.1. PIC

PIC version 1.1.0 (Li et al., 2020) was retrieved from Physionet (Goldberger et al., 2000).

A.1.2. MELODY-KISPI

The PICU dataset from the University Children's Hospital Zurich, referred to as MELODY-KISPI, includes patients younger than 18 years who were admitted between 1 January 2015 and 31 August 2023. The retrospective use of these data was approved by the Cantonal Ethics Committee of Zurich under the MELODY project ("Machine Learning in the Paediatric Intensive Care Unit: Development of Risk Prediction Models"; BASEC ID 2023-02077). In accordance with the ethics approval, patients with a documented refusal of general consent were not included.

A.2. Extracted data

Table 1: List of variables extracted from raw datasets and aggregated features included in the patient-day dataset.

Raw variable	**Source Table**	**Aggregated features**	**Notes**
Patient and admission context			
Age at admission (years)	PATIENTS		
Gender	PATIENTS		One-Hot Encoding (OHE)
Ethnicity	ADMISSIONS		OHE
Admission type	ADMISSIONS		OHE
Insurance	ADMISSIONS		OHE
Current LOS hours at day start	ADMISSIONS		Visit trajectory context
Cumulative ICU hours	ADMISSIONS	icu_hours_cum	Visit trajectory context
Physiological measurements *Vital signs*			
Temperature	CHARTEVENTS	sum, mean, n, min, max	PIC ITEMID: 1001
Heart Rate	CHARTEVENTS	sum, mean, n, min, max	PIC ITEMID: 1003
Respiratory Rate	CHARTEVENTS	sum, mean, n, min, max	PIC ITEMID: 1004
Oxygen Saturation	CHARTEVENTS	sum, mean, n, min, max	PIC ITEMID: 1006
Diastolic Pressure	CHARTEVENTS	sum, mean, n, min, max	PIC ITEMID: 1015
Systolic Pressure	CHARTEVENTS	sum, mean, n, min, max	PIC ITEMID: 1016
Height	CHARTEVENTS	sum, mean, n, min, max	PIC ITEMID: 1013
Weight	CHARTEVENTS	sum, mean, n, min, max	PIC ITEMID: 1014

Continued on next page

Table 1 – continued

Raw variable	Source Table	Aggregated features	Notes
Trend features			
Vital trend features	CHARTEVENTS	delta_prev_day_max, delta_abx_start_max, var_3d	Built for TEMP, RR, HR, SBP, DBP, SPO2. Day-to-day differences, deviations from antibiotic start, 3-day variance.
Laboratory values			
Albumin	LABEVENTS	sum, mean, n	LOINC: 1751-1
Alanine Aminotransferase (ALT)	LABEVENTS	sum, mean, n	LOINC: 1742-6
Lactate Dehydrogenase	LABEVENTS	sum, mean, n	LOINC: 2532-0
Hemoglobin	LABEVENTS	sum, mean, n	LOINC: 718-7
Red Blood Cells	LABEVENTS	sum, mean, n	LOINC: 789-8
Lactate	LABEVENTS	sum, mean, n	LOINC: 2532-0
Sodium Whole Blood	LABEVENTS	sum, mean, n	LOINC: 2947-0
PCO2	LABEVENTS	sum, mean, n	LOINC: 11557-6
Ph	LABEVENTS	sum, mean, n	LOINC: 11558-4
PO2	LABEVENTS	sum, mean, n	LOINC: 11556-8
Oxygen Saturation	LABEVENTS	sum, mean, n	LOINC: 20564-1
White Blood Count (WBC)	LABEVENTS	sum, mean, n, recent_90d	LOINC: 26464-8
C-Reactive Protein (CRP)	LABEVENTS	sum, mean, n, recent_90d	LOINC: 32264
Procalcitonin	LABEVENTS	sum, mean, n, recent_90d	LOINC: 33959-8
Antibiotic exposure features			
Primary ATC code	PRESCRIPTIONS		Longest running antibiotic course class
Antibiotic medication features	PRESCRIPTIONS	distinct_atc_ever, concurrent_atc_day, any_oral_rx_last_3d	Antibiotic exposure dynamics: classes, concurrent count, recent oral use.

A.3. Description of data and features

Patient and admission context. Static and semi-static features include age at admission, sex, ethnicity, admission type, and insurance status (Table 1). These features capture baseline patient characteristics and care context. These are encoded using one-hot representations where applicable. We also include visit-level context such as length of stay at the start of the day and cumulative ICU time to capture disease progression and care trajectory.

Physiological measurements. Time-varying clinical variables include vital signs (e.g., temperature, heart rate, respiratory rate, blood pressure, and oxygen saturation) and laboratory measurements (e.g., C-reactive protein (CRP), white blood cell count (WBC), and lactate). These features capture the evolving physiological state of the patient. For the 24h aggregated data, we compute summary statistics over the pre-day window, including sum, mean, count, and where appropriate, minimum and maximum. For sparsely recorded inflammatory markers (e.g. CRP, procalcitonin), we additionally include longer-term summaries (e.g., most recent value in last 90 days) to capture baseline trends. To incorporate short-term dynamics within the tabular representation, we also compute engineered trend features for key vital signs (e.g. temperature, heart rate, respiratory rate, blood pressure, oxygen saturation), including day-to-day differences, deviations from antibiotic start, and short-window variability (e.g. 3-day (lookback) variance).

Antibiotic exposure features. We include features describing antibiotic treatment patterns, including current and prior antibiotic classes, route of administration, and concurrent therapies. These features provide context for the AMS interventions. We derive features describing antibiotic usage patterns from prescription data, including current and historical antibiotic classes (ATC codes), number of concurrent antibiotics, prior exposure to high-priority antibiotics, and recent oral antibiotic use. These features aim to capture treatment dynamics relevant for stewardship decisions. To support the cross-cohort comparison and the definition of the de-escalation outcome, we map antibiotics to ATC codes, product names, and their ASI scores (Gerber et al., 2017) in Table 2 below.

A.4. Antibiotics mapping

Table 2: Antibiotics tracked for creating patient-days dataset. Mappings to the Anatomical Therapeutic Chemical (ATC) standard were used for extraction from the hospital information system of the University Children's Hospital Zurich. A mapping to the Antibiotic Spectrum Index (ASI) score (Gerber et al., 2017) and the corresponding product names in the PRESCRIPTIONS table of *PIC* is included.

ATC	Label	ASI	PIC product name
J02AX01	5-fluorocytosine		—
J01FA09	acetylspiramycin	4	Clarithromycin Granule for Oral Suspension; Clarithromycin Tablets
J01GB06	amikacin	5	Amikacin Sulfate Injection
J01CA04	amoxicillin	2	Amoxicillin Sodium and Clavulanate Potassium for Injection; Amoxicillin and Clavulanate Potassium Granules; Amoxicillin Sodium and Sulbactam Sodium for Injection; Amoxicillin and Clavulanate Potassium Tablets
J01CR02	amoxicillin/clavulanic acid	6	—
J02AA01	amphotericin b		Amphotericin B Liposome for Injection

Continued on next page

Table 2 — continued

ATC	Label	ASI	PIC product name
J01CA01	ampicillin	2	Ampicillin Sodium and Sulbactam Sodium for Injection; Ampicillin Sodium for Injection
J01CR01	ampicillin/sulbactam	6	—
J01FA10	azithromycin	4	Azithromycin for Injection; Azithromycin for Suspension; Azithromycin Lactobionate for Injection; Azithromycin Tablets
J01DF01	aztreonam	5	Aztreonam for Injection
J01DC04	cefaclor	4	Cefaclor for Suspension; Cefaclor Capsules
J01DB04	cefazolin	3	—
J01DE01	cefepime	6	Cefepime Hydrochloride For Injection
J01DD62	cefoperazone/sulbactam	6	Cefoperazone Sodium and Sulbactam Sodium for Injection
J01DD01	cefotaxime	5	Cefotaxime Sodium for Injection
J01DC05	cefotetan	4	—
J01DC01	cefoxitin	4	—
	cefoxitin screen		—
J01DD02	ceftazidime	4	Ceftazidime for Injection
J01DD04	ceftriaxone	5	Ceftriaxone Sodium for Injection
J01DC02	cefuroxime	4	Cefuroxime Sodium For Injection; Cefuroxime Sodium for Injection; Cefuroxime Axetil Tablets
J01BA01	chloramphenicol	5	Chloramphenicol Injection
J01MA02	ciprofloxacin	8	Ciprofloxacin and Sodium Chloride Injection
J01FA09	clarithromycin	4	Clarithromycin Granule for Oral Suspension; Clarithromycin Tablets
J01FF01	clindamycin	4	Clindamycin Phosphate for Injection
J01EE01	compound sulfamethoxazole	4	Compound Sulfamethoxazole Tablets; Compound Sulfamethoxazole Injection
J01AA02	doxycycline	5	—
J01DH03	erbepenem	9	Ertapenem for Injection
J01FA01	erythromycin	2	Erythromycin Lactobionate for Injection; Erythromycin Ointment; Erythromycin Eye Ointment
	extended spectrum beta-lactamase detection		—
J02AC01	fluconazole		Fluconazole Capsules; Fluconazole and Sodium Chloride Injection; Fluconazole Injection
J01MA16	gatifloxacin	9	—

Continued on next page

Table 2 — continued

ATC	Label	ASI	PIC product name
J01GB03	gentamicin	5	Gentamycin Sulfate and Procaine Hydrochloride Capsules; Gentamycin Sulfate,Procaine Hydrochloride and Vitamin B
J01GB03	gentamicin 500	5	Gentamycin Sulfate and Procaine Hydrochloride Capsules; Gentamycin Sulfate,Procaine Hydrochloride and Vitamin B
J01DH51	imipenem	10	Imipenem and Cilastatin Sodium for Injection
	induced clindamycin resistance		—
J02AC02	itraconazole		—
J01FA07	josamycin	4	—
	koalaranin		—
J01MA12	levofloxacin	9	Levofloxacin Eye Drops; Levofloxacin and Sodium Chloride Injection; Levofloxacin Hydrochloride Ear Drops
J01XX08	linezolid	5	Linezolid Injection; Linezolid Tablets
J01DH02	meropenem	10	Meropenem for Injection
	methicillin-resistant staphylococcus		—
J01AA08	minocycline	5	—
J01MA14	moxifloxacin	9	—
J01XE01	nitrofurantoin	2	Nitrofurantoin Enteric-coated Tablets
J01MA01	ofloxacin	9	Ofloxacin Eye Ointment; Ofloxacin Gel
J01CF04	oxacillin	1	Oxacillin Sodium for Iniection
J01CE01	penicillin	2	Benzylpenicillin Sodium for Injection; Benzathine Benzylpenicillin for Injection
J01CA12	piperacillin	5	Piperacillin Sodium and Tazobactam Sodium Injection; Piperacillin Sodium and Tazobactam Sodium for Injection; Piperacillin Sodium And Sulbactam Sodium For Injection
J01CR05	piperacillin/tazobactam	8	—
J01XB02	polymyxin b	4	Polymyxin B Sulfate for injection
J01FG02	quinupristin/dalfopristin	5	—
J04AB02	rifampin	3	Rifampicin Capsules; Rifampicin for Eye Use
J01FA06	roxithromycin	4	—
J01MA09	sparfloxacin	8	—
J01GA01	streptomycin 2000	5	—
J01FA15	telithromycin	4	—
J01AA07	tetracycline	5	—
J01CA13	ticarcillin	5	—
J01AA12	tigecycline	13	Tigecycline for Injection

Continued on next page

Table 2 — continued

ATC	Label	ASI	PIC product name
J01GB01	tobramycin	5	Tobramycin and Dexamethasone Ophthalmic Ointment; Tobramycin Sulfate Injection; Tobramycin Dexamethasone Eye Drops; Tobramycin Eye Drops
J01XA01	vancomycin	5	Vancomycin Hydrochloride for Intra Venous; Norvancomycin Hydrochloride for Injection
J02AC03	voriconazole		Voriconazole for Injection; Voriconazole Tablets
	beta-lactamase		—

Appendix B. Implementation details

B.1. Data preprocessing

Numeric variables are clipped at the 1st and 99th percentile thresholds using Winsorization (Wilcox, 2011) to reduce the impact of erroneous measurements and recording artifacts while preserving clinically plausible ranges. Undefined aggregates (e.g., no measurements within a window) are assigned neutral default values (e.g., zero counts), consistent with feature design. Data is further standardized using sklearn's (Pedregosa et al., 2011) StandardScaler unless stated otherwise.

B.2. Construction of antibiotic patient-day data

The datasets are organized at the patient-day level: each row is one calendar day on the PICU with an overlapping antibiotic course. Antibiotic prescriptions within 24 hours are merged into clinically contiguous courses for each admission and antibiotic. Patient-day rows are enumerated only from those merged intervals. Multi-day antibiotic coverage therefore yields multiple rows; days without qualifying overlap are excluded. Merged 'same-day' antibiotic courses shorter than 24 hours are dropped as these are mostly prophylactic and the duration is too short for prospective audit and feedback. If several courses overlap the same day, all are considered when attaching flags (or across overlapping courses) rather than collapsing to a single agent. Prediction is defined at the start of each patient-day (DAY_START), default 8:00AM, and all features are constructed using only information available up to this time point to prevent information leakage (DAY_START-W,DAY_START), with a default 24h aggregation window.

B.2.1. 24h resolution

Time-varying clinical variables, including vital signs, laboratory measurements, and treatment features, are aggregated within a fixed temporal windows anchored to the start of each patient-day. We compute clinically interpretable summary statistics (e.g., mean, extrema, and most recent value) to represent the current patient state. Treatment-related features capture antibiotic regimen characteristics such as route, number of concurrent agents, and days since antibiotic initiation. The final feature representation combines these compo-

nents using one-hot encoding and engineered summary features, resulting in a total of 194 features.

B.2.2. 1 hour bins

We augment each patient-day timestep with a fixed-resolution grid of pre-anchor chart and laboratory measurements. For day index d with anchor time `DAY_START`(d), numeric vitals and labs with timestamps strictly in the leak-safe window [`DAY_START`(d) - W, `DAY_START`(d)) are aggregated to an hourly grid of 24 bins (B) per day. Empty bins are the forward-filled using LOCF (Engels and Diehr, 2003), yielding a dense matrix of shape ($B \times S$) with S timeseries variables, as a simple multivariate representation of irregularly sampled vitals and labs on a common time base.

Encoding the temporal matrix. This grid is encoded by a small `EventGridEncoder` module comprising a linear projection per bin followed by two one-dimensional convolutions with global pooling over the bin axis, yielding a fixed-size embedding vector `event_embed_dim` per patient-day timestep. This separation avoids the sequence-length explosion that would result from unrolling hourly bins directly into the admission timeline (expanding sequence length by a factor of B).

Fusion of 1h embeddings to aggregated data for sequence models. The resulting `event_embed_dim` vector is concatenated with the scaled tabular feature vector for that day to form a fused representation of dimension `tab_dim` + `event_embed_dim`, which is then passed through a LayerNorm before entering the sequence trunk and task heads. Where applicable, columns from the tabular feature vector whose information overlaps the longitudinal window were omitted from the tabular branch so coarse window-level summaries are not duplicated. The `EventGridEncoder` is not pre-trained or trained as a standalone classifier; it is optimized end-to-end within the full prediction network comprising the tabular pathway, event encoder, sequence trunk, and prediction head. The training objective is masked timestep binary cross-entropy with logits over valid sequence positions, with per-task positive-class reweighting, minimized with AdamW (Kingma and Ba, 2017). Gradients backpropagate through the prediction head, sequence trunk, fusion concatenation, and the Conv1d layers of the `EventGridEncoder`, so the convolutional filters learn whatever within-day temporal structure improves the final supervised prediction objective, subject to gradient clipping and early stopping on validation loss.

Fusion of 1h embeddings and aggregated data for tabular models. For the tabular model [1h] benchmark, the same `EventGridEncoder` is trained with a lightweight per-day MLP trunk over the fused tabular and event embedding. The exported day-level event embeddings are then appended to the aggregated tabular feature representation and used by standard tabular classifiers. With this setup we can give the tabular models a fair temporal dimension, and understand if the difference between tabular and sequential models is not just because of access to more data.

B.2.3. Sub-day irregular sequence representations

We collect the physiological timeseries data (Table 1) in the same leak-safe pre-anchor window used throughout the study, [`DAY_START` $- W$, `DAY_START`) with W=24 h. Rather than

discretizing this window onto a fixed hourly grid and forward filling, we construct an irregular multivariate sequence whose timesteps are the unique observation times within the window. Each sample is represented by a value tensor, an observation mask, and a relative-time vector representing time from window start. This yields an irregular-time representation of EHR measurements that preserves asynchronous sampling and true missingness patterns within the day.

B.3. Targets

For the outcomes we first derive course-level stewardship events from merged antibiotic prescription courses, then attach them to the patient-day unit.

IV-to-oral. An IV-to-oral event is defined when a course contains an earlier intravenous prescription followed by a non-intravenous formulation. The timestamp of the first qualifying non-intravenous prescription is recorded as the event time. The corresponding patient-day is labeled positive only on the calendar day of the switch. Clinically, this target is best interpreted as a proxy for route-change review opportunity rather than a recommendation that a switch was appropriate.

De-escalation. A de-escalation event is defined when a subsequent antibiotic prescription within the same admission has a different Anatomic Therapeutic Classification (ATC) (WHO Collaborating Centre for Drug Statistics Methodology, 2024) code and a strictly lower antibiotic spectrum index (ASI) (Gerber et al., 2017) than the current course (Table 2). Only the patient-day corresponding to the first qualifying de-escalation event is labeled positive. This target captures a prescription-based spectrum-narrowing proxy rather than adjudicated stewardship intent.

Discontinuation. A discontinuation event is defined when an antibiotic course ends before discharge, with at least 72 hours remaining in the admission and no overlapping antibiotic prescriptions in the subsequent interval. The label is assigned to the final day of the course. The label therefore depends on an operational drug-free-window threshold rather than direct clinician annotation.

Short-course. A short-course event is defined when the duration of an antibiotic course is less than 96 hours. Note that since we exclude stays shorter than 24 hours, this means that a short-course antibiotic target is actually a course with a length between 24 and 96 hours. Additionally, Cefazolin (ATC: J01DB04) is excluded from this definition due to its common prophylactic use in PICU. As with discontinuation, only the final day of the course is labeled as positive. This is an auxiliary duration-based target and should likewise be viewed as a stewardship-relevant proxy rather than a gold-standard decision label.

B.4. Models

B.4.1. Tabular Models

Tabular models operate on a flattened patient-day representation, where each instance corresponds to a single day of a patient's stay. These models do not explicitly model temporal dependencies across timesteps. Instead, temporal information is incorporated indirectly

through engineered features, such as length-of-stay variables and short-term trend summaries. As a result, each prediction is made independently based on the current patient state and aggregated context of the previous day, without access to the full temporal trajectory. We do not specifically handle class imbalance for the tabular models.

Logistic regression. Logistic regression uses L2 regularization with standard solver `lbfgs` and `max_iter=3000` and default inverse of regularization strength `C=1.0`)

LightGBM. LightGBM uses 500 trees with `learning_rate=0.05` and `num_leaves=31` (Ke et al., 2017). We don't standardize nor scale the features for this model.

MLP classifier. Scikit-learn's `MLPClassifier` (Pedregosa et al., 2011) was used with `hidden_layer_size=(64,0)`, ReLu activation function, batch_size of 256, initial learning rate of 0.001, and 200 max iteration with early stopping.

B.4.2. Tabular foundation model TabPFN

TabPFN-v2.6 weights were downloaded locally from the official GitHub https://github.com/PriorLabs/TabPFN (Hollmann et al., 2025). Feature values were converted to floating point and non-finite values replaced with zeros before inference. No standardization or scaling is applied to the input data. For the *MELODY-KISPI* cohort we trained/tested on the full sets (14,573/3,573 patient-days), for PIC we trained on the full training set (64,051), and tested on a stratified subset of 1000 patient because of compute limitations.

B.4.3. Sequence-based temporal models

Sequence-based temporal models train one network per target. Missing data are encoded using indicator variables in the sequence representations. Rows are grouped by admission id, ordered by timestamp of day start, truncated to at most 512 timesteps, and trained with masked per-timestep logistic losses so padded days do not contribute to optimization. Class imbalance is handled with a positive weight scalar for the target-specific loss, and the readout is one logit per patient-day rather than a pooled admission-level prediction.

Recurrent architectures. The recurrent models are Gated Recurrent Unit (GRU) (Cho et al., 2014) and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) trunks with two stacked layers, hidden size 128, dropout 0.2 between recurrent layers, and a linear per-timestep head (unidirectional recurrent trunk).

Temporal Convolutional Network. The Temporal Convolutional Network (TCN) (Bai et al., 2018) is implemented as a residual stack of left-padded causal dilated temporal convolutions over the admission trajectory. The causal structure ensures that predictions at timestep t depend only on observations up to and including t. By using convolutions, the TCN is able to capture longer-range temporal dependencies while maintaining a fixed computational cost per timestep. The output at each timestep is passed to a linear prediction head.

Transformer. We use a Transformer encoder (Vaswani et al., 2017) with learned positional embeddings to model temporal dependencies. The architecture employs multi-head

self-attention with GELU activations and a norm_first configuration. To handle variable-length sequences, key-padding masks are applied to ignore padded timesteps. A causal attention mask is used to ensure that predictions at timestep t depend only on observations up to and including t, preventing information leakage from future timesteps.

State Space Model. The state space model (SSM) is a lightweight Mamba-inspired selective SSM (Gu and Dao, 2024), implemented as a residual stack of gated, input-conditioned causal state-update blocks over the admission trajectory. Four input-dependent projections are computed from the hidden representation $\mathbf{x} \in \mathbb{R}^{B\times T\times H}$ at each timestep t. This gating mechanism allows the model to interpolate between propagating the updated latent state and passing the input through directly. Residual connections with dropout and a final layer normalisation are applied across the block stack, and padding safety is enforced by masking state updates and zeroing outputs beyond each sequence's valid length. The causal structure ensures that predictions at timestep t depend only on observations up to and including t. The output at each timestep is passed to a linear prediction head. This implementation is intentionally lightweight: it does not require a CUDA scan kernel and omits the convolutional mixer branch and full SSD parameterisation of the canonical Mamba release (Gu and Dao, 2024), prioritising ease of integration and fair comparison to the other simple sequence-based temporal architectures.

B.4.4. Multi-task learning setting

In the multi-task learning (MTL) setting of the sequence models, all four targets are predicted jointly using a shared encoder, with task-specific output heads. The models use the same per-timestep inputs as in the single task learning setting, i.e., tabular patient-day features alone, or fused with learnable features when applicable. We consider several MTL strategies. In hard parameter sharing, a single encoder is shared across tasks with separate linear heads. In uncertainty-weighted MTL (Kendall et al., 2018), task losses are combined using learned homoscedastic uncertainty parameters. We also evaluate a multi-gate mixture-of-experts (MMoE) architecture (Ma et al., 2018) with PCGrad (Yu et al., 2020, projecting conflicting gradients), where the shared sequence representation is passed through a small set of expert networks with task-specific gating, and gradient updates are adjusted to reduce destructive interference between tasks. All MTL variants can be paired with GRU (Cho et al., 2014), LSTM (Hochreiter and Schmidhuber, 1997), TCN (Bai et al., 2018), Transformer (Vaswani et al., 2017), or SSM sequence encoders.

B.4.5. Sequence foundation model CLMBR-T-Base.

As a pretrained adult EHR sequence-model comparison, we used CLMBR-T-Base (Clinical Language-Model-Based Representations with a Transformer backbone), a 141M-parameter autoregressive model pretrained on structured EHR code sequences from approximately 2.57 million adult patients at Stanford Medicine (Wornow et al., 2023). CLMBR-T-Base was trained to predict the next medical code in a patient's timeline given prior codes and produces fixed-dimensional patient representations intended for downstream prediction (Guo et al., 2024). We used the publicly released CLMBR-T-Base checkpoint without fine-tuning the transformer weights.

Event representation and code mapping. For each hospital admission, we constructed a chronological sequence of coded clinical events in Medical Event Data Standard (MEDS, version 0.1.3) (Arnrich et al., 2024). This input schema is expected by the Framework for Electronic Medical Records (FEMR, version 0.2.3) (Shah Lab, 2026) inference stack used to run CLMBR-T-Base. Events included date of birth, admission visit type, sex, time-stamped laboratory and vital-sign measurements, and antibiotic prescription starts.

Local PICU chart/lab item identifiers and antibiotic ATC codes were mapped to CLMBR-compatible vocabulary tokens (e.g., LOINC for measurements, RxNorm/ATC-derived drug codes for prescriptions, and visit/gender concept codes for static admission attributes). Unmapped local codes were omitted from each patient's timeline rather than imputed. We report the code mappings used below, which could contain errors as no specific specific documentation the e.g. admission types is provided by the *PIC* database. Date of birth (DOB) was mapped to MEDS_BIRTH, Sex to *M* or *F* for males and females, respectively. Because our PICU data only contains inpatients, we mapped the admission department to *Visit/IP*. Further mappings for the admission type (Table 3), the measurement data (Table 4), and the antibiotics (Table 5) are specified below.

Table 3: Mapping of Admission type to Visit.

Admission type	**Visit code**
AMBULATORY OBSERVATION	Visit/OP
DIRECT OBSERVATION	Visit/IP
ELECTIVE	Visit/IP
EMERGENCY	Visit/ER
EU OBSERVATION	Visit/ER
NEWBORN	Visit/IP
OBSERVATION	Visit/OP
SURGICAL SAME DAY ADMISSION	Visit/OP
URGENT	Visit/IP
__NA__	Visit/IP

Table 4: Mapping of Chartevents and Labevents Item IDs to LOINC (Forrey et al., 1996).

Item ID	CLMBR code	Description
1001	LOINC/8310-5	Temperature
1003	LOINC/8867-4	Heart rate
1004	LOINC/9279-1	Respiratory rate
1006	LOINC/59408-5	Oxygen saturation
1013	LOINC/8302-2	Height
1014	LOINC/29463-7	Weight
1015	LOINC/8462-4	Diastolic BP
1016	LOINC/8480-6	Systolic BP
5024	LOINC/1751-7	Albumin
5026	LOINC/1742-6	ALT
5057	LOINC/2532-0	LDH
5099	LOINC/718-7	Hemoglobin
5132	LOINC/789-8	RBC
5141	LOINC/6690-2	WBC
5227	LOINC/2524-7	Lactate
5230	LOINC/2947-0	Sodium whole blood
5235	LOINC/2019-8	pCO2 ABG
5237	LOINC/2746-6	pH ABG
5239	LOINC/2703-7	pO2 ABG
5252	LOINC/59408-5	O2 sat blood gas
5626	LOINC/1988-5	CRP
6085	LOINC/33959-8	PCT

Table 5: Antibiotics mapping of ATC (WHO Collaborating Centre for Drug Statistics Methodology, 2024) to RxNorm (Nelson et al., 2011).

ATC	CLMBR code	Drug / notes
J01AA02	RxNorm/3640	doxycycline
J01CA04	RxNorm/723	Amoxicillin
J01CE01	RxNorm/7980	Benzathine penicillin
J01CE02	RxNorm/7984	penicillin V
J01CE08	RxNorm/7984	Penicillin G
J01CF01	RxNorm/7987	Dicloxacillin
J01CF04	RxNorm/7989	Oxacillin
J01CF05	RxNorm/4448	floxacillin
J01CG01	RxNorm/10395	Sulbactam
J01CR02	RxNorm/1659141	Amoxicillin-clavulanate
J01CR05	RxNorm/1659139	Piperacillin-tazobactam
J01DB01	RxNorm/7984	Penicillin (benzathine proxy)
J01DB04	RxNorm/7984	Penicillin G procaine

Continued on next page

Table 5 – *Continued from previous page*

ATC	CLMBR code	Drug / notes
J01DC02	RxNorm/1664986	Cefazolin
J01DD01	RxNorm/1665050	Cefotaxime
J01DD02	RxNorm/1665052	Ceftazidime
J01DD04	RxNorm/1665054	Ceftriaxone
J01DD08	RxNorm/1665056	Cefepime
J01DE01	RxNorm/1665060	Meropenem
J01DH02	RxNorm/1665062	Imipenem
J01FA01	RxNorm/4053	erythromycin
J01FA09	RxNorm/3640	Clarithromycin
J01FA10	RxNorm/3642	Azithromycin
J01FF01	RxNorm/3644	Clindamycin
J01GB01	RxNorm/10627	tobramycin
J01GB03	RxNorm/10395	Gentamicin
J01GB06	RxNorm/10396	Amikacin
J01MA02	RxNorm/7052	Ciprofloxacin
J01MA12	RxNorm/7054	Levofloxacin
J01MA14	RxNorm/7056	Moxifloxacin
J01XA01	RxNorm/11124	Vancomycin
J01XA02	RxNorm/57021	teicoplanin
J01XD01	RxNorm/7980	Metronidazole
J01XE01	RxNorm/7052	Nitrofurantoin proxy
J01XX08	RxNorm/11124	Linezolid proxy
J02AA01	RxNorm/732	amphotericin B
J02AC01	RxNorm/1908	Fluconazole
J02AC02	RxNorm/1909	Itraconazole
J02AC03	RxNorm/1910	Voriconazole
J02AX04	RxNorm/140108	caspofungin
J04AB02	RxNorm/9384	rifampin
J04AC01	RxNorm/6038	isoniazid
J05AB01	RxNorm/281	acyclovir
J05AB06	RxNorm/4678	ganciclovir
J05AB11	RxNorm/73645	valacyclovir
J05AB12	RxNorm/83171	cidofovir
J05AB14	RxNorm/275891	valganciclovir
J05AB16	RxNorm/2284718	remdesivir
J05AD01	RxNorm/33562	foscarnet
J05AH02	RxNorm/260101	oseltamivir

Anchor-time representation extraction. Predictions were made at the same patient-day anchors used for all other models in our study: the start of each antibiotic stewardship day at `DAY_START` (default 8:00AM). For each anchor, only events occurring strictly before that timestamp were included, ensuring a causal, leakage-free input window. For each admission, we ran one forward pass through the frozen CLMBR-T-Base model over the

full preprocessed event timeline and extracted, at each anchor, the 768-dimensional hidden representation corresponding to the last model state at or before the anchor time. This yields one embedding vector per patient-day. We did not train a recurrent or temporal head across days; each day was scored from its anchor embedding alone.

Downstream prediction head. We fit a separate L2-regularized logistic regression model for each forward-looking stewardship target using the CLMBR-T-Base embeddings as the sole input features. Models used balanced class weights to mitigate label imbalance.

Evaluation protocol. CLMBR-T-Base results were evaluated under the same patient-level temporal split and evaluated on the same fully held-out test set. CLMBR-T-Base was included as a reference baseline to assess how much signal an adult pretrained structured-EHR foundation model provides for PICU antibiotic stewardship tasks.

B.4.6. Graph-based temporal model RAINDROP

We apply RAINDROP (Zhang et al., 2022) to the sub-day irregular sequence representations to learn a fixed-dimensional temporal embedding for each patient-day. RAINDROP treats sensors as nodes of a graph and uses structured message passing to exchange information across variables at each observation time, thereby modeling inter-sensor dependencies. We use a standard RAINDROP configuration with 2 Transformer encoder layers, 2 attention heads, dropout of 0.2, and 2 observation-propagation layers over the sensor graph. Temporal self-attention is applied causally over the irregular timeline, with time encoded via continuous positional features. The resulting per-timestep representations are aggregated by mean pooling over valid (non-padded) steps and concatenated with static tabular patient-context features. The fused vector is passed through the default two layer MLP classification head to produce binary predictions for each stewardship outcome. In this way, RAINDROP complements binned sequence models by operating directly on irregular observation times while remaining aligned with the same pre-anchor window.

Appendix C. Additional Results

C.1. Cross-cohort prevalence

Quantity	PIC	MELODY-KISPI
Patient-days (all)	79,831	18,148
Patient-days (train)	64,051	14,573
Patient-days (test)	15,780	3,575
Distinct patients (all)	5,515	2,059
Distinct admissions (all)	5,671	2,789
Test-set prevalence patient-days		
IV-to-oral	0.0035	0.0098
de-escalation	0.0392	0.0520
discontinuation	0.0169	0.0568
short-course	0.0577	0.1860
Test-set positive-course duration, days (median [Q1, Q3])		
IV-to-oral	6.62 [5.62, 9.25]	3.19 [2.67, 5.07]
de-escalation	6.62 [3.68, 10.02]	1.99 [1.00, 3.30]
discontinuation	6.76 [3.98, 9.71]	2.00 [1.33, 4.33]
short-course	2.60 [1.81, 3.07]	1.88 [1.01, 2.50]
Age at treatment initiation, months (median [Q1, Q3])		
Intravenous	10.8 [2.0, 48.4]	5.0 [0.5, 42.3]
Oral	14.6 [2.2, 60.4]	8.7 [0.1, 26.8]

Table 6: Cross-cohort summary for *PIC* and the *MELODY-KISPI*. Test-set prevalences of the patient-day dataset and antibiotics course duration metrics split by targets are also included. We additionally report age metrics at treatment start.

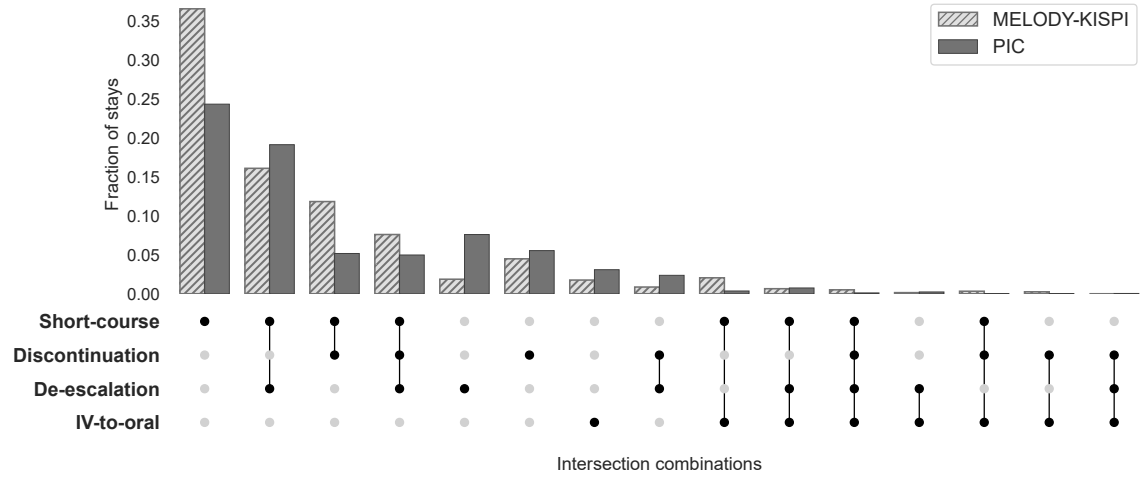

Figure 6: Intersections of AMS intervention targets on stay level across cohorts.

C.2. Comparison to adults

Table 7: Comparison of tabular models trained on 24-hour aggregated data with results reported by Tran-The et al. (2024) on a private Korean cohort and Bolton et al. (2024) on MIMIC-IV and eICU. For comparison, the early- and late-de-escalation target metrics of Tran-The et al. (2024) were averaged. We report AUROC, AUPRC, the true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV).

Task	Cohort	Model	Prev.	AUROC	AUPRC	TPR	TNR	PPV	NPV
IV-to-oral	MELODY-KISPI	LGBM	0.010	0.808	0.033	0.000	1.000	0.000	0.992
		MLP	0.010	0.555	0.009	0.083	0.845	0.004	0.991
	PIC	LGBM	0.003	0.883	0.102	0.077	0.999	0.214	0.997
		MLP	0.003	0.852	0.053	0.282	0.993	0.118	0.998
	Tran The	LGBM	0.030	0.810	0.160	0.780	0.710	0.070	0.990
	MIMIC-IV†	MLP	–	0.80	0.37	0.85	0.25	–	–
	eICU†	MLP	–	0.77	0.33	0.90	0.35	–	–
De-escalation	MELODY-KISPI	LGBM	0.052	0.932	0.396	0.147	0.991	0.490	0.952
		MLP	0.052	0.873	0.272	0.552	0.906	0.256	0.972
	PIC	LGBM	0.040	0.922	0.337	0.263	0.980	0.373	0.966
		MLP	0.040	0.910	0.314	0.429	0.955	0.308	0.973
	Tran The	LGBM	0.035	0.750	0.130	0.640	0.760	0.080	0.985
Discontinuation	MELODY-KISPI	LGBM	0.057	0.673	0.120	0.040	0.999	0.600	0.951
		MLP	0.057	0.612	0.074	0.403	0.750	0.079	0.959
	PIC	LGBM	0.017	0.835	0.110	0.063	0.998	0.293	0.986
		MLP	0.017	0.816	0.066	0.137	0.982	0.102	0.987
	Tran The	LGBM	0.090	0.800	0.360	0.710	0.720	0.210	0.960

† Not directly comparable to other cohorts due to differences in cohort selection and label definition (e.g., restricted to patients receiving both IV and oral antibiotics and evaluated at the per-day level).

C.3. Performance of models on finer resolution for MELODY-KISPI

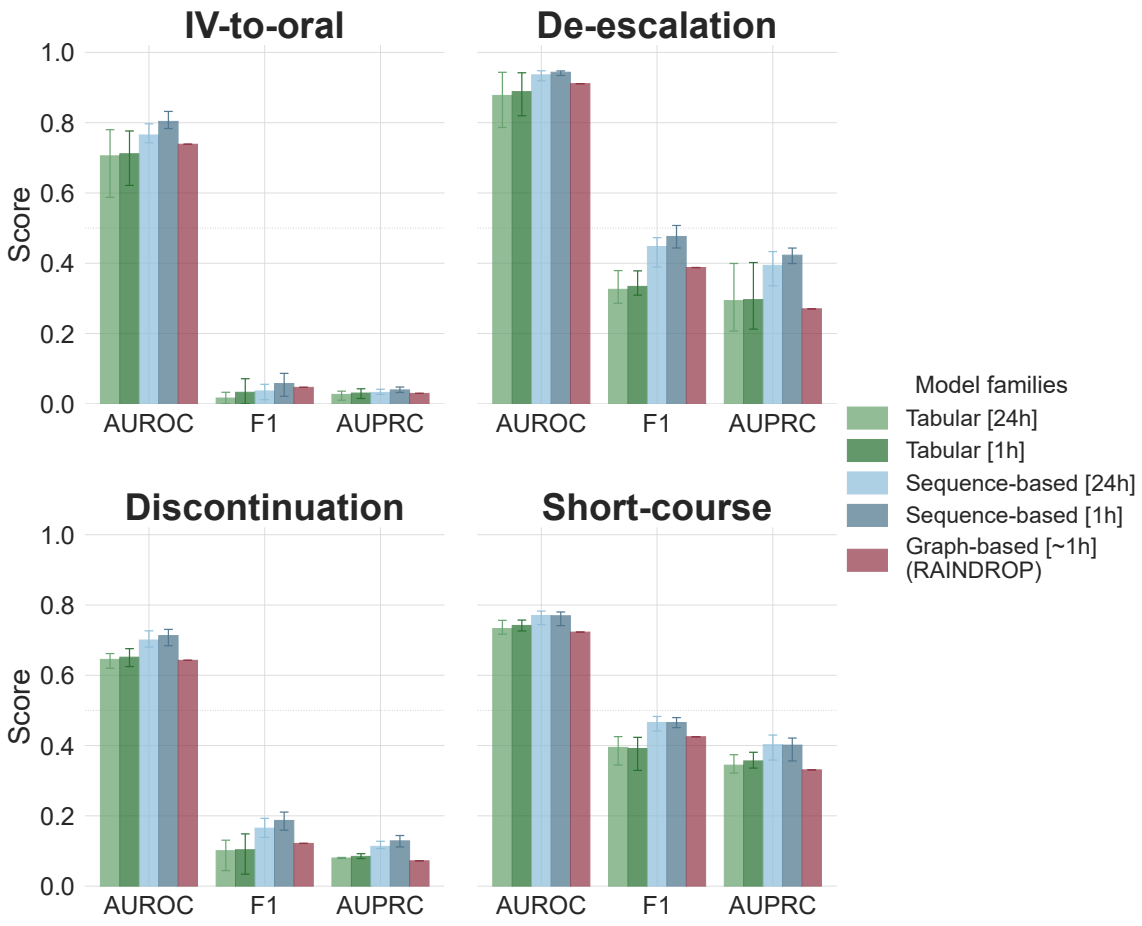

Figure 7: Performance of tabular and sequence-based models at 24-hour resolution and with augmented 1-hour representation, together with the graph-based model RAINDROP on irregular time series on the *MELODY-KISPI* cohort. Bars and error bars show the model class mean and minimum-maximum range, respectively.

C.4. Detailed results per target

Table 8: Full results for target IV-to-oral.

	IV-to-oral									
	MELODY-KISPI					PIC				
Model	**AUROC**	**F1@0.5**	**F1@Threshold**	**Threshold**	**AUPRC**	**AUROC**	**F1@0.5**	**F1@Threshold**	**Threshold**	**AUPRC**
Tabular [24h]										
Logistic regression	0.780 ± 0.000	0.000 ± 0.000	0.033 ± 0.000	0.078	0.034 ± 0.000	0.930 ± 0.000	0.000 ± 0.000	0.224 ± 0.000	0.089	0.137 ± 0.000
LightGBM	0.749 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.484	0.036 ± 0.000	0.942 ± 0.000	0.174 ± 0.000	0.159 ± 0.000	0.759	0.177 ± 0.000
MLP	0.588 ± 0.051	0.000 ± 0.000	0.016 ± 0.010	0.088	0.010 ± 0.002	0.909 ± 0.012	0.000 ± 0.000	0.220 ± 0.035	0.071	0.120 ± 0.009
TabPFN	0.810 ± 0.005	0.000 ± 0.000	0.030 ± 0.052	0.500	0.042 ± 0.002	0.978 ± 0.030	0.000 ± 0.000	0.000 ± 0.000	0.500	0.395 ± 0.096
Tabular [1h]										
Logistic regression	0.776 ± 0.000	0.000 ± 0.000	0.071 ± 0.000	0.066	0.034 ± 0.000	0.918 ± 0.000	0.000 ± 0.000	0.269 ± 0.000	0.097	0.144 ± 0.000
LightGBM	0.737 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.489	0.043 ± 0.000	0.930 ± 0.000	0.222 ± 0.000	0.100 ± 0.000	0.876	0.172 ± 0.000
MLP	0.621 ± 0.040	0.000 ± 0.000	0.027 ± 0.001	0.039	0.015 ± 0.001	0.879 ± 0.021	0.000 ± 0.000	0.184 ± 0.026	0.069	0.091 ± 0.013
Sequence-based [24h]										
GRU	0.745 ± 0.001	0.049 ± 0.003	0.028 ± 0.026	0.947	0.027 ± 0.003	0.949 ± 0.007	0.074 ± 0.007	0.182 ± 0.009	0.965	0.093 ± 0.002
LSTM	0.747 ± 0.011	0.046 ± 0.003	0.047 ± 0.006	0.897	0.028 ± 0.002	0.931 ± 0.011	0.075 ± 0.005	0.217 ± 0.035	0.962	0.116 ± 0.011
TCN	0.743 ± 0.002	0.050 ± 0.008	0.012 ± 0.020	0.930	0.027 ± 0.001	0.954 ± 0.002	0.081 ± 0.011	0.202 ± 0.028	0.955	0.129 ± 0.032
Transformer	0.794 ± 0.025	0.051 ± 0.005	0.056 ± 0.049	0.952	0.041 ± 0.007	0.930 ± 0.012	0.093 ± 0.024	0.262 ± 0.034	0.992	0.167 ± 0.047
SSM	0.797 ± 0.012	0.060 ± 0.005	0.041 ± 0.037	0.876	0.038 ± 0.006	0.951 ± 0.006	0.089 ± 0.002	0.182 ± 0.031	0.965	0.112 ± 0.023
CLMBR-T-base	0.723 ± 0.029	0.053 ± 0.002	0.036 ± 0.031	0.692	0.022 ± 0.002	0.877 ± 0.003	0.087 ± 0.010	0.114 ± 0.011	0.983	0.049 ± 0.002
Sequence-based [1h]										
GRU	0.789 ± 0.007	0.050 ± 0.008	0.062 ± 0.031	0.908	0.032 ± 0.005	0.946 ± 0.004	0.082 ± 0.011	0.205 ± 0.024	0.981	0.106 ± 0.017
LSTM	0.783 ± 0.003	0.056 ± 0.002	0.042 ± 0.016	0.908	0.037 ± 0.007	0.924 ± 0.036	0.091 ± 0.028	0.193 ± 0.011	0.973	0.151 ± 0.004
TCN	0.800 ± 0.013	0.050 ± 0.004	0.022 ± 0.037	0.936	0.036 ± 0.007	0.949 ± 0.007	0.098 ± 0.003	0.204 ± 0.003	0.967	0.135 ± 0.033
Transformer	0.814 ± 0.002	0.057 ± 0.008	0.076 ± 0.018	0.936	0.044 ± 0.007	0.929 ± 0.005	0.093 ± 0.022	0.249 ± 0.028	0.993	0.155 ± 0.005
SSM	0.832 ± 0.009	0.052 ± 0.006	0.087 ± 0.009	0.862	0.048 ± 0.003	0.942 ± 0.011	0.084 ± 0.003	0.216 ± 0.025	0.993	0.133 ± 0.013
Graph [~1h]										
RAINDROP	0.738 ± 0.015	0.011 ± 0.019	0.047 ± 0.029	0.288	0.030 ± 0.002	0.908 ± 0.015	0.070 ± 0.121	0.171 ± 0.023	0.296	0.119 ± 0.012

Table 9: Full results for target De-escalation.

	De-escalation									
	MELODY-KISPI					PIC				
Model	**AUROC**	**F1@0.5**	**F1@Threshold**	**Threshold**	**AUPRC**	**AUROC**	**F1@0.5**	**F1@Threshold**	**Threshold**	**AUPRC**
Tabular [24h]										
Logistic regression	0.903 ± 0.000	0.159 ± 0.000	0.379 ± 0.000	0.108	0.275 ± 0.000	0.910 ± 0.000	0.177 ± 0.000	0.356 ± 0.000	0.123	0.264 ± 0.000
LightGBM	0.943 ± 0.000	0.352 ± 0.000	0.312 ± 0.000	0.555	0.399 ± 0.000	0.936 ± 0.000	0.293 ± 0.000	0.390 ± 0.000	0.315	0.410 ± 0.000
MLP	0.786 ± 0.059	0.055 ± 0.050	0.286 ± 0.097	0.178	0.207 ± 0.065	0.917 ± 0.005	0.237 ± 0.044	0.354 ± 0.015	0.209	0.304 ± 0.006
TabPFN	0.952 ± 0.001	0.378 ± 0.027	0.430 ± 0.072	0.500	0.479 ± 0.006	0.942 ± 0.024	0.222 ± 0.071	0.222 ± 0.071	0.500	0.469 ± 0.137
Tabular [1h]										
Logistic regression	0.904 ± 0.000	0.160 ± 0.000	0.378 ± 0.000	0.108	0.275 ± 0.000	0.909 ± 0.000	0.155 ± 0.000	0.359 ± 0.000	0.119	0.260 ± 0.000
LightGBM	0.942 ± 0.000	0.356 ± 0.000	0.309 ± 0.000	0.545	0.402 ± 0.000	0.934 ± 0.000	0.283 ± 0.000	0.378 ± 0.000	0.300	0.387 ± 0.000
MLP	0.820 ± 0.085	0.013 ± 0.023	0.315 ± 0.052	0.110	0.213 ± 0.059	0.915 ± 0.014	0.180 ± 0.065	0.352 ± 0.022	0.196	0.279 ± 0.034
Sequence-based [24h]										
GRU	0.948 ± 0.001	0.428 ± 0.013	0.467 ± 0.017	0.817	0.433 ± 0.033	0.962 ± 0.001	0.423 ± 0.026	0.524 ± 0.014	0.908	0.524 ± 0.015
LSTM	0.941 ± 0.005	0.399 ± 0.035	0.473 ± 0.038	0.823	0.415 ± 0.040	0.962 ± 0.002	0.409 ± 0.022	0.532 ± 0.014	0.877	0.519 ± 0.016
TCN	0.941 ± 0.008	0.391 ± 0.032	0.466 ± 0.015	0.842	0.392 ± 0.017	0.961 ± 0.004	0.382 ± 0.029	0.524 ± 0.030	0.866	0.483 ± 0.025
Transformer	0.919 ± 0.027	0.310 ± 0.062	0.390 ± 0.083	0.807	0.336 ± 0.088	0.956 ± 0.000	0.352 ± 0.016	0.473 ± 0.011	0.885	0.437 ± 0.013
SSM	0.932 ± 0.008	0.360 ± 0.007	0.442 ± 0.031	0.858	0.395 ± 0.035	0.959 ± 0.001	0.349 ± 0.006	0.496 ± 0.011	0.914	0.477 ± 0.008
CLMBR-T-base	0.585 ± 0.004	0.110 ± 0.001	0.120 ± 0.016	0.576	0.088 ± 0.007	0.686 ± 0.001	0.122 ± 0.001	0.129 ± 0.005	0.738	0.070 ± 0.001
Sequence-based [1h]										
GRU	0.947 ± 0.003	0.411 ± 0.008	0.508 ± 0.035	0.885	0.441 ± 0.051	0.966 ± 0.002	0.397 ± 0.024	0.536 ± 0.011	0.891	0.543 ± 0.015
LSTM	0.946 ± 0.002	0.393 ± 0.016	0.487 ± 0.006	0.902	0.443 ± 0.037	0.965 ± 0.002	0.391 ± 0.020	0.533 ± 0.012	0.869	0.537 ± 0.019
TCN	0.946 ± 0.004	0.406 ± 0.024	0.480 ± 0.022	0.862	0.418 ± 0.038	0.965 ± 0.002	0.381 ± 0.004	0.530 ± 0.017	0.871	0.524 ± 0.021
Transformer	0.934 ± 0.004	0.295 ± 0.027	0.443 ± 0.043	0.850	0.413 ± 0.050	0.955 ± 0.003	0.335 ± 0.035	0.469 ± 0.010	0.870	0.443 ± 0.039
SSM	0.944 ± 0.001	0.376 ± 0.001	0.463 ± 0.017	0.883	0.399 ± 0.017	0.962 ± 0.002	0.367 ± 0.032	0.508 ± 0.006	0.933	0.504 ± 0.007
Graph [~1h]										
RAINDROP	0.907 ± 0.025	0.383 ± 0.025	0.388 ± 0.027	0.557	0.273 ± 0.008	0.896 ± 0.005	0.308 ± 0.010	0.331 ± 0.007	0.669	0.233 ± 0.007

Table 10: Full results for target Discontinuation.

	Discontinuation									
	MELODY-KISPI					PIC				
Model	**AUROC**	**F1@0.5**	**F1@Threshold**	**Threshold**	**AUPRC**	**AUROC**	**F1@0.5**	**F1@Threshold**	**Threshold**	**AUPRC**
Tabular [24h]										
Logistic regression	0.654 ± 0.000	0.000 ± 0.000	0.131 ± 0.000	0.123	0.080 ± 0.000	0.796 ± 0.000	0.000 ± 0.000	0.119 ± 0.000	0.054	0.060 ± 0.000
LightGBM	0.662 ± 0.000	0.000 ± 0.000	0.044 ± 0.000	0.289	0.082 ± 0.000	0.841 ± 0.000	0.014 ± 0.000	0.098 ± 0.000	0.188	0.097 ± 0.000
MLP	0.620 ± 0.023	0.000 ± 0.000	0.128 ± 0.015	0.102	0.079 ± 0.007	0.740 ± 0.020	0.000 ± 0.000	0.105 ± 0.007	0.065	0.053 ± 0.003
TabPFN	0.721 ± 0.004	0.000 ± 0.000	0.116 ± 0.102	0.500	0.122 ± 0.006	0.903 ± 0.025	0.000 ± 0.000	0.000 ± 0.000	0.500	0.278 ± 0.093
Tabular [1h]										
Logistic regression	0.654 ± 0.000	0.000 ± 0.000	0.149 ± 0.000	0.109	0.083 ± 0.000	0.775 ± 0.000	0.000 ± 0.000	0.112 ± 0.000	0.051	0.052 ± 0.000
LightGBM	0.676 ± 0.000	0.000 ± 0.000	0.034 ± 0.000	0.281	0.092 ± 0.000	0.828 ± 0.000	0.014 ± 0.000	0.117 ± 0.000	0.168	0.112 ± 0.000
MLP	0.625 ± 0.013	0.000 ± 0.000	0.128 ± 0.011	0.067	0.078 ± 0.007	0.744 ± 0.034	0.000 ± 0.000	0.110 ± 0.012	0.058	0.057 ± 0.008
Sequence-based [24h]										
GRU	0.726 ± 0.004	0.159 ± 0.003	0.193 ± 0.015	0.743	0.128 ± 0.007	0.835 ± 0.009	0.095 ± 0.004	0.151 ± 0.012	0.857	0.093 ± 0.011
LSTM	0.704 ± 0.003	0.154 ± 0.010	0.156 ± 0.009	0.731	0.107 ± 0.010	0.836 ± 0.008	0.094 ± 0.004	0.164 ± 0.001	0.889	0.105 ± 0.004
TCN	0.706 ± 0.011	0.159 ± 0.004	0.157 ± 0.016	0.706	0.113 ± 0.015	0.819 ± 0.015	0.083 ± 0.003	0.127 ± 0.019	0.814	0.068 ± 0.011
Transformer	0.685 ± 0.011	0.140 ± 0.004	0.139 ± 0.007	0.663	0.107 ± 0.009	0.823 ± 0.011	0.087 ± 0.003	0.135 ± 0.012	0.857	0.081 ± 0.008
SSM	0.680 ± 0.018	0.140 ± 0.010	0.174 ± 0.022	0.701	0.111 ± 0.005	0.809 ± 0.002	0.086 ± 0.002	0.124 ± 0.001	0.826	0.067 ± 0.003
CLMBR-T-base	0.614 ± 0.006	0.135 ± 0.009	0.122 ± 0.007	0.544	0.077 ± 0.002	0.720 ± 0.002	0.071 ± 0.001	0.083 ± 0.009	0.863	0.041 ± 0.000
Sequence-based [1h]										
GRU	0.728 ± 0.010	0.160 ± 0.005	0.191 ± 0.033	0.742	0.132 ± 0.020	0.824 ± 0.011	0.086 ± 0.009	0.141 ± 0.014	0.850	0.081 ± 0.007
LSTM	0.723 ± 0.016	0.155 ± 0.005	0.178 ± 0.009	0.690	0.137 ± 0.012	0.823 ± 0.005	0.085 ± 0.002	0.147 ± 0.007	0.831	0.085 ± 0.003
TCN	0.731 ± 0.018	0.163 ± 0.008	0.211 ± 0.010	0.720	0.144 ± 0.015	0.826 ± 0.009	0.087 ± 0.008	0.149 ± 0.009	0.855	0.081 ± 0.012
Transformer	0.684 ± 0.011	0.145 ± 0.003	0.159 ± 0.012	0.592	0.111 ± 0.007	0.808 ± 0.007	0.072 ± 0.006	0.132 ± 0.012	0.878	0.073 ± 0.006
SSM	0.699 ± 0.012	0.145 ± 0.012	0.196 ± 0.020	0.706	0.119 ± 0.009	0.803 ± 0.006	0.086 ± 0.005	0.133 ± 0.017	0.856	0.076 ± 0.007
Graph [~1h]										
RAINDROP	0.627 ± 0.026	0.130 ± 0.007	0.109 ± 0.033	0.520	0.069 ± 0.007	0.739 ± 0.007	0.074 ± 0.019	0.094 ± 0.004	0.379	0.047 ± 0.001

Table 11: Full results for target Short-course.

	Short-course									
	MELODY-KISPI					PIC				
Model	**AUROC**	**F1@0.5**	**F1@Threshold**	**Threshold**	**AUPRC**	**AUROC**	**F1@0.5**	**F1@Threshold**	**Threshold**	**AUPRC**
Tabular [24h]										
Logistic regression	0.726 ± 0.000	0.048 ± 0.000	0.425 ± 0.000	0.185	0.337 ± 0.000	0.745 ± 0.000	0.034 ± 0.000	0.216 ± 0.000	0.141	0.152 ± 0.000
LightGBM	0.756 ± 0.000	0.193 ± 0.000	0.345 ± 0.000	0.358	0.374 ± 0.000	0.805 ± 0.000	0.183 ± 0.000	0.326 ± 0.000	0.224	0.287 ± 0.000
MLP	0.717 ± 0.026	0.086 ± 0.059	0.414 ± 0.028	0.220	0.322 ± 0.028	0.766 ± 0.006	0.133 ± 0.018	0.257 ± 0.009	0.181	0.194 ± 0.009
TabPFN	0.779 ± 0.001	0.144 ± 0.010	0.363 ± 0.188	0.500	0.412 ± 0.003	0.840 ± 0.007	0.173 ± 0.108	0.173 ± 0.108	0.500	0.342 ± 0.062
Tabular [1h]										
Logistic regression	0.726 ± 0.000	0.060 ± 0.000	0.422 ± 0.000	0.196	0.336 ± 0.000	0.742 ± 0.000	0.030 ± 0.000	0.218 ± 0.000	0.149	0.149 ± 0.000
LightGBM	0.757 ± 0.000	0.193 ± 0.000	0.329 ± 0.000	0.371	0.381 ± 0.000	0.794 ± 0.000	0.181 ± 0.000	0.300 ± 0.000	0.247	0.272 ± 0.000
MLP	0.742 ± 0.005	0.121 ± 0.022	0.423 ± 0.010	0.236	0.352 ± 0.002	0.766 ± 0.003	0.138 ± 0.019	0.263 ± 0.004	0.179	0.195 ± 0.014
Sequence-based [24h]										
GRU	0.776 ± 0.003	0.471 ± 0.002	0.471 ± 0.018	0.584	0.411 ± 0.006	0.848 ± 0.003	0.249 ± 0.008	0.365 ± 0.005	0.859	0.343 ± 0.013
LSTM	0.780 ± 0.006	0.471 ± 0.004	0.483 ± 0.006	0.622	0.419 ± 0.010	0.832 ± 0.005	0.236 ± 0.007	0.356 ± 0.005	0.823	0.332 ± 0.011
TCN	0.783 ± 0.011	0.473 ± 0.001	0.465 ± 0.018	0.588	0.430 ± 0.021	0.828 ± 0.006	0.247 ± 0.019	0.343 ± 0.005	0.797	0.286 ± 0.005
Transformer	0.768 ± 0.001	0.468 ± 0.008	0.459 ± 0.015	0.578	0.394 ± 0.008	0.808 ± 0.005	0.228 ± 0.005	0.289 ± 0.012	0.809	0.252 ± 0.013
SSM	0.744 ± 0.003	0.444 ± 0.006	0.442 ± 0.007	0.561	0.359 ± 0.002	0.769 ± 0.005	0.207 ± 0.006	0.260 ± 0.017	0.788	0.204 ± 0.020
CLMBR-T-base	0.650 ± 0.008	0.353 ± 0.005	0.338 ± 0.023	0.490	0.279 ± 0.006	0.645 ± 0.002	0.138 ± 0.003	0.137 ± 0.001	0.620	0.085 ± 0.001
Sequence-based [1h]										
GRU	0.780 ± 0.008	0.474 ± 0.014	0.473 ± 0.008	0.586	0.417 ± 0.006	0.855 ± 0.005	0.246 ± 0.007	0.363 ± 0.025	0.821	0.359 ± 0.011
LSTM	0.779 ± 0.009	0.470 ± 0.004	0.480 ± 0.001	0.583	0.421 ± 0.010	0.842 ± 0.005	0.241 ± 0.019	0.355 ± 0.004	0.823	0.354 ± 0.008
TCN	0.780 ± 0.014	0.473 ± 0.007	0.470 ± 0.013	0.557	0.421 ± 0.015	0.855 ± 0.004	0.263 ± 0.008	0.366 ± 0.015	0.810	0.347 ± 0.012
Transformer	0.767 ± 0.005	0.464 ± 0.004	0.454 ± 0.008	0.573	0.391 ± 0.013	0.806 ± 0.001	0.230 ± 0.018	0.300 ± 0.007	0.744	0.260 ± 0.015
SSM	0.741 ± 0.002	0.449 ± 0.009	0.451 ± 0.008	0.519	0.356 ± 0.005	0.788 ± 0.003	0.212 ± 0.011	0.289 ± 0.005	0.796	0.233 ± 0.007
Graph [~1h]										
RAINDROP	0.725 ± 0.010	0.434 ± 0.013	0.420 ± 0.013	0.567	0.332 ± 0.009	0.752 ± 0.007	0.228 ± 0.007	0.219 ± 0.009	0.683	0.151 ± 0.010

C.5. Calibration plots *MELODY-KISPI*

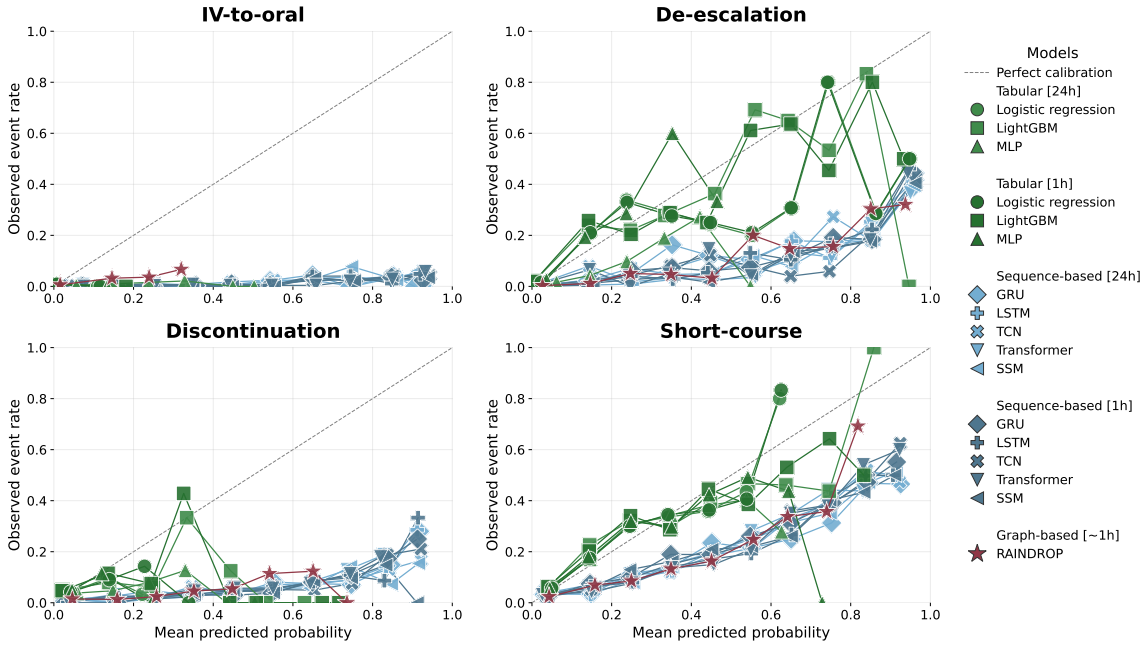

Figure 8: Calibration plots for tabular, sequence-based, and graph-based temporal models across temporal representations for the four AMS intervention predictions tasks in the *MELODY-KISPI* cohort.

C.6. Multi-task learning for AMS intervention prediction

We finally evaluate whether multi-task learning can improve performance by jointly modeling multiple stewardship targets. Since these interventions are clinically related and sometimes considered jointly (e.g., intervention vs no intervention), we hypothesize that sharing representations across tasks may improve learning. We therefore compare STL and MTL on targets de-escalation, discontinuation, and short-course as these co-occur (Figure 6) as we expect them to be more related to each other than to IV-to-oral. We use sequence models using 24-hour representations. AUROC performance across cohorts and training both settings are shown in Figures 9 and 10. Full results are described in section C.6.1.

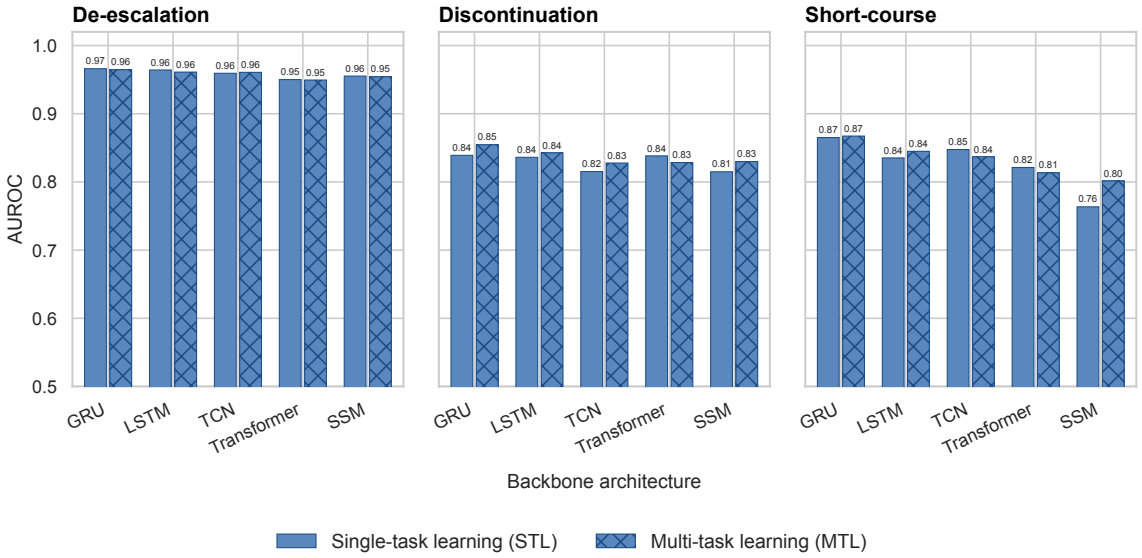

Figure 9: Comparison of single-task and multi-task learning for sequence models across three AMS targets on the *PIC* cohort.

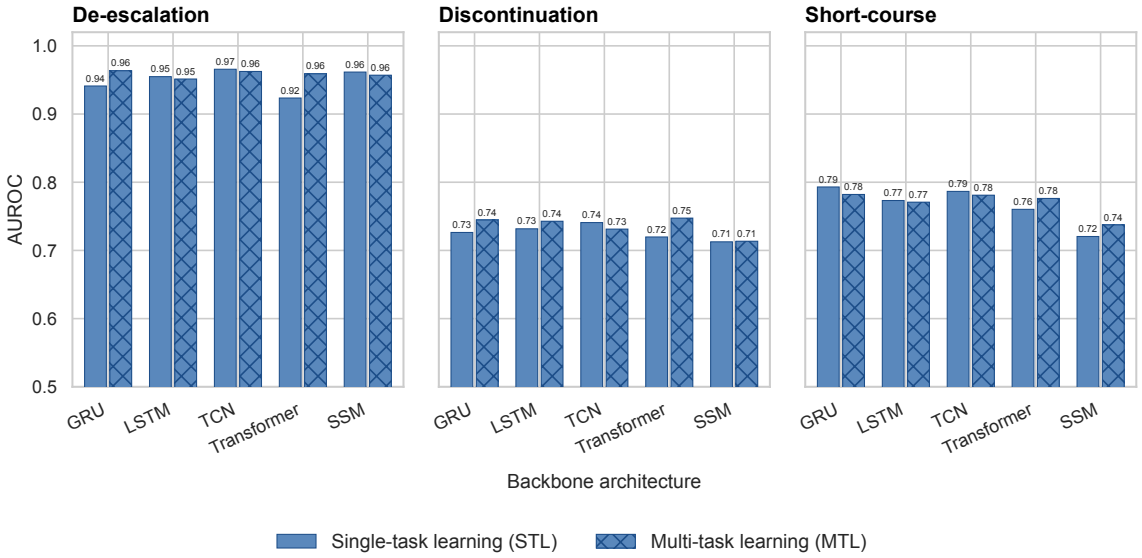

Figure 10: Comparison of single-task and multi-task learning for sequence models across three AMS targets on the *MELODY-KISPI* cohort.

C.6.1. MTL modeling detailed results

Table 12: MTL results for de-escalation.

		De-escalation					
		MELODY-KISPI			PIC		
Architecture	**Approach**	**AUROC**	**F1**	**AUPRC**	**AUROC**	**F1**	**AUPRC**
Sequential [24h]							
GRU	hard-sharing	0.955	0.464	0.520	0.964	0.433	0.549
GRU	uncertainty-weighted	0.956	0.469	0.525	0.965	0.447	0.550
GRU	MMoE+PCGrad	0.962	0.458	**0.595**	0.965	0.443	0.560
LSTM	hard-sharing	0.951	0.447	0.540	0.961	0.404	0.511
LSTM	uncertainty-weighted	0.949	0.423	0.516	0.962	0.403	0.511
LSTM	MMoE+PCGrad	0.958	0.456	0.558	0.964	0.410	0.537
TCN	hard-sharing	0.962	0.499	0.525	0.961	0.395	0.512
TCN	uncertainty-weighted	0.959	0.511	0.510	0.962	0.434	0.536
TCN	MMoE+PCGrad	0.960	0.482	0.515	0.964	0.424	0.538
Transformer	hard-sharing	0.959	0.465	0.548	0.949	0.331	0.397
Transformer	uncertainty-weighted	0.955	0.463	0.507	0.951	0.351	0.432
Transformer	MMoE+PCGrad	0.956	0.423	0.545	0.948	0.349	0.389
SSM	hard-sharing	0.957	0.469	0.511	0.954	0.352	0.443
SSM	uncertainty-weighted	0.958	0.477	0.536	0.951	0.328	0.416
SSM	MMoE+PCGrad	0.953	0.465	0.493	0.943	0.304	0.391

Table 13: MTL results for discontinuation.

		Discontinuation					
		MELODY-KISPI			PIC		
Architecture	**Approach**	**AUROC**	**F1**	**AUPRC**	**AUROC**	**F1**	**AUPRC**
Sequential [24h]							
GRU	hard-sharing	0.742	0.234	0.234	0.855	0.099	0.106
GRU	uncertainty-weighted	0.742	0.233	0.237	0.855	0.102	0.104
GRU	MMoE+PCGrad	0.743	0.229	0.212	0.860	0.102	0.115
LSTM	hard-sharing	0.743	**0.248**	0.199	0.843	0.106	0.100
LSTM	uncertainty-weighted	0.730	0.226	0.187	0.844	0.105	0.100
LSTM	MMoE+PCGrad	0.743	0.243	0.180	0.854	0.105	0.121
TCN	hard-sharing	0.731	0.238	**0.228**	0.828	0.091	0.075
TCN	uncertainty-weighted	0.743	0.248	0.209	0.841	0.096	0.095
TCN	MMoE+PCGrad	0.740	0.226	0.188	0.849	0.092	0.101
Transformer	hard-sharing	0.747	0.239	0.208	0.828	0.096	0.092
Transformer	uncertainty-weighted	0.732	0.242	0.201	0.844	0.086	0.111
Transformer	MMoE+PCGrad	0.722	0.226	0.208	0.822	0.083	0.081
SSM	hard-sharing	0.713	0.238	0.173	0.830	0.101	0.092
SSM	uncertainty-weighted	0.717	0.246	0.177	0.815	0.094	0.086
SSM	MMoE+PCGrad	0.699	0.217	0.164	0.824	0.087	0.093

Table 14: MTL results for short-courses.

		Short-course					
		MELODY-KISPI			PIC		
Architecture	**Approach**	**AUROC**	**F1**	**AUPRC**	**AUROC**	**F1**	**AUPRC**
Sequential [24h]							
GRU	hard-sharing	0.774	0.501	0.457	0.866	0.287	0.423
GRU	uncertainty-weighted	0.775	0.504	0.459	0.865	0.285	0.422
GRU	MMoE+PCGrad	0.779	0.501	**0.474**	0.867	0.280	0.427
LSTM	hard-sharing	0.771	0.497	0.454	0.845	**0.274**	0.367
LSTM	uncertainty-weighted	0.767	0.482	0.458	0.845	0.271	0.366
LSTM	MMoE+PCGrad	0.780	0.511	0.458	0.864	0.260	0.402
TCN	hard-sharing	0.781	**0.519**	0.460	0.837	0.264	0.322
TCN	uncertainty-weighted	0.776	0.516	0.450	0.848	0.267	0.343
TCN	MMoE+PCGrad	0.768	0.497	0.459	0.855	0.254	0.360
Transformer	hard-sharing	0.776	0.496	0.451	0.813	0.229	0.263
Transformer	uncertainty-weighted	0.763	0.494	0.456	0.827	0.223	0.305
Transformer	MMoE+PCGrad	0.762	0.491	0.446	0.810	0.227	0.253
SSM	hard-sharing	0.737	0.481	0.417	0.801	0.235	0.282
SSM	uncertainty-weighted	0.739	0.482	0.421	0.790	0.227	0.248
SSM	MMoE+PCGrad	0.724	0.465	0.393	0.788	0.232	0.249

C.7. When do sequence models help

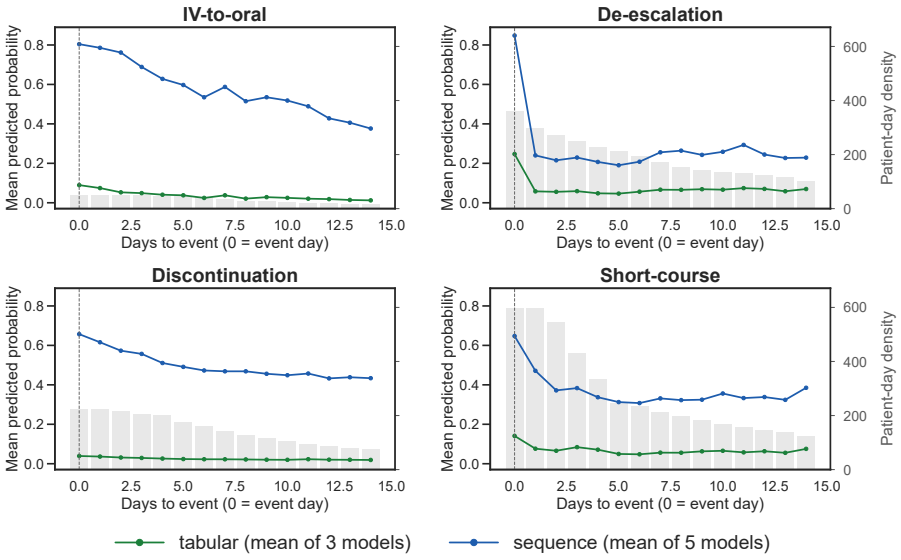

Figure 11: Average event aligned probabilities of tabular and sequence models of *PIC* targets. Tabular models include logistic regression, LightGBM, and the MLP. Sequence models include the GRU, LSTM, TCN, Transformer, and SSM. Grey bars behind show the density of as the number of patient-days.

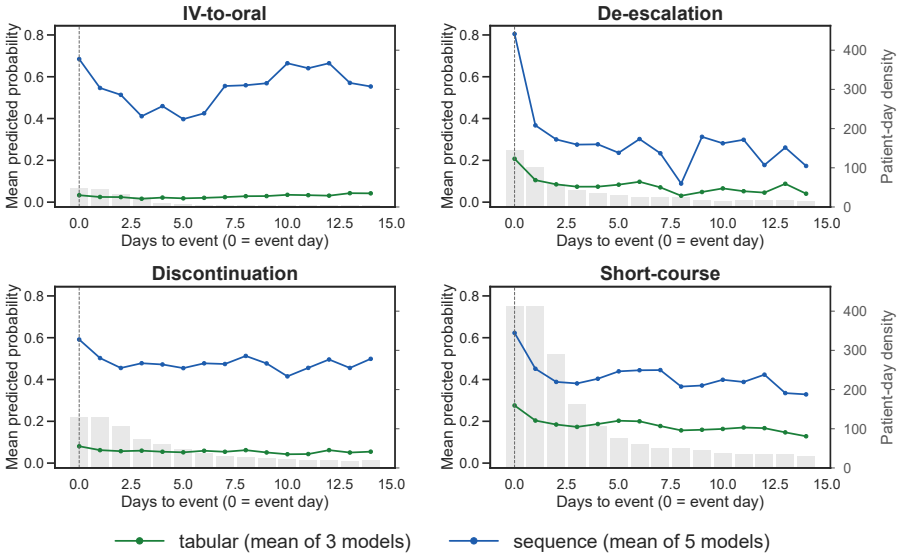

Figure 12: Average event aligned probabilities of tabular and sequence models of *MELODY-KISPI* targets.

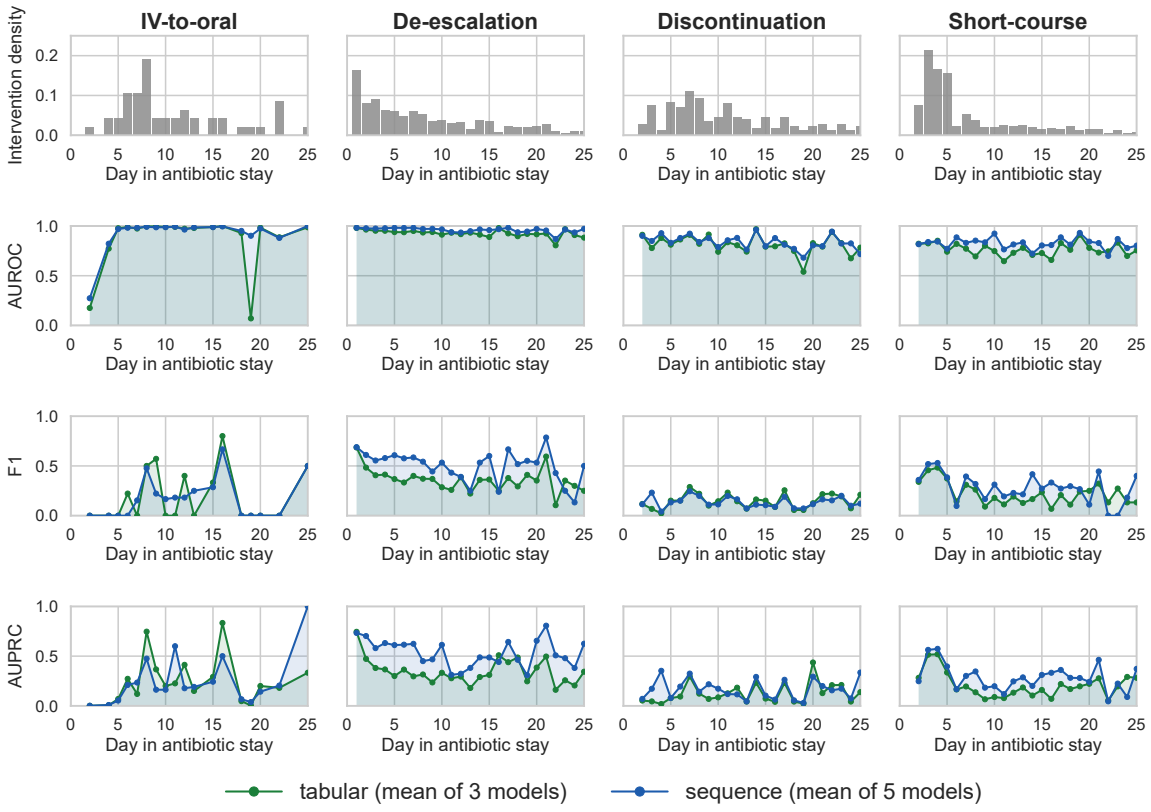

Figure 13: Average prediction performance of tabular and sequence model groups across antibiotic days in a stay in *PIC*. Tabular models include Logistic regression, LightGBM, and the MLP. Sequence models include the GRU, LSTM, TCN, Transformer, and SSM.

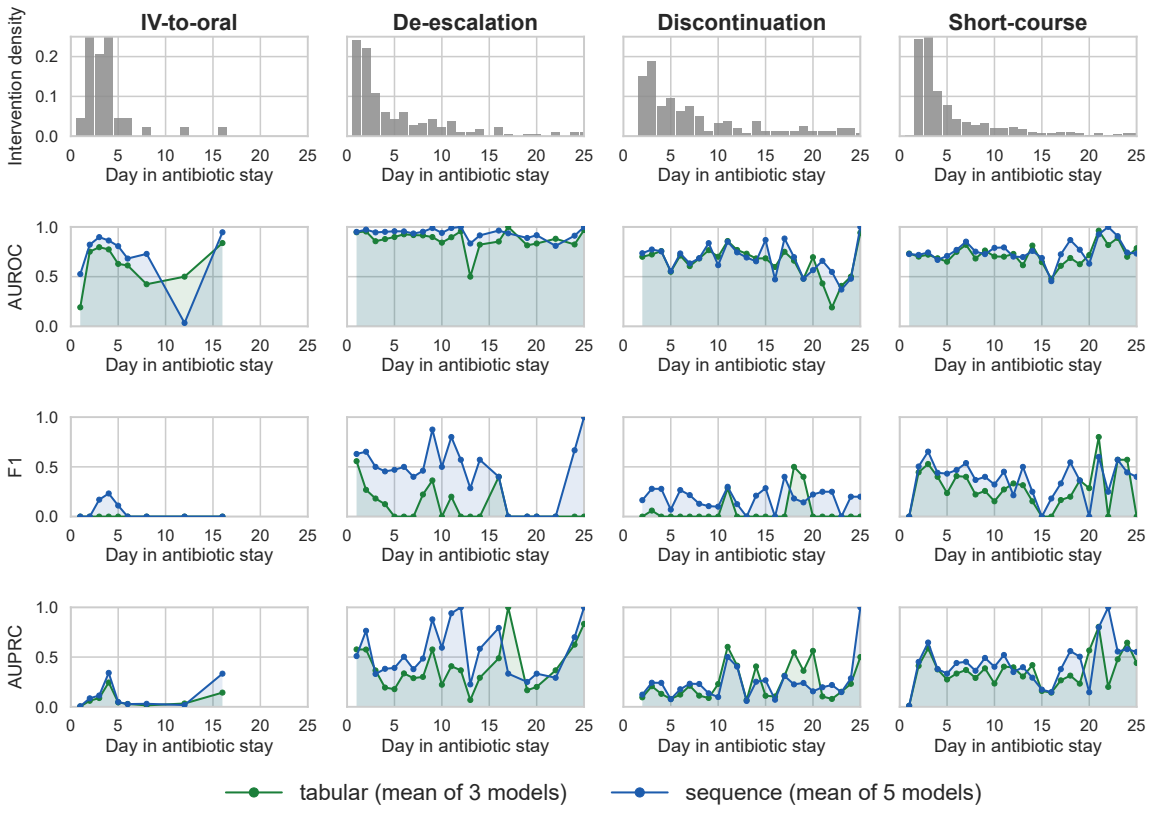

Figure 14: Average prediction performance of tabular and sequence model groups across antibiotic days in a stay in *MELODY-KISPI*. Tabular models include Logistic regression, LightGBM, and the MLP. Sequence models include the GRU, LSTM, TCN, Transformer, and SSM.

C.8. 1h encoder ablations

To better understand the added benefit of finer-grained 1h data and the learnable event grid encoder we performed a set of ablations (Figures 15 and 16 below). We here compare the daily sequence-based [24h] model family on daily aggregated data, adding the 1h CNN learnable encoder such as described above in Section B.2.2, and compare to the following two ablations;

1h linear projection encoder. This encoder keeps the day-level sequence architecture used by the CNN longitudinal model and changes only how each day's hourly event grid is compressed into an embedding. We use the default hourly chart/lab measurement set with LOCF, binary indicators of whether a new observation occurred in that bin, and the hours since the most recent observation. The resulting high-dimensional tensor is flattened through a linear projection into the 64-dimensional embedding that is concatenated to the day's tabular features.

1h direct encoder. This encoder does not use a day-level event-grid encoder. Instead it models the recent history as an hourly sequence and fuses tabular features only after the sequence trunk (late fusion). Starting from the same per-day 24-bin grids (value / observation mask / hours-since-last-observation planes, we stack up to 5 consecutive prior days and reshape them into a single sequence. All grid planes and tabular features are standardized on the training split. A sequence backbone (same five architectures as above; hidden size 128, 2 layers, dropout 0.2) reads the raw hourly feature vectors directly. The hidden state at the last valid hour is concatenated with the current day's tabular vector and mapped to a single patient-day logit. Relative to the longitudinal CNN/linear models, the direct hourly modeling differs in three structural ways: (1) the sequence unit is an hour, not a day; (2) multi-day context is a fixed 5-day hourly stack, not a day-level sequence over the admission; (3) there is no learned day-level grid encoder, tabular information enters only after temporal encoding.

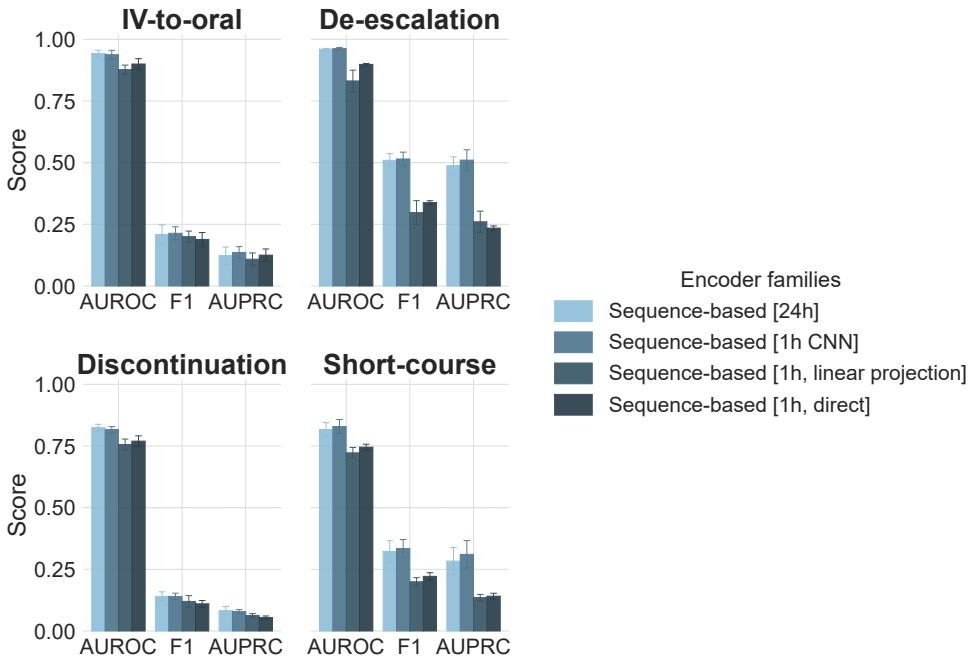

Figure 15: Encoder ablations of 1h sequence models on *PIC* data. Bars represent the average performance metrics and error bars show standard deviation across models and seeds.

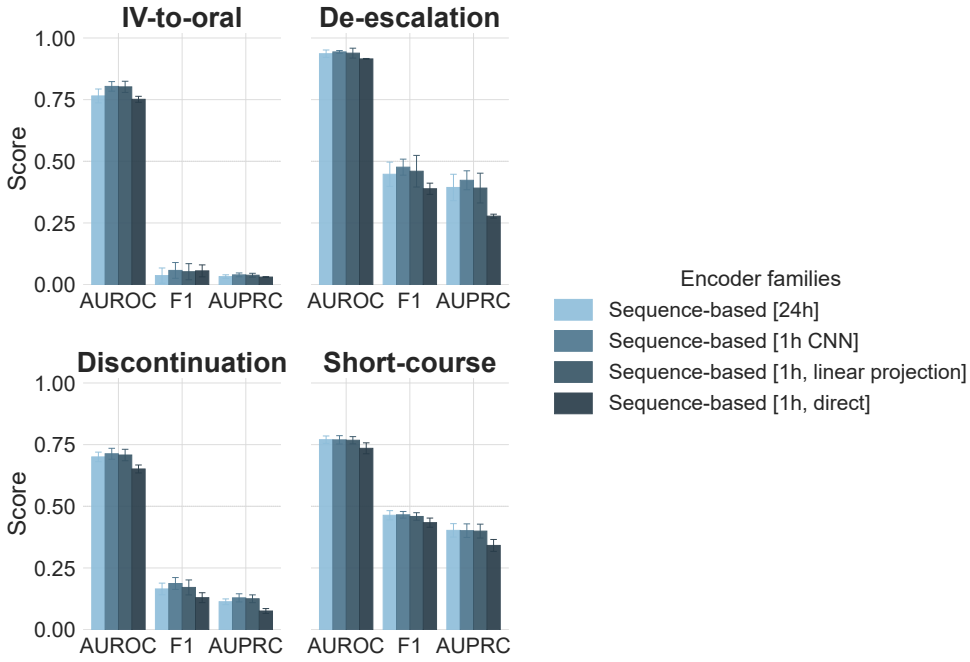

Figure 16: Encoder ablations of 1h sequence models on *MELODY-KISPI* data. Bars represent the average performance metrics and error bars show standard deviation across models and seeds.

C.9. Patient-temporal splitting

Table 15: Comparing patient-random and patient-temporal splitting in *MELODY-KISPI*. For patient-temporal splitting, patients are ordered by admission time, and a hard cutoff time T is set at the last 20% of patients. Train $< T$, test $\geq T$. AUROC, F1 score (threshold-optimized), and AUPRC are reported with mean scores and standard deviation across three seeds.

Target	Model	AUROC		F1		AUPRC	
		random	temporal	random	temporal	random	temporal
IV-to-oral	Logistic regression	0.847 ± 0.000	0.780 ± 0.000	0.083 ± 0.000	0.032 ± 0.000	0.065 ± 0.000	0.034 ± 0.000
IV-to-oral	LightGBM	0.797 ± 0.000	0.756 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.097 ± 0.000	0.035 ± 0.000
IV-to-oral	MLP	0.547 ± 0.070	0.583 ± 0.047	0.025 ± 0.009	0.019 ± 0.006	0.014 ± 0.002	0.010 ± 0.002
IV-to-oral	GRU	0.836 ± 0.013	0.753 ± 0.018	0.090 ± 0.014	0.033 ± 0.031	0.064 ± 0.008	0.026 ± 0.001
IV-to-oral	LSTM	0.838 ± 0.008	0.743 ± 0.001	0.096 ± 0.015	0.031 ± 0.027	0.064 ± 0.015	0.027 ± 0.001
IV-to-oral	TCN	0.836 ± 0.008	0.739 ± 0.013	0.115 ± 0.004	0.000 ± 0.000	0.070 ± 0.006	0.028 ± 0.002
IV-to-oral	Transformer	0.787 ± 0.059	0.781 ± 0.013	0.069 ± 0.034	0.039 ± 0.040	0.058 ± 0.017	0.039 ± 0.009
IV-to-oral	SSM	0.831 ± 0.005	0.791 ± 0.014	0.100 ± 0.042	0.041 ± 0.039	0.079 ± 0.013	0.038 ± 0.004
De-escalation	Logistic regression	0.900 ± 0.000	0.903 ± 0.000	0.388 ± 0.000	0.377 ± 0.000	0.282 ± 0.000	0.274 ± 0.000
De-escalation	LightGBM	0.932 ± 0.000	0.945 ± 0.000	0.241 ± 0.000	0.315 ± 0.000	0.432 ± 0.000	0.415 ± 0.000
De-escalation	MLP	0.723 ± 0.064	0.812 ± 0.090	0.255 ± 0.070	0.308 ± 0.088	0.172 ± 0.054	0.226 ± 0.073
De-escalation	GRU	0.943 ± 0.002	0.946 ± 0.000	0.491 ± 0.018	0.472 ± 0.013	0.413 ± 0.019	0.412 ± 0.010
De-escalation	LSTM	0.937 ± 0.004	0.943 ± 0.006	0.485 ± 0.014	0.483 ± 0.022	0.411 ± 0.024	0.410 ± 0.029
De-escalation	TCN	0.937 ± 0.003	0.945 ± 0.005	0.454 ± 0.024	0.483 ± 0.012	0.405 ± 0.008	0.398 ± 0.029
De-escalation	Transformer	0.933 ± 0.005	0.938 ± 0.011	0.418 ± 0.035	0.440 ± 0.035	0.386 ± 0.015	0.405 ± 0.035
De-escalation	SSM	0.932 ± 0.002	0.936 ± 0.006	0.400 ± 0.021	0.437 ± 0.035	0.378 ± 0.031	0.427 ± 0.040
Discontinuation	Logistic regression	0.665 ± 0.000	0.654 ± 0.000	0.139 ± 0.000	0.126 ± 0.000	0.074 ± 0.000	0.080 ± 0.000
Discontinuation	LightGBM	0.670 ± 0.000	0.665 ± 0.000	0.084 ± 0.000	0.065 ± 0.000	0.085 ± 0.000	0.090 ± 0.000
Discontinuation	MLP	0.623 ± 0.007	0.603 ± 0.013	0.121 ± 0.004	0.126 ± 0.015	0.073 ± 0.002	0.074 ± 0.006
Discontinuation	GRU	0.721 ± 0.024	0.739 ± 0.007	0.170 ± 0.022	0.199 ± 0.017	0.113 ± 0.025	0.139 ± 0.003
Discontinuation	LSTM	0.707 ± 0.019	0.711 ± 0.007	0.151 ± 0.018	0.173 ± 0.012	0.100 ± 0.007	0.126 ± 0.006
Discontinuation	TCN	0.714 ± 0.014	0.717 ± 0.013	0.168 ± 0.017	0.180 ± 0.025	0.119 ± 0.002	0.126 ± 0.003
Discontinuation	Transformer	0.702 ± 0.010	0.703 ± 0.006	0.153 ± 0.012	0.165 ± 0.019	0.101 ± 0.007	0.115 ± 0.003
Discontinuation	SSM	0.711 ± 0.023	0.676 ± 0.013	0.177 ± 0.022	0.165 ± 0.015	0.110 ± 0.019	0.113 ± 0.010
Short-course	Logistic regression	0.732 ± 0.000	0.726 ± 0.000	0.406 ± 0.000	0.427 ± 0.000	0.335 ± 0.000	0.338 ± 0.000
Short-course	LightGBM	0.762 ± 0.000	0.754 ± 0.000	0.312 ± 0.000	0.327 ± 0.000	0.364 ± 0.000	0.383 ± 0.000
Short-course	MLP	0.712 ± 0.036	0.707 ± 0.044	0.406 ± 0.030	0.400 ± 0.037	0.321 ± 0.028	0.314 ± 0.035
Short-course	GRU	0.791 ± 0.007	0.773 ± 0.005	0.463 ± 0.011	0.467 ± 0.018	0.419 ± 0.007	0.405 ± 0.006
Short-course	LSTM	0.796 ± 0.003	0.772 ± 0.008	0.485 ± 0.009	0.475 ± 0.011	0.422 ± 0.001	0.403 ± 0.010
Short-course	TCN	0.799 ± 0.003	0.766 ± 0.017	0.485 ± 0.007	0.458 ± 0.024	0.438 ± 0.012	0.405 ± 0.025
Short-course	Transformer	0.778 ± 0.006	0.756 ± 0.015	0.451 ± 0.016	0.454 ± 0.019	0.395 ± 0.007	0.370 ± 0.019
Short-course	SSM	0.765 ± 0.004	0.743 ± 0.003	0.441 ± 0.008	0.448 ± 0.003	0.381 ± 0.008	0.356 ± 0.005

C.10. Within-admission correlation of performance metrics

Table 16: Comparing AUROCs of models across targets for naive to clustered bootstrapping of test patient days in *PIC*.

Target	Model	AUROC [cluster 95% CI]	AUROC [naive 95% CI]
IV-to-oral	Logistic regression	0.934 [0.892–0.971]	0.934 [0.892–0.968]
IV-to-oral	LightGBM	0.932 [0.897–0.959]	0.932 [0.901–0.961]
IV-to-oral	MLP	0.926 [0.873–0.972]	0.926 [0.877–0.970]
IV-to-oral	GRU	0.956 [0.925–0.980]	0.956 [0.929–0.979]
IV-to-oral	LSTM	0.938 [0.893–0.970]	0.938 [0.895–0.974]
IV-to-oral	TCN	0.950 [0.913–0.977]	0.950 [0.917–0.977]
IV-to-oral	Transformer	0.950 [0.916–0.976]	0.950 [0.917–0.976]
IV-to-oral	SSM	0.954 [0.924–0.977]	0.954 [0.925–0.976]
De-escalation	Logistic regression	0.911 [0.899–0.923]	0.911 [0.901–0.920]
De-escalation	LightGBM	0.936 [0.928–0.945]	0.936 [0.929–0.942]
De-escalation	MLP	0.920 [0.908–0.931]	0.920 [0.910–0.928]
De-escalation	GRU	0.964 [0.957–0.969]	0.964 [0.959–0.971]
De-escalation	LSTM	0.963 [0.957–0.970]	0.963 [0.957–0.968]
De-escalation	TCN	0.964 [0.957–0.970]	0.964 [0.958–0.969]
De-escalation	Transformer	0.953 [0.946–0.959]	0.953 [0.947–0.957]
De-escalation	SSM	0.959 [0.953–0.965]	0.959 [0.954–0.964]
Discontinuation	Logistic regression	0.799 [0.777–0.821]	0.799 [0.775–0.821]
Discontinuation	LightGBM	0.842 [0.823–0.860]	0.842 [0.822–0.862]
Discontinuation	MLP	0.734 [0.708–0.760]	0.734 [0.708–0.759]
Discontinuation	GRU	0.841 [0.821–0.860]	0.841 [0.821–0.862]
Discontinuation	LSTM	0.843 [0.821–0.863]	0.843 [0.823–0.863]
Discontinuation	TCN	0.837 [0.817–0.858]	0.837 [0.817–0.855]
Discontinuation	Transformer	0.829 [0.805–0.851]	0.829 [0.809–0.848]
Discontinuation	SSM	0.815 [0.797–0.837]	0.815 [0.792–0.838]
Short course	Logistic regression	0.746 [0.729–0.763]	0.746 [0.728–0.762]
Short course	LightGBM	0.805 [0.788–0.820]	0.805 [0.787–0.819]
Short course	MLP	0.761 [0.744–0.782]	0.761 [0.744–0.779]
Short course	GRU	0.858 [0.845–0.871]	0.858 [0.847–0.869]
Short course	LSTM	0.844 [0.830–0.858]	0.844 [0.830–0.857]
Short course	TCN	0.843 [0.829–0.856]	0.843 [0.829–0.856]
Short course	Transformer	0.818 [0.803–0.833]	0.818 [0.804–0.832]
Short course	SSM	0.784 [0.768–0.801]	0.784 [0.769–0.802]