Expediting Switches from Intravenous to Oral Antimicrobial Therapy with Neural Processes

Magnus Ross[1*], **Nel Swanepoel**[2], **Akish Luintel**[3], **Emma McGuire**[3], **Ingemar J. Cox**[1,4], **Steve Harris**[2], **Vasileios Lampos**[1†]
[1] *Department of Computer Science, Centre for AI, UCL, UK*
[2] *Institute of Health Informatics, UCL, UK*
[3] *Department of Medical Microbiology, Infection Division, University College Hospital London, UK*
[4] *Department of Computer Science, University of Copenhagen, Denmark*

Abstract

Intravenous (IV) antimicrobial therapy often continues after an oral alternative would be safe, harming patients and increasing costs. The key decision—when to switch from IV to oral (IVOS)—is complex. Past machine-learning approaches are trained on labels constructed from prescription records that indicate when switches actually happened, not when they were first clinically appropriate; they therefore reproduce delays and suboptimal practice. We take a different approach, forecasting each patient's physiological state using a probabilistic neural process and applying established IVOS criteria to those forecasts to estimate switch-readiness. The system ranks patients currently on IV therapy by the probability they will meet all criteria within the next 12 hours, highlighting candidates for clinician review while keeping the final decision with the prescriber. By anchoring recommendations to clinical criteria rather than past actions, the tool targets a reduction in unnecessary IV days instead of reinforcing the status quo.

Keywords: clinical decision support, antimicrobial therapy, antibiotics, time series forecasting, neural processes

Data and Code Availability We use the MIMIC-IV dataset, which is publicly available (Johnson et al., 2023). The source code is available at github.com/SAFEHR-data/ivos-model.

Institutional Review Board (IRB) This project has been approved by the UCL Computer Science Low Risk Ethics Committee (Ref. CS LREC-2025-1347).

* Email: magnus.ross@ucl.ac.uk
† Email: v.lampos@ucl.ac.uk

1. Introduction

The timely transition of hospital patients from intravenous (IV) to oral antimicrobial therapy, known as the IV-to-oral switch (IVOS), has positive effects for both patients and hospitals (Cyriac and James, 2014). The benefits range from clinical, e.g. shorter hospital stays or a lower risk of catheter-related bloodstream infections (Wald-Dickler et al., 2022; Zanella et al., 2025), to operational, such as freeing up nursing time and reduced healthcare costs (Jenkins, 2023). However, identifying the optimal moment for this switch depends on a holistic assessment of the patient's physiological state. The decision can be overlooked, resulting in the IV course being maintained unnecessarily. Clinicians rely on a range of criteria, such as trends in vital signs (e.g. temperature), to determine a patient's suitability for IVOS (Harvey et al., 2023). In many cases, however, these switches happen later than recommended by guidelines—estimates indicate 19% of patients in England still receive IV antibiotics despite switching criteria being met (Office for Health Improvement & Disparities, 2025)—or not at all (Li et al., 2019), resulting in avoidable costs for both patients and hospitals.

Recently, clinical decision support systems (CDSSs) have been proposed to prompt timely IVOS (Quintens et al., 2022; Kan et al., 2019; Berrevoets et al., 2017; Beeler et al., 2015). Notably, Bolton et al. (2024) propose an AI-enabled CDSS and predict "readiness" for switching using labels derived from routine prescription records. These labels record when a switch actually occurred, not the earliest clinically appropriate time. Models trained on such labels learn historical delays and suboptimal practice. As a result, they risk preserving current behaviour rather than

reducing unnecessary IV days. Additionally, the automatic labelling of these events from raw prescription data is a complex and challenging task, which we show results in incorrectly labelled training examples, and misleading estimates of the length of antibiotic courses.

In this work, we outline a plan for an alternative AI-enabled CDSS which resolves these issues. At the core of our approach is a probabilistic model which can effectively forecast the patient's physiological state with well-calibrated uncertainty. Instead of modelling the past decisions of clinicians we apply established switching criteria to these forecasts to determine if a patient will be ready for switching in the near future. We present this information to clinicians in the form of a ranked list, from which the top patients are selected for manual review. This allows us to sidestep the fraught process of constructing switching labels from electronic health records (EHRs), and produce a system that can effectively reduce the number of unnecessary IV therapy days, whilst keeping a human in the loop.

2. Previous approaches

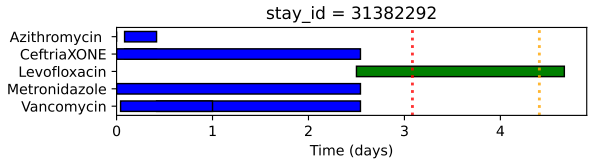
Bolton et al. (2024) build a model to assess the suitability of a patient for IVOS. Their approach is to train a model to predict, for a given day of a patient's encounter, whether that patient received IV antibiotics. The model's prediction, specifically the probability that the patient will *not* be on IV antibiotics, is used as a proxy for the probability that the patient is suitable for a switch to oral therapy.

2.1. Learning from prior practice

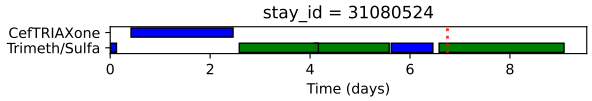
By training a model to predict historical clinical actions (i.e., the administration of IV antibiotics), the resulting system will necessarily learn to replicate the existing patterns and frequency of IVOS decisions present in the training data. Clinical practice is changing, for many conditions there is growing evidence that long courses of IV antibiotics are unnecessary (Wald-Dickler et al., 2022; Li et al., 2019; Iversen et al., 2019). If the goal of a CDSS is to improve upon current practice—in this case, by encouraging earlier switches and thus reducing the total number of IV therapy days—then a model trained to mimic historical, suboptimal behaviour is inherently limited. This limitation is particularly acute for the complex or non-standard cases where clinical decision support is most needed.

Although performance of such a model on held-out retrospective data might appear encouraging, this approach inevitably results in a brittle CDSS, frozen in time and incapable of providing continuous support to clinicians participating in a resilient Learning Health System (Friedman, 2022; Kilbourne et al., 2024). Our approach of modelling patient physiology keeps the prescribing clinician front and centre and results in a CDSS sufficiently flexible to adapt to clinical evidence and changing guidelines.

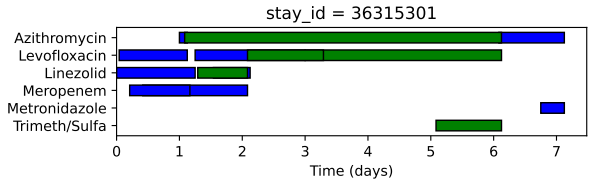
2.2. Labelling switches



(a) Patient dies ($\cdots$) shortly after being labelled as ready for switch ($\cdots$).



(b) Initial switch incorrectly labelled on day 6 ($\cdots$), when switch truly occurred on day 3.



(c) Encounter dropped from the dataset despite containing a valid switch.

Figure 1: Examples of incorrectly processed antibiotic prescribing timelines from the MIMIC-IV dataset, with blue bars (■) representing IV prescriptions and green bars (■) oral.

Even setting aside the problem of learning from past practice, constructing reliable ground-truth IVOS labels from EHRs is difficult, and label quality caps model performance and clinical utility. The approach to label construction taken by Bolton et al. (2024) contains several methodological issues that result in a dataset that is noisy and systematically biased.

Firstly, the methodology does not incorporate patient outcomes into the labelling process. Our analysis of the MIMIC-IV dataset revealed cases where a patient was labelled as being suitable for a switch on the same day that they died. As shown in Figure 1(*a*), this represents a severe mislabelling.

Secondly, IV and oral therapy periods are defined by taking the earliest start time and the latest end time across all antibiotic prescriptions of a given administration route within a single hospital stay. This can cause switches to be labelled incorrectly and systematically excludes more complex, but common, clinical scenarios. Figure 1(*b*) shows an example where the switch label only appears for the second time the patient was switched in the encounter on day 6, where a switch initially occurred on day 3. Figure 1(*c*) shows an encounter containing a valid switch event that has been dropped because the patient is later returned to IV antibiotics. The cumulative effect is to bias the final dataset towards unrealistically simple and linear treatment trajectories.

Finally, all antibiotics are treated as interchangeable, with no check to ensure that the oral drug is a clinically appropriate replacement for the IV drug, or that the prescription indications are for the same condition. For example, treatment with broad-spectrum IV antibiotics can cause *C. difficile* infection (Cymbal et al., 2024). In serious cases, this requires the IV course to be halted, and an oral antibiotic, often vancomycin, used to control the *C. difficile*. Disregarding indications, this would be labelled as a switch, when one has not in fact occurred.

3. Proposed AI-enabled IVOS approach

The approach outlined in the present work seeks to decouple the modelling from past, subjective clinical practice. Instead of modelling the decision itself, we predict the underlying physiological state of the patient by forecasting the clinical variables required to assess IVOS criteria. This formulation allows for the possibility of recommending a switch even in cases where one would not have been performed historically, provided the patient's forecasted physiological state meets the established criteria for a safe transition to oral therapy (Harvey et al., 2023).

Forecasting model The core of our approach is a probabilistic model that directly forecasts the patient's physiological state from their history of sparse and irregularly-sampled clinical measurements, using the framework of *neural processes* (Garnelo et al., 2018; Jha et al., 2023). Specifically, we use a Convolutional Conditional Neural Process (ConvCNP), a model well suited for this time series task due to its inclusion of the inductive bias of translation equivariance (Bruinsma, 2024). Neural processes treat each patient encounter as a realisation of an underlying stochastic process, allowing the model to learn a global prior over patient dynamics from a large dataset of encounters. Unlike many traditional time-series models, they naturally handle irregular sampling without requiring imputation and produce probabilistic predictions. We use the trained model to forecast each patient's physiological trajectory over the subsequent day. We forecast 5 variables relating to the patient's physiological state: temperature, pulse rate, respiration rate, systolic blood pressure, and oxygen saturation. These variables were chosen because they are widely recorded, however our method can trivially be extended to forecast other variables commonly part of IVOS criteria, such as C-reactive protein levels and white cell counts.

Probabilistic switch-readiness prediction To obtain an actionable assessment of IVOS suitability, we leverage the model's full probabilistic output rather than rely on a single point prediction. We employ a sampling-based approach to estimate a 'switch readiness probability'. For each patient, we draw multiple complete trajectories of their future physiological state from the model's joint predictive distribution. Each sampled trajectory is then evaluated against the full set of established clinical IVOS criteria (e.g. temperature within normal range and stable vital signs). The final probability is calculated as the proportion of these sampled trajectories that meet all criteria throughout the forecast window.

CDSS After consultation with clinical experts, we determined that the optimal way to present the model predictions is a ranked list of patients currently receiving IV antimicrobial therapy sorted by predicted likelihood of switch-readiness over the next day. This list will show other relevant information about the patient, such as how long they have been on IV and the most recent values of important physiological variables. This will allow clinicians to review and determine follow-up actions more efficiently, by presenting them with the patients most likely to need a switch, whilst keeping the final switch decision in their hands.

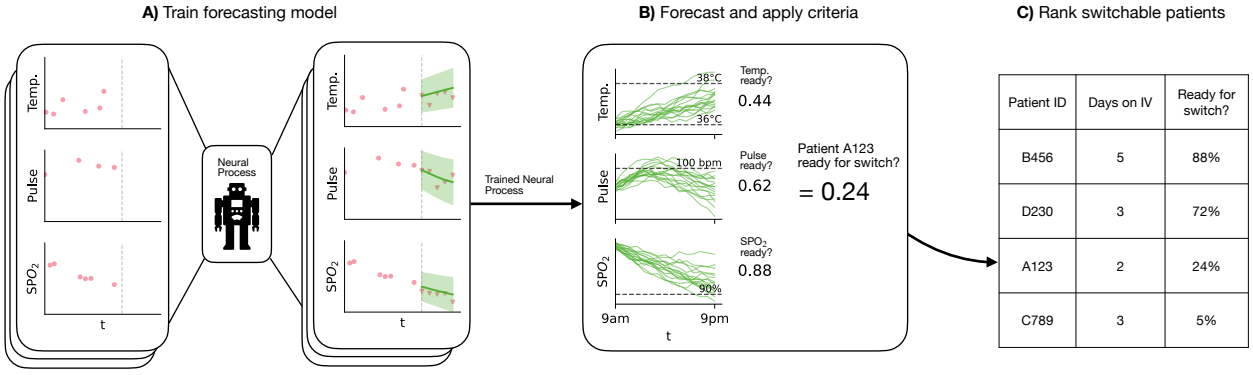


Figure 2: The proposed AI-enabled IVOS system. **A)** Training a probabilistic neural process model to forecast patient trajectories. **B)** Using the trained model to generate likely trajectory samples over the following day and applying clinical criteria to determine the proportion of samples ready to be switched. **C)** Presentation of switch-readiness probabilities as a ranked list of patients for review.

	Baselines		
Metric	Majority	Last val.	ConvCNP
Brier ↓	0.16	0.47	**0.10**
Precision@5 ↑	0.13	0.22	**0.56**

Table 1: Performance metrics for the model and baselines. Note the Precision@5 for a model with perfect accuracy is 0.95 in this case.

4. Results

We provide provisional results by fitting a ConvCNP model to physiological trajectories extracted from the MIMIC-IV dataset (Johnson et al., 2023). We split the dataset temporally into train and test sets using reconstructed admission times. We train the forecasting model using a look-back window of 48 hours— generally the minimum length of time before a prescription is reviewed—and a prediction window of 12 hours, on all encounters in the training set, regardless of whether they have received an antibiotic prescription. In order to evaluate the model's ability to predict switch readiness, we filter the test set for encounters that have had an IV prescription at any point. For this set, we mirror the intended deployment by tasking the model to predict the patient's physiological trajectory over the next 12 hours for each day of their encounter. We sample 200 trajectories and apply a set of criteria representing the normal healthy range for each variable; the proportion of samples meeting the criteria is the predicted probability of switch-readiness. The criteria are also applied to the ground truth data in the prediction window to generate target labels. We obtain a total of 34,885 test days from 9,590 encounters; 16% of these days are of the positive class, which is to say the patient is ready to be switched according to the criteria. Further experimental and model setup details are provided in Appendix A.

We compare to two baselines: *majority class* simply predicts the dominant class for every instance and *last value* applies the criteria to the last measured value for each variable before the prediction time. Given the deployment context discussed in Section 3 we are interested in measuring the model's ability to produce well-calibrated probabilities and to rank patients accurately, instead of accuracy at a set decision threshold. We report the Brier score (Brier, 1950) and Precision@5. The Precision@5 is computed by producing a ranked list of predictions for each day in the test set, for all the encounters that span that day, excluding days that have fewer than 20 encounters (25% of the test days); the Precision@5 is the proportion of the top 5 ranked encounters that truly meet criteria. This metric is intended to simulate a doctor selecting patients to review from the ranked list. We additionally report the AUROC for the ConvCNP model. Further discussion of metrics is provided in Appendix B. Table 1 shows the metrics for the NP model and the two baseline methods. We can see that the NP model outperforms each of the baselines on both metrics. In particular, the Precision@5 metric indicates a more than two-fold improvement in the number of relevant patients likely to be selected for review. The

AUROC for the ConvCNP model is 0.86, above the threshold generally considered to be clinically useful (Çorbacıoğlu and Aksel, 2023). Appendix C contains some additional results, with Table 5 demonstrating the forecasting performance of the ConvNP model relative to a simple persistence baseline, and Figure 3 showing model predictions for a number of test days.

5. Conclusions

Our goal is to create a clinically useful tool that helps get patients off IV antibiotics sooner. By forecasting patient physiology instead of past decisions, we can meaningfully reduce the number of unnecessary days of IV therapy, leading to better outcomes for patients and hospitals. We are currently working on implementing a prospective evaluation of the proposed system, which we hope to present in future work.

Acknowledgments

MR, IJC, and VL are supported by the EPSRC Digital Health Hub for AMR (EP/X031276/1). SH is supported by the National Institute for Health and Care Research University College London Hospitals Biomedical Research Centre.

References

Patrick E. Beeler, Stefan P. Kuster, Emmanuel Eschmann, Rainer Weber, and Jürg Blaser. Earlier switching from intravenous to oral antibiotics owing to electronic reminders. *International Journal of Antimicrobial Agents*, 46(4):428–433, 2015. doi: 10.1016/j.ijantimicag.2015.06.013.

Marvin A. H. Berrevoets, Johannes (Hans) L. W. Pot, Anne E. Houterman, Anton (Ton) S. M. Dofferhoff, Marrigje H. Nabuurs-Franssen, Hanneke W. H. A. Fleuren, Bart-Jan Kullberg, Jeroen A. Schouten, and Tom Sprong. An electronic trigger tool to optimise intravenous to oral antibiotic switch: a controlled, interrupted time series study. *Antimicrobial Resistance & Infection Control*, 6(1):81, 2017. doi: 10.1186/s13756-017-0239-3.

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), 2024. doi: 10.1038/s41467-024-44740-2.

Glenn W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1, 1950. ISSN 0027-0644. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Wessel P. Bruinsma. *Convolutional Conditional Neural Processes*. PhD thesis, 2024.

Michael Cymbal, Arjun Chatterjee, Brian Baggott, and Moises Auron. Management of Clostridioides difficile Infection: Diagnosis, Treatment, and Future Perspectives. *The American Journal of Medicine*, 137(7):571–576, 2024. doi: 10.1016/j.amjmed.2024.03.024.

Jissa Maria Cyriac and Emmanuel James. Switch over from intravenous to oral therapy: A concise overview. *Journal of Pharmacology & Pharmacotherapeutics*, 5(2):83–87, 2014. doi: 10.4103/0976-500X.130042.

Charles P. Friedman. What is unique about learning health systems? *Learning Health Systems*, 6(3), 2022. ISSN 2379-6146. doi: 10.1002/lrh2.10328.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional Neural Processes. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018.

Eleanor J Harvey, Monsey McLeod, Caroline De Brún, and Diane Ashiru-Oredope. Criteria to achieve safe antimicrobial intravenous-to-oral switch in hospitalised adult populations: a systematic rapid review. *BMJ Open*, 13(7):e068299, 2023. doi: 10.1136/bmjopen-2022-068299.

Kasper Iversen, Nikolaj Ihlemann, Sabine U. Gill, Trine Madsen, Hanne Elming, Kaare T. Jensen, Niels E. Bruun, Dan E. Høfsten, Kurt Fursted, Jens J. Christensen, Martin Schultz, Christine F. Klein, Emil L. Fosbøll, Flemming Rosenvinge, Henrik C. Schønheyder, Lars Køber, Christian Torp-Pedersen, Jannik Helweg-Larsen, Niels Tønder, Claus Moser, and Henning Bundgaard. Partial Oral versus Intravenous Antibiotic Treatment of Endocarditis. *New England Journal of Medicine*, 380(5):415–424, 2019. doi: 10.1056/NEJMoa1808312.

Abi Jenkins. IV to oral switch: a novel viewpoint. *The Journal of Antimicrobial Chemotherapy*, 78(10): 2603–2604, 2023. doi: 10.1093/jac/dkad239.

Saurav Jha, Dong Gong, Xuesong Wang, Richard E. Turner, and Lina Yao. The Neural Process Family: Survey, Applications and Perspectives, 2023.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1): 1, 2023. doi: 10.1038/s41597-022-01899-x.

Tiffany Kan, Derrick Kwan, Thomas Chan, Pavani Das, and Sumit Raybardhan. Implementation of a Clinical Decision Support Tool to Improve Antibiotic IV-to-Oral Conversion Rates at a Community Academic Hospital. *The Canadian Journal of Hospital Pharmacy*, 72(6):455–461, 2019.

Amy M. Kilbourne, Amanda E. Borsky, Robert W. O'Brien, Melissa Z. Braganza, and Melissa M. Garrido. The foundational science of learning health systems. *Health Services Research*, 59(6), 2024. ISSN 1475-6773. doi: 10.1111/1475-6773.14374.

Ho-Kwong Li, Ines Rombach, Rhea Zambellas, A. Sarah Walker, Martin A. McNally, Bridget L. Atkins, Benjamin A. Lipsky, Harriet C. Hughes, Deepa Bose, Michelle Kümin, Claire Scarborough, Philippa C. Matthews, Andrew J. Brent, Jose Lomas, et al. Oral versus Intravenous Antibiotics for Bone and Joint Infection. *New England Journal of Medicine*, 380(5):425–436, 2019. doi: 10.1056/NEJMoa1710926.

Office for Health Improvement & Disparities. Public health profiles: Percentage of patients still receiving intravenous (iv) antibiotics past the point at which they meet iv-oral switch criteria. https://fingertips.phe.org.uk/profile/amr-local-indicators, 2025. [Accessed 08 September 2025].

Charlotte Quintens, Marie Coenen, Peter Declercq, Minne Casteels, Willy E Peetermans, and Isabel Spriet. From basic to advanced computerised intravenous to oral switch for paracetamol and antibiotics: an interrupted time series analysis. *BMJ Open*, 12(4):e053010, 2022. doi: 10.1136/bmjopen-2021-053010.

Noah Wald-Dickler, Paul D. Holtom, Matthew C. Phillips, Robert M. Centor, Rachael A. Lee, Rachel Baden, and Brad Spellberg. Oral Is the New IV. Challenging Decades of Blood and Bone Infection Dogma: A Systematic Review. *The American Journal of Medicine*, 135(3):369–379.e1, 2022. doi: 10.1016/j.amjmed.2021.10.007.

Marie-Céline Zanella, Gaud Catho, Holly Jackson, Nasim Lotfinejad, Valérie Sauvan, Marie-Noëlle Chraiti, Walter Zingg, Stephan Harbarth, and Niccolò Buetti. Dwell Time and Risk of Bloodstream Infection With Peripheral Intravenous Catheters. *JAMA Network Open*, 8(4):e257202, 2025. doi: 10.1001/jamanetworkopen.2025.7202.

Şeref Kerem Çorbacıoğlu and Gökhan Aksel. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine*, 23(4):195–198, 2023. ISSN 2452-2473. doi: 10.4103/tjem.tjem_182_23.

Appendix A. Experimental setup

A.1. Data Preprocessing

All experiments use the MIMIC-IV v2.2 dataset (Johnson et al., 2023).

Cohort Selection We first select a cohort of hospital encounters corresponding to patients admitted to the ICU, with a length of stay greater than one day and a recorded age. For this cohort, we extract all measurements for the five physiological variables relevant to our switching criteria: body temperature, pulse rate, respiratory rate, oxygen saturation, and systolic blood pressure. These are sourced from the `chartevents` table.

Data Cleaning and Filtering The timestamp of each measurement is converted to a duration in hours relative to the patient's admission time. We discard encounters with no measurement data beyond 24 hours from admission and truncate all encounters at a maximum length of 14 days. To handle erroneous entries common in EHR data, we filter measurements to within clinically plausible ranges, as detailed in Table A.1. This filtering step removed a maximum of 0.05% of measurements for any single variable.

Variable	Range	Units
Pulse Rate	10 – 400	bpm
Systolic Blood Pressure	0 – 400	mmHg
Oxygen Saturation	0 – 100	%
Respiration Rate	0 – 120	breaths/min
Temperature	50 – 120	°F

Table 2: Ranges of plausible values used for data cleaning.

Train/Validation/Test Split A key step in our experimental design is a strict temporal split of the data, which is necessary to simulate a realistic deployment scenario and prevent look-ahead bias. The de-identification process in MIMIC-IV shifts all dates for a given patient by a random offset, which preserves intra-patient timelines but scrambles the chronological order of admissions between different patients. A random split on this scrambled data would incorrectly mix past and future data, leading to an over-optimistic evaluation.

To address this, we reconstruct an approximate, chronologically-sortable admission date for each encounter. Following the methodology described by the MIMIC-IV authors (Johnson et al., 2023), we combine the real admission year, available in the `admissions` table as `anchor_year`, with the month and day from the shifted `admittime`. While this does not recover the exact date, it allows us to correctly order all encounters by year. This is a significant improvement over prior work that ignores the temporal nature of the data.

Using these reconstructed dates, we create our splits: all encounters admitted after January 1, 2019, form the test set. From the remaining pre-2019 encounters, we randomly sample 10% to create the validation set, with the rest forming the training set. Finally, all variables are standardised (zero mean, unit variance) using statistics computed solely from the training set.

A.2. Forecasting Model & Training

We use a Convolutional Conditional Neural Process (ConvCNP) implemented with the `neuralprocesses` package.[1]

Task Construction The forecasting model is trained on tasks constructed from the preprocessed encounters. A task consists of a context set and a target set. For each encounter, we create tasks by uniformly sampling prediction start times. Each task uses a 48-hour look-back window (context) to predict all measurements within a subsequent 12-hour forecast horizon (target). To create a large training set and manage computational load, we sample tasks with a frequency proportional to the length of the encounter, generating one task for every 24 hours of data. Any sampled task with fewer than 10 total measurements across both context and target sets is discarded.

Model Selection and Training We performed a grid search to select the best hyperparameters based on validation set performance. The search space is detailed in Table A.2. The best-performing model used a 4-layer U-Net with 512 channels, a discretisation length scale of 1.0 hour, and a learning rate of 10^{-5}. For more detail on neural process architectures, see Bruinsma (2024). All models were trained with the Adam optimizer and a batch size of 64. We defined a training epoch as 2^{14} tasks and used an early stopping mechanism with a patience of 100 epochs, saving the model checkpoint with the lowest validation loss.

1. github.com/wesselb/neuralprocesses

Hyperparameter	Values
Learning Rate	{1e-4, 1e-5}
U-Net Channels	{256, 512}
U-Net Layers	{3, 4}
Discretisation scale (hours)	{0.5, 1.0}

Table 3: Grid of hyperparameters explored for the ConvCNP model.

A.3. Switch-Readiness Evaluation

To evaluate the model's utility in a clinical setting, we simulate its deployment on the test set.

Evaluation Task Construction We filter the test set to include only encounters with at least one IV antibiotic prescription. For each of these encounters, we create one evaluation task per day, starting from the third day of the stay. Each task is anchored at 09:00, uses the preceding 48 hours of data as context, and has a forecast horizon of 12 hours (from 09:00 to 21:00).

Label Generation A ground-truth label for each evaluation day is generated by checking if *all* recorded measurements within the 12-hour forecast window fall within the switch-readiness criteria defined in Table A.3. If any measurement violates its criterion, the day is labelled as not ready (negative class). If a variable has no measurements on a given day, it is considered to have met the criteria. This process yielded 34,885 test days, of which 16% were positive (switch-ready).

Measurement	Switch Criterion
Temperature (°F)	96.8 – 100.4
Pulse Rate (bpm)	< 100
Respiratory Rate (breaths/min)	8 – 24
Systolic BP (mmHg)	90 – 220
SpO2 (%)	> 90

Table 4: Physiological criteria for switch readiness.

Probabilistic Prediction For each evaluation task, we generate a switch-readiness probability. We draw 200 full trajectory samples from the ConvCNP's predictive distribution over the 12-hour forecast horizon. For each sample, we check if all forecasted values across all five variables meet the criteria in Table A.3 at all points in time. The final predicted probability is the proportion of the 200 samples that successfully meet all criteria.

Appendix B. Metrics

Here we provide additional details on the computation of the evaluation metrics, particularly for the baseline models.

Brier Score The Brier score measures the accuracy of probabilistic predictions, penalising both miscalibration and incorrectness (Brier, 1950). It is the mean squared error between the predicted probability and the actual outcome. For a set of N predictions, it is calculated as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$$

where p_i is the predicted probability for the i-th instance and o_i is the actual outcome (1 if the patient meets the switch criteria, 0 otherwise). A lower Brier score indicates a better model. For the ConvCNP model, p_i is the continuous probability output by the model. For the baseline models, which produce a binary prediction, we interpret their output as a probability of 100% (for a positive prediction) or 0% (for a negative prediction).

Precision@5 Precision@k is a ranking metric that evaluates the fraction of positive instances within the top k predictions of a ranked list. We use Precision@5 (P@5) to simulate the clinical utility of our system, where a clinician might review the top 5 patients flagged by the CDSS each day. Generating the ranked list for the different models is done as follows:

- ConvCNP: Instances are ranked in descending order of their predicted switch-readiness probability.
- Baselines: These models produce a binary output. To generate a ranked list, all instances predicted as positive (1) are ranked above all instances predicted as negative (0). Within each group (all 1s or all 0s), the order is random, as the models provide no further information for ranking. The top 5 instances are then taken from this list to calculate P@5.

We compute the Precision@5 on each day in the test set that has more than 20 patients, and report the mean value over all such days.

AUROC The Area Under the Receiver Operating Characteristic (AUROC) curve evaluates a model's ability to discriminate between positive and negative classes. The ROC curve is created by plotting the true positive rate against the false positive rate at various classification thresholds. For our ConvCNP model, this threshold is applied to the continuous probability output. A higher AUROC, with a maximum of 1, indicates better discriminative power. This metric is only reported for our main model, as the deterministic nature of the baselines does not allow for varying a decision threshold to generate a curve.

Appendix C. Further results

C.1. Forecasting performance

Measurement	Last val.	ConvCNP
Pulse (bpm)	9.13 ± 0.02	8.23 ± 0.02
Respiratory Rate (bpm)	4.07 ± 0.01	3.45 ± 0.01
SpO_2 (%)	2.03 ± 0.00	1.75 ± 0.00
Systolic BP (mmHg)	4.2 ± 0.0	11.9 ± 0.01
Temperature (°F)	0.635 ± 0.003	0.557 ± 0.002

Table 5: Forecast mean absolute error on the test set for repeat last value baseline, and proposed ConvCNP model. The mean and standard error over the individual task error are shown. In the case where the variable is missing the last value, the mean value of the training set is used as the prediction

C.2. Model predictions

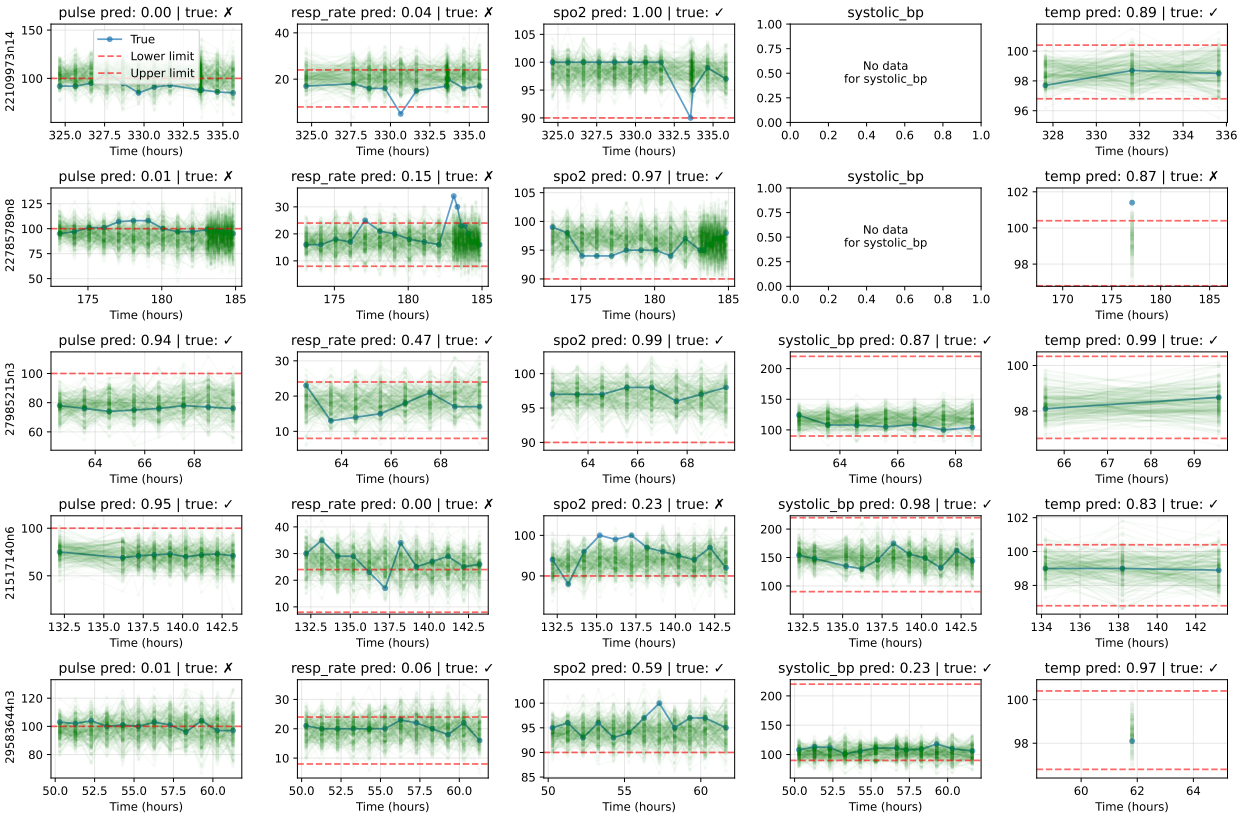


Figure 3: Plots showing forecasts of different variables (columns) for 5 days from the test set (rows). Each plot shows predicted model samples (translucent green), ground truth measurements (solid blue), and the criteria limits (dotted red) from Table A.3. In the title of each subplot is the predicted probability of criteria fulfilment, as well as the ground truth for each variable. All examples except the middle row have ground truth "no switch". Examples in the first two rows indicate it is possible for variables to be missing data in the prediction window