Early Prediction of Causes (not Effects) in Healthcare by Long-Term Clinical Time Series Forecasting

Michael Staniek STANIEK@CL.UNI-HEIDELBERG.DE
Marius Fracarolli FRACAROLLI@CL.UNI-HEIDELBERG.DE
Michael Hagmann HAGMANN@CL.UNI-HEIDELBERG.DE
Stefan Riezler RIEZLER@CL.UNI-HEIDELBERG.DE
Department of Computational Linguistics and
Interdisciplinary Center for Scientific Computing (IWR)
Heidelberg University, Germany

Abstract

Machine learning for early syndrome diagnosis aims to solve the intricate task of predicting a ground truth label that most often is the outcome (effect) of a medical consensus definition applied to observed clinical measurements (causes), given clinical measurements observed several hours before. Instead of focusing on the prediction of the future effect, we propose to directly predict the causes via time series forecasting (TSF) of clinical variables and determine the effect by applying the gold standard consensus definition to the forecasted values. This method has the invaluable advantage of being straightforwardly interpretable to clinical practitioners, and because model training does not rely on a particular label anymore, the forecasted data can be used to predict any consensus-based label. We exemplify our method by means of long-term TSF with Transformer models, with a focus on accurate prediction of sparse clinical variables involved in the SOFA-based Sepsis-3 definition and the new Simplified Acute Physiology Score (SAPS-II) definition. Our experiments are conducted on two datasets and show that contrary to recent proposals which advocate set function encoders for time series and direct multi-step decoders, best results are achieved by a combination of standard dense encoders with iterative multi-step decoders. The key for success of iterative multi-step decoding can be attributed to its ability to capture cross-variate dependencies and to a student forcing training strategy that teaches the model to rely on its own previous time step predictions for the next time step prediction.

1. Introduction

Early detection of syndromes like Sepsis is key to prevent a rapid progression to deadly stages by timely clinical invervention (Ferrer et al., 2014; Rudd et al., 2020). Machine learning from electronic health records (EHRs) bears the big promise of enabling early prediction of syndromes from historic measurements of vital and physiological parameters and laboratory test results. Inspired by supervised machine learning, early syndrome prediction is often framed as predicting a future label given clincial variables observed so far. Crucial prerequisites in this regime are the availability of syndrome labels and the exact time of syndrome onset. However, such labels are often not routinely documented by clinicians. Furthermore, in the rare case where they are part of the information provided in EHRs, caution is advised in interpreting the chart time of diagnosis as the true time of syndrome

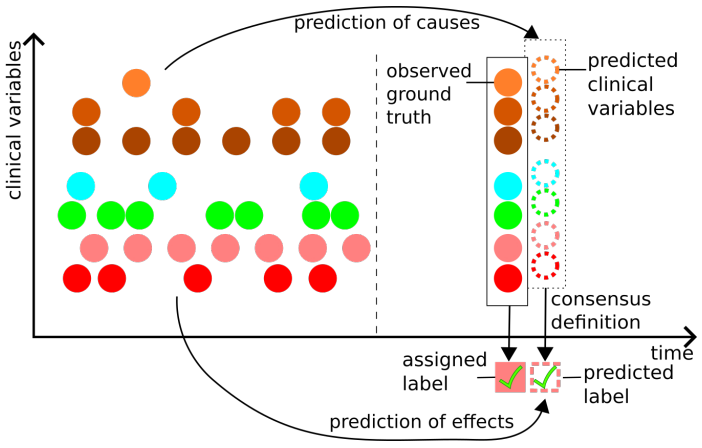

Figure 1: Based on clinical variables observed several hours before the prediction time point (part left of the dashed line), a prediction of effects (bottom arrow) aims to predict the label that the consensus definition assigns to unseen future observations. In contrast, a prediction of causes (top arrow) directly predicts future values of clinical variables and then deterministically applies the consensus definition to these predictions.

onset.[1] Most machine learning approaches to early prediction of syndromes thus resort to automatically creating ground truth labels by applying consensus definitions, for example, the Sepsis-3 definition (Seymour et al., 2016; Singer et al., 2016). Such consensus definitions are widely accepted as the basis of ICD codes and build on standard clinical measurements in EHR databases. For example, the Sepsis-3 label is assigned by determining an infection accompanied by an organ function deterioration according to the SOFA (Sepsis-related Organ Failure Assessment) criteria (Vincent et al., 1996), which are themselves based on thresholding various fundamental clinical measurements. The time of sepsis onset is determined by the onset of salient organ dysfunction or the onset of suspicion of infection, or by the earlier of these two events (Cohen et al., 2024). Examples for early prediction of sepsis using this automatic labeling scheme are the approaches described in the overview papers of Reyna et al. (2019) or Moor et al. (2021).

From a machine learning perspective, the setup of predicting a label that is the outcome of a consensus definition can be abstractly viewed as the *prediction of an effect* that is caused by applying a consensus definition to future values of clinical variables. We propose to turn this setup around: Instead of predicting an effect based on measurements taken at an earlier time than the values causing the ground truth label, we *directly predict the*

1. For example, in the two datasets that to our awareness provide expert annotations of the exact time of sepsis diagnosis, one dataset only reports a normalized daily diagnosis time of 2 p.m. (Schamoni et al., 2019; Lindner et al., 2022), and the other (Pollard et al., 2018) explicitly warns that the "diagnoses that were documented in the ICU stay by clinical staff [...] may or may not be consistent with diagnoses that were coded and used for professional billing or hospital reimbursement purposes." (Documentation of diagnosis field of eICU Collaborative Research Database, available at https://eicu-crd.mit.edu/eicutables/diagnosis/.)

causes by forecasting clinical variables, and determine the effect by applying the consensus definition to the forecasted values. This setup is illustrated in Figure 1.

Prediction of causes has a different goal than prediction of effects in that it aims at the general task of forecasting clinical variables to which arbitrary consensus definitions can be applied, without being restricted to a specific definition that is used to annotate ground truth labels. This flexibility comes at the cost of constituting an arguably harder task since several irregularly and sparsely sampled clinical variables need to be predicted instead of a single non-sparse effect label. However, despite its challenges, it has several intrinsic benefits. First and foremost, a prediction of clinical causes has the invaluable advantage of being immediately interpretable by clinical practitioners. Instead of requiring post-hoc interpretability techniques to understand a SOFA or Sepsis prediction, the predicted clincial causes determining the effect of SOFA or Sepsis can be directly inspected. Furthermore, if a consensus definition is applied to predicted values of clinical variables, there is no room for shortcut learning (Geirhos et al., 2020), information leakage (Kaufmann et al., 2011) or circularity (Hagmann et al., 2023; Riezler and Hagmann, 2024) that can skew a machine learning task. Last, the deterministic application of a consensus definition directly reveals aspects like the choice of onset time that can have a great impact on predictive performance (Cohen et al., 2024). It allows a direct investigation of such subtle variations during testing, while such variations are much harder to explore in approaches where consensus definitions are hard-coded and hidden in the training labels.

In the following, we present an exemplification of prediction of causes by means of long-term time series forecasting (TSF) of clinical variables, with a focus on accurate prediction of the sparse clinical variables involved in the SOFA-based Sepsis-3 definition (Seymour et al., 2016; Singer et al., 2016) and the new Simplified Acute Physiology Score (SAPS-II) (Gall et al., 1993) definition. We present an extensive investigation of optimal network architectures and training procedures to perform long-term clinical TSF on two critical care databases, MIMIC-III (Johnson et al., 2016) and eICU (Pollard et al., 2018). We build upon the expressive family of Transformers (Vaswani et al., 2017) that has been shown to be able to model dynamical or even chaotic systems (Geshkovski et al., 2023; Inoue et al., 2022) and has been applied competitively to various TSF tasks (Ahmed et al., 2023; Wen et al., 2023). We compare *sparse encoders* specialized to modeling irregularly sampled time series data as sets of observation triplets, containing clinical variable, time of measurement, and measurement value (Tipirneni and Reddy, 2022) to a standard encoding of time series as a *dense matrix* of input features times time-steps, using binning and mean-value imputation. These encoders are combined with a *direct multi-step (DMS)* forecasting decoder (Zhou et al., 2021; Wu et al., 2021; Zeng et al., 2023) that has been proposed specifically for long-term TSF, and a standard autoregressive Transformer decoder performing long-term TSF in an *iterative multi-step (IMS)* fashion. Our experiments show that contrary to recent proposals, best results are achieved by a combination of a standard dense encoder with an IMS decoder. This can be attributed to a training strategy called *student forcing* that supplies the model's own previously predicted time steps as context for next time step prediction. Student forcing outperforms the generic *teacher forcing* which relies solely on ground truth contexts (Williams and Zipser, 1989) as well as scheduled sampling (Bengio et al., 2015; Teutsch and Mäder, 2023) which mixes both context types.

We present a thorough evaluation of all combinations of the encoders and decoders implemented by us, together with an evaluation of Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), and (D)Linear (Zeng et al., 2023) models, on progressive and increasingly complex TSF tasks on clinical data. As shown in Figure 2, our evaluation extends the standard mean squared error (MSE) calculation of 131 clinical variables on the MIMIC-III dataset and 98 clinical variables on the eICU dataset to an evaluation of the influence of forecasting accuracy on downstream clinical tasks such as early prediction of SOFA (Vincent et al., 1996), Sepsis-3 (Seymour et al., 2016; Singer et al., 2016), and SAPS-II (Gall et al., 1993). Furthermore, we present a study of cross-variate effects of drug administration on other clinical variables.

Generalizable Insights about Machine Learning in the Context of Healthcare

The contributions of our work to machine learning in the context of healthcare are as follows:

- We present a method for early syndrome prediction that is straightforwardly interpretable by medical practitioners. It directly predicts the clinical causes of a diagnosis by long-term TSF of clinical variables, and identifies the label by applying the known consensus definition that has been used to determine the ground truth to the forecasted clinical measurements. This technique is general and flexible since it is applicable to arbitrary consensus-based prediction tasks.
- From a machine learning perspective on long-term TSF, we find that contrary to recent proposals which advocate set function encoders for time series and direct multi-step decoders, best results are achieved by a combination of standard dense encoders with iterative multi-step decoders. The latter can be attributed to a student forcing training strategy that supplies the model's own predictions as context for next time step prediction, and outperforms standard teacher forcing or scheduled sampling.
- Our experiments are conducted on two critical care databases for SOFA-based Sepsis-3 prediction and prediction of the SAPS-II score. We find consistent wins for the combination of dense encoders and iterative multi-step decoders for both consensus-based prediction tasks on both datasets.

Code and data of this work are available at https://github.com/StatNLP/mlhc_2024_prediction_of_causes.

2. Related Work

TSF of clinical variables is a challenging task in itself. First, clinical data constitute multivariate time series that are irregularly sampled, both in time and across dimensions. Irregular sampling in time series is commonly addressed by data imputation methods. These include simple imputation of unobserved values by last observed or mean values, or by learned values using machine learning, e.g., regression of unobserved values against observed neighboring values (see Fang and Wang (2020) for an overview). Recently, neural ODEs have been proposed as a natural solution for irregular sampling in time series by integrating the dynamics over the time interval (Chen et al., 2018; De Brouwer et al., 2019;

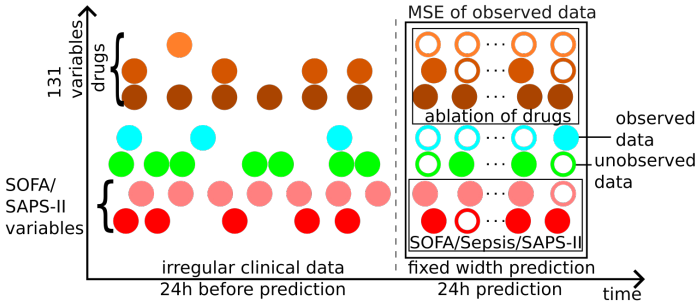

Figure 2: Evaluation setup for TSF experiments: Given irregularly distributed real-world clinical input data, we predict values for 24 hours in the future. We evaluate MSE loss on all variables, SOFA, SAPS-II, and Sepsis-3 accuracy, and perform a study on cross-variate effects of drugs.

Kidger et al., 2020). While these techniques show impressive results, they are incompatible with the parallelization idea of Transformers that explicitly avoid recurrence in the time dimension. For our work, we compare the so-called set functions embedding technique (Horn et al., 2020; Tipirneni and Reddy, 2022) that encodes time series as sets of observations with a dense encoder that compresses the input time series into 24 hourly bins and uses mean-value imputation.

Second, clinical patient data require long-term TSF capabilities from machine learning models. Recently, there has been a surge of works on long-term TSF using Transformers (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022) and work that explicitly questions the effectiveness of transformers for long-term TSF by conjecturing that the transformer's input permutation invariance may cause an ignorance of temporal input relations (Zeng et al., 2023). The proposed approaches agree in their argumentation that long-term TSF requires sophisticated DMS techniques in order to overcome the error propagation of IMS techniques. We show that a standard autoregressive Transformer decoder can be tuned to achieve comparable and sometimes even better long-term behavior than DMS decoders. The crucial ingredient is here to teach the Transformer to trust its own predictions as context in prediction of the next time step. We call this training technique student-forcing to differentiate it from teacher-forcing (Williams and Zipser, 1989). Combinations of both techniques by scheduled sampling have been shown to be useful in the context of sequence learning and TSF with RNNs (Bengio et al., 2015; Ranzato et al., 2016; Teutsch and Mäder, 2023), however, an investigation of scheduled sampling for TSF with autoregressive Transformers has so far been missing.

Third, multivariate clinical time series exhibit dependencies such as administration of certain medications resulting in changes in related clinical variables. This requires accurate modeling of cross-variate dependencies between clinical variables. It seems intuitive that models like Transformers that explicitly learn cross-variate connections should be more effective than univariate models like those proposed by Zeng et al. (2023) or Nie et al. (2023). As shown in Chen et al. (2023), the advantages of univariate models come into play only for certain types of benchmarks, but not for datasets that contain complex cross-variate

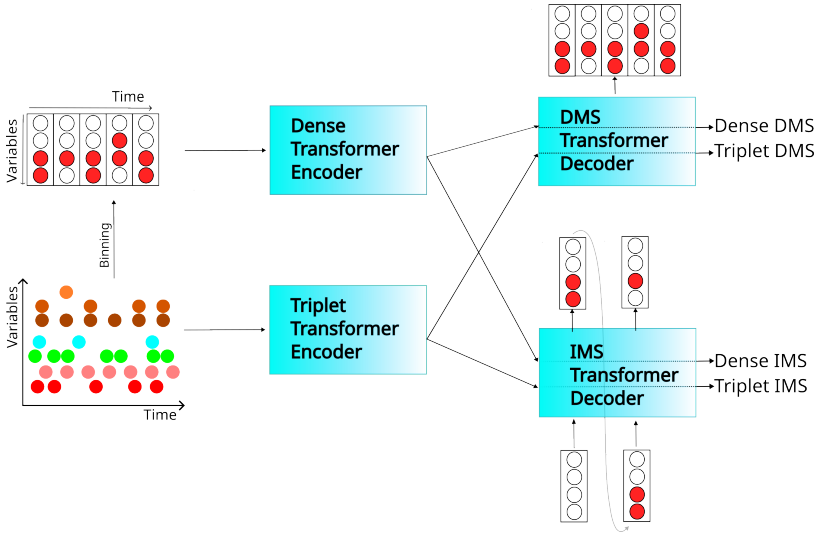

Figure 3: Overview of combination of encoders and decoders at inference time. Irregularly sampled time series (lower left) are encoded by a sparse triplet representation (lower middle), or they are binned into a representation with each vector representing an hour of observations (upper left), and encoded into a dense matrix (upper middle). The encoders are combined with different decoding strategies, either DMS (non-autoregressive) (upper right) or IMS (autoregressive) (lower right), yielding four different encoder-decoder pipelines.

information and auxiliary features. We believe that our investigation of cross-variate effects of drug administration on other clinical variables in critical care data proves the ability of Transformers to model complex cross-variate information in real-world clinical data.

3. Methods

3.1. Neural Architecture

Figure 3 gives an overview of the neural architectures used in our experiments to encode irregularly sampled time series and performing long-term TSF.

A sparse multivariate input time series comprised of n irregular measurements of $|F|$ clinical variables (lower left) is first variable-wise standardized, and then represented by n triplets $\mathrm{S} = \{(t_i, f_i, v_i)\}_{i=1}^{n}$, where $t_i \in \mathbb{R}_{\geq 0}$ is a time index, $f_i \in F$ is a clinical variable identifier, and $v_i \in \mathbb{R}$ the observed value of f_i at t_i. For sparse triplet (set function) encoding, each component of a triplet — time, variable, and value — receives a separate embedding of length m. These embeddings are summed up for each triplet, and an $m \times n$ matrix is fed into a Transformer architecture for encoding (lower middle), following the setup of Tipirneni and Reddy (2022).

Alternatively to this sparse triplet encoder, we implement a standard dense encoding architecture. To achieve this, the irregularly sampled input data (lower left) is binned into hourly buckets by recording the first observed value of each variable (upper left) and

applying mean imputation for unobserved values (effectively resulting in imputation of zeros because of standardization). Furthermore, we generate a masking matrix of the same size that informs the model if a corresponding value in the binned matrix is imputed or observed and append this matrix to the first one. A time series of length $D := \max(t_i)$[2] then results in a dense $\lceil D \rceil \times 2|F|$ matrix that is fed into a standard Transformer encoder (Vaswani et al., 2017) (upper middle).

We implement two decoder types to generate dense target time series of length T for an encoded input x. The direct multi-step (DMS) decoder uses T randomly initialized self-attention modules to perform T prediction steps at once. This non-autoregressive decoder predicts time steps independent of each other, and the inferred output $\hat{y}_t$ depends only on the encoded input x and the model parameters θ:

$$\hat{y}_t = f_\theta(x) \tag{1}$$

This method is depicted as the DMS strategy (upper right) in Figure 3.

Alternatively to the DMS decoder, the iterative multi-step (IMS) decoder generates an output vector $\hat{y}_t \in \mathbb{R}^{|F|}$ using a standard autoregressive model (Vaswani et al., 2017). The inferred output $\hat{y}_t$ is a function of the history $\hat{y}_{<t}$ of predicted tokens until time t, the encoded input x, and the model parameters θ:

$$\hat{y}_t = f_\theta(\hat{y}_{<t}, x) \tag{2}$$

To perform long-term TSF using the IMS setup, the outputs $\hat{y}_t$ from each time step $t = 1, \ldots, T$ are concatenated, yielding the IMS strategy (lower right) in Figure 3.

3.2. Training Methods

It is hypothesized in the literature that DMS decoders are superior to IMS decoders for long-term TSF because of potential error propagation in the latter (Zhou et al., 2021; Wu et al., 2021; Zeng et al., 2023). In our experiments, we show that an IMS decoder can be tuned to achieve comparable and even better long-term TSF behavior than DMS decoders by applying a strategy called *student-forcing* during training. Student forced training of an autoregressive decoder means to apply equation 2 during training in contrast to *teacher-forcing* (Williams and Zipser, 1989) where during training the history for next-token prediction is the ground truth value $y_{<t}$:

$$\hat{y}_t = f_\theta(y_{<t}, x) \tag{3}$$

Additionally, we combine teacher forcing and student forcing via scheduled sampling (Bengio et al., 2015; Teutsch and Mäder, 2023). Following Teutsch and Mäder (2023), we implemented linear curricula determining the teacher forcing ratio $C_{lin}(e)$ per epoch e as

$$C_{lin}(e) = \epsilon_{start} + (\epsilon_{end} - \epsilon_{start}) \cdot \frac{\min(e, Ł)}{Ł}. \tag{4}$$

We varied between an increasing ($\epsilon_{start} = 0.25$ and $\epsilon_{end} = 1$) and a decreasing curriculum ($\epsilon_{start} = 1$ and $\epsilon_{end} = .25$), both with a curriculum length Ł of 200. We also varied the

2. For simplicity, we assume that t_i is measured in hours.

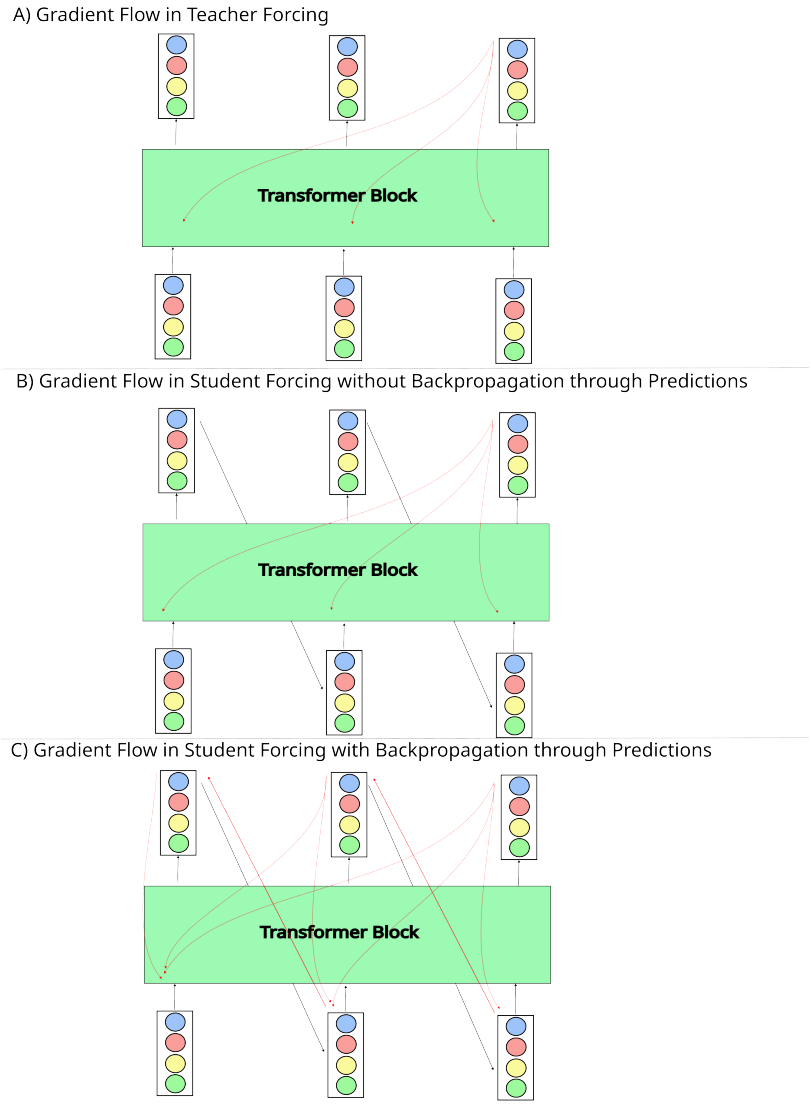

Figure 4: Gradient flow in time for different styles of training. Gradients (in red) are only shown for a loss generated at the third timestep.

selection of time steps for teacher forcing, either in a randomized or deterministic way. When teacher forcing is performed randomly, a time step is teacher forced with probability $C_{lin}(e)$. When time steps are selected in a deterministic way, teacher forcing depends not only on $C_{lin}(e)$, but also on the current position within the predicted sequence of length l where only the first $\lfloor l \cdot C_{lin}(e) \rfloor$ positions are teacher forced.

Different to natural language processing, the Transformer architecture, when applied to TSF, is used to predict continuous values. This allows us to enhance the expressive power of the model by back-propagating through the continuously differentiable prediction

Table 1: Main evaluation results on MIMIC-III for combinations of sparse and dense encoders with direct multi-step (DMS) and iterative multi-step (IMS) decoders, trained by teacher forcing (TF), student forcing (SF), or scheduled sampling (SS), with or without backpropagation (BP) through predictions. Curricula for scheduled sampling are denoted as deterministic increasing (DI) or decreasing (DD), randomized increasing (RI) or decreasing (RD). Evaluation is done according to MSE of all variables (Equation 5), MSE for SOFA (Equation 8), and Accuracy for Sepsis (Equation 9). Numbers in subscripts denote the 95% confidence interval for the estimation of the respective evaluation score on the test set. Best results are shown in bold face.

Model	Enc	Dec	Train	BP	MSE	MSE-SOFA	Acc-Sepsis
1	Triplet	DMS	-	-	$7.326_{[7.320,7.332]}$	$2.825_{[2.822,2.828]}$	$90.11_{[88.08,92.14]}$
2	Triplet	IMS	TF	-	$12.371_{[12.360,12.382]}$	$10.785_{[10.776,10.794]}$	$75.14_{[72.19,78.08]}$
3	Triplet	IMS	SF	No	$6.881_{[6.875,6.887]}$	$2.872_{[2.869,2.876]}$	$89.56_{[87.48,91.64]}$
4	Triplet	IMS	SF	Yes	$7.065_{[7.058,7.073]}$	$2.917_{[2.912,2.922]}$	$89.29_{[87.18,91.39]}$
5	Triplet	IMS	SS-DI	No	$8.764_{[8.753,8.775]}$	$3.466_{[3.457,3.475]}$	$82.85_{[80.28,85.42]}$
6	Triplet	IMS	SS-DI	Yes	$8.993_{[8.982,9.004]}$	$4.150_{[4.141,4.159]}$	$80.43_{[77.73,83.14]}$
7	Triplet	IMS	SS-DD	No	$24.360_{[24.349,24.371]}$	$16.345_{[16.336,16.353]}$	$70.65_{[67.55,73.75]}$
8	Triplet	IMS	SS-DD	Yes	$22.046_{[22.035,22.057]}$	$11.055_{[11.046,11.063]}$	$63.53_{[60.25,66.81]}$
9	Triplet	IMS	SS-RI	No	$7.068_{[7.062,7.073]}$	$2.689_{[2.685,2.693]}$	$85.63_{[83.24,88.02]}$
10	Triplet	IMS	SS-RI	Yes	$6.967_{[6.960,6.973]}$	$2.609_{[2.605,2.613]}$	$86.47_{[84.14,88.80]}$
11	Triplet	IMS	SS-RD	No	$7.761_{[7.752,7.771]}$	$3.092_{[3.085,3.100]}$	$86.11_{[83.76,88.47]}$
12	Triplet	IMS	SS-RD	Yes	$7.389_{[7.382,7.395]}$	$2.857_{[2.852,2.862]}$	$86.71_{[84.40,89.03]}$
13	Dense	DMS	-	-	$8.789_{[8.780,8.798]}$	$6.024_{[6.016,6.031]}$	$85.87_{[83.50,88.24]}$
14	Dense	IMS	TF	-	$11.907_{[11.896,11.918]}$	$9.242_{[9.233,9.251]}$	$66.18_{[62.96,69.41]}$
15	Dense	IMS	SF	No	$7.456_{[7.447,7.466]}$	$3.376_{[3.368,3.384]}$	$88.96_{[86.66,91.25]}$
16	Dense	IMS	SF	Yes	$\mathbf{6.216}_{[6.211,6.221]}$	$\mathbf{2.497}_{[2.495,2.500]}$	$\mathbf{90.34}_{[88.33,92.35]}$
17	Dense	IMS	SS-DI	No	$8.861_{[8.850,8.871]}$	$3.598_{[3.590,3.607]}$	$83.09_{[80.54,85.64]}$
18	Dense	IMS	SS-DI	Yes	$9.159_{[9.148,9.170]}$	$4.008_{[3.999,4.017]}$	$80.19_{[77.48,82.91]}$
19	Dense	IMS	SS-DD	No	$10.466_{[10.455,10.478]}$	$6.392_{[6.382,6.401]}$	$75.60_{[72.68,78.53]}$
20	Dense	IMS	SS-DD	Yes	$10.513_{[10.502,10.524]}$	$6.422_{[6.413,6.431]}$	$76.57_{[73.68,79.46]}$
21	Dense	IMS	SS-RI	No	$6.929_{[6.924,6.935]}$	$2.830_{[2.826,2.834]}$	$86.47_{[84.14,88.80]}$
22	Dense	IMS	SS-RI	Yes	$6.909_{[6.904,6.915]}$	$2.770_{[2.766,2.773]}$	$85.27_{[82.85,87.68]}$
23	Dense	IMS	SS-RD	No	$11.749_{[11.738,11.760]}$	$9.667_{[9.658,9.676]}$	$64.86_{[61.60,68.11]}$
24	Dense	IMS	SS-RD	Yes	$11.697_{[11.686,11.709]}$	$9.525_{[9.516,9.534]}$	$64.37_{[61.11,67.63]}$
Informer	Dense	DMS	-	-	$6.354_{[6.350,6.358]}$	$2.616_{[2.614,2.617]}$	$85.75_{[83.37,88.13]}$
Autoformer	Dense	DMS	-	-	$7.438_{[7.432,7.444]}$	$3.352_{[3.348,3.356]}$	$89.61_{[87.54,91.69]}$
DLinear	Dense	DMS	-	-	$8.487_{[8.479,8.495]}$	$5.019_{[5.012,5.025]}$	$76.81_{[73.94,79.69]}$
Linear	Dense	DMS	-	-	$8.486_{[8.477,8.494]}$	$5.001_{[4.995,5.007]}$	$77.29_{[74.44,80.15]}$

history. The gradient flow of teacher forcing, student forcing, and student forcing with backpropagation through predictions is shown in Figure 4. As shown in our experiments, this modified gradient flow improves the model further.

Table 2: Evaluation results on MIMIC-III showing MSE at beginning 8 hours ($MSE_{1:8}$) and last 16 hours ($MSE_{9:24}$) of 24 hour forecasting period. Numbers in subscript in square brackets denote the 95% confidence interval for the estimation of the respective evaluation score on the test set. Best results are shown in bold face.

Model	Enc	Dec	Train	BP	$MSE_{1:8}$	$MSE_{9:24}$
1	Triplet	DMS	-	-	$7.687_{[7.676,7.697]}$	$7.160_{[7.154,7.165]}$
2	Triplet	IMS	TF	-	$12.753_{[12.734,12.773]}$	$12.203_{[12.195,12.212]}$
3	Triplet	IMS	SF	No	$6.686_{[6.676,6.696]}$	$6.991_{[6.986,6.997]}$
4	Triplet	IMS	SF	Yes	$6.899_{[6.886,6.911]}$	$7.162_{[7.156,7.168]}$
5	Triplet	IMS	SS-DI	No	$9.335_{[9.316,9.354]}$	$8.495_{[8.487,8.503]}$
6	Triplet	IMS	SS-DI	Yes	$9.593_{[9.574,9.612]}$	$8.710_{[8.702,8.718]}$
7	Triplet	IMS	SS-DD	No	$28.394_{[28.374,28.413]}$	$22.390_{[22.382,22.399]}$
8	Triplet	IMS	SS-DD	Yes	$22.163_{[22.144,22.183]}$	$22.033_{[22.024,22.041]}$
9	Triplet	IMS	SS-RI	No	$6.914_{[6.903,6.925]}$	$7.158_{[7.153,7.162]}$
10	Triplet	IMS	SS-RI	Yes	$6.705_{[6.694,6.717]}$	$7.110_{[7.104,7.116]}$
11	Triplet	IMS	SS-RD	No	$7.654_{[7.638,7.671]}$	$7.829_{[7.822,7.837]}$
12	Triplet	IMS	SS-RD	Yes	$7.205_{[7.192,7.217]}$	$7.494_{[7.488,7.500]}$
13	Dense	DMS	-	-	$9.116_{[9.099,9.133]}$	$8.641_{[8.634,8.648]}$
14	Dense	IMS	TF	-	$12.627_{[12.608,12.646]}$	$11.570_{[11.562,11.578]}$
15	Dense	IMS	SF	No	$7.427_{[7.410,7.445]}$	$7.484_{[7.477,7.491]}$
16	Dense	IMS	SF	Yes	$\mathbf{6.117}_{[6.109,6.124]}$	$\mathbf{6.278}_{[6.273,6.283]}$
17	Dense	IMS	SS-DI	No	$9.328_{[9.309,9.347]}$	$8.643_{[8.635,8.651]}$
18	Dense	IMS	SS-DI	Yes	$9.744_{[9.725,9.763]}$	$8.884_{[8.876,8.892]}$
19	Dense	IMS	SS-DD	No	$11.258_{[11.239,11.278]}$	$10.090_{[10.082,10.098]}$
20	Dense	IMS	SS-DD	Yes	$11.331_{[11.311,11.350]}$	$10.124_{[10.116,10.133]}$
21	Dense	IMS	SS-RI	No	$6.752_{[6.742,6.763]}$	$7.031_{[7.027,7.036]}$
22	Dense	IMS	SS-RI	Yes	$6.587_{[6.578,6.597]}$	$7.084_{[7.078,7.089]}$
23	Dense	IMS	SS-RD	No	$12.554_{[12.535,12.574]}$	$11.369_{[11.361,11.377]}$
24	Dense	IMS	SS-RD	Yes	$12.575_{[12.556,12.595]}$	$11.281_{[11.272,11.289]}$
Informer	Dense	DMS	-	-	$6.349_{[6.343,6.354]}$	$6.369_{[6.365,6.373]}$
Autoformer	Dense	DMS	-	-	$7.232_{[7.224,7.241]}$	$7.556_{[7.550,7.561]}$
DLinear	Dense	DMS	-	-	$8.360_{[8.346,8.373]}$	$8.568_{[8.561,8.575]}$
Linear	Dense	DMS	-	-	$8.359_{[8.345,8.372]}$	$8.566_{[8.559,8.573]}$

4. Experiments

4.1. Datasets

In our experiments, we use the Medical Information Mart for Intensive Care III (MIMIC-III) data (Johnson et al., 2016). The data were collected from the Beth Israel Deaconess Medical Center between 2001 and 2012 and contain over 40k patients. After filtering for patients with an ICU stay of at least 24 hours with reported gender and age of at least 18 years, our dataset contains 44,858 ICU stays with 56 million data points. We split the data into partitions for training (28,708), development (7,270), and testing (8,880). For our study, we used 131 different clinical variables, including 15 clinical markers for the calculation of SOFA and 13 for SAPS-II, and the demographic variables gender and age. This selection

Table 3: SAPS-II evaluation results on MIMIC-III. Numbers in subscripts denote the 95% confidence interval for the estimation of the respective evaluation score on the test set. Best results are shown in bold face.

Model	Enc	Dec	Train	BP	MSE	MSE-SAPS-II
1	Triplet	DMS	-	-	$7.326_{[7.320,7.332]}$	$222.426_{[222.417,222.435]}$
2	Triplet	IMS	TF	-	$12.371_{[12.360,12.382]}$	$270.245_{[270.236,270.254]}$
3	Triplet	IMS	SF	No	$6.881_{[6.875,6.887]}$	$187.826_{[187.817,187.835]}$
4	Triplet	IMS	SF	Yes	$7.065_{[7.058,7.073]}$	$187.826_{[187.817,187.835]}$
5	Triplet	IMS	SS-DI	No	$8.764_{[8.753,8.775]}$	$167.347_{[167.338,167.356]}$
6	Triplet	IMS	SS-DI	Yes	$8.993_{[8.982,9.004]}$	$166.765_{[166.756,166.774]}$
7	Triplet	IMS	SS-DD	No	$24.360_{[24.349,24.371]}$	$147.559_{[147.550,147.567]}$
8	Triplet	IMS	SS-DD	Yes	$22.046_{[22.035,22.057]}$	$377.004_{[376.996,377.013]}$
9	Triplet	IMS	SS-RI	No	$7.068_{[7.062,7.073]}$	$102.445_{[102.442,102.449]}$
10	Triplet	IMS	SS-RI	Yes	$6.967_{[6.960,6.973]}$	$110.174_{[110.170,110.178]}$
11	Triplet	IMS	SS-RD	No	$7.761_{[7.752,7.771]}$	$101.756_{[101.749,101.764]}$
12	Triplet	IMS	SS-RD	Yes	$7.389_{[7.382,7.395]}$	$103.756_{[103.751,103.761]}$
13	Dense	DMS	-	-	$8.789_{[8.780,8.798]}$	$132.819_{[132.811,132.826]}$
14	Dense	IMS	TF	-	$11.907_{[11.896,11.918]}$	$183.693_{[183.684,183.702]}$
15	Dense	IMS	SF	No	$7.456_{[7.447,7.466]}$	$109.381_{[109.373,109.389]}$
16	Dense	IMS	SF	Yes	$\mathbf{6.216}_{[6.211,6.221]}$	$\mathbf{89.115}_{[89.112,89.117]}$
17	Dense	IMS	SS-DI	No	$8.861_{[8.850,8.871]}$	$169.390_{[169.381,169.399]}$
18	Dense	IMS	SS-DI	Yes	$9.159_{[9.148,9.170]}$	$175.625_{[175.616,175.634]}$
19	Dense	IMS	SS-DD	No	$10.466_{[10.455,10.477]}$	$192.777_{[192.768,192.786]}$
20	Dense	IMS	SS-DD	Yes	$10.513_{[10.502,10.524]}$	$166.208_{[166.199,166.217]}$
21	Dense	IMS	SS-RI	No	$6.929_{[6.923,6.934]}$	$107.428_{[107.424,107.432]}$
22	Dense	IMS	SS-RI	Yes	$6.909_{[6.904,6.915]}$	$105.098_{[105.095,105.101]}$
23	Dense	IMS	SS-RD	No	$10.466_{[10.455,10.478]}$	$169.432_{[169.423,169.440]}$
24	Dense	IMS	SS-RD	Yes	$10.513_{[10.502,10.524]}$	$206.098_{[206.089,206.107]}$
Informer	Dense	DMS	-	-	$6.354_{[6.350,6.358]}$	$92.008_{[92.006,92.010]}$
Autoformer	Dense	DMS	-	-	$7.438_{[7.432,7.444]}$	$95.306_{[95.302,95.310]}$
DLinear	Dense	DMS	-	-	$8.487_{[8.479,8.495]}$	$129.476_{[129.470,129.482]}$
Linear	Dense	DMS	-	-	$8.486_{[8.477,8.494]}$	$129.570_{[129.563,129.576]}$

comprises all vital signs and laboratory values used in the PhysioNet challenge for early prediction of sepsis (Reyna et al., 2019). The full list of extracted MIMIC-III features is given in Appendix A. As shown in Appendix D, the sparsity of the dataset is very high, meaning that the average number of observations per patient per hour is at most 20 and rapidly declines with longer length of stay. Converting the data to a dense one hour representation yields 89.08% missing data, changing per variable from under 15% (HR, RR, SBP, DBP, MBP, and O2 Saturation) to more than 90% for 101 variables, and exceeding 99% for 42 variables. On the other side we are loosing 17.73% of the data points through the densification procedure. Demoscopic data are complete for all patients.

Furthermore, we conducted some experiments on the larger eICU dataset (Pollard et al., 2018). These data were collected from over 200 US hospitals and comprise over 200,000 ICU stays. After filtering for patients with an ICU stay of at least 48 hours, reported gender and

aged 18 years or older, we arrived at 77,704 ICU stays with 415 million datapoints. This set was partitioned in subsets for training (49,730), development (12,433), and testing (15,541). As shown in Appendix A, we extracted 98 clinical variables and 17 demographic markers for our experiments. As shown in Appendix D, the measurements in the eICU are denser than in MIMIC-III since the number of observations per patient per hour is three times higher than for MIMIC-III and decreases at a slower rate with length of stay. In a dense encoding, this reduces to one sixth of the measurements (16.86%). To densify the data like before yields even 89.85% missing data. The same six variables as before are quite complete while 84 variables are missing more than 90% and 29 variables 99%. The demographic data are almost complete.

4.2. Evaluation Measures

We use the following measures to evaluate our models for long-term TSF. Given N time series in our dataset, with T hours prediction for TSF, the mean squared error (MSE) over hourly prediction vectors $\hat{y}_t^n$ is defined as follows:

$$\text{MSE} = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} ||(y_t^n - \hat{y}_t^n) \odot m_t^n||_2^2 \tag{5}$$

where $m_t^n \in \{0,1\}^{|F|}$ is a mask indicating if the variables in y_t^n were observed or not, and $\odot$ is a component-wise product. In our experiments, T is set to 24 hours.

For a closer inspection of the long-term TSF error, we compute MSE at the beginning and at the end of the forecasting task. We calculate $\text{MSE}_{1:8}$ for the first 8 hours, and $\text{MSE}_{9:24}$ for the last 16 hours of a 24 hour forecasting period:

$$\text{MSE}_{1:8} = \frac{1}{N \cdot 8} \sum_{n=1}^{N} \sum_{t=1}^{8} ||(y_t^n - \hat{y}_t^n) \odot m_t^n||_2^2, \tag{6}$$

$$\text{MSE}_{9:24} = \frac{1}{N \cdot 16} \sum_{n=1}^{N} \sum_{t=9}^{24} ||(y_t^n - \hat{y}_t^n) \odot m_t^n||_2^2. \tag{7}$$

Prediction of SOFA is evaluated by the MSE between the SOFA score computed on the ground truth values and on the forecasted variable values relevant for the consensus definition (Vincent et al., 1996). The SOFA score is defined as the sum of six organ system subscores ranging from 0-4 depending for their part on thresholded fundamental clinical variables observed during a 24h window (see Appendix B). We compared the sub-scores obtained from forecasted data ($\widehat{\text{SOFA}_6}$) to those obtained from the corresponding gold data (SOFA_6) masking forecasted values that have no corresponding observed gold value by the following MSE calculation:

$$\text{MSE-SOFA} = \frac{1}{N} \sum_{n=1}^{N} ||\text{SOFA}_6^n - \widehat{\text{SOFA}_6}^n||_2^2. \tag{8}$$

Accuracy of sepsis prediction is evaluated based on the Sepsis-3 consensus definition (Singer et al., 2016; Seymour et al., 2016). According to this definition, a patient is classified

as septic if two criteria are met. Firstly, a patient must have a verified or suspected infection, and secondly, a SOFA score showing an increase of 2 or more points in the following 24 hours after the infection. Similar to Seymour et al. (2016), we identify a suspected infection by a combination of antibiotics treatment and blood culture, starting within the first 24 hours after admission, and gather all infected patients in the set I. For those patients, $\text{SOFA}_{1:24}$ denotes the SOFA score for the first 24 hours. For the following 24 hours, $\text{SOFA}_{25:48}$ denotes the SOFA score based on the ground truth data, and $\widehat{\text{SOFA}}_{25:48}$ denotes the SOFA score based on the forecasted data. The accuracy of the predicted sepsis label is calculated as the average match with the ground truth label, where labels are assigned by a check whether a change in SOFA ≥ 2 happens between the first 24 hours and the prediction window of 24 hours. We use the following notation for indicator functions: $[\![a]\!] = 1$ if a is true, 0 otherwise. Accuracy of Sepsis-labeling can then be defined as follows:

$$\begin{aligned} \text{Acc-Sepsis} &= \frac{1}{|I|}\sum_{i\in I}[\![\chi_i = \hat{\chi}_i]\!], \\ \text{where } \chi_i &:= [\![(\text{SOFA}^i_{25:48} - \text{SOFA}^i_{1:24}) \geq 2]\!], \\ \text{and } \hat{\chi}_i &:= [\![(\widehat{\text{SOFA}}^i_{25:48} - \text{SOFA}^i_{1:24}) \geq 2]\!]. \end{aligned} \tag{9}$$

The new Simplified Acute Physiology Score (SAPS-II) (Gall et al., 1993) scores the illness severity of an ICU patient based on a moderate number of routine clinical measurements collected during a 24 hour period. Our implementation is reduced to all non-static variables involved in the calculation of SAPS-II scores (see Appendix C). To evaluate predictions for SAPS-II, we calculate the MSE between the integer point scores ranging between 0 and 120 for ground truth SAPS-II and predicted $\widehat{\text{SAPS-II}}$ scores:[3]

$$\text{MSE-SAPS-II} = \frac{1}{N}\sum_{n=1}^{N}(\text{SAPS-II}^n - \widehat{\text{SAPS-II}}^n)^2 \tag{10}$$

4.3. Experimental Results

All training runs in our experiments used a 24 hour observation window, followed by a 24 hour prediction window. In order to best exploit the training data, we used sliding windows of 24 hours observations and 24 hours prediction that were shifted in 4 hour steps from admission time up to five days. The learning objective is to minimize the MSE by applying Equation 5 to the training data. Extensive metaparameter search was conducted on the MIMIC-III and eICU development sets (see Appendix B) where the best result on the development set was chosen for final evaluation on the test set.

Table 1 shows the main results of our evaluation of various combinations of encoders and decoders, and competing current approaches, on increasingly complex TSF tasks. The first task is the standard evaluation of the MSE (Equation 5) for TSF of 131 clinical variables on the MIMIC-III testset. The best result is obtained for model 16 — a dense encoder, combined with an IMS decoder, trained with student forcing and backpropagation. In general, student forcing is far better than teacher forcing (models 2 and 14) or scheduled

3. We ignore the standard probability conversion for the MSE calculation since the comparison of untransformed scores provides a more accurate evaluation of the utility of the forecasted data.

sampling, which underperforms especially when a decreasing schedule is used (models 7, 8, 19, 20, 23, 24), irrespective of the encoding method. Our evaluation progresses from MSE of all clinical variables to a subset of 15 clinical variables that are relevant for the computation of the SOFA score (see the variables marked with * in Appendix A). MSE-SOFA is computed according to Equation 8. Table 1 shows that MSE-SOFA is well correlated with MSE over all variables, with similar rankings in both columns. Best results are again obtained for the combination of a dense encoder with an IMS decoder, trained with student forcing and backpropagation (model 16). The last column in Table 1 presents the accuracy of Sepsis prediction. The best results are obtained again for model 16. Since Sepsis accuracy is computed according to Equation 9, by filtering patients with infection, the data filtering resulted in a small test set of only 345 patients. Because of the high variance in evaluation scores, many confidence intervals overlap, so that we cannot conclude statistical significance of result differences. In Table 10 in Appendix F, we report similar results on the larger eICU dataset. There the number of patients with suspected infection is larger (3,789 patients), and the confidence intervals are accordingly smaller. Furthermore, in Appendix G, we report a breakdown of Sepsis prediction results according to true positives, false positives, true negative, false negatives, and F1 score. Model 16 outperforms all competitors in these tables as well. Out of the competing models shown in the last block of Table 1, the Informer model achieves a competitive MSE and MSE-SOFA, however, these advantages do not carry over to Sepsis accuracy comparable with model 16.

The results presented in Table 2 allow us to draw a more nuanced comparison between student- and teacher-forcing. We see indications of error propagation for student-forcing, shown by a slight MSE increase for later time-steps for models 3, 4, 15, and 16. However, this effect is magnitudes smaller than the beneficial consequences of student-forcing. Teacher-forced models 2 and 14 does not suffer from error propagation. The interpolation between pure student- and teacher-forcing that is done via scheduled sampling shows that teacher forcing is exceedingly detrimental in the early stages of training and when applied non-randomly. Thus, scheduled sampling cannot offer a way to mitigate the small error propagation happening in student forcing.

An evaluation of the prediction of the SAPS-II prediction is given in Table 3. The ranking of models according to MSE results for all variables is similar to the ranking in Table 1, and consistent with the prediction of SOFA and Sepsis, the combination of a dense encoder with an iterative multi-step decoder and backpropagation through predictions achieves the best results.

Similar experiments were conducted on the eICU dataset (full result tables are given in Appendix F). Consistent with the results on MIMIC-III, best results for MSE-SOFA and accuracy of Sepsis prediction are obtained for the combination of a dense encoder with an IMS decoder, trained with student forcing and backpropagation (model 16). This model also achieves best results for SAPS-II prediction. Since eICU is larger than MIMIC-III, confidence intervals are smaller, yielding increased significance of result differences.

Lastly, we performed an experiment where compared direct prediction of the SOFA score (called effect prediction) our best TSF model (called prediction of causes). We conducted this experiment by adding a regression head to our dense encoder, and trained this model on automatically assigned SOFA scores (ranging from 0 to 24) in order to use the whole training set. The evaluation was done by computing MSE of the predicted scores. The resulting MSE

scores on eICU were 2.2869 for model 16 compared to 2.6748 for the regression model. On MIMIC-III, we obtained MSE scores of 2.6973 for model 16 and of 2.5865 for the regression model. This shows that prediction of causes is comparable to prediction of effects. A possible explanation for the slightly worse result of direct effect prediction on eICU is the high sparsity of target variables for prediction of causes, leading to many masked values that are ignored in the evaluation of the latter approach. Masking for TSF is consistent with the SOFA consensus definition where unobserved values are set to default values of 0, and with clinical practice where data are collected in a panel-wise fashion with dedicated time slots causing sparsity of observed data.

4.4. Cross-variate Effects of Cardio-vascular Medications

In order to assess the cross-variate effects of cardio-vascular drug administration for our best model on MIMIC-III (model 16), we created synthetic inputs based on test data where we separately altered the value of dopamine, dobutamine, and norepinephrine during decoding, keeping everything else unchanged. For each drug, we decoded an input two times, setting the respective drug level to the first (Q1) and third quartile (Q3) of sampled drug doses during decoding for each time step. For analysis, the resulting predictions were averaged over 24 time points for all variables (except for the manipulated drug, which was excluded from analysis), and the two groups (Q1 and Q3) were compared by t-tests and Mann-Whitney-U-Test applying the Bonferroni correction (Bonferroni, 1935) to account for multiple testing. There are only two statistically significant results in both tests and their effect size according to Cohen's d (Cohen, 1988) is reported. The results of this ablation of drugs show that model 16 has learned that larger doses of dobutamine are associated with lower doses of amidarone ($d = .11$). This result is in line with a contraindication of dobutamine for patients that suffer from arrhythmia. We also observe a small positive association ($d = .09$) between dobutamine dose and midazolam dose, gastric meds and fiber.

5. Discussion

Most machine learning approaches to early syndrome diagnosis define ground truth labels as the effect of an application of a known medical consensus definition to future clinical measurements. Knowledge of this construction of the ground truth suggests to predict the clinical causes to which the known consensus definition can be deterministically applied. This leads to a prediction that is straightforwardly interpretable by clinicians and can be used for arbitrary consensus-based prediction tasks. The machine learning focus is then shifted to accurate long-term TSF of clinical variables that are fundamental to consensus definitions of syndromes. Since consensus definitions such as the SOFA-based Sepsis-3 and SAPS-II scores are mostly based on sparsely observed laboratory measurements, a proper encoding of sparse inputs together with a decoding strategy that exploits dependencies between multivariate inputs is key. Our experiments on two datasets show that a combination of a standard dense encoder using data imputation with an iterative multi-step forecasting outperforms specialized set function encoders and direct multi-step decoders. We conjecture that the accuracy advantage of the dense encoder is attributable to the compression of long input time series into 24 hourly bins that record the most important observations, while the advantage of the IMS decoder lies in its ability to capture cross-variate dependencies.

Limitations The experiments in our work were conducted by choosing the best metaparameter setting on the development set for final evaluation on the test set, and reporting confidence intervals for evaluation scores. This corresponds only to conservative significance testing and hides the variance induced by metaparameter variation. Time constraints prohibited the use of more sophisticated techniques for significance testing and variance component analysis.

Acknowledgments

This work was partially funded by the Helmholtz Information & Data Science School for Health (HIDSS4Health) and the German Research Foundation (DFG) through Germany's Excellence Strategy EXC 2181/1 – 390900948 (STRUCTURES).

References

Sabeen Ahmed, Ian E. Nielsen, Aakash Tripathi, Shamoon Siddiqui, Ravi P. Ramachandran, and Ghulam Rasool. Transformers in time-series analysis: A tutorial. *Circuits, Systems, and Signal Processing*, 2023. URL https://doi.org/10.1007/s00034-023-02454-8.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf.

Carlo E Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in onore del professore salvatore ortu carboni*, pages 13–60. Tip. del Senato, 1935.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems (NeurIPS)*, Montreal, Canada, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *arXiv*, 2023. URL https://doi.org/10.48550/arXiv.2303.06053.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, New York, 1988. URL https://doi.org/10.4324/9780203771587.

Samuel N. Cohen, James Foster, Peter Foster, Hang Lou, Terry Lyons, Sam Morley, James Morrill, Hao Ni, Edward Palmer, Bo Wang, Yue Wu, Lingyi Yang, and Weixin Yang. Subtle variation in sepsis-iii definitions markely influences predictive performance within and across methods. *scientific reports*, 14(1920):1–10, 2024. URL https://doi.org/10.1038/s41598-024-51989-6.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In

Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2019. URL https://papers.nips.cc/paper_files/paper/2019/file/455cb2657aaa59e32fad80cb0b65b9dc-Paper.pdf.

Chenguang Fang and Chen Wang. Time series data imputation: A survey on deep learning approaches. *arXiv*, abs/2011.11347, 2020. URL https://doi.org/10.48550/arXiv.2011.11347.

Ricard Ferrer, Ignacio Martin-Loeches, George P. Phillips, Tiffany Medlin Osborn, Sean Townsend, R. Phillip Dellinger, Antonio Artigas, Christa A. Schorr, and M. M. Robert Lévy. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Critical Care Medicine*, 42(8):1749–1755, 2014. URL https://doi.org/10.1097/ccm.0000000000000330.

Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. *JAMA*, 270(24):2957–2963, 1993. URL https://doi.org/10.1001/jama.270.24.2957.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv*, abs/2312.10794, 2023. URL https://doi.org/10.48550/arXiv.2312.10794.

Michael Hagmann, Shigehiko Schamoni, and Stefan Riezler. Validity problems in clinical machine learning by indirect data labeling using consensus definitions. In *Extended Abstract presented at Machine Learning for Health (ML4H) symposium*, New Orleans, United States, 2023. URL https://doi.org/10.48550/arXiv.2311.03037.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Online, 2020. URL https://proceedings.mlr.press/v119/horn20a.html.

Katsuma Inoue, Soh Ohara, Yasuo Kuniyoshi, and Kohei Nakajima. Transient chaos in bidirectional encoder representations from transformers. *Phys. Rev. Res.*, 4, 2022. doi: 10.1103/PhysRevResearch.4.013204. URL https://link.aps.org/doi/10.1103/PhysRevResearch.4.013204.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. URL https://doi.org/10.1038/sdata.2016.35.

Shachar Kaufmann, Saharon Rosset, and Claudia Perlich. Leakage in data mining: Formulation, detection, and avoidance. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)*, San Diego, CA, USA, 2011. URL http://dx.doi.org/10.1145/2020408.2020496.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2020. URL https://proceedings.neurips.cc/paper/2020/file/4a5876b450b45371f6cfe5047ac8cd45-Paper.pdf.

Holger A. Lindner, Shigehiko Schamoni, Thomas Kirschning, Corinna Worm, Bianka Hahn, Franz-Simon Centner, Jochen J. Schoettler, Michael Hagmann, Jörg Krebs, Dennis Mangold, Stephanie Nitsch, Stefan Riezler, Manfred Thiel, and Verena Schneider-Lindner. Ground truth labels challenge the validity of sepsis consensus definitions in critical illness. *Journal of Translational Medicine*, 20(27), 2022. URL https://doi.org/10.1186/s12967-022-03228-7.

Michael Moor, Bastian Rieck, Max Horn, Catherine Jutzeler, and Karsten Borgwardt. Early prediction of sepsis in the ICU using machine learning: A systematic review. *Frontiers in Medicine*, 8, 2021. URL https://doi.org/10.3389/fmed.2021.607952.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multicenter database for critical care research. *Scientific Data*, 5(180178), 2018. URL https://doi.org/10.1038/sdata.2018.178.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representation (ICLR)*, San Juan, Puerto Rico, 2016. URL https://doi.org/10.48550/arXiv.1511.06732.

Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P.B. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019. *Critical Care Medicine*, 48(2):210–217, 2019. URL https://doi.org/10.1097/CCM.0000000000004145.

Stefan Riezler and Michael Hagmann. *Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science*. Springer, second edition, 2024. URL https://doi.org/10.1007/978-3-031-57065-0.

Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan

Kissoon, Simon Finfer, Carolin Fleischmann-Struzek, Flavia R Machado, Konrad K Reinhart, Kathryn Rowan, Christopher W Seymour, R Scott Watson, T Eoin West, Fatima Marinho, Simon I Hay, Rafael Lozano, Alan D Lopez, Derek C Angus, Christopher J L Murray, and Mohsen Naghavi. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219): 200–211, 2020. URL https://doi.org/10.1016/S0140-6736(19)32989-7.

Shigehiko Schamoni, Holger A. Lindner, Verena Schneider-Lindner, Manfred Thiel, and Stefan Riezler. Leveraging implicit expert knowledge for non-circular machine learning in sepsis prediction. *Journal of Artificial Intelligence in Medicine*, 100:1–9, 2019. URL https://doi.org/10.1016/j.artmed.2019.101725.

Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of clinical criteria for sepsis for the third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):762–774, 2016. URL https://doi.org/10.1001/jama.2016.0288.

Mervyn Singer, Clifford S. Deutschman, and Christopher Warren Seymour. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8): 801–810, 2016. URL 10.1001/jama.2016.0287.

Philipp Teutsch and Patrick Mäder. Flipped classroom: Effective teaching for time series forecasting. *Transactions on Machine Learning Research*, 9, 2023. URL https://doi.org/10.48550/arXiv.2210.08959.

Sindhu Tipirneni and Chandan K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data*, 16(6), 2022. doi: 10.1145/3516367. URL https://doi.org/10.1145/3516367.

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 (NIPS)*, Long Beach, CA, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

J.L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 1996. URL https://doi.org/10.1007/BF01709751.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2023. URL https://doi.org/10.24963/ijcai.2023/759.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bcc0d400288793e8bdcd7c19a8ac0c2b-Paper.pdf.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, Washington DC, USA, 2023. URL https://ojs.aaai.org/index.php/AAAI/article/view/26317/26089.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, virtual, 2021. URL https://doi.org/10.1609/aaai.v35i12.17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning (ICLR)*, 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

Appendix A. Features for Prediction Task

Table 4: Feature list for MIMIC-III: Besides the following 131 dynamic variables, only age and gender were extracted. The 15 variables marked with an asterisk are directly used for calculating the SOFA score.

ALP	Epinephrine*	LDH	Packed RBC
ALT	Famotidine	Lactate	Pantoprazole
AST	Fentanyl	Lactated Ringers	Phosphate
Albumin	FiO2*	Levofloxacin	Piggyback
Albumin 25%	Fiber	Lorazepam	Piperacillin
Albumin 5%	Free Water	Lymphocytes	Platelet Count*
Amiodarone	Fresh Frozen Plasma	Lymphocytes (Absolute)	Potassium
Anion Gap	Furosemide	MBP	Pre-admission Intake
BUN	GCS_eye*	MCH	Pre-admission Output
Base Excess	GCS_motor*	MCHC	Propofol
Basophils	GCS_verbal*	MCV	RBC
Bicarbonate	GT Flush	Magnesium	RDW
Bilirubin (Direct)	Gastric	Magnesium Sulfate (Bolus)	RR
Bilirubin (Indirect)	Gastric Meds	Magnesium Sulphate	Residual
Bilirubin (Total)*	Glucose (Blood)	Mechanically ventilated	SBP*
CRR	Glucose (Serum)	Metoprolol	SG Urine
Calcium Free	Glucose (Whole Blood)	Midazolam	Sodium
Calcium Gluconate	HR	Milrinone	Solution
Calcium Total	Half Normal Saline	Monocytes	Sterile Water
Cefazolin	Hct	Morphine Sulfate	Stool
Chest Tube	Heparin	Neosynephrine	TPN
Chloride	Hgb	Neutrophils	Temperature
Colloid	Hydralazine	Nitroglycerine	Total CO2
Creatinine Blood*	Hydromorphone	Nitroprusside	Ultrafiltrate
Creatinine Urine	INR	Norepinephrine*	Urine*
D5W	Insulin Humalog	Normal Saline	Vancomycin
DBP*	Insulin NPH	O2 Saturation	Vasopressin
Dextrose Other	Insulin Regular	OR/PACU Crystalloid	WBC
Dobutamine*	Insulin largine	PCO2	Weight
Dopamine*	Intubated	PO intake	pH Blood
EBL	Jackson-Pratt	PO2*	pH Urine
Emesis	KCl	PT	
Eoisinophils	KCl (Bolus)	PTT	

Table 5: Feature list for eICU: Besides the following 98 dynamic variables, there are 17 static variables covering age, gender, admission information, and ICU type. The 15 variables marked with an asterisk are directly used for calculating the SOFA score.
On the right column, there are 35 drug-related variables. Some of them seem redundant due to different hospitals but can not be merged because of different or not standardized concentrations.

ALP	Lactate	Amiodarone
ALT	Lymphocytes	Dobutamine dose
AST	MBP	Dobutamine ratio*
Albumin	MCH	Dopamine dose
Anion Gap	MCHC	Dopamine ratio*
BUN	MCV	Epinephrine dose
Base Deficit	MPV	Epinephrine ratio*
Base Excess	Magnesium	Fentanyl 1
Basophils	Monocytes	Fentanyl 2
Bedside Glucose	Neutrophils	Fentanyl 3
Bicarbonate	O2 L/%	Furosemide
Bilirubin (Direct)	O2 Saturation	Heparin 1
Bilirubin (Total)*	PT	Heparin 2
Bodyweight (kg)	PTT	Heparin 3
CO2 (Total)	PaCO2	Heparin vol
Calcium	PaO2*	Insulin 1
Chloride	Phosphate	Insulin 2
Creatinine (Blood)*	Platelets*	Insulin 3
Creatinine (Urine)	Potassium	Midazolam 1
DBP*	Protein (Total)	Midazolam 2
Eoisinophils	RBC	Milrinone 1
EtCO2	RDW	Milrinone 2
FiO2*	RR	Nitroglycerin 1
Fibrinogen	SBP*	Nitroglycerin 2
GCS eye*	Sodium	Nitroprusside
GCS motor*	Stool	Norepinephrine 1
GCS verbal*	Temperature	Norepinephrine 2
Glucose	Troponin - I	Norepinephrine ratio*
HR	Urine*	Pantoprazole
Hct	WBC	Propofol 1
Hgb	pH	Propofol 2
INR		Propofol 3
		Vasopressin 1
		Vasopressin 2
		Vasopressin 3

Appendix B. SOFA score

The Sepsis-related Organ Failure Assessment (SOFA) is calculated by summing six subscores ranging from 0 to 4. In our setting, we had to recalculate MAP (mean arterial pressure) by SBP and DBP (systolic and diastolic blood pressure), the Horowitz coefficient PaO2/FiO2 by PaO2 and FiO2, but ignored the kind of mechanical ventilation. If no value for calculation in a SOFA subsystem was available, we took a value of 0.

Table 6: SOFA score (Vincent et al., 1996). Abbreviations: CNS = Central nervous system; GCS = Glasgow Coma Scale; MV = mechanically ventilated including CPAP; MAP = mean arterial pressure, UO = Urine output.

Score	CNS	Cardiovascular	Respiratory	Coagulation	Liver	Renal
	GCS	MAP or vasopressors	PaO2/FiO2 (mmHg)	Platelets ($\times 10^3/\mu$l)	Bilirubin (mg/dl)	Creatinine (mg/dl) or UO
+0	15	MAP $\geq$ 70 mmHg	$\geq$ 400	$\geq$ 150	$<$ 1.2	$<$ 1.2
+1	13–14	MAP $<$ 70 mmHg	$<$ 400	$<$ 150	1.2–1.9	1.2–1.9
+2	10–12	dopamine $\leq$ 5 μg/kg/min OR dobutamine (any dose)	$<$ 300	$<$ 100	2.0–5.9	2.0–3.4
+3	6–9	dopamine $>$ 5 μg/kg/min OR epinephrine $\leq$ 0.1 μg/kg/min OR norepinephrine $\leq$ 0.1 μg/kg/min	$<$ 200 AND MV	$<$ 50	6.0–11.9	3.5–4.9 OR $<$ 500 ml/day
+4	$<$ 6	dopamine $>$ 15 μg/kg/min OR epinephrine $>$ 0.1 μg/kg/min OR norepinephrine $>$ 0.1 μg/kg/min	$<$ 100 AND MV	$<$ 20	$>$ 12.0	$>$ 5.0 OR $<$ 200 ml/day

Appendix C. SAPS-II score

Table 7: List of non-static components of SAPS-II score (Gall et al., 1993). The total score is computed by summing the individual contributions. Abbreviations: HR = heart rate; SBP = systolic blood pressure; GCS = Glasgow Coma Scale; BUN = blood urea nitrogen; WBC = white blood cell count, MV = mechanically ventilated including CPAP.

Score	HR[1] [1/s]	SBP[1] [mmHG]	Temp[2] [°C]	GCS[3]	PaO2/Fi02[3] [mmHg/%]	BUN[2] [mg/dl]	Urine[4] [ml/24h]	Sodium[1] [mmol/L]	Potassium[1] [mEq/L]	Bicarbonate[3] [mEq/L]	WBC[1] [1000/mm³]	Bilirubin[2] [mg/dL]
+0	70-119	100-199	else	14-15	not MV	<28	≥1000	125-144	3.0-4.9	≥20	1.0-19.9	<4.0
+1								≥145				
+2	40-69	>199										
+3			≥39							15-19	≥20	
+4	120-159						500-599					4.0-5.9
+5		70-99		11-13				<125	<3 or ≥5			
+6					>199	28-83				<15		
+7	>159			9-10								
+9					100-199							≥6
+10						>83						
+11	<40				<100		<500					
+12											<1.0	
+13		<70		6-8								
+26				<6								

[1] worst (min or max) value in 24h
[2] largest (max) value in 24h
[3] smallest (min) value in 24h
[4] total (sum) value in 24h

Appendix D. Dataset densities

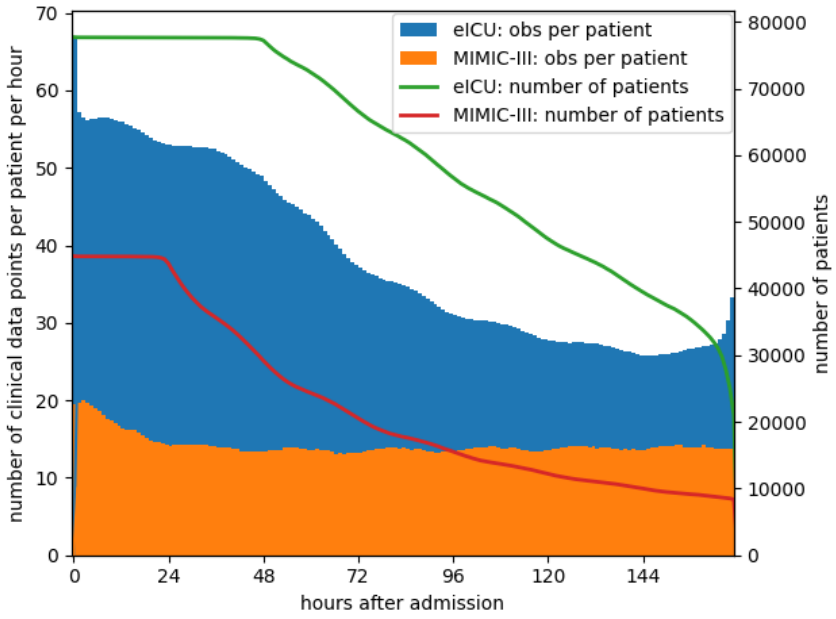

Figure 5: Comparison of the length of stay of patients (lines) and observations per hour per patient (bars) for MIMIC-III and eICU.

Appendix E. Metaparameter settings

Table 8: Metaparameter settings for training runs on MIMIC-III. Best settings chosen on development data are shown in bold face.

	MIMIC-III	
Parameter	Triplet	Dense
Embedding Size	25, **50**, 100	128, 256, **512**
Hidden Size Encoder	25, **50**, 100	128, 256, **512**
Hidden Size DMS Decoder	**50**, 100, 200	128, 256, **512**
Hidden Size IMS Decoder	Output Dimensionality	Output Dimensionality
# Encoder Layers	1, **2**	1, **2**
# DMS Decoder Layers	**1**, 2	**1**, 2
# IMS Decoder Layers	**1**, 2	**1**, 2
Learning Rate	0.0001, **0.0005**, 0.001	0.0001, **0.0005**, 0.001
Batch Size	8, 16, **32**	8, 16, **32**
Attention Heads Encoder	2, **4**, 8	2, 4, **8**
Attention Heads DMS Decoder	Prediction Timesteps	Prediction Timesteps
Attention Heads IMS Decoder	**1**, 2, 4	**1**, 2, 4
Dropout	0.05, 0.1, **0.2**	**0.05**, 0.1, 0.2
Epochs	100	100
Patience	6	6
Random Seed	Unixtime variation	Unixtime variation

Table 9: Metaparameter settings for training runs on eICU. Best settings chosen on development data are shown in bold face.

	eICU	
Parameter	Triplet	Dense
Embedding Size	25, **50**, 100	**256**, 512, 1024
Hidden Size Encoder	25, **50**, 100	**256**, 512, 1024
Hidden Size DMS Decoder	25, **50**, 100	**256**, 512, 1024
Hidden Size IMS Decoder	Output Dimensionality	Output Dimensionality
# Encoder Layers	1, **2**, 3	1, **2**, 3
# DMS Decoder Layers	**1**	**1**
# IMS Decoder Layers	**1**	**1**
Learning Rate	**0.0005**	**0.0005**
Batch Size	**32**	**32**
Attention Heads Encoder	2, **4**, 8	**8**
Attention Heads DMS Decoder	Prediction Timesteps	Prediction Timesteps
Attention Heads IMS Decoder	**1**,2,4	**1**,2,4
Dropout	**0.2**	**0.05**
Epochs	600	600
Patience	6	6
Random Seed	Unixtime variation	Unixtime variation

Appendix F. Experiments on eICU

Table 10: Main evaluation results on eICU for combinations of sparse and dense encoders with direct multi-step (DMS) and iterative multi-step (IMS) decoders, trained by teacher forcing (TF), student forcing (SF), or scheduled sampling (SS), with or without backpropagation (BP) through predictions. Only curricula for randomized increasing (RI) scheduling are shown. Evaluation is done according to MSE of all variables (Equation 5), MSE for SOFA (Equation 8), and Accuracy for Sepsis (Equation 9). Numbers in subscripts denote the 95% confidence interval for the estimation of the respective evaluation score on the test set. Best results are shown in bold face.

Model	Enc	Dec	Train	BP	MSE	MSE-SOFA	Acc-Sepsis
1	Triplet	DMS	-	-	$5.748_{[5.747,5.749]}$	$2.673_{[2.672,2.673]}$	$87.59_{[86.54,88.64]}$
2	Triplet	IMS	TF	-	$9.175_{[9.174,9.176]}$	$4.429_{[4.429,4.429]}$	$83.97_{[82.80,85.14]}$
3	Triplet	IMS	SF	No	$\mathbf{5.134}_{[5.133,5.135]}$	$2.113_{[2.113,2.114]}$	$87.05_{[86.09,88.00]}$
4	Triplet	IMS	SF	Yes	$5.160_{[5.159,5.161]}$	$2.295_{[2.295,2.295]}$	$87.40_{[86.42,88.38]}$
9	Triplet	IMS	SS-RI	No	$5.256_{[5.255,5.257]}$	$2.233_{[2.232,2.233]}$	$86.98_{[86.02,87.94]}$
10	Triplet	IMS	SS-RI	Yes	$5.331_{[5.330,5.332]}$	$2.044_{[2.243,2.244]}$	$86.27_{[85.28,87.26]}$
13	Dense	DMS	-	-	$5.352_{[5.351,5.353]}$	$2.321_{[2.321,2.321]}$	$87.88_{[87.18,88.58]}$
14	Dense	IMS	TF	-	$9.104_{[9.102,9.105]}$	$4.292_{[4.291,4.292]}$	$84.92_{[83.92,85.91]}$
15	Dense	IMS	SF	No	$5.528_{[5.527,5.529]}$	$2.125_{[2.125,2.126]}$	$90.02_{[89.07,90.98]}$
16	Dense	IMS	SF	Yes	$5.395_{[5.394,5.396]}$	$\mathbf{2.024}_{[2.024,2.024]}$	$\mathbf{90.66}_{[89.73,91.58]}$
21	Dense	IMS	SS-RI	No	$5.322_{[5.321,5.323]}$	$2.273_{[2.273,2.274]}$	$88.34_{[87.45,89.24]}$
22	Dense	IMS	SS-RI	Yes	$5.399_{[5.398,5.300]}$	$2.208_{[2.208,2.209]}$	$88.95_{[88.03,89.86]}$
Informer	Dense	DMS	-	-	$5.578_{[5.677,5.679]}$	$2.305_{[2.304,2.305]}$	$87.99_{[87.29,88.68]}$
Autoformer	Dense	DMS	-	-	$5.984_{[5.983,5.986]}$	$2.254_{[2.254,2.255]}$	$85.21_{[84.22,86.19]}$
DLinear	Dense	DMS	-	-	$5.974_{[5.973,5.975]}$	$3.340_{[3.340,3.340]}$	$85.10_{[84.11,86.09]}$
Linear	Dense	DMS	-	-	$5.975_{[5.974,5.976]}$	$3.331_{[3.330,3.331]}$	$85.07_{[84.08,86.07]}$

Table 11: SAPS-II evaluation results on eICU for combinations of sparse and dense encoders with direct multi-step (DMS) and iterative multi-step (IMS) decoders, trained by teacher forcing (TF), student forcing (SF), or scheduled sampling (SS), with or without backpropagation (BP) through predictions. Only curricula for randomized increasing (RI) scheduling are shown. Evaluation is done according to MSE of all variables (Equation 5), MSE for SOFA (Equation 8), and Accuracy for Sepsis (Equation 9). Numbers in subscripts denote the 95% confidence interval for the estimation of the respective evaluation score on the test set. Best results are shown in bold face.

Model	Enc	Dec	Train	BP	MSE	MSE-SAPS-II
1	Triplet	DMS	-	-	$5.748_{[5.747,5.749]}$	$93.799_{[93.799,93.799]}$
2	Triplet	IMS	TF	-	$9.175_{[9.174,9.176]}$	$123.362_{[123.362,123.363]}$
3	Triplet	IMS	SF	No	$\mathbf{5.134}_{[5.133,5.135]}$	$91.337_{[91.337,91.337]}$
4	Triplet	IMS	SF	Yes	$5.160_{[5.159,5.161]}$	$94.379_{[94.379,94.380]}$
9	Triplet	IMS	SS-RI	No	$5.256_{[5.255,5.257]}$	$96.613_{[96.613,96.614]}$
10	Triplet	IMS	SS-RI	Yes	$5.331_{[5.330,5.332]}$	$91.571_{[91.571,91.571]}$
13	Dense	DMS	-	-	$5.352_{[5.351,5.353]}$	$89.429_{[89.429,89.429]}$
14	Dense	IMS	TF	-	$9.104_{[9.102,9.105]}$	$121.478_{[121.478,121.479]}$
15	Dense	IMS	SF	No	$5.528_{[5.527,5.529]}$	$87.598_{[87.598,87.599]}$
16	Dense	IMS	SF	Yes	$5.395_{[5.394,5.396]}$	$\mathbf{86.279}_{[86.279,86.280]}$
21	Dense	IMS	SS-RI	No	$5.322_{[5.321,5.323]}$	$92.525_{[92.525,92.526]}$
22	Dense	IMS	SS-RI	Yes	$5.399_{[5.398,5.300]}$	$91.827_{[91.827,91.828]}$
Informer	Dense	DMS	-	-	$5.578_{[5.577,5.579]}$	$89.139_{[89.138,89.139]}$
Autoformer	Dense	DMS	-	-	$5.984_{[5.983,5.985]}$	$89.513_{[89.513,89.514]}$
DLinear	Dense	DMS	-	-	$5.974_{[5.973,5.975]}$	$94.177_{[94.176,94.177]}$
Linear	Dense	DMS	-	-	$5.975_{[5.974,5.976]}$	$93.743_{[93.743,93.743]}$

Appendix G. Detailed Results for Sepsis Prediction

Table 12: F1 score and the values of the confusion matrix (TP, FP, FN, TN) used to compute it for the MIMIC-III dataset.

Model	Enc	Dec	Train	BP	TP	FP	FN	TN	F1
1	Triplet	DMS	-	-	84	41	41	662	67.2
2	Triplet	IMS	TF	-	84	165	41	538	44.92
3	Triplet	IMS	SF	No	84	46	41	657	65.88
4	Triplet	IMS	SF	Yes	84	48	41	655	65.37
9	Triplet	IMS	SS-RI	No	84	78	41	625	58.54
10	Triplet	IMS	SS-RI	Yes	84	71	41	632	60.0
13	Dense	DMS	-	-	84	76	41	627	58.95
14	Dense	IMS	TF	-	81	236	44	467	36.65
15	Dense	IMS	SF	No	80	46	45	656	63.75
16	Dense	IMS	SF	Yes	85	40	40	663	68.0
21	DENSE	IMS	SS-RI	No	85	72	40	631	60.28
22	DENSE	IMS	SS-RI	Yes	85	82	40	621	58.22
Informer	Dense	DMS	-	-	77	70	48	633	56.62
Autoformer	Dense	DMS	-	-	89	50	36	653	67.42
DLinear	Dense	DMS	-	-	77	149	48	554	43.87
Linear	Dense	DMS	-	-	77	140	48	563	45.03

Table 13: F1 score and the values of the confusion matrix (TP, FP, FN, TN) used to compute it for the eICU dataset.

Model	Enc	Dec	Train	BP	TP	FP	FN	TN	F1
1	Triplet	DMS	-	-	83	181	289	3236	26.1
2	Triplet	IMS	TF	-	46	281	326	3136	13.16
3	Triplet	IMS	SF	No	71	190	301	3227	22.43
4	Triplet	IMS	SF	Yes	75	180	297	3237	23.92
9	Triplet	IMS	SS-RI	No	71	192	301	3215	22.36
10	Triplet	IMS	SS-RI	Yes	76	194	326	3223	22.62
13	Dense	DMS	-	-	84	171	288	3246	26.79
14	Dense	IMS	TF	-	69	269	303	3148	19.44
15	Dense	IMS	SF	No	98	104	274	3313	34.15
16	Dense	IMS	SF	Yes	90	72	282	3345	33.71
21	Dense	IMS	SS-RI	No	75	145	297	3272	25.34
22	Dense	IMS	SS-RI	Yes	86	132	286	3285	29.15
Informer	Dense	DMS	-	-	84	167	288	3250	26.97
Autoformer	Dense	DMS	-	-	72	260	300	3157	20.45
DLinear	Dense	DMS	-	-	57	249	315	3168	16.81
Linear	Dense	DMS	-	-	57	250	315	3167	16.79