

Benchmarking imputation methods on real-world clinical time series with simulated spatio-temporal missingness

Niklas Giesa

niklas.giesa@charite.de

Charité <https://orcid.org/0000-0003-0808-3966>

Rustam Zhumagambetov

Physikalisch-Technische Bundesanstalt

Maria Sekutowicz

Charité - Universitätsmedizin Berlin

Victor Sikora

Charité <https://orcid.org/0009-0001-6554-2316>

Sophie Piper

Charité

Felix Balzer

Charité - Universitätsmedizin Berlin

Stefan Haufe

Technische Universität Berlin <https://orcid.org/0000-0003-1470-9195>

Sebastian Daniel Boie

<https://orcid.org/0000-0001-7993-7901>

Article

Keywords: Missingness Patterns, Electronic Health Records, Imputation Methods, Clinical Time Series

Posted Date: September 1st, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7263571/v1>

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

[Read Full License](#)

Additional Declarations: Yes there is potential Competing Interest. Sebastian Boie reports salaried employment at Pfizer Pharma GmbH. Pfizer had no involvement in the conception, design, execution, or

interpretation of the study, nor in the preparation or decision to submit the manuscript for publication. We do not report further conflicts of interest.

Version of Record: A version of this preprint was published at Nature Communications on October 5th, 2026. See the published version at <https://doi.org/10.1038/s41467-026-77303-8>.

Benchmarking imputation methods on real-world clinical time series with simulated spatio-temporal missingness

Niklas Giesa^{1,*}, Rustam Zhumagambetov², Maria Sekutowicz¹, Victor Sikora¹, Sophie K. Piper¹, Felix Balzer^{1,+}, Stefan Haufe^{1,2,3,+}, and Sebastian D. Boie^{1,+}

¹Charité - Universitätsmedizin Berlin, Berlin, 10115, Germany

²Physikalisch-Technische Bundesanstalt, Berlin, 10587, Germany

³Technische Universität Berlin, Berlin, 10623, Germany

⁺authors contributed equally

^{*}niklas.giesa@charite.de

ABSTRACT

Existing studies evaluating imputation methods on clinical time series do not jointly account for spatial (e.g., across features) and temporal (across time) structure in missingness. We present a novel simulation model that induces realistic spatio-temporal missingness in three real-world clinical datasets by sampling from Markov chains. On the resulting data, we apply a variety of imputation methods including Last observation Carried Forward (LOCF), linear interpolation, and spatio-temporal autoencoder (STAE). Finally, we evaluate the influence of time series imputations on a downstream prediction task. Deep STAE outperformed simpler baseline methods only when the missingness patterns lacked spatio-temporal structure. In contrast, linear interpolations performed best on such spatio-temporal missingness, contributing to sufficient downstream performances. This study is the first to demonstrate that simple linear imputations attain robust performance for clinical real-world data with spatio-temporal missingness. We provide an open benchmark to evaluate the effect of missingness in clinical studies and prediction tasks.

Missingness Patterns, Electronic Health Records, Imputation Methods, Clinical Time Series

Introduction

Routinely collected electronic health records (EHRs) offer extensive opportunities for clinical machine learning (ML) or retrospective clinical studies but present inherent challenges due to poor data quality¹⁻³. Among these challenges, missingness in longitudinal data hampers the application of ML caused by suboptimal model fits⁴⁻⁶. It also remains an open question whether current data handling strategies faithfully reconstruct patient trajectories.

Rubin⁷ introduced patterns of missing completely at random (MCAR) where data with missing values form an arbitrary subset of the full data which is unobserved, missing at random (MAR) which occurs when the probability of a missing value depends on observed data, and missing not at random (MNAR) where the probability of a missing value depends on unobserved data. These patterns may also occur jointly⁸. As an example, MCAR possibly occurs due to a temperature sensor failure which randomly misses recordings, MAR might be caused by difficulties in measuring blood oxygen saturation (Spo2) for hypothermic patients^{9,10}. In most cases, the exact reason for missing data in clinical time series remains unresolved.

To estimate missing clinical data, various imputation strategies have been established over the past years. Multiple imputations by chained equations (MICE) iteratively learns spatial relationships from covariances and is mostly used for population-based studies¹¹. K-nearest neighbor (KNN) imputation uses available data from the most similar observed subjects, based on a defined similarity metric¹². In 2020, Ma et al. proposed a multiple imputation denoising autoencoder (MIDA) model which comprises stacked multi-layer perceptrons (MLPs) in an encoder-decoder architecture, reconstructing tabular MCAR data¹³. All these methods omit temporal dynamics, which are of particular interest to ML algorithms using time series of features¹⁴. Linear or quadratic interpolation, last observation carried forward (LOCF), or running median techniques are commonly used to easily impute time series, but they lack in capturing spatial cross-feature dependencies¹⁵⁻¹⁷. Deep learning (DL) techniques, like spatio-temporal denoising autoencoder (STAE)¹⁸⁻²⁰, are claimed to overcome these issues by jointly learning temporal - as well as spatial aspects across features, time, and between time series.

Previous studies have applied imputation techniques on a wide range of clinical data modalities like EHRs²¹, omics²², electrocardiograph²⁰, or electroencephalography²³ recordings. Existing EHR imputation benchmarks often rely on the Medical

Information Mart for Intensive Care (MIMIC) providing open accessible clinical data from intensive care units (ICUs)^{15,24–26}. Effects of the imputation quality on downstream classification tasks, like predicting acute kidney injury (AKI) or mortality, have been additionally investigated^{18,19,27}. While temporal aspects were addressed by improved versions of KNN or MICE algorithms^{12,28}, further studies also proposed DL autoencoder architectures. Lia et al.²⁹ and Li et al.³⁰ proposed STAE models, based on long-short term memory (LSTM) modules, which were enhanced by a temporal smoothing function. Other STAE models used combinations of convolutional neural networks (CNNs), gated-recurrent units (GRUs), and attention layers^{24,31} to impute both, spatial and temporal missingness. Fan et al. proposed a CNN autoencoder architecture for processing time series via windowing, which involved a trainable graph for static attributes³².

A comprehensive evaluation by Jäger et al.³³ demonstrated that MICE or KNN methods are favorable over DL methods for imputing tabular data. Similarly, Zamanzadeh et al.¹⁸ found that KNN outperformed STAE in more complex missing scenarios like MNAR. Results from additional studies raised the question whether complex and highly tailored DL architectures translate into generalizable methods^{19,24,29}. Specifically, previous studies evaluated STAE against mean or median imputation, MICE, or additional DL models, rather than considering LOCF or linear imputations^{19–21,24,26,29}.

For simulating missing data, the majority of previous authors simply created synthesized or real-world CCD sets inducing MCAR under varying missing rates^{13,15,19,20,23–25,29,31,34}. Missingness patterns in clinical time series are, however, characterized by spatio-temporal correlations which are not addressed by these MCAR inductions³⁵. As an example, systolic and diastolic blood pressure (BP) values are likely either both present or absent at any time, since they rely on the same sensor and are measured simultaneously. These spatio-temporal relations have not been simulated by previous works. For comparing imputed data and CCD as ground truth (GT) data, the mean squared error (MSE) and the mean absolute error (MAE) are established performance metrics^{24,29,31}. However, these metrics are not necessarily well-suited to measure the faithful reproduction of aforementioned spatio-temporal data correlations.

Related studies report heterogeneous results, advocating various imputation methods, which have not been systematically evaluated under realistic missingness patterns. This study addresses these issues by introducing a novel benchmark for simulating realistic spatio-temporal missingness by sampling from Markov chains. We evaluate 20 different imputation methods, covering MICE, KNN, LOCF, and DL models, and apply them on commonly available vital sign data from three real-world datasets. Based on Li et al.³⁰, we hypothesize that more complex STAE outperform common baseline imputation techniques like LOCF. We further define a clinical downstream prediction task for each dataset (e.g., predicting delirium) to assess the impact of imputations. Standard evaluation metrics like MSE and MAE are complemented by novel metrics assessing the reconstruction of spatio-temporal properties, overall offering a comprehensive benchmarking toolkit.

Methods

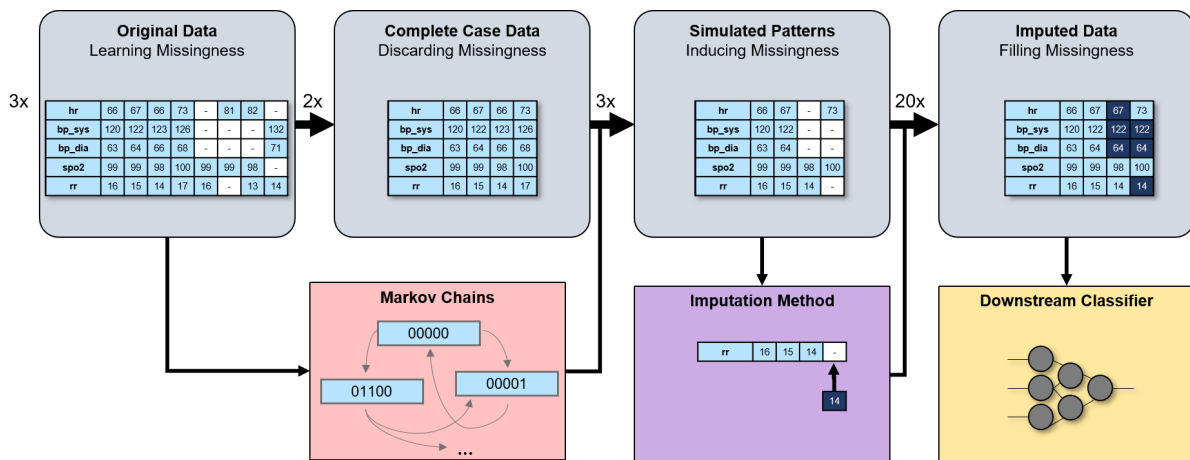


Figure 1. Experimental setup. We analyzed three datasets holding clinical time series as original data, applied two cleaning strategies for building complete case datasets (CCD) and simulated three types of missingness. We then applied twenty methods for imputing data and train downstream classifiers, resulting in a total of 360 numerical experiments.

Experimental Setup

Figure 1 outlines our experimental setup. All methods are explained in the following paragraphs. We estimated transition probabilities between consecutive states of spatio-temporal missingness from three original datasets. Hereby, we built simulation

models for three different types of spatio-temporal missingness either reflecting the original patterns or additionally amplifying MAR or MNAR effects using Markov chains. After two types of data cleaning approaches, we then discarded missing values to generate CCD for each dataset. Afterwards, we induced missingness patterns on CCD based on the learned probabilistic models, resulting in data with simulated missingness. Twenty different methods were applied to fill in the missing values and, thereby, generated imputed datasets. Imputation performances were assessed using the known CCD as the GT. Finally, the imputed datasets were used to solve a clinical downstream classification task. Downstream performances were compared with respect to previously applied imputation methods.

Data Retrieval

This study leveraged three different clinical datasets consisting of pre-processed EHRs. Two public datasets (MIMIC and HIRID) contain medical records collected during ICU stays, while one in-hospital dataset (ICDEP) contains intraoperative variables recorded during surgeries.

MIMIC We applied our methods on the third version of MIMIC³⁶. This public database from the U.S. contains data of 38,597 patients admitted to ICUs between 2001 and 2012 at the Beth Israel Deaconess Medical Center in Boston. We used data which were pre-processed with the pipeline provided by Wang et al.³⁷ The pipeline semantically integrates features that describe the same physiological parameters (e.g., invasive and non-invasive BP) while resampling time series to an hourly equidistance grid. Wang et al. selected frequently used features related to vital signs, clinical interventions, and laboratory results for clinical prediction tasks. They limited their cohort to adult patients (≥ 18 years) with a minimum of 12 hours of recorded time series. Although multiple ICU admissions for the same patient could be observed, the authors included one distinct time series per patient. The openly accessible pre-processed data on PhysioNet³⁸ cover 34,472 patients with 104 time-dynamic features³⁹.

HIRID The high time-resolution ICU dataset HIRID contains data of 36,098 adult patients admitted to the Bern University Hospital in Switzerland between 2008 and 2016^{40,41}. The original dataset stores 681 variables representing bedside monitoring data, settings of medical devices, laboratory values, and administered drugs. In the original publication, Hyland et al.⁴¹ selected and semantically integrated 209 clinical variables for the prediction of circulatory failure focusing on cardiac patients. We acquired data from Physionet³⁸ and manually selected 101 of these 209 features, omitting non-numeric and static ones while integrating identical physiological features similar to Wang et al.³⁷ (see **Table B.1 in Supplement B**). We mean-aggregated and resampled the data into a ten minutes equidistant grid as suggested by the original publication⁴¹. HIRID treats each ICU admission individually and assigns a unique patient identifier to each hospital stay increasing patient privacy.

ICDEP We used data from the internal Intraoperative Care Delirium Prediction dataset ICDEP acquired at the German medical center Charité - University Medicine Berlin. During surgical interventions, patients receive high frequency monitoring for the time of a surgical procedure⁴². ICDEP was originally created on the basis of monitoring data for the prediction of postoperative delirium^{14,43,44}. The dataset comprises 41,832 adult patients admitted between 2017 and 2020 at Charité. We identified the most predictive features that are available for at least 10% of all patients. We semantically integrated features that described the same physiological parameter as mentioned before. As a result, we selected 61 intraoperative time series and mean-aggregated these features into a five minutes equidistant time grid. Due to some patients undergoing multiple surgeries, we reduced our time-series data to the longest sequence per patient to minimize the loss of temporal dynamics.

All three datasets contain demographic information like sex and age. After temporal resampling longitudinal data to increase their availability, we explored the joint missingness of different feature sets. We then selected five vital signs that are highly available across all three datasets to conduct further analysis: heart rate (hr), systolic (bp_sys) and diastolic (bp_dia) BP, blood oxygen saturation (spo2), and respiratory rate (rr) (see **Results**).

Spatio-Temporal Missingness

The data pre-processing pipelines, which were introduced by the previous section, yielded one data matrix $X_n = (x_{j,t})_n$ with $X_n \in \mathbb{R}^{F \times l_n}$ for each n -th patient with $f = \{1, 2, \dots, F\}$ features at time indexes $t = \{1, 2, \dots, l_n\}$ where the time sequence length l_n varies across patients (see colorized tables in **Figure 2a**). Thus, pre-processed data exhibited spatial relationships across features and temporal relationships across time for a population of N patients. For each X_n , we derived a binary missing indicator matrix $R_n = (r_{j,t})_n$ where

$$r_{j,t} = \begin{cases} 1 & \text{if } x_{j,t} \text{ is missing} \\ 0 & \text{if } x_{j,t} \text{ is observed} \end{cases} \quad (\text{see colorized tables in Figure 2b}).$$

The quantity $Q_{n,j} = \sum_{t=1}^{l_n} r_{j,t}$ calculates the number of missing values per patient n for the j -th feature. We define feature missingness (FM) as the fraction of missing values for feature j across all time steps and patients:

$$FM_j = \frac{\sum_{n=1}^N Q_{n,j}}{\sum_{n=1}^N l_n}. \quad (1)$$

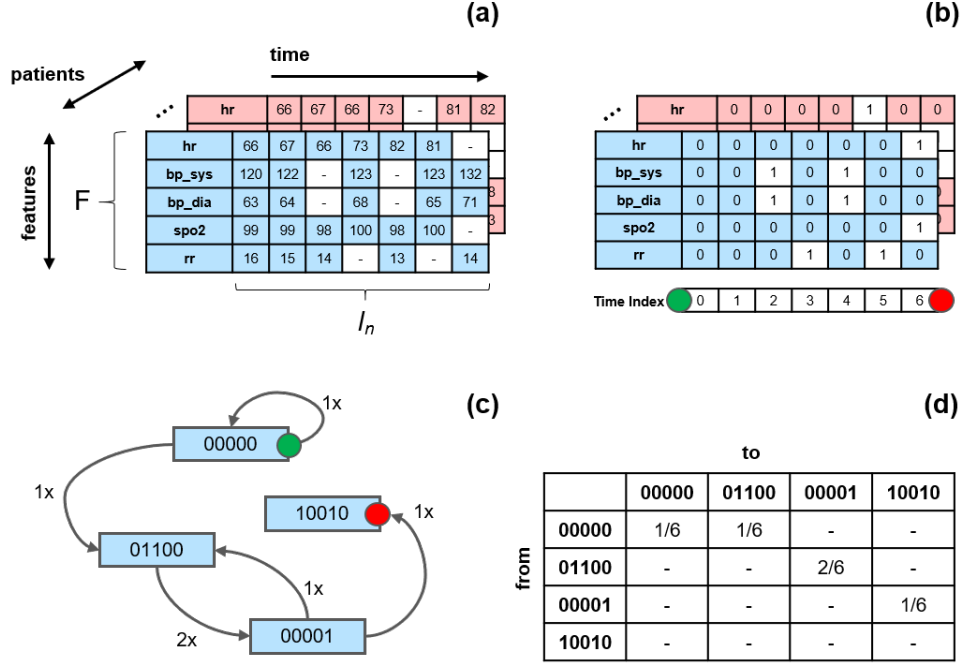


Figure 2. Illustration of spatio-temporal missingness and Markov chain model. We show an illustrative example with $F = 5$ features: heart rate (hr), systolic (bp_sys) - and diastolic (bp_dia) blood pressure, blood oxygen saturation (spo2), and respiratory rate (rr). A data matrix $X_n \in \mathbb{R}^{F \times I_n}$ is stored for each patient n holding values for F features at time indices $i = 0, \dots, I_n$. Panel **a** shows two examples for X_n (colored), dimensions across features, patients, and time are highlighted. Panel **b** shows corresponding binary missing indicator matrices $R_n \in \{0, 1\}^{F \times I_n}$ for each X_n . In panel **c**, we illustrated all 4 distinct temporal slices (building the Markov states) that occurred in the blue R_n (of dimensions 5×7) as missing vectors in boxes. The occurrences of transitions from one temporal slice to another are counted (shown as numbers next to arrows), resulting into a Markov chain. In panel **d**, the relative frequencies of transitions are inserted into a transition table. For our simulation model, we retrieved a transition table across all patient time series covered by our population where rows (all probabilities for a consecutive state) sum up to one.

The defined missingness metric FM_j was used to explore and select features that were available across all datasets and to quantified differences between original and simulated spatio-temporal missingness patterns. Values of FM_j range between 0 and 1 where levels of 0.2 are considered as low due to the inclusion of multiple sequences with varying lengths in **Equation 1**. We also report additional missingness metrics, like block missingness, in **Supplementary A** to further prove the quality of our simulated missingness patterns (see **Additional Methods A.1**).

Complete Case Data

For generating CCD without missingness, the resampled original data were initially cleaned with two separate cleansing strategies. We observed some outliers across all datasets holding physiologically impossible values, e.g., Spo2 exceeding 100%². Similar to previous studies^{14,37,45}, we used data ranges of [5, 400] bpm for hr, [2, 300] bpm for rr, [10, 100] % for spo2, [20, 400] mmHg for bp_sys, and [10, 300] mmHg for bp_dia to detect outliers. Feature values that fall outside these defined ranges are marked as outliers. The first cleaning strategy preserved the original spatio-temporal missingness by setting outliers to the nearest edges of the valid ranges, e.g, 120% Spo2 was set to 100%. Thereby, outliers were corrected instead of removed from the dataset. We report results from this strategy in the main manuscript. In contrast, the second strategy set outliers to missing, also altering spatio-temporal missingness patterns that were learned by our proposed simulation model. At this point, no patients were excluded from the cleaned dataset under either of the data cleaning strategies.

For the generation of CCD, we selected the largest continuous block of available data in each data matrix. We therefore reduced every binary missing indicator matrix R_n to the longest sequence of consecutive zeros and kept the corresponding values of X_n . As another constraint, resulting sequences should exhibit a minimum length of three values. All other data that did not comply with these conditions were discarded from the resulting CCD. Hence, patients that did not have a continuous block of complete data with at least three time steps were dropped out from the dataset. The generation of CCD also resulted in the overall reduction of sequence lengths, we further describe the impact of these modifications in the **Results** section.

Simulation Model

We define a state space S where each state s^k in $S = \{s^1, s^2, \dots, s^K\}$ represents a distinct temporal slice $(r_{:,t})_n$ of the binary missing indicator matrix R_n (see **Figure 2c**). We stored the occurrences of unique states in a frequency table $FT = \{f^1, f^2, \dots, f^K\} \in [0, 1]^K$ where each f^k denotes the marginal probability of the occurrence of each s^k across all patients. Assuming the Markov property⁴⁶, stating that the probability for moving to the next state depends only on the current state, we estimated transition probabilities between states in S yielding a transition matrix $P \in [0, 1]^{K \times K}$, with rows summing up to one (see **Figure 2d**).

For the induction of learned spatio-temporal missingness patterns on complete time series, we simulated one binary missing indicator matrix $\hat{R}_n \in \{0, 1\}^{F \times \hat{l}_n}$, with $\hat{l}_n$ as the novel (potentially) reduced sequence length, for each patient n that remained in the CCD, using the following strategy:

1. **Initialization** We sampled the initial state s_0 that yields the first temporal slice of $\hat{R}_n$ as $(\hat{r}_{:,0})_n$ by drawing one state s^k out of S with the probability f^k from FT .
2. **Iterative Sampling** For generating the next state s_{i+1} , we draw a sample using the transition probabilities defined by P .
3. **Termination** Step 2 was repeated until all $\hat{l}_n$ temporal slices are filled resulting in the simulated $\hat{R}_n$.

Due to the state space being finite, we could estimate occurrence as well as transition probabilities for all included states s^k , building the basis of our discrete Markov chain model⁴⁷. We describe the change of original and simulated spatio-temporal patterns in the **Results** section. Due to the artificial induction of missingness on CCD, we have access to the GT values, which allows us to quantitatively assess the quality of imputation methods.

Amplification of Missingness Patterns

Extracted clinical time series possibly inherent all three missingness types MCAR, MAR, and MNAR⁷. If values are MCAR, the available data is independent from co-variables or other (un)observed data. While we assume that MCAR occurs inherently in our detected spatio-temporal missingness patterns in the original datasets (e.g., random malfunction of Spo2 sensor), we did not further artificially increase these effects unlike previous authors^{15, 19, 25, 31}.

We estimated existing age-related MAR effects as $p(R|age)$ in the original data that were artificially amplified. To this end, we introduced three equally sized age bins across all datasets (age_bins = {(18,51], (51,69], (69,120]} years]). We then modified the probability for the state $s = [1, 1, 1, 1, 1]$ where all features are missing, denoted as total absent state (TAS), in step 1 and 2 described in **Simulation Model**. For a positive MAR effect, the occurrence - as well as transition probabilities for the TAS were increased for elderly patients by multiplying probabilities with a factor that corresponds to the previously defined age bins (age_factors = {1.2, 2.2, 3.2}, higher factors were applied to bins of higher age). For a negative MAR effect, probabilities were increased for younger patients by applying factors in a reverse order (age_factors⁻¹ = {3.2, 2.2, 1.2}). We decreased the probabilities of the total present state (TPS), encoded as $s = [0, 0, 0, 0, 0]$ where all features are observed, by the margin that was added to the TAS so that probabilities sum up to one.

Since we know the GT underlying values for the induced spatio-temporal missingness on CCD, we further conditioned missingness on GT value levels after all steps in **Simulation Model** to amplify MNAR. We repeated the previously described binning procedure for GT feature values resulting into a $F \times 3$ matrix holding one of three bins per feature (see **Table A.1** in **Supplement A**). For a positive MNAR effect, we induced more FM for values falling into higher value bins by multiplying probabilities with increased factors (value_factors = {1.8, 2.8, 3.8}). Unlike MAR, MNAR effects could not be assessed in our original datasets since, by definition, the reasons for missing values were not documented.

Imputation Methods

Simple Imputation Models

We applied 20 imputation methods (see **Table 1**) via a 3-fold cross-validation (CV) approach where statistics are derived from 66% as training data and applied to 33% as testing data⁴⁸. The CV splitting of time series and fittings of imputation methods was repeated three times to distinct chunks of simulated patient data. We estimated median and mean statistics per feature j on training data across all patients and time (PMEAN, PMEDIAN) as well as at the initial time step ($t = 0$) across patients (1MEAN, 1MEDIAN). Additionally, we retrieved both statistics as running means or medians (RMEAN, RMEDI) within one single patient time series on previously recorded observed values (from time 0, 1, ..., t) to impute the next missing value (at time $t + 1$) per feature, thus time-causal aspects were preserved.

Inter- and extrapolation methods⁴⁹ were furthermore applied for one single patient time series per feature. Linear (LINPOL) and nearest neighbor (NEAPOL) imputation required two observed values (at t and $t + m$ with $0 < m < \hat{l}_n$) to impute missing values in between. To fit a quadratic interpolation (QUAPOL) term, at least three observed values were required within one time series. If less values were observed, interpolation fell back to a lower-order linear term. As the most prominent method

ID	Name	Imputation Method	Cross Population	Cross Features	Time Causal	Static Features
Averaging Methods						
1	PMEAN	Population Mean	TRUE	FALSE	FALSE	FALSE
2	PMEDI	Population Median	TRUE	FALSE	FALSE	FALSE
3	1MEAN	Mean Initial Time Steps	TRUE	FALSE	TRUE	FALSE
4	1MEDI	Median Initial Time Steps	TRUE	FALSE	TRUE	FALSE
5	RMEAN	Running Mean	FALSE	FALSE	TRUE	FALSE
6	RMEDI	Running Median	FALSE	FALSE	TRUE	FALSE
Interpolation Methods						
7	LINPOL	Linear Interpolation	FALSE	FALSE	FALSE	FALSE
8	NEAPOL	Nearest Interpolation	FALSE	FALSE	FALSE	FALSE
9	QUAPOL	Quadratic Interpolation	FALSE	FALSE	FALSE	FALSE
Extrapolation Methods						
10	LOCF	Last Observation Carried Forward	FALSE	FALSE	TRUE	FALSE
11	LINEXT	Linear Temporal Extrapolation	FALSE	FALSE	TRUE	FALSE
12	QUAEXT	Quadratic Temporal Extrapolation	FALSE	FALSE	TRUE	FALSE
Multiple Imputations by Chained Equations						
13	MICTRE	MICE Random Forest	TRUE	TRUE	FALSE	TRUE
14	MICLRE	MICE Linear Regression	TRUE	TRUE	FALSE	TRUE
15	MICKNN	MICE K-Nearest Neighbour	TRUE	TRUE	FALSE	TRUE
Denoising Autoencoder						
16	SAE	Stacked DAE	TRUE	TRUE	FALSE	TRUE
17	MIDA	Multiple Imputations DAE	TRUE	TRUE	FALSE	TRUE
18	BILSTAE	Spatio-Temporal BiLSTM DAE	TRUE	TRUE	FALSE	TRUE
19	LSTAE/GRUAE	Spatio-Temporal GRU/LSTM DAE	TRUE	TRUE	TRUE	TRUE
20	STGAE	Spatio-Temporal Graph DAE	TRUE	TRUE	TRUE	TRUE

Table 1. Overview about applied imputation methods. We imputed data with 20 different methods configuring several model variants. We fitted averaging methods like running means, inter- and extrapolation methods, and denoising autoencoder (DAE). Methods could infer statistics across patients (cross population), cross features, and time-causal as preserving the timely order not taking values from future time points. Models optionally used static feature attributes like age.

Autoencoder Model	Hyperparameter
SAE	$d_{sae} = \{32, 16, 8, 4\}, \{128, 64, 32, 16\}, \{256, 128, 64, 32\}$
MIDA	$d_{mida} = F + \Theta \times q, \Theta = \{4, 6, 9\}$
STAE	$d_{stae} = \{32, 64, 128, 256\}$
STGAE	$d_{stgae} = \{32, 16\}, \{64, 32\}, \{128, 64\}$
All	$lr = \{1e^{-4}, 1e^{-5}, 1e^{-6}\}$
	$bz = \{32, 64, 128\}$
	$dr = \{0.5, 0.75, 0.9\}$
	$wd = \{0, 1e^{-1}, 1e^{-2}, 1e^{-3}\}$

Table 2. Learnable Hyperparameters for Autoencoder Models. The dimensions d is learned per specific model as the size of the latent space in an encoder-decoder architecture. For SAE and MIDA, d determines the dimensions of stacked multi-layer perceptrons (MLPs), where F defines the number of input features. For MIDA, the dimension of each MLP on the q -th layer is calculated with factors of Θ . STAEs comprise either one or two modules of LSTM, BiLSTM, or GRU with hidden dimensions of d_{stae} . The learning rate lr , dropout rate dr , and weight decay wd are optimized for all autoencoders independent from their architecture. We trained STAEs with batch size one and accumulate gradients after bz steps. For all other autoencoders, the batch size was directly equal to bz .

from related work, LOCF was applied^{16,50}, which simply uses the last observed value (at t) to impute the next missing one (at $t + 1$). Similar to RMEAN and RMEDI, linear- and quadratic extrapolation methods relied only on previously observed values. If initial values (at $t = 0$) within one time series were unobserved, IMEDIAN was applied to replace them.

The multiple imputations method MICE^{11,28}, originally not designed for learning temporal correlations, fills initial missing values for $F - 1$ features with placeholders like PMEAN. Then, the method approximates the unobserved values for one feature j from all other features by fitting a function over the whole population. This approach is iteratively repeated across features until imputed values converge. We deployed three MICE models using random forests (MICTRE), multivariable linear regression (MICKRE), and KNN (MICKNN) as approximation functions, respectively (see Table 1). Models were fitted on additional static features which were always observed to potentially improve imputation performances.

Autoencoder Architectures

Table 2 displays the hyperparameter search space for autoencoder architectures. The stacked autoencoder (SAE) and multiple imputation denoising autoencoder (MIDA) follow an encoder-decoder architecture^{13,51} to reconstruct missing data. SAE was configured with four layers of stacked MLPs with configurable dimensions d_{sae} , where the initial input dimension was always set to the number of selected features $F = 5$. MIDA stacks three MLPs sequentially and uses a fixed factor Θ to determine the dimension d_{mida} per q -th layer with $F + \Theta \times q$ (see Table 2). SAE uses logistic sigmoid and MIDA hyperbolic tangent functions as activation functions⁵².

SAE and MIDA infer spatial relations, rather than temporal ones, globally for the entire population. In contrast, STAEs learn from both relations using DL models like LSTMs to reconstruct time series (see Figure A.1 in Supplement A). We configured the STAE architecture, motivated by similar studies^{20,29,30}, with one or two stacked GRUs (GRUAE), LSTMs (LSTAE), or bidirectional LSTMs (BiLSTAE) (see Table 1). The size of the latent space d_{stae} determined the number of hidden neurons (see Table 2). The latent space was optionally enhanced in the temporal dimension by the output of one perceptron, combining a linear layer with logistic sigmoid activation, to process static features such as a patient’s age. Due to the processing of time series with varying lengths, we used a batch size of one and accumulated gradients after $ag = bz$ steps⁵³ (see Table 2).

Fan et al.³² developed the so-called STGAE architecture with two stacked CNN modules for the encoding and decoding process. We trained the model with dimensions of d_{stgae} . In addition to CNNs, the proposed architecture comprises a graph for static attributes. We configured this graph with demographics where Euclidean distances (edges) were calculated between patients (nodes) for the attributes age and sex. For applying the CNN to unequal lengths of sequences, we used padding where the padding token did not contribute to the loss.

We z-standardized train and test time series by calculating mean and standard deviations on the three training folds only for all autoencoders. Afterwards, missing $\bar{x} \in ts$ were filled with ones or zeros as noise. Training statistics were also used for de-standardization after decoding. As a reconstruction loss, we configured a MSE function⁵⁴ as l_{MSE} for SAE, MIDA, and STGAE. For our configured STAE models (GRUAE, LSTAE, BiLSTAE), we either used an unmodified MSE loss or configured a combined loss cl as a linear combination of l_{MSE} and the WSD⁵⁵ loss l_{WSD} as $cl = \lambda_1 \times l_{MSE} + \lambda_2 \times l_{WSD}$ (also see Evaluation Metrics). Empirical scaling set $\lambda_1 = 1.0, \lambda_2 = 0.445$ so that both loss terms contributed equally. The l_{WSD} was realized by integrating the Sinkhorn loss⁵⁶ that adds an entropy term to the WSD to accelerate differentiability.

We employed hyperparameter tuning via grid search⁵⁷ and hyperband search⁵⁸ with asynchronous scheduling. For SAE,

MIDA, and STGAE, dimensions $d_{sae}, d_{mida}, d_{stgae}$ were optimized with grid search alongside learning rate lr , batch size bz , drop out dr , and weight decay for adam optimizer wd . Due to higher computation costs of STAEs, we used hyperband optimization for lr, wd, dr like before in conjunction with the latent sizes d_{sae} and gradient accumulation steps ag . Data with simulated missingness were filled with LOCF to retrieve the GT for our models to offer a fair comparison. Using available real GT data for fitting models was possible, but other approaches like MICE also assume simple imputations as a starting point. Using LOCF enable STAE to learn spatio-temporal dependencies while generalizing well. For a comparison, we also used the known CCD for GT validation while training, these known values are otherwise concealed for all methods. Final model parameters are presented in **Table B.2** in **Supplementary B**.

Evaluation Metrics

We assume GT data $Z_n = (z_{j,t})_n$ and imputed data $\bar{X}_n = (\bar{x}_{j,t})_n$ per patient n , feature j , and time t where imputations just affected simulated missing data for $\hat{R}_n = (\hat{r}_{j,t})_n = 1$. We calculated the normalized root-mean squared error (NRMSE) and the normalized mean absolute error (NMAE)⁵⁹ for imputed values only with $I = \sum_{t=1}^{\hat{l}_n} \hat{r}_{j,t}$ as

$$NRMSE_{j,n} = \frac{\sqrt{\frac{1}{I} \sum_{t=1}^{\hat{l}_n} (z_{j,t} - \bar{x}_{j,t})^2}}{\mu_n(z_j)} \quad (2)$$

$$NMAE_{j,n} = \frac{\frac{1}{I} \sum_{t=1}^{\hat{l}_n} |z_{j,t} - \bar{x}_{j,t}|}{\mu_n(z_j)} \quad (3)$$

where $\mu(z_j)$ in the denominator calculates the mean over GT feature values from one patient time series as a normalization term. In contrast to these two errors, we applied the normalized Wasserstein distance (NWSW)⁶⁰ as a metric comparing two distributions over all data (including non-imputed data) across all patients per single feature with

$$NWSW_j = \frac{\int_{-\infty}^{\infty} |CDF(\bar{X}'_j) - CDF(Z'_j)|}{\max(Z'_j) - \min(Z'_j)} \quad (4)$$

where Z'_j and $\bar{X}'_j$ hold GT and imputed values per feature for the entire population with lengths of $\sum_{n=1}^N \hat{l}_n$. The cumulative density function $CDF(x) = P(X \leq x)$ projects the probability of a random variable X being less or equal to x . We additionally calculated NRMSE and NMAE globally for the entire population (see **Additional Methods A.1** in **Supplement A**). To further quantify the changes in temporal dynamics by our imputations, we used auto-correlation error (ACorE) and cross-correlation error (CCorE) terms⁶¹. When the function $Cov(x, y)$ calculates the covariance for two random numbers, we derived errors between ACor coefficients, calculated on GT and imputed data, for L time shifts (lags) per patient and feature as

$$ACorE_{j,n} = \frac{1}{L} \sum_{lag=1}^L |AC(z_{j,t}, z_{j,t-lag}) - AC(\bar{x}_{j,t}, \bar{x}_{j,t-lag})|, \text{ with } AC(x, y) = \frac{Cov(x, y)}{\sigma(x)^2}. \quad (5)$$

In contrast to the ACorE that describes the temporal correlations of single feature time series values with a time-shifted version of themselves, the CCorE further quantifies differences of spatial correlations between features j and j' as

$$CCorE_n = \frac{1}{F^2} \sum_{j=0}^F \sum_{j'=0}^F \frac{1}{L} \sum_{lag=1}^L |CC(z_j, z_{j'}) - CC(\bar{x}_j, \bar{x}_{j'})|, \text{ with } CC(x_t, y_t) = \frac{Cov(x_t, y_{t-1})}{\sigma(x_t)\sigma(y_t)}. \quad (6)$$

ACorE and CCorE range between 0 and 1. Metrics that were calculated per patient, such as NMAE, were called local errors. We refer to metrics that were calculated for the entire population, such as the NWSW, as global errors.

Downstream Classifier

For each dataset, we defined a downstream classification task. Specifically, we predicted in-hospital mortality for MIMIC and HIRID and postoperative delirium for ICDEP. In our previous work, we showed that an encoder-only transformer (TRAN) architecture successfully predicts binary outcomes on the basis of clinical time series, outperforming LSTM or MLP

approaches¹⁴. The TRAN model applies a linear layer on model inputs to expand the learnable latent space with a dimension of $d_{tran} = \{32, 64, \text{ or } 128\}$. After the linear layer, a single-head self-attention module learns from the dimensionally expanded input space using a feed forward neural network (FFNN) with dimension $d_{fn} = \{32, \text{ or } 64\}$ to combine query, key, and value results. Afterwards, another FFNN is applied to focus on temporal aspects with fixed dimensions of 4 nodes on each of the 4 layers. Hyperparameters d_{tran} and d_{fn} were tuned with Hyperband⁵⁸ including batch sizes, learning rates, and weight decays into a searchable parameter space as before mentioned in the STAE training procedure. We determined initial hyperparameters separately for MIMIC, HIRID, and ICDEP on CCD training data and applied the TRAN architecture with a fixed configuration on various reconstructed sets to measure the importance of imputation methods. We trained and evaluated the model three times on the three imputed cross-validation hold-out sets, splitting them again into 66% train and 33% testing data. We applied 1000x bootstrapping on the testing sets to estimate 95% confidence intervals (CI). Models were trained for 150 epochs while early stopping interrupted the training process when the testing loss increased over more than 10 epochs. Downstream classification performances were evaluated across amplification scenarios with the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC)⁶². In contrast to the previous study¹⁴, TRAN was configured as the main downstream classifier without comparison with MLPs or additional model variants. Implementation details are provided in our previous work¹⁴, final tuned parameters are included in **Table B.3** in **Supplementary B**.

Code and Data Availability

We share code, trained downstream models, and instructions for the application of our simulation model via GitHub⁶³. The pre-processed MIMIC³⁷ and HIRID⁶⁴ data are openly accessible via Physionet⁶⁵ where a research proposal must be submitted for HIRID. ICDEP data is not published openly, due to German data privacy regulations, but can be requested from the corresponding author to establish a research collaboration. After data retrieval, the GitHub code can be used to reproduce our results.

Ethics Statement

For processing ICDEP data, the IRB EA4/254/21 was granted by the ethics committee at Charité - Universitätsmedizin Berlin. For MIMIC and HIRID, data privacy and sharing regulations apply which are further defined at the PhysioNet platform.

Results

Missingness Simulation

To construct CCD from original data, we searched for a set of common features that jointly examined low missingness (high availability). Time series were resampled with seven multiples $2, 3, \dots, 8\times$ of the original sampling intervals (e.g., $\{120, 180, \dots, 480\}$ min. for a 60 min. interval in MIMIC) to decrease their FM rates (see **Figure 3a**). While hr and spo2 expressed low FM < 0.15 across all datasets for original sampling intervals already, $2\times$ resampling especially decreased the missingness for rr in HIRID (FM reduction from 0.403 at 10 min. to 0.222 at 20 min.) and BP in ICDEP (FM reduction from 0.422 at 5 min. to 0.218 at 10 min). Even after $8\times$ resampling, temperature (temp) remained highly missing in some datasets (e.g., FM > 0.7 in HIRID for 80 min. sampling) (see **Table A.2** in **Supplement A**). We continued our analysis without temp, due to high missingness, and without bp_mean, since the feature is a linear combination of bp_sys and bp_dia.

We resampled HIRID and ICDEP with intervals of 20 and 10 min., kept the originally 60 min. sampled MIMIC data, and selected the highly available features hr, bp_sys, bp_dia, spo2, and rr. After correcting outliers, mean statistics were marginally shifted while the standard-deviation (SD) decreased for most of the features (e.g., 86.36 ± 17.46 v.s. 86.11 ± 16.80 bpm as mean $\pm$ SD for hr in HIRID) (see **Table A.3** in **Supplement A**). Instead of correcting outliers, we also removed them (by setting outliers to missing) changing spatio-temporal missingness patterns. We aimed to simulate missingness closest to the original data, so we report these additional results in the **Supplement A** (see **Additional Results A.1**).

In the original resampled data, the cross-correlation (CCor) (lag=0) of missingness, was perfect (CCor=1) between bp_sys and bp_dia across all datasets and ranged on higher levels in MIMIC (CCor > 0.71) than in other sets (see **Figure 3b**). The auto-correlation (ACor) (lag=1) was highest for rr in HIRID (ACor = 0.73) as well as in ICDEP (ACor = 0.72) and relatively decreased for both BP features (ACor = 0.20 for bp_sys and bp_dia) in ICDEP. FM rates were consistently low between features in MIMIC (FM = $[0.10 - 0.12]$), more rr measures were missing in HIRID (FM = 0.27) while BP was more absent in ICDEP (FM = 0.24 for bp_sys and bp_dia) compared to other features (see **Figure 3b**).

For the generation of CCD, resampled time series were reduced to blocks of at least three consecutive available values across all features. 14,286, 11,286, and 17,811 patients who complied with these conditions remained in MIMIC, HIRID, and ICDEP, respectively. Original patient characteristics were preserved (e.g., age of 64.1 ± 17.68 v.s. 63.88 ± 17.59 years in MIMIC)

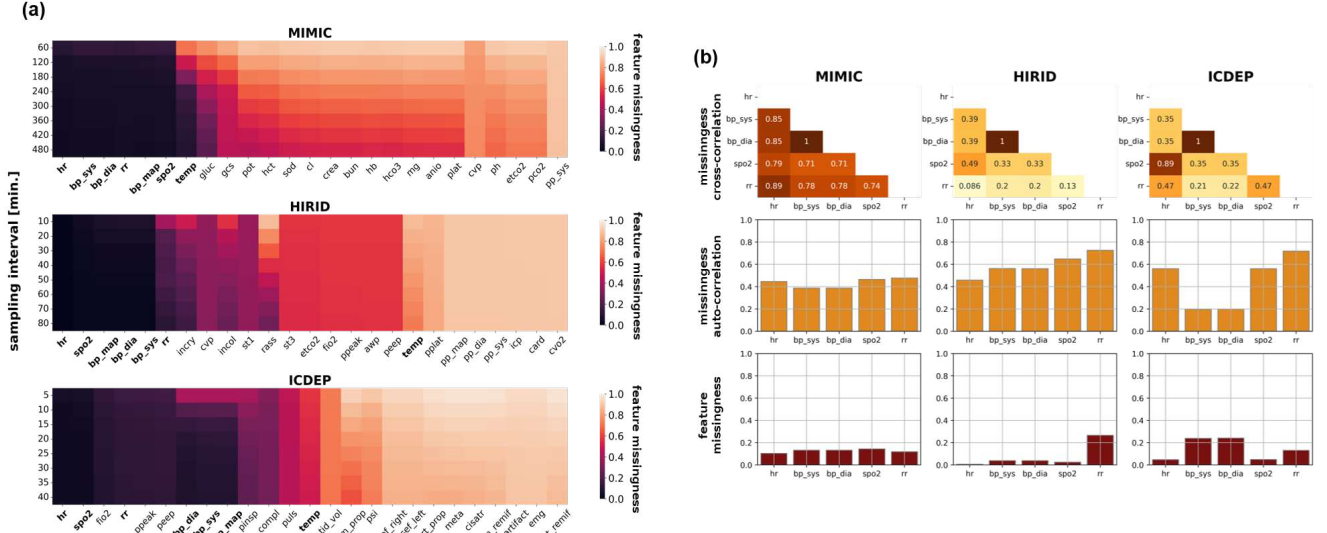


Figure 3. Feature missingness for varying sampling intervals (a) and missing properties for resampled datasets (b). Datasets were resampled with seven multiples of the original sampling interval of 60, 10, 5 min. as $\{120, 180, \dots, 480\}, \{20, 30, \dots, 80\}, \{10, 15, \dots, 40\}$ min. for MIMIC, ICDEP, and HIRID, respectively. Panel **a** contains top 25 features per dataset expressing the lowest feature missingness (FM) in the original data, sorted ascending (from left to right) by the mean FM across intervals. Features that are available in all three datasets are emphasized in bold. Panel **b** shows missingness properties in form of cross- (lag=0) and auto-correlation (lag=1), which correlate binary missing indicators R_n , as well as the FM. Properties are derived for the 60 min. (originally) sampled MIMIC, 20 min. resampled HIRID, and 10 min. resampled ICDEP data for which our simulation model learned spatio-temporal missingness.

while the lengths of sequences were clearly reduced (e.g. 250.75 ± 456.49 v.s. 129.35 ± 191.04 for HIRID) (see **Table 3** and **Figure A.2** in Supplement A).

To induce missingness in CCD, a Markov chain simulation model learned occurrence - and transition probabilities for the selected five features from original resampled data. The state space covered $2^5 = 32$ possible states where each state vector of length 5 indicated if one of each feature was present (0) or absent (1) at a single time step. For each state, binary indicators were concatenated in the order of hr, bp_sys, bp_dia, spo2, rr. We observed the actual occurrence of 28, 23, and 24 states for MIMIC, HIRID, and ICDEP, respectively (see **Figure 4a**). The TPS ($s = [0, 0, 0, 0, 0]$), encoding that all features were available, was most frequent ($f > 0.70$) across all original datasets while the isolated absence of rr measures ($s = [0, 0, 0, 0, 1]$) was occurring predominantly in HIRID ($f = 0.23$) than in other sets ($f < 0.07$). The simultaneous missingness of bp_dia and bp_sys alone ($s = [0, 1, 1, 0, 0]$) occurred more often in ICDEP ($f = 0.17$) than in MIMIC or HIRID ($f < 0.024$). The transition from one TPS to another TPS was most frequent for all datasets ($f > 0.55$). The transitions from the TPS to the simultaneous absence of bp_sys and bp_dia values and vice versa were about equally frequent in ICDEP ($f = 0.11, 0.12$) and occurred rarely in other datasets ($f < 0.018$). We found minor differences of occurrence and transition probabilities (frequencies) between original and simulated missingness (see **Figure 4a**). We further detected simulated missingness patterns that were similar to patterns presented by the original resampled data (see **Figure A.3, Table A.4** in **Supplementary A**).

When analyzing existing MAR effects of original ICU data from MIMIC and HIRID, we observed that clinical parameters were more absent for younger (age bin (18,51] years: FM= 0.15, 0.09 in MIMIC, HIRID) than for elderly patients (age bin (69,120] years: FM= 0.12, 0.07 in MIMIC, HIRID) across all features (see **Figure 4b**). In ICDEP, we detected a reversed effect (FM= 0.15, 0.13 for age bins (18,51], (69,120] years). After the multiplication of occurrence and transition probabilities with factors according to three equally sized age bins (see **Methods**), we could amplify these effects with our simulation model (age bin (18,51] years: FM= 0.15, 0.64 before and after MAR amplification in MIMIC for all features). We found no ad-hoc association of GT feature values and the simulated missingness by our Markov model that was also initially not conditioned on these relations. Thus, we simply assumed a positive relationship between GT value levels and their missingness to amplify MNAR effects (FM= 0.11, 0.60 for the value bin 0 before and after amplification) (see **Figure 4b**). We also show further implications of amplifications on simulated occurrence states (see **Figure A.4** in **Supplement A**).

	MIMIC		HIRID		ICDEP	
	Original data	CCD	Original data	CCD	Original data	CCD
#patients	34,472	14,286	36,098	11,286	41,832	17,811
sampling interval	1 h	1 h	10 min	20 min	5 min	10 min
sequence length	57.21±45.19	21.32±19.64	250.75±456.49	129.35±191.04	14.06±7.39	9.48±3.71
age [years]	64.1±17.68	63.88±17.59	63.52±15.20	64.24±14.74	53.45±18.47	54.25±18.11
In-hospital death	9.58%	10.44%	6.78%	9.98%	1.29%	1.32%
sex [F:female, M:male]	43.47% F 56.53% M	44.46% F 55.54% M	35.80% F 64.20% M	31.46% F 68.54% M	46.56% F 53.44% M	49.76% F 50.24% M

Table 3. Patient characteristics for original - and complete case data. The number of patients (#patients), the sampling interval, and further statistics are provided per dataset for the original data and the complete case data (CCD) after resampling and removing missingness. Data were cleaned so that original spatio-temporal aspects are preserved. Numerical statistics are reported as mean± standard-deviation.

	MIMIC			HIRID			ICDEP		
Name	$CCorE_{loc} \downarrow$	$NRMSE_{loc} \downarrow$	$NWSD_{glob} \downarrow$	$CCorE_{loc} \downarrow$	$NRMSE_{loc} \downarrow$	$NWSD_{glob} \downarrow$	$CCorE_{loc} \downarrow$	$NRMSE_{loc} \downarrow$	$NWSD_{glob} \downarrow$
Non Time-Causal Methods									
PMEAN	9.769±0.195	16.256±0.135	0.529±0.105	6.610±0.105	43.133±0.596	0.379±0.334	8.789±0.155	15.561±0.128	0.438±0.216
PMEDI	9.666±0.192	15.967±0.131	0.527±0.105	4.893±0.091	27.152±0.305	0.322±0.272	8.388±0.151	15.359±0.130	0.432±0.213
LINPOL	6.581±0.164	10.578±0.103	0.090±0.016	3.384±0.090	17.824±0.296	0.026±0.024	7.555±0.156	12.539±0.128	0.184±0.069
NEAPOL	7.383±0.175	12.280±0.116	0.106±0.019	3.596±0.093	20.892±0.337	0.021±0.019	8.108±0.156	14.244±0.135	0.215±0.078
QUAPOL	8.060±0.179	13.429±0.132	0.096±0.013	4.726±0.097	30.984±0.567	0.218±0.246	8.012±0.157	13.951±0.135	0.197±0.072
MICTRE	12.738±0.211	23.687±0.184	0.014±0.003	8.288±0.115	94.193±1.828	0.008±0.007	12.211±0.184	22.102±0.182	0.017±0.005
MICLRE	11.965±0.203	23.408±0.184	0.015±0.003	6.597±0.104	55.903±0.933	0.008±0.006	10.795±0.172	21.942±0.173	0.014±0.003
MICKNN	10.048±0.198	16.641±0.135	0.471±0.104	6.060±0.103	47.676±0.777	0.145±0.115	9.484±0.170	16.223±0.135	0.301±0.143
SAE	9.858±0.198	16.251±0.135	0.529±0.105	6.597±0.105	43.020±0.595	0.379±0.333	8.769±0.154	15.578±0.128	0.438±0.216
MIDA	9.846±0.197	16.265±0.135	0.518±0.103	6.662±0.107	43.127±0.596	0.378±0.333	9.250±0.164	15.538±0.128	0.436±0.215
BILSTAE	6.581±0.159	10.840±0.097	0.234±0.041	3.348±0.091	18.685±0.364	0.337±0.385	8.166±0.156	13.040±0.119	0.281±0.132
Time-Causal Methods									
IMEAN	9.852±0.195	16.695±0.139	0.541±0.112	4.719±0.092	26.046±0.276	0.333±0.281	9.064±0.153	17.423±0.133	0.491±0.242
IMEDI	9.696±0.192	16.333±0.134	0.535±0.108	4.618±0.093	25.769±0.277	0.338±0.285	8.735±0.149	16.783±0.131	0.472±0.228
RMEAN	6.663±0.162	11.798±0.104	0.131±0.023	3.370±0.079	19.542±0.233	0.116±0.112	7.945±0.152	14.870±0.134	0.210±0.075
RMEDI	6.692±0.162	11.873±0.107	0.117±0.020	3.617±0.082	20.406±0.256	0.141±0.144	7.970±0.153	14.815±0.136	0.196±0.067
LOCF	7.041±0.164	11.862±0.114	0.062±0.012	3.164±0.081	17.280±0.242	0.010±0.008	7.764±0.154	14.247±0.139	0.161±0.050
LINEXT	9.390±0.181	17.145±0.173	0.131±0.023	4.188±0.090	25.686±0.399	0.137±0.148	8.609±0.157	18.455±0.180	0.182±0.066
QUAEXT	13.658±0.223	29.242±0.331	0.459±0.092	5.870±0.101	46.389±0.796	0.393±0.423	9.898±0.164	25.231±0.288	0.324±0.163
LSTAE	7.107±0.168	11.775±0.105	0.265±0.045	4.588±0.096	23.791±0.317	0.177±0.168	8.553±0.161	13.877±0.122	0.319±0.146
STGAE	9.843±0.194	16.544±0.138	0.454±0.096	5.892±0.100	35.975±0.471	0.452±0.439	9.125±0.159	16.225±0.135	0.417±0.209
Method Fit with Static Features									
MICTRE	12.738±0.211	23.687±0.184	0.014±0.003	8.288±0.115	94.193±1.828	0.008±0.007	12.211±0.184	22.102±0.182	0.017±0.005
MICLRE	11.854±0.202	23.110±0.180	0.016±0.004	6.531±0.102	55.026±0.908	0.007±0.005	10.672±0.174	21.553±0.172	0.019±0.006
MICKNN	10.629±0.208	17.368±0.144	0.014±0.003	6.055±0.104	47.294±0.770	0.144±0.116	9.429±0.169	16.183±0.135	0.282±0.150
SAE	9.843±0.197	16.248±0.135	0.529±0.105	6.670±0.107	43.110±0.596	0.382±0.338	8.756±0.154	15.545±0.128	0.437±0.215
MIDA	9.833±0.196	16.263±0.135	0.522±0.104	6.659±0.106	43.817±0.609	0.377±0.332	9.220±0.163	15.563±0.128	0.432±0.213
BILSTAE	6.442±0.156	10.765±0.096	0.216±0.036	4.548±0.094	24.204±0.335	0.183±0.175	8.683±0.160	13.862±0.124	0.298±0.150
LSTAE	7.278±0.170	12.010±0.106	0.269±0.047	4.783±0.096	25.560±0.338	0.224±0.214	8.302±0.156	13.699±0.121	0.310±0.146

Table 4. Performance of 20 different imputation methods on spatio-temporal missingness. Averaging methods, including imputation with a population mean (PMEAN), linear (or additional) inter- and extrapolations (LINPOL/LINEXT), multiple imputations by chained equations (MICE), and more complex denoising autoencoder (LSTAE) were applied with “no amplification” (see Table 1). Methods were either non time-causal, consuming time points from the future, or time-causal, being limited to previous measurements only. As a separate analysis, we investigated the change in imputation performance when fitting methods with static attributes like age. All spatio-temporal autoencoder such as (BI)LSTAE or STGAE were trained on LOCF pre-imputed GT values. Cross-correlation errors ($CCorE_{loc}$) and the normalized root mean squared error ($NRMSE_{loc}$) were calculated per time series (locally), the normalized Wasserstein distance ($NWSD_{glob}$) was calculated across all measurements and time series (globally). We report all errors in % with mean±95%CI across time series, features, and validation folds. Best results, according to smallest metrics, are marked in bold.

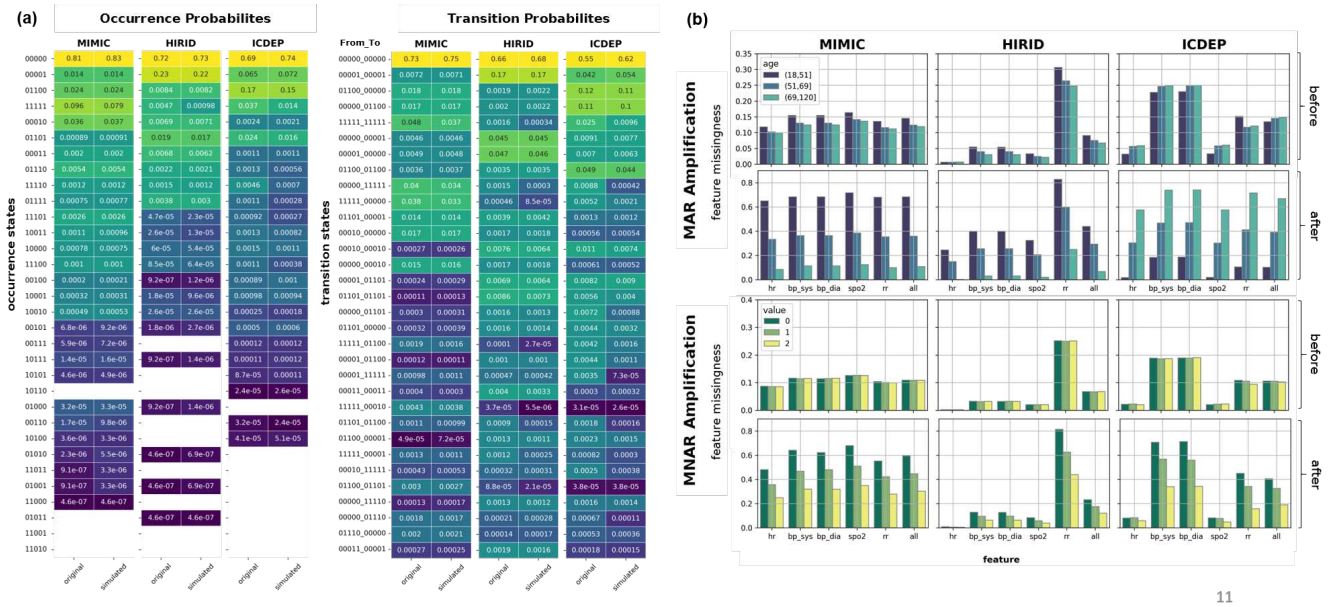


Figure 4. Occurrence and transition probabilities learned by a Markov chain simulation model (a), amplification of missing at random and missing not at random effects (b). In panel **a**, we display the occurrence probabilities of 32 states encoding if one of the five features are missing (1) or observed (0) at one time step. In addition, transition probabilities are displayed from one state to a consecutive one (as From_To) for the top 32 most frequent transitions. Probabilities are displayed for the original resampled data (original) and the data with simulated missingness (simulated) to visualize the quality of our simulation model. In panel **b**, the feature missingness (per age or value bin) before and after missing at random (MAR) or missing not at random (MNAR) amplification is shown for each feature as well as across all features. In MAR, missingness depends on observed data (here age). In a MNAR setting, missingness depends on unobserved data (here the GT values of vital signs which are not known after missing induction). Equally sized age bins are used in the MAR setup. Value edges for bins (0,1,2 in MNAR setting) are included in **Table A.3** in Supplement A.

After simulating spatio-temporal missingness on CCD, we applied 20 imputation methods comprising averaging techniques like running median (RMEDI), interpolation methods like linear interpolation (LINPOL), extrapolation methods like LOCF, multiple imputations by chained equations (e.g., MICLRE using linear estimations), and DL architectures like (Bi)LSTM autoencoder (see **Methods**). We initially conducted our analysis with "no amplification" for MAR and MNAR. Techniques were categorized into time-causal and non time-causal methods, with the latter having the advantage of accessing future data points (see **Table 4**).

Among time-causal methods, LOCF imputations performed best across all datasets and yielded low error margins in HIRID ($CCorE_{loc} = 3.164 \pm 0.081$, $NRMSE_{loc} = 17.280 \pm 0.242$, $NWSD_{glob} = 0.010 \pm 0.008$) compared to other imputation methods (see **Table 4**). The autoencoder LSTAE resulted in smallest NRMSE for MIMIC ($NRMSE_{loc} = 11.775 \pm 0.105$) as well as ICDEP ($NRMSE_{loc} = 13.877 \pm 0.122$). Running mean imputation (RMEAN) was particularly preserving cross-correlations in MIMIC ($CCorE_{loc} = 6.663 \pm 0.162$). Non time-causal models yielded overall better performances than time-causal ones, LOCF and linear imputation resulted in overall small errors across additional performance metrics (see **Table A.5** in **Supplement A**).

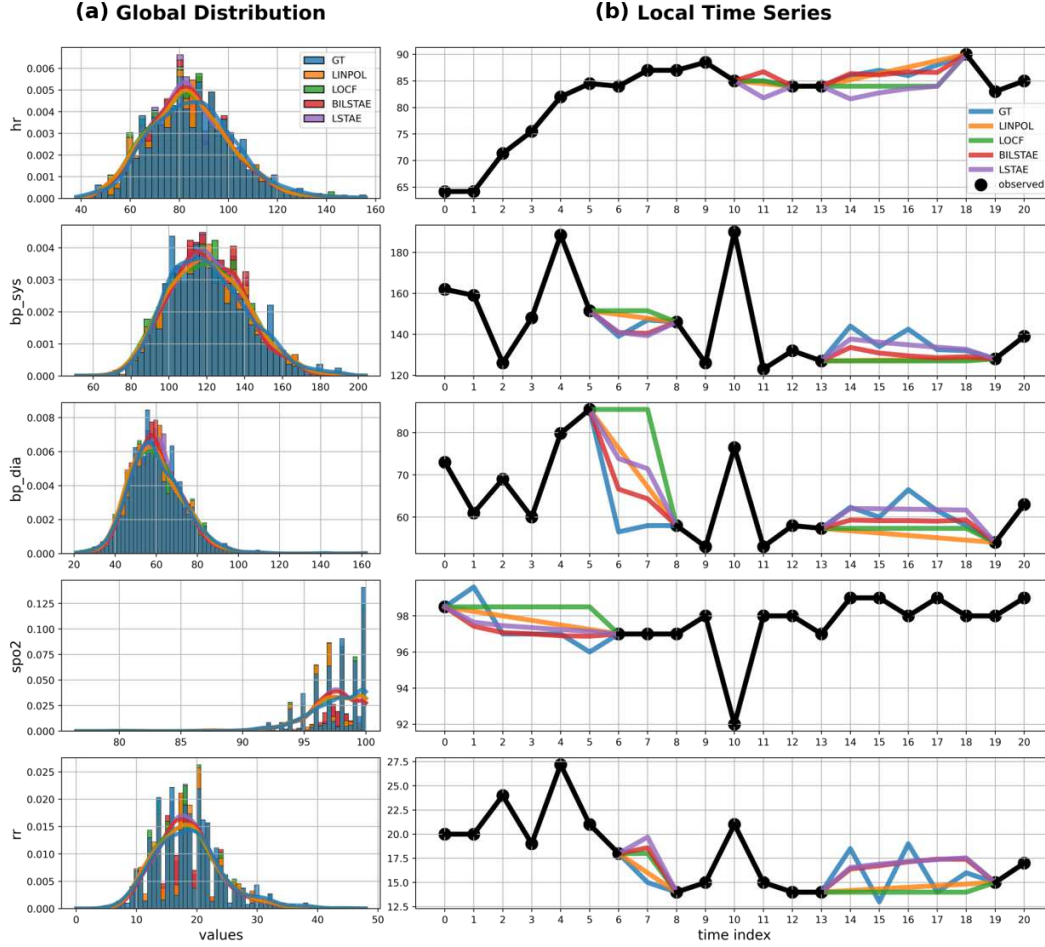


Figure 5. Global distributions (a) and one selected local time series (b) comparing four imputation methods per feature. We display imputation results from linear interpolation (LINPOL), last observation carried forward (LOCF), bidirectional (BILSTM) and unidirectional (LSTAE) autoencoder with LSTM modules compared to the ground truth (GT) data for MIMIC. In panel **a**, the global distribution of feature values across all patient time series is shown. Panel **b** shows an example of one local time series for one patient where observed values that did not require imputation are marked as black dots.

imputations ($NWSD_{glob}=0.007\pm0.005$). Further ablations of trained STAE are integrated in the **Supplement A** (see **Table A.6** and **Additional Results A.2**). Here, we note that performances varied across ablation parameters not highlighting a single one that boosted performances across all datasets.

We visualize results from LINPOL and LOCF as strong baseline models, alongside LSTAE and BILSTAE architectures, as these performed well across all datasets (see **Figure 5**). The distribution of imputed heart rate values was shifted to the left (median: 83) for all four methods in comparison to GT data (median: 91) in MIMIC. For bp_sys, BILSTAE introduced two peaks around 112 and 133 mmHg in contrast to the GT distribution expressing one single peak at 119 mmHg (see **Figure 5a**). For one example patient time series, BILSTAE and LSTAE improved temporal signals of bp_sys between 5 and 8 hours as well as between 13 and 19 hours for bp_dia and rr outperforming baselines (see **Figure 5b**). BILSTAE alone could capture the temporal spike between 5 and 8 hours for bp_dia closest to GT data. For other occasions (e.g., between 6 and 8 hours for rr), simple baselines like LINPOL yielded better estimates than autoencoders which over- or underestimated temporal trends. Here, though not exclusively, we found differences in distributions and value errors between baselines and autoencoders explaining divergences in previously reported metrics.

We further fitted imputation methods on data with amplified MAR effects (see **Table A.7** in **Supplementary A**). Similar to simulated missingness without amplification, LINPOL imputations performed best for most of the performance metrics across all datasets (e.g., $CCorD_{loc}=15.088\pm0.272$, $NRMSE_{loc}=11.727\pm0.089$ for MIMIC) outperforming BILSTAE (e.g., $CCorD_{loc}=18.293\pm0.293$, $NRMSE_{loc}=14.108\pm0.098$ for MIMIC). Unlike previous results, running mean or median techniques resulted in a majority of performance scores superior to LOCF but presented overlapping confidence intervals (e.g., $NRMSE_{loc}=$

17.201 $\pm$ 0.172, 18.485 $\pm$ 0.264 for RMEAN, LOCF in HIRID). All autoencoders (LSTAE and BILSTAE) that were trained with static features could improve their stand-alone version (without these attributes) across performance metrics for amplified MAR effects (e.g., $NRMSE_{loc}$ = 13.058 $\pm$ 0.087, 13.422 $\pm$ 0.088 for LSTAE_STAT, LSTAE in ICDEP) (see **Figure A.5** and **Table A.7** in **Supplementary A**). For amplified MNAR effects, LINPOL and LOCF showed overall best imputation performance in MIMIC and HIRID (see **Table A.8** in **Supplementary A**). Autoencoders outperformed these corresponding baselines in ICDEP (e.g., $NRMSE_{loc}$ = 3.944 $\pm$ 0.070, 4.131 $\pm$ 0.077 for LINPOL, BILSTAE) for most of the performance metrics.

After observing decreased imputation performances of autoencoders under our simulated missingness patterns, we conducted additional analyses on missingness restricted to the MCAR assumption. These data did not preserve the spatio-temporal dependencies explicitly simulated by our proposed Markov model. Following prior work, we generated random subsets of CCD by inducing missingness based on original FM rates (see **Table A.4** in **Supplement A**). In contrast to all previous results, autoencoder architectures clearly outperformed simpler models across datasets and performance metrics on this kind of simulated missingness (e.g., $NMAE_{loc}$ = 1.153 $\pm$ 0.033, 3.179 $\pm$ 0.077, 1.951 $\pm$ 0.030 for BILSTAE and $NMAE_{loc}$ = 2.133 $\pm$ 0.048, 4.506 $\pm$ 0.118, 5.802 $\pm$ 0.064 for LINPOL in MIMIC, HIRID, ICDEP). Results indicate that our Markov model created a more demanding imputation benchmark for DL imputation methods than previous work.

Downstream Performance

We investigated the influence of imputation performances on downstream classification tasks across amplification scenarios. For our transformer-based downstream prediction model (TRAN), which requires fixed-length sequences, we used the first 12 hours after ICU admission for MIMIC and HIRID patients and the first 1.5 hours for the perioperative ICDEP cohort, leading to the inclusion of 9233, 8696, and 16020 patient time series, respectively. Prevalence rates were at 8.71% and 9.73% for in-hospital death in MIMIC and HIRID, 9.23% patients suffered from POD in the selected ICDEP cohort.

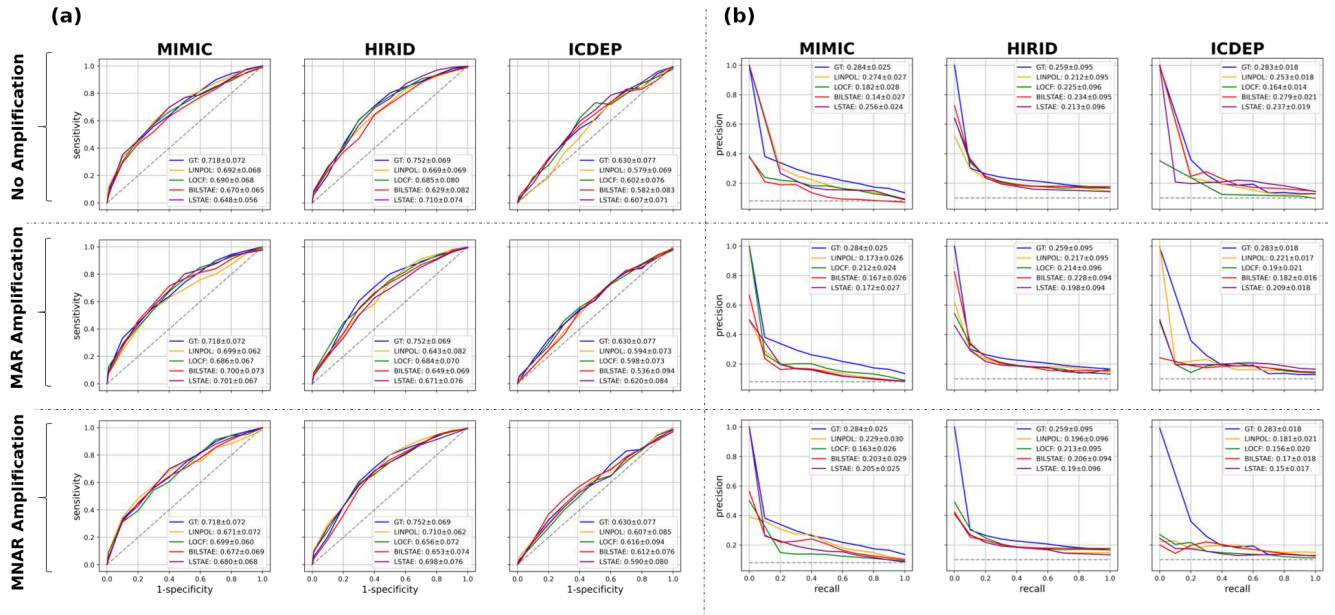


Figure 6. Area under the receiver operating characteristics - (a) and area under the precision-recall (b) curves across amplification scenarios. In panel a, we display the area under the receiver operating characteristics curve (AUROC) for ground truth (GT) and imputed datasets. Linear interpolation (LINPOL), last observation carried forward (LOCF), LSTM autoencoder (LSTAE), and bidirectional LSTM autoencoder (BILSTM) were applied for imputations. In panel b, we draw the area under the precision-recall curve (AUPRC). All results are gained from a downstream transformer model classifying in-hospital death (MIMIC, HIRID) or postoperative delirium (ICDEP).

We trained our downstream model on GT data in addition to imputed spatio-temporal missingness patterns by LINPOL, LOCF, and more complex autoencoders BILSTAE, LSTAE. We evaluated downstream classification performances by drawing receiver-operating characteristic curves and precision-recall curves (see **Figure 6**). Performance evaluations were best for TRAN trained on GT data across all datasets, amplification scenarios, showing elevated AUROC as well as AUPRC metrics. Overall, the prediction of in-hospital mortality (e.g., AUROC = 0.752 $\pm$ 0.069 for GT in HIRID) was more easy than the postoperative delirium task (AUROC = 0.630 $\pm$ 0.077 in ICDEP). We observed minor differences in receiver-operating curves across imputation

models and amplification scenarios, with no single method showing a clearly superior performance. Differences were more pronounced for precision-recall curves (e.g., AUPRC= 0.274 ± 0.027 , 0.140 ± 0.027 for LINPOL, BILSTAE in MIMIC with "no amplification"). In MIMIC, downstream models which used LINPOL imputed data expressed higher AUPRC scores than models trained on BILSTAE data. In HIRID, LOCF imputations resulted in better downstream AUPRC metrics than LSTAE (e.g., AUPRC= 0.225 ± 0.096 , 0.213 ± 0.096 for LOCF, LSTAE in HIRID with "no amplification"). Detected trends were widely dispersed across datasets. Most imputation approaches yielded comparable results in downstream classification tasks, suggesting no clear advantage for one method over the others.

Discussion

Here, we present results from a novel simulation model that induces realistic spatio-temporal missingness patterns in three real-world clinical longitudinal datasets by sampling from Markov chains, with transition probabilities estimated from real data. Enhancements of simulations led to the amplification of detected MAR effects between missingness and age as well as for the artificial induction of MNAR effects between a feature's missingness with its value levels. The application of 20 imputation methods and their variations showed that commonly used approaches like linear imputations, applicable for study data, or LOCF, eligible for clinical prediction settings, yield good performances on spatio-temporal missingness patterns. MICE imputations estimated global distributions of vital signs within our study population closest to GT data. When benchmarking imputation performances on simpler MCAR patterns without spatio-temporal aspects, the superior performance of simpler methods diminishes for the benefit of DL models. Finally, we trained a downstream classifier on a variety of datasets where spatio-temporal missingness was imputed by different methods. Classification metrics varied across these datasets while imputation methods expressed similarly good downstream performances. High performances of classifying clinical outcomes highlight the potentials in temporal vital sign analysis.

Akin to Jäger et al.³³, we developed an open accessible benchmarking suite while evaluating temporal imputation methods on clinical time series instead of tabular data, similar to Bertsimas et al.¹². For baseline comparisons, we replicated previous studies that limited missing simulations on MCAR scenarios with varying missingness rates^{15, 19, 23, 27, 29, 31}. In contrast, our proposed Markov chain model further reflects original spatio-temporal patterns and is capable of simulating MAR or MNAR scenarios that condition missingness on demographics or the underlying value levels^{18, 21, 51}. Li et al.⁶⁶ and Tan et al.⁶⁷ detected MAR patterns of laboratory values related to illness severity or hospital sites. In our data, age was positively correlated with the fraction of overall missing values per feature. We extensively describe original missingness patterns, such as the simultaneous absence or presence of systolic and diastolic BP values, and show key differences between real-world clinical datasets. While some of these patterns have been analyzed in MIMIC^{15, 26, 35, 68}, we are not aware of any work that built upon HIRID or an intraoperative dataset like ICDEP. Marginal differences in descriptive statistics between original and simulated spatio-temporal patterns underscores the quality of our simulation model.

Lai et al.²⁹ and Qian et al.²⁴ presented DL autoencoders that improved temporal imputations of clinical time series over standard statistics. Hence, we hypothesized that trained spatio-temporal autoencoder could outperform simpler methods like linear imputation. Inspired by previous studies^{22, 24, 29, 68}, we designed a spatio-temporal autoencoder architecture using (Bi)LSTM modules. Surprisingly, simple approaches such as linear imputation, LOCF, and MICE, widely used by medical studies^{11, 15, 19}, outperformed more complex autoencoders in most scenarios. MICE could always reconstruct value distributions across all patient time series closest to the known distributions, similar claims have been made before^{51, 69}. While selected studies claim superior imputation performance of proposed DL methods^{18, 19, 24}, we found more generalizable studies that covered a larger variety of datasets highlighted the benefit of simpler linear imputation models over DL^{19, 33, 51}.

Additional benchmarks which we restricted to MCAR suggest that previous studies^{22, 24, 24, 29} may have overestimated the performance of DL models. We argue that our model introduces more realistic missingness patterns in clinical time series (e.g., simultaneous absence of bp_sys and bp_dia). In previous studies, DL models may infer missing values more easily without these patterns (e.g., missing bp_sys from observed bp_dia), especially since clinical data expressed strong spatial correlation (Pearson correlation for bp_sys and bp_dia of 0.5, 0.6, 0.7 in MIMIC, HIRID, ICDEP). In our study, deep STAE learned spatio-temporal characteristics of clinical time series when the missingness itself lacks a spatio-temporal pattern.

Findings suggest the usage of linear interpolation for retrospective clinical studies since clinical values were reconstructed sufficiently across datasets and missingness scenarios. If smaller errors from one patient time series can account for larger errors from other patients, the popular method MICE reconstructs global distributions of clinical vital signs⁷⁰. LOCF demonstrated robust imputation performance and remains widely used in clinical ML studies, despite ongoing criticism for its potential to introduce bias and underestimate variability^{16, 71}. Our findings suggest that using LOCF remains a reasonable approach for imputations prior to prediction tasks. When designing more complex autoencoder architectures, we recommend incorporating static attributes, such as age, as they improved model performances, particularly under more complex MAR scenarios.

A variety of related studies have evaluated downstream classification tasks, such as ours, predicting in-hospital death^{12, 21, 24}, heart failure^{21, 23}, or kidney disease¹⁸. Psychogios et al. found that DL imputed data improved the prediction of coronary heart

disease²¹ while KNN or MICE performed better for small missingness rates. Zamanzadeh et al. showed that the accuracy of kidney disease classification varies a lot across prior imputed sets independent from their imputation evaluation¹⁸. Luo et al. trained autoencoders which yielded best imputation as well as best in-hospital death classification results¹⁵. Qian et al. argued that results from downstream classifiers are decoupled from prior imputation performances²⁴. In our study, we also found such decoupled behavior so that some BiLSTM autoencoders in HIRID improved the downstream performance while imputing time series with more deficits. In this study, we see the downstream performances of simpler methods like LOCF as sufficient.

Architectures like GRU-D or BRITS have been designed to learn both, downstream- and imputation tasks simultaneously^{6,35,72,73}. We did not apply these models but rather handled these two aspects as independent problems (e.g., no restriction to ML setting). A majority of such proposed architectures are not capable of processing sequences of unequal lengths. We see an extension of models like BRITS⁷³ or attention networks²⁹ to meet this requirement as potential future work. Our study is further limited to five vital signs that are commonly used for clinical analysis. Prior work evaluated imputation methods on laboratory values⁶⁶ which we did not consider here. We described selection biases due to preprocessing for original, complete case, and imputed sets. Constraints for creating such analyses with complete data, however, introduced some shifts in demographics or correlations between features. We also cannot trace simulated missingness back to individual patients in original sets, potentially affecting clinical data analysis which is applied on our missing data. The inclusion of three different datasets increased the generalizability of our results, but other datasets might result in the detection of different spatio-temporal patterns. We foster replicability and transparency of our results by making models and processed data publicly available⁶³.

Conclusion

We present a novel Markov Chain simulation model for successfully inducing spatio-temporal missingness patterns in clinical time series. We evaluate commonly used linear imputation methods as an overall robust way to imputed unobserved values in longitudinal clinical data expressing sufficient downstream performances. Previous studies that found superior performance of deep learning methods lack in spatio-temporal aspects which were simulated by our proposed model. We provide an open source benchmarking suite that enables future ML projects to increase their robustness and transparency in the presence of different realistic missingness scenarios.

References

1. Sauer, C. M. *et al.* Leveraging electronic health records for data science: common pitfalls and how to avoid them. **4**, e893–e898, DOI: [10.1016/S2589-7500\(22\)00154-6](https://doi.org/10.1016/S2589-7500(22)00154-6).
2. Jetley, G. & Zhang, H. Electronic health records in IS research: Quality issues, essential thresholds and remedial actions. **126**, 113137, DOI: [10.1016/j.dss.2019.113137](https://doi.org/10.1016/j.dss.2019.113137).
3. Honeyford, K. *et al.* Challenges and recommendations for high quality research using electronic health records. **4**, 940330, DOI: [10.3389/fdgth.2022.940330](https://doi.org/10.3389/fdgth.2022.940330).
4. Yang, S., Zhu, F., Ling, X., Liu, Q. & Zhao, P. Intelligent health care: Applications of deep learning in computational medicine. **12**, 607471, DOI: [10.3389/fgene.2021.607471](https://doi.org/10.3389/fgene.2021.607471).
5. Chai, X. *et al.* Deep learning for irregularly and regularly missing data reconstruction. **10**, 3302, DOI: [10.1038/s41598-020-59801-x](https://doi.org/10.1038/s41598-020-59801-x).
6. Ruan, X., Wang, L., Thongprayoon, C., Cheungpasitporn, W. & Liu, H. GRU-d-weibull: A novel real-time individualized endpoint prediction. **146**, 102696, DOI: [10.1016/j.artmed.2023.102696](https://doi.org/10.1016/j.artmed.2023.102696).
7. Rubin, D. B. Inference and missing data. **63**, 581–592, DOI: [10.1093/biomet/63.3.581](https://doi.org/10.1093/biomet/63.3.581).
8. Afshar, A. S. *et al.* An exploratory data quality analysis of time series physiologic signals using a large-scale intensive care unit database. **4**, ooab057, DOI: [10.1093/jamiaopen/ooab057](https://doi.org/10.1093/jamiaopen/ooab057).
9. Khan, M. *et al.* Analysing the effects of cold, normal, and warm digits on transmittance pulse oximetry. **26**, 34–41, DOI: [10.1016/j.bspc.2015.12.006](https://doi.org/10.1016/j.bspc.2015.12.006).
10. Heymans, M. W. & Twisk, J. W. Handling missing data in clinical research. **151**, 185–188, DOI: [10.1016/j.jclinepi.2022.08.016](https://doi.org/10.1016/j.jclinepi.2022.08.016).
11. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: what is it and how does it work? **20**, 40–49, DOI: [10.1002/mp.329](https://doi.org/10.1002/mp.329).
12. Bertsimas, D., Orfanoudaki, A. & Pawlowski, C. Imputation of clinical covariates in time series. **110**, 185–248, DOI: [10.1007/s10994-020-05923-2](https://doi.org/10.1007/s10994-020-05923-2).

13. Ma, Q., Lee, W.-C., Fu, T.-Y., Gu, Y. & Yu, G. MIDIA: exploring denoising autoencoders for missing data imputation. **34**, 1859–1897, DOI: [10.1007/s10618-020-00706-8](https://doi.org/10.1007/s10618-020-00706-8).
14. Giesa, N. *et al.* Applying a transformer architecture to intraoperative temporal dynamics improves the prediction of postoperative delirium. **4**, 251, DOI: [10.1038/s43856-024-00681-x](https://doi.org/10.1038/s43856-024-00681-x).
15. Luo, Y. Evaluating the state of the art in missing data imputation for clinical data. **23**, bbab489, DOI: [10.1093/bib/bbab489](https://doi.org/10.1093/bib/bbab489).
16. Lachin, J. M. Fallacies of last observation carried forward analyses. **13**, 161–168, DOI: [10.1177/1740774515602688](https://doi.org/10.1177/1740774515602688).
17. Cesare, N. & Were, L. P. O. A multi-step approach to managing missing data in time and patient variant electronic health records. **15**, 64, DOI: [10.1186/s13104-022-05911-w](https://doi.org/10.1186/s13104-022-05911-w).
18. Zamanzadeh, D. J. *et al.* Autopopulus: A novel framework for autoencoder imputation on large clinical datasets. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2303–2309, DOI: [10.1109/EMBC46164.2021.9630135](https://doi.org/10.1109/EMBC46164.2021.9630135) (IEEE).
19. Luo, Y., Cai, X., Zhang, Y., Xu, J. & others. Multivariate time series imputation with generative adversarial networks. **31**.
20. Jia, Y., Zhou, C. & Motani, M. Spatio-temporal autoencoder for feature learning in patient data with missing observations. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 886–890, DOI: [10.1109/BIBM.2017.8217773](https://doi.org/10.1109/BIBM.2017.8217773) (IEEE).
21. Psychogyios, K., Ilias, L., Ntanos, C. & Askounis, D. Missing value imputation methods for electronic health records. **11**, 21562–21574, DOI: [10.1109/ACCESS.2023.3251919](https://doi.org/10.1109/ACCESS.2023.3251919).
22. Song, M. *et al.* An autoencoder-based deep learning method for genotype imputation. **5**, 1028978, DOI: [10.3389/frai.2022.1028978](https://doi.org/10.3389/frai.2022.1028978).
23. Bashir, F. & Wei, H.-L. Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm. **276**, 23–30, DOI: [10.1016/j.neucom.2017.03.097](https://doi.org/10.1016/j.neucom.2017.03.097).
24. Qian, L. *et al.* Beyond random missingness: Clinically rethinking for healthcare time series imputation, DOI: [10.48550/arXiv.2405.17508](https://doi.org/10.48550/arXiv.2405.17508). 2405.17508[cs].
25. Xu, D., Hu, P. J.-H., Huang, T.-S., Fang, X. & Hsu, C.-C. A deep learning-based, unsupervised method to impute missing values in electronic health records for improved patient management. **111**, 103576, DOI: [10.1016/j.jbi.2020.103576](https://doi.org/10.1016/j.jbi.2020.103576).
26. Bignoumba, N., Mellouli, N. & Yahia, S. B. A new efficient ALignment-driven neural network for mortality prediction from irregular multivariate time series data. **238**, 122148, DOI: [10.1016/j.eswa.2023.122148](https://doi.org/10.1016/j.eswa.2023.122148).
27. Huang, L., Huang, J., Li, H. & Cui, J. Robust spatial temporal imputation based on spatio-temporal generative adversarial nets. **279**, 110919, DOI: [10.1016/j.knosys.2023.110919](https://doi.org/10.1016/j.knosys.2023.110919).
28. Luo, Y., Szolovits, P., Dighe, A. S. & Baron, J. M. 3d-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. **25**, 645–653, DOI: [10.1093/jamia/ocx133](https://doi.org/10.1093/jamia/ocx133).
29. Lai, X., Yao, Y., Mu, J., Lu, W. & Zhang, L. Ratai: recurrent autoencoder with imputation units and temporal attention for multivariate time series imputation. **58**, 37, DOI: [10.1007/s10462-024-11039-z](https://doi.org/10.1007/s10462-024-11039-z).
30. Li, D., Li, L., Li, X., Ke, Z. & Hu, Q. Smoothed LSTM-AE: A spatio-temporal deep model for multiple time-series missing imputation. **411**, 351–363, DOI: [10.1016/j.neucom.2020.05.033](https://doi.org/10.1016/j.neucom.2020.05.033).
31. Li, Y., Du, M. & He, S. Attention-based sequence-to-sequence model for time series imputation. **24**, 1798, DOI: [10.3390/e24121798](https://doi.org/10.3390/e24121798).
32. Fan, Y. *et al.* Spatio-temporal denoising graph autoencoders with data augmentation for photovoltaic data imputation. **1**, 1–19, DOI: [10.1145/3588730](https://doi.org/10.1145/3588730).
33. Jäger, S., Allhorn, A. & Bießmann, F. A benchmark for data imputation methods. **4**, 693674, DOI: [10.3389/fdata.2021.693674](https://doi.org/10.3389/fdata.2021.693674).
34. Song, X. *et al.* Longitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine: retrospective cohort study. **8**, e15510. Publisher: JMIR Publications Inc., Toronto, Canada.
35. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. **8**, 6085, DOI: [10.1038/s41598-018-24271-9](https://doi.org/10.1038/s41598-018-24271-9).
36. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. **3**, 160035, DOI: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35).

37. Wang, S. *et al.* MIMIC-extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 222–235, DOI: [10.1145/3368555.3384469.1907.08322\[cs,stat\]](https://doi.org/10.1145/3368555.3384469.1907.08322[cs,stat]).
38. Moody, G., Mark, R. & Goldberger, A. PhysioNet: a web-based resource for the study of physiologic signals. **20**, 70–75, DOI: [10.1109/51.932728](https://doi.org/10.1109/51.932728).
39. Johnson, A. MIMIC-III clinical database.
40. Faltys, M. *et al.* HiRID, a high time-resolution ICU dataset, DOI: [10.13026/NKWC-JS72](https://doi.org/10.13026/NKWC-JS72). Version Number: 1.1.1.
41. Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. **26**, 364–373, DOI: [10.1038/s41591-020-0789-4](https://doi.org/10.1038/s41591-020-0789-4).
42. Dvorak, J., Sutter, M., Eggspuehler, A., Szpalski, M. & Aebi, M. Multimodal intraoperative monitoring: towards a routine use in surgical treatment of severe spinal disorders. **16**, 113–114, DOI: [10.1007/s00586-007-0415-x](https://doi.org/10.1007/s00586-007-0415-x).
43. Giesa, N. *et al.* Predicting postoperative delirium assessed by the nursing screening delirium scale in the recovery room for non-cardiac surgeries without craniotomy: A retrospective study using a machine learning approach. *PLOS Digit. Heal.* **3**, e0000414 (2024).
44. Boie, S. D. *et al.* A scalable approach for critical care data extraction and analysis in an academic medical center. *Int. J. Med. Informatics* **192**, 105611 (2024).
45. Hinrichs, N. *et al.* Short-term vital parameter forecasting in the intensive care unit: A benchmark study leveraging data from patients after cardiothoracic surgery. **3**, e0000598, DOI: [10.1371/journal.pdig.0000598](https://doi.org/10.1371/journal.pdig.0000598).
46. Frydenberg, M. The chain graph markov property. **17**, 333–353. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].
47. Rabiner, L. & Juang, B. An introduction to hidden markov models. **3**, 4–16, DOI: [10.1109/MASSP.1986.1165342](https://doi.org/10.1109/MASSP.1986.1165342).
48. Browne, M. W. Cross-validation methods. **44**, 108–132, DOI: [10.1006/jmps.1999.1279](https://doi.org/10.1006/jmps.1999.1279).
49. Wahab, M. A. Interpolation and extrapolation. **17**, 1–6.
50. Safi, A. A., Beyer, C., Unnikrishnan, V. & Spiliopoulou, M. Multivariate time series as images: Imputation using convolutional denoising autoencoder. In Berthold, M. R., Feelders, A. & Kreml, G. (eds.) *Advances in Intelligent Data Analysis XVIII*, vol. 12080, 1–13, DOI: [10.1007/978-3-030-44584-3_1](https://doi.org/10.1007/978-3-030-44584-3_1) (Springer International Publishing). Series Title: Lecture Notes in Computer Science.
51. Costa, A. F., Santos, M. S., Soares, J. P. & Abreu, P. H. Missing data imputation via denoising autoencoders: The untold story. In Duivesteijn, W., Siebes, A. & Ukkonen, A. (eds.) *Advances in Intelligent Data Analysis XVII*, vol. 11191, 87–98, DOI: [10.1007/978-3-030-01768-2_8](https://doi.org/10.1007/978-3-030-01768-2_8) (Springer International Publishing). Series Title: Lecture Notes in Computer Science.
52. Dubey, S. R., Singh, S. K. & Chaudhuri, B. B. Activation functions in deep learning: A comprehensive survey and benchmark. **503**, 92–108, DOI: [10.1016/j.neucom.2022.06.111](https://doi.org/10.1016/j.neucom.2022.06.111).
53. Hermans, J. R., Spanakis, G. & Möckel, R. Accumulated gradient normalization. In Zhang, M.-L. & Noh, Y.-K. (eds.) *Proceedings of the Ninth Asian Conference on Machine Learning*, vol. 77 of *Proceedings of Machine Learning Research*, 439–454 (PMLR).
54. Nguyen, V. A., Shafieezadeh-Abadeh, S., Kuhn, D. & Mohajerin Esfahani, P. Bridging bayesian and minimax mean square error estimation via wasserstein distributionally robust optimization. **48**, 1–37, DOI: [10.1287/moor.2021.1176](https://doi.org/10.1287/moor.2021.1176).
55. Panaretos, V. M. & Zemel, Y. Statistical aspects of wasserstein distances. **6**, 405–431, DOI: [10.1146/annurev-statistics-030718-104938](https://doi.org/10.1146/annurev-statistics-030718-104938).
56. Patrini, G. *et al.* Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, 733–743 (PMLR).
57. Shekar, B. H. & Dagneu, G. Grid search-based hyperparameter tuning and classification of microarray cancer data. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, 1–8, DOI: [10.1109/ICACCP.2019.8882943](https://doi.org/10.1109/ICACCP.2019.8882943) (IEEE).
58. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. **52**.
59. Hodson, T. O. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. **2022**, 1–10. Publisher: Göttingen, Germany.

60. Ramdas, A., Garcia, N. & Cuturi, M. On wasserstein two sample testing and related families of nonparametric tests. [1509.02237\[math,stat\]](#).
61. Felekyan, S., Kalinin, S., Sanabria, H., Valeri, A. & Seidel, C. A. M. Filtered FCS: Species auto- and cross-correlation functions highlight binding and dynamics in biomolecules. **13**, 1036–1053, DOI: [10.1002/cphc.201100897](#).
62. Bradley, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition* **30**, 1145–1159 (1997).
63. Giesa, N. https://github.com/ngiesa/mi_icu.
64. Faltys, M. *et al.* HiRID, a high time-resolution ICU dataset, DOI: [10.13026/NKWC-JS72](#). Version Number: 1.1.1.
65. Johnson, A. *et al.* MIMIC-IV, DOI: [10.13026/HXP0-HG59](#).
66. Li, J. *et al.* Imputation of missing values for electronic health record laboratory data. **4**, 147, DOI: [10.1038/s41746-021-00518-0](#).
67. Tan, A. L. *et al.* Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? **139**, 104306, DOI: [10.1016/j.jbi.2023.104306](#).
68. Restrepo, D. *et al.* Representation learning of lab values via masked AutoEncoder, DOI: [10.48550/arXiv.2501.02648.2501.02648\[cs\]](#).
69. Gao, S. *et al.* Enhanced data imputation framework for bridge health monitoring using wasserstein generative adversarial networks with gradient penalty. **57**, 105277, DOI: [10.1016/j.istruc.2023.105277](#).
70. Luo, S., Lawson, A. B., He, B., Elm, J. J. & Tilley, B. C. Bayesian multiple imputation for missing multivariate longitudinal data from a parkinson's disease clinical trial. **25**, 821–837, DOI: [10.1177/0962280212469358](#).
71. Mavridis, D. *et al.* Allowing for uncertainty due to missing and LOCF imputed outcomes in meta-analysis. **38**, 720–737, DOI: [10.1002/sim.8009](#).
72. Chen, B. *et al.* A joint learning im-BiLSTM model for incomplete time-series sentinel-2a data imputation and crop classification. **108**, 102762, DOI: [10.1016/j.jag.2022.102762](#).
73. Cao, W. *et al.* BRITS: Bidirectional recurrent imputation for time series. [1805.10572\[cs,stat\]](#).

Author contributions

NG extracted and processed EHRs, induced missing patterns, and evaluated imputation methods. SB, SH, and RZ provided methodological guidance. SP ensured statistical correctness. FB gave clinical feedback. All authors read and proofed the manuscript.

Competing Interests

Sebastian Boie reports salaried employment at Pfizer Pharma GmbH. Pfizer had no involvement in the conception, design, execution, or interpretation of the study, nor in the preparation or decision to submit the manuscript for publication. We do not report further conflicts of interest.

Supplementary Files

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

- [supplementA.pdf](#)
- [supplementB.xlsx](#)