TIME-DEPENDENT ITERATIVE IMPUTATION FOR MULTIVARIATE LONGITUDINAL CLINICAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing data is a major challenge in clinical research. In electronic medical records, often a large fraction of the values in laboratory tests and vital signs are missing. The missingness can lead to biased estimates and limit our ability to draw conclusions from the data. Additionally, many machine learning algorithms can only be applied to complete datasets. A common solution is data imputation, the process of filling-in the missing values. However, some of the popular imputation approaches perform poorly on clinical data. We developed a simple new approach, Time-Dependent Iterative imputation (TDI), which offers a practical solution for imputing time-series data. It addresses both multivariate and longitudinal data, by integrating forward-filling and Iterative Imputer. The integration employs a patient, variable, and observation-specific dynamic weighting strategy, based on the clinical patterns of the data, including missing rates and measurement frequency. We tested TDI on randomly masked clinical datasets. When applied to a cohort consisting of more than 500,000 patient observations from MIMIC III, our approach outperformed state-of-the-art imputation methods for 25 out of 30 clinical variables, with an overall root-mean-squared-error of 0.63, compared to 0.85 for SoftImpute, the second best method. MIMIC III and COVID-19 inpatient datasets were used to perform prediction tasks. Importantly, these tests demonstrated that TDI imputation can lead to improved risk prediction.

1 INTRODUCTION

In many practical domains, missing data is a major challenge that has a significant effect on data interpretation and analysis. Missing data can lead to biased estimates (Ayilara et al., 2019) and limit our ability to study and draw conclusions. In addition, most machine learning models (ML) can be applied to complete datasets only. Hence, techniques for dealing with missing data are necessary. A simple approach is *omission*, in which observations with missing values are discarded from further analysis. However, in most cases, the proportion of missing data is not negligible. A retrospective electronical medical record (EMR) dataset is likely to have a large amount of missing values in patients' records. The omission would lead to a significant loss of valuable information, resulting in an unrepresentative sample and strong bias. In practice, data *imputation* methods are used to deal with missing data by filling in the missing data with artificial values. Such methods offer powerful tools for addressing missing data in large datasets with complex data patterns.

Data imputation is usually done by inferring the missing values based on observed data. Imputation of EMR remains a major challenge, as it includes irregular individualized time-series data. Although various imputation methods are available (Batista & Monard, 2003; Mazumder et al., 2010; Van Buuren & Groothuis-Oudshoorn, 2011), many of these are not well designed for clinical data, as they fail to account for inherent individual longitudinal patterns and clinical characteristics. In recent years, novel deep learning approaches were developed for data imputation (Che et al., 2018; Yoon et al., 2018). However, such approaches usually require massive amounts of data and not always practical.

We developed a new approach, Time-Dependent Iterative imputation (TDI), for imputing individualized time-series data. TDI addresses both multivariate and longitudinal data, by integrating forward-filling and Iterative Imputer, a version of MICE (Multivariate Imputation by Chained Equation, Van Buuren & Groothuis-Oudshoorn, 2011). The integration employs a patient, variable, and

observation-specific dynamic weighting strategy, based on the clinical patterns of the data including missing rates and measurement frequency. TDI offers a simple and practical solution, that utilizes the underlying metadata available, without requiring massive amounts of data. Experiments on real-world clinical datasets demonstrated that our model outperforms state-of-the-art imputation, in both producing better value estimates, and improving predictive performance.

2 Problem Formulation

Let $X \in \mathbb{R}^{N \times T \times D}$ be a time-series clinical dataset, where N is the number of patients, D is the number of covariates, and each patient's trajectory is represented by discrete time points indexed as $t \in \{1, ..., T\}$. We use the term *observation* for the vector of the D covariates of a subject at a particular time-point. The covariates can be either time-dependent (longitudinal) or time-independent (static). The number of time points (observations) available for patient i is denoted by $t_i \leq T$. Hence, the longitudinal data of subject i is represented by $X^i = (X_1^i, \ldots, X_{t_i}^i)$, where X_t^i is the observation of subject i at time point t. The actual time when the t-th observation is obtained is denoted by s_t^i. X_t^i may be incomplete. We define $M \in \{0,1\}^{N \times T \times D}$ as the mask matrix of X, where $m_{t,d}^i = 1$ if $x_{t,d}^i$ is observed and zero otherwise (See Figure A.1).

Our goal is to impute the missing values in X. The imputed dataset $\tilde{X} \in \mathbb{R}^{N \times T \times D}$ is:

$$\tilde{X} = M \odot X + (1 - M) \odot \hat{X}$$

Where $\hat{X}$ is an estimate of X and $\odot$ is the element-wise product of matrices.

3 Time-Dependent Iterative imputation (TDI)

3.1 Preliminaries

Our method relies on two established imputation methods that are applied to the data matrix X. In Forward-filling, each missing value of a patient's variable is imputed by using its last observed value. Let $\tilde{X}_F$ be the resulting dataset. Notably, $\tilde{X}_F$ could remain incomplete, due to missing values before the first measurement of a variable, or variables that were not recorded for a patient at all. Additionally, let $\tilde{X}_I$ be the imputed dataset after applying the multivariate Iterative Imputer algorithm to X. Specifically, we applied *scikit-learn*'s IterativeImputer (Pedregosa et al.), which was inspired by MICE (Van Buuren & Groothuis-Oudshoorn, 2011). The Iterative Imputer uses a regression model to predict the missing values of each feature based on the other features, in a round-robin fashion. It is applied on the entire dataset X, independently of $\tilde{X}_F$.

For each patient i, time point t, and variable d, we define the time passed since the last observation as $\Delta t_d^i = s_t^i - s_{t-1,d}^i$, where $s_{t-1,d}^i$ is the time of the previous record of variable d of patient i ($\Delta t_d^i \geq 0$). $\Delta t_d^i = 0$ corresponds to the current time. $\Delta t_d^i = \infty$ if there is no past measurement. Additionally, let r_t^i denote the fraction of available values of patient i at time point t: $r_t^i = \frac{1}{D} \sum_{d=1}^{D} m_{t,d}^i$. Finally, denote the measurement frequency of variable d by f_d. Namely, f_d is the inverse value of the cohort average time (in hours) between measurements of variable d.

3.2 The proposed model

We propose a new time-dependent approach that considers both multivariate and longitudinal data in missing values imputation. At each time-point, the imputed value is the weighted sum of two imputed values estimated by the Iterative Imputer ($\tilde{X}_I$) and forward-filling ($\tilde{X}_F$). The integration employs a **patient**, **variable**, and **observation**-specific dynamic weighting strategy, based on the clinical patterns of the data. The imputed value for variable d of patient i at time point t is:

$$\tilde{x}_{t,d}^i = w_{t,d}^i \cdot \tilde{x}_{F_{t,d}}^i + (1 - w_{t,d}^i) \cdot \tilde{x}_{I_{t,d}}^i$$

Where $w : \mathbb{R}^+ \to [0,1]$ is the following weight function:

$$w_{t,d}^i = \frac{1}{1 + f_d \cdot r_t^i \cdot \Delta t_d^i}$$

In other words, $w(\Delta t_d^i)$ assigns weights to the imputation estimations as a function of three factors:

(i) Time elapsed since past measurement (Δt_d^i): The key idea is that higher weights are assigned to Forward-filling with more recent observed values (lower Δt), in a similar manner to the inference by a caregiver in the clinical practice, when there are recent available values. In contrast, for large Δt, relying on the multivariate distributions using the observed data of other covariates is plausible. Importantly, while $\tilde{X}_I$ is a complete dataset, $\tilde{X}_F$ can still contain missing values (see Section 3.1). Such values are imputed solely based on the Iterative Imputer algorithm.

(ii) Variable sampling rates (f_d): For variables measured less often than others, Forward-filling is assigned more weight, covering a longer period back.

(iii) The observational availability rates (r_t^i): The Iterative Imputer initializes missing values with a naive estimate (e.g., mean). In time points with high missing rates, the imputed values might be estimated mainly based on this initial strategy. Adding r_t^i to the weight function penalizes the Iterative Imputer at time points with higher missing rates.

While we require that w will be a decay function, we do not limit it to a specific function family. It can be either chosen a priori or treated as a hyper-parameter and selected via cross-validation. The weights should remain between 0 and 1 to preserve clinically plausible imputed values.

4 EVALUATION

We empirically evaluated our method and six other simple and state-of-the-art imputation methods, on real-world clinical datasets (see Section 5.1). Section A.1 describes the benchmark imputation methods. We evaluated the methods using two approaches:

Masking: This test compares imputed values to their known true values. Here, in addition to the originally missing data, we randomly mask a fraction of the available values of each variable in the dataset (i.e., values $x_{t,d}^i$ such that $m_{t,d}^i = 1$). Then, we impute the resulting masked dataset and compare the imputed values of the masked data to the ground truth values using different metrics (see Table S3).

Prediction: Missing data imputation does not always provide substantial improvement to predictive models. Hence, this test aims to assess the impact of the imputation on the quality of predicting a clinical outcome. We consider two different types of predictive tasks: (1) *Baseline* prediction: A single prediction for each patient, based on baseline data obtained on admission or a few hours thereafter. (2) *Longitudinal* predictions: Repeated predictions during hospitalization (e.g., at every time point t), using the patient's baseline and longitudinal data up to the current time point (Figure A.2).

5 RESULTS

5.1 DATASETS

MIMIC-III: A public dataset of de-identified clinical care data with over 40,000 patients admitted to the Beth Israel Deaconess Medical Center, Boston, between 2001 to 2012 (Johnson et al., 2016). We extracted 30 longitudinal features, including vital signs, basic metabolic panel (BMP), hematology panel, and coagulation panel (Table S1).

COVID-19: A dataset containing EMRs of 3,293 COVID-19 inpatients admitted to two hospitals (detailed hidden for double-blind review).

Preprocessing details can be found in Section A.2.

5.2 MASKING

The masking experiment was conducted on a subsample of $559,837$ observations from $N = 8000$ patients ($D = 30$). Data standardization was performed, as required for some imputation methods (e.g., KNN). We randomly masked 10% of the available values of each variable, and then imputed the data. We used three metrics to measure the difference between the imputed and real values (Table

S3). Figure 1 summarizes the performance of seven imputation methods, in terms of normalized root mean squared error (NRMSE). TDI outperformed the other methods for 25 out of 30 variables. It also had the best overall score in all performance metrics, with an overall RMSE of 0.63, compared to 0.85 for SoftImpute, the second best method (see Table S4).

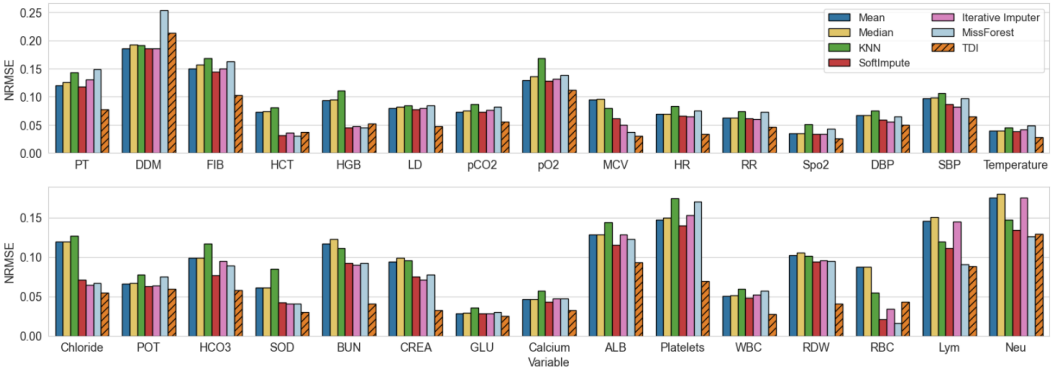

Figure 1: **Masking performace.** A comparison of imputation models for each variable, using NRMSE (lower is better). TDI imputation (orange dashed bar) does best in 25 out of 30 variables.

5.3 Prediction

We performed a baseline binary classification task on MIMIC-III for predicting mortality in hospitalization or 30 days thereafter, based on data obtained in the first 48 hours after admission. We evaluated ten ML models for this prediction task, including feedforward neural network (Haykin, 1994), SVM (Cortes & Vapnik, 1995), Logistic Regression, Random Forest (Breiman, 2001), CatBoost (Dorogush et al., 2018), XGBoost (Chen & Guestrin, 2016) and GRU-D (Che et al., 2018). Note that GRU-D does not use an external imputation method. We measured the predictive performance in 5-fold cross-validation. In each iteration, an imputation method was first applied, and then each of the predictive models was trained on the imputed data. The imputation was done separately on the train and test folds. Each model's performance was measured using the area under the receiver-operator characteristics curve (AUROC) and the area under the precision-recall curve (AUPR) over the test folds. When applied to a subsample of N=2,000 patients, TDI improved the mean AUROC and AUPR in all classifiers, six of which perfomed better than GRU-D (Figures 2, A.3). For a subsample of N=10,000 patients, GRU-D and CatBoost with TDI achieved comparable results (A difference of 0.1% in favor of GRU-D in both metrics) (Table S5). Similar results were achieved in a longitudinal setting on the COVID-19 datasets, for predicting mortality in the next 1-7 days (Supplementary A.4). Section A.3 describes the experimental design of this benchmark.

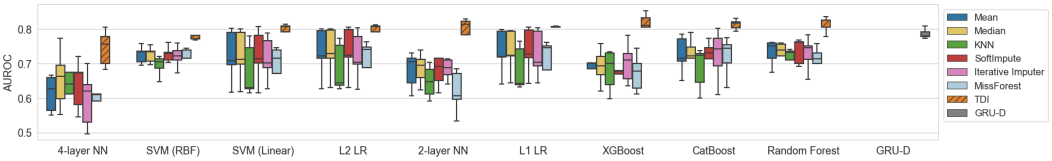

Figure 2: **Performance of baseline mortality prediction in MIMIC-III, using 5-fold cross-validation**. TDI (orange dashed) improves the AUROC (y-axis) in all classifiers. Horizonal line: median.

6 Discussion

We presented TDI Imputation, a practical approach for missing data imputation designed for time-series clinical datasets. Our method imputes missing data by integrating forward-filling and the Iterative Imputer (Pedregosa et al.). The integration employs a patient, variable, and observation-specific dynamic weighting strategy, based on the clinical characteristics of the data. Compared to state-of-the-art imputation methods, our model achieved the best performance in the estimation of values of masked clinical variables with known ground truth. While being simple and applicable, TDI imputation led to an improved risk prediction in different ML models on real-world datasets. Future work will study additional prediction tasks using samples of varying sizes.

CODE AND DATA AVAILABILITY

The code for our method and for data preprocessing will be published in a GitHub repository upon publicity. The MIMIC-III dataset is available upon request at http://dx.doi.org/10.13026/C2XW26. Requests for access to the COVID-19 datasets should be directed to the hospitals.

REFERENCES

Olawale F Ayilara, Lisa Zhang, Tolulope T Sajobi, Richard Sawatzky, Eric Bohm, and Lisa M Lix. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and quality of life outcomes*, 17(1):1–9, 2019.

Gustavo EAPA Batista and Maria Carolina Monard. An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6):519–533, 2003.

Gustavo EAPA Batista, Maria Carolina Monard, et al. A study of k-nearest neighbour as an imputation method. *His*, 87(251-260):48, 2002.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

Alistair EW 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):1–9, 2016.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, et al. a.(2011). scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–283.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Stef Van Buuren and Karin Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pp. 5689–5698. PMLR, 2018.

A APPENDIX

A.1 BENCHMARK IMPUTATION METHODSS

Our benchmark includes the following imputation methods:

- *Mean, Median*: The missing values of a certain variable are replaced by the mean (or median) of the available ones.
- *SoftImpute* (Mazumder et al., 2010): Missing values are imputed using matrix completion by iterative soft thresholding of Singular Value Decomposition (SVD).
- *KNN* (Batista et al., 2002): Missing values are imputed using the mean value of the k nearest neighbors (closest samples) found in the training set.
- *Iterative Imputer (Pedregosa et al.)*: A version of the MICE algorithm (Van Buuren & Groothuis-Oudshoorn, 2011). See 3.1 for additional information.
- *MissForest* (Stekhoven & Bühlmann, 2012): A non-parametric imputation method that uses random forests to predict the missing values, based on the observed values.
- *GRU-D* (Che et al., 2018): A deep learning model, based on Gated Recurrent Unit (GRU), that incorporates two representations of missing patterns, *masking* and *time interval*. Note that GRU-D (Che et al., 2018) is not explicitly designed for missing values imputation, and cannot be directly used in an unsupervised setting. Hence, it was used for prediction and excluded from the masking experiment.

A.2 DATASETS PREPROCESSING DETAILS

A.2.1 MIMIC-III

This section describes the preprocessing of the MIMIC-III dataset. MIMIC-III contains clinical care data for N=46,520 patients (patients available in both ADMISSIONS and PATIENTS data tables).

Variable extraction and mapping. We extracted 30 longitudinal features, including vital signs, basic metabolic panel (BMP), hematology panel, and coagulation panel (Supplementary Table S1). Variables that were measured with different monitoring modalities were merged (for example, fingerstick glucose and blood glucose). We note that laboratory tests that were measured through other body fluids than blood (e.g., urine) were removed. Unit conversions were done when required. For the prediction tasks, we also utilized the patients' age and gender information.

Outlier removal. To remove incorrect measurement values (due to typos for example), we manually defined with clinicians ranges of possible values for each longitudinal variable (including pathological ones). Values outside these ranges were excluded.

Time discretization. The temporal data was discretized into a 15-minute grid.

Inclusion and exclusion criteria. We included only patients aged between ages 18 and 89 (patients aged over 89 are masked in MIMIC) (N=36,564). In cases of several hospital admissions, we focused on the first admission of each patient. Only patients with at least one measurement of any of the longitudinal variables during the first hospitalization were included. The resulting cohort consists of 2,575,254 observations of N=35,968 patients,

A.2.2 COVID-19

We evaluated our proposed model on an additional retrospective cohort comprising two datasets from hospitals A and B. The A dataset consisted of all COVID-19 patients admitted to A between March 2020 and March 2021 ($N_A = 782$). The B dataset consisted of all COVID-19 patients admitted to B between March 2020 and April 2021 $N_B = 2,511$). The study was reviewed and approved by the Institutional Review Boards (details removed for blinded refereeing).

The data used for this study included age, gender, and 13 longitudinal features, including laboratory test results and vital signs (See Supplementary Table S2). We mapped variables from the two hospitals and performed unit conversions when required. Similarly to the MIMIC-III preprocessing, we removed clinical outliers, namely, variable values that are out of predefined valid ranges. The

temporal data was discretized to an hourly time grid, and multiple values of a test measured within the same hour were aggregated by mean.

A.3 Prediction tasks details

A.4 Baseline prediction

We performed a baseline binary classification task on MIMIC-III for predicting mortality in hospitalization or 30 days thereafter, based on data obtained in the first 48 hours after admission. We evaluated ten ML models for this prediction task, including feedforward neural network (Haykin, 1994), SVM (Cortes & Vapnik, 1995), Logistic Regression, Random Forest (Breiman, 2001), CatBoost (Dorogush et al., 2018), XGBoost (Chen & Guestrin, 2016) and GRU-D (Che et al., 2018). Note that GRU-D doesn not use an external imputation method. While the considered classifiers cannot directly handle time series of different lengths for predicting a single target in time (e.g., mortality), the GRU-D uses a time-series sequence input for a single prediction. To perform a fair comparison, after data imputation, we sampled the time-series data to get a fixed-length input. Specifically, we used the two last observations in the first 48 hours for each patient. We also concatenated the masking vector along with the measurements, similarly as done in the GRU-D architecture. The GRU-D in contrast was fed initially with the full time-series dataset, giving it a potential advantage. To estimate the effect of the imputations on the predictive performance in different subsamples, we used 5-fold cross-validation. In each iteration, an imputation method was first applied, and then each of the predictive models was trained on the imputed data. The imputation was done separately on the train and test folds. Finally, the model performance was measured using the area under the receiver-operator characteristics curve (AUROC) and the area under the precision-recall curve (AUPR) over the test folds.

To create the baseline prediction setting, exclusion/inclusion criteria were applied. The number of patients and observations remaining (out of N=2000 subsamples) after each criterion are listed:

- Exclude patients who died less than 60 hours since admission (N=1,941 patients, 145,287 observations), as we wish to predict at least 12 hours in advance.
- Include observations from the first 48 hours (N=1,938 patients, 45,492 observations).
- Exclude empty observations, namely, such that all 30 longitudinal features are missing (N=1938 patients, 44,659 observations).
- Include patients with at least three time-points (N=1,740 patients, 44,312 observations).

The resulting cohort contains 1,740 patients, of which 156 patients had a positive outcome (8.9$\%$).

A.5 Longitudinal predictions

We performed a binary classification task for every hourly observation to predict mortality in one to seven days. Patient observations where mortality was reported in the next seven days were called positive, and the rest were called negative. Observations from the 24-hours prior to the target event were excluded.

The initial cohort included 3,293 (195,237 observations) patients from both datasets. To create the longitudinal prediction setting, the following exclusion/inclusion criteria were applied:

- Exclude pregnant women and patients aged ≤ 18 (N=2,940 patients).
- Include only patients admitted for more than 24 hours (N=2,563 patients).
- Exclude patients who died less than 48 hours since admission (N=2,534 patients).
- Exclude observations recorded in the 24-h gap prior to the target, as we wish to predict in advance.

The resulting cohort consists of 43,812 hourly observations of 2,534 patients. 3,487 (7.9$\%$) observations of 453 patients were labeled positive.

A.6 SUPPLEMENTAL TABLES AND FIGURES

$x_{1,1}$			...	$x_{1,D}$
	$x_{2,2}$	$x_{2,3}$	...	
...	...	...	...	...
	$x_{t_i,2}$		...	$x_{t_i,D}$

(a) Data Matrix

1	0	0	1	1
0	1	1	1	0
...	...	...	...	...
0	1	0	1	1

(b) Mask Matrix

Figure A.1: **Patient data**. (a) The longitudinal data matrix of patient i, denoted as X^i. Each patient's matrix consists of D covariates (columns) and a different number of time-points t_i (rows). Missing values are shaded. (b) A mask matrix that indicates which values of X are observed.

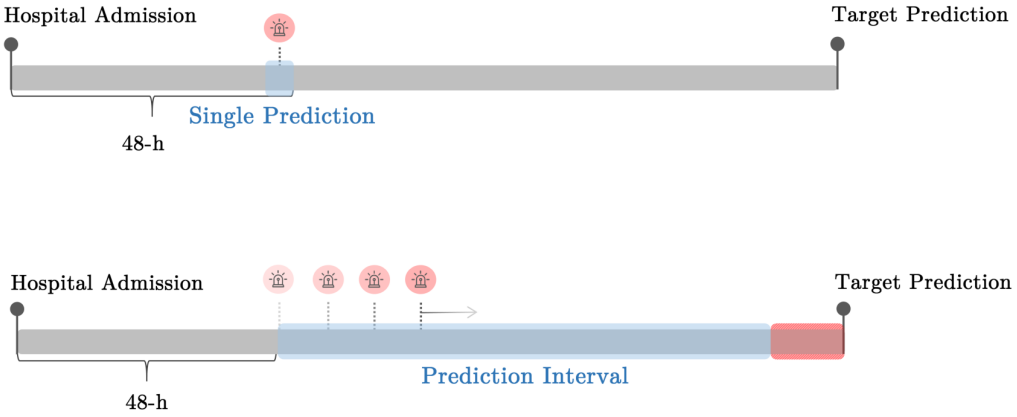

Figure A.2: **Predictive tasks along the patient timeline. Top: Baseline prediction.** The model generates a single prediction for each patient, based on baseline data from the first 48-h after admission. **Bottom: Longitudinal prediction.** The model generates longitudinal predictions for each patient, based on data from the entire hospitalization period. The blue areas refer to the time interval when the predictions are made. The Red dashed area represents blocked prediction periods during which no predictions are made.

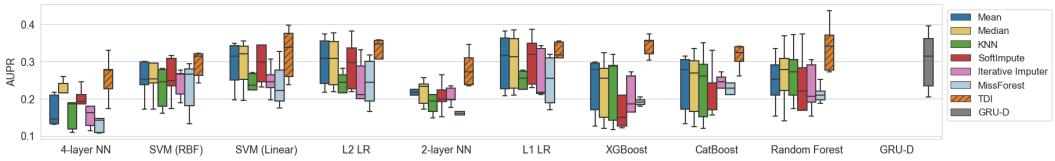

Figure A.3: **Performance of baseline mortality prediction in MIMIC-III, using 5-fold cross-validation**. TDI (orange dashed) improves the AUPR (y-axis) in all classifiers. Horizonal line: median.

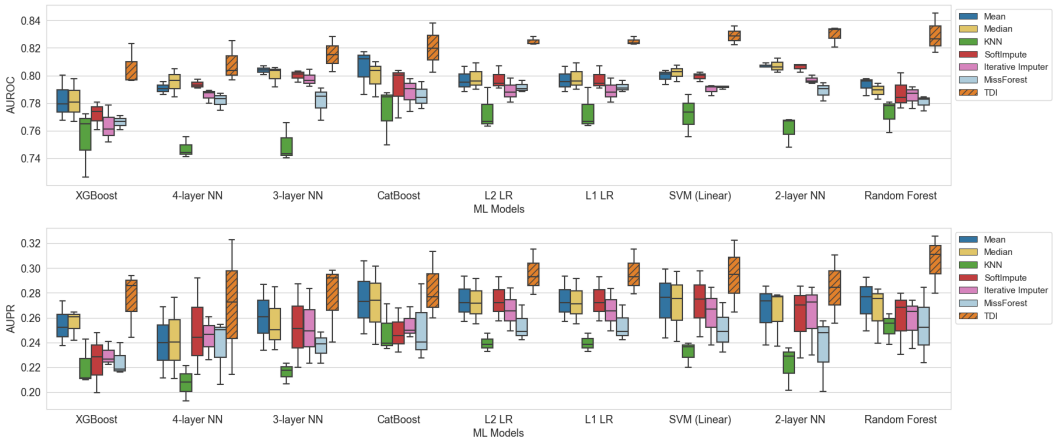

Figure A.4: **Performance of longitudinal mortality predictions on the COVID-19 data**. Comparison of machine learning models using 3-fold cross-validation. TDI (orange dashed) improves AUROC (top) and AUPR (bottom) in all classifiers. Horizonal line: median.

Variable	Unit	N	Mean $\pm$ SD	Missing Rate (%)
Heart Rate (HR)	BPM	1483322	86.73 $\pm$ 17.83	42.4
Respiratory Rate (RR)	BPM	1482838	20.07 $\pm$ 5.83	42.42
Spo2	%	1459840	96.93 $\pm$ 3.15	43.31
Systolic Blood Pressure (SBP)	mmHG	1427439	121.47 $\pm$ 22.38	44.57
Diastolic Blood Pressure (DBP)	mmHG	1426266	62.41 $\pm$ 14.55	44.62
Glucose (GLU)	mg/dL	822380	136.43 $\pm$ 55.98	68.07
Hematocrit (HCT)	%	464369	30.33 $\pm$ 4.88	81.97
Potassium (POT)	mEq/L	443183	4.08 $\pm$ 0.59	82.79
Hemoglobin (HGB)	g/dL	439220	10.3 $\pm$ 1.77	82.94
Sodium (SOD)	mEq/L	422425	138.69 $\pm$ 5.03	83.6
Temperature	Celsius	414649	36.98 $\pm$ 0.83	83.9
Chloride	mmol/L	414014	104.05 $\pm$ 5.97	83.92
Creatinine (CREA)	mg/dL	414084	1.41 $\pm$ 1.41	83.92
Urea Nitrogen (BUN)	mg/dL	412401	28.61 $\pm$ 23.06	83.99
Bicarbonate (HCO3)	mEq/L	404863	25.33 $\pm$ 4.76	84.28
Platelet Count	10e3/μL	396073	233.38 $\pm$ 148.75	84.62
White Blood Cells (WBC)	10e3/μL	379718	11.37 $\pm$ 6.93	85.26
Red Blood Cells (RBC)	m/μL	377993	3.42 $\pm$ 0.6	85.32
MCV	fL	377547	89.86 $\pm$ 6.2	85.34
Red Cell Distribution Width (RDW)	%	377251	15.54 $\pm$ 2.32	85.35
pCO2	mmHg	349334	41.85 $\pm$ 10.11	86.43
pO2	mmHg	349387	141.19 $\pm$ 90.66	86.43
Calcium	mg/dL	326887	8.39 $\pm$ 0.8	87.31
PT	Sec	248814	16.02 $\pm$ 4.97	90.34
Albumin (ALB)	g/dL	63613	2.9 $\pm$ 0.67	97.53
Lactate Dehydrogenase (LD)	IU/L	54444	530.19 $\pm$ 1045.02	97.89
Lymphocytes (Lym)	%	49035	15.57 $\pm$ 16.88	98.1
Neutrophils (Neu)	%	48818	74.39 $\pm$ 18.88	98.1
Fibrinogen (FIB)	mg/dL	29644	312.39 $\pm$ 192.28	98.85
D-Dimer (DDM)	ng/mL	1826	4128.61 $\pm$ 3435.9	99.93

Table S1: List of 30 variables after preprocessing the MIMIC-III dataset.

Variable	Unit	N	Mean ± SD	Missing Rate (%)
Heart Rate (HR)	BPM	129026	85.88 ± 18.94	33.91
Temperature	Celsius	80513	36.9 ± 0.82	58.76
SBP	mmHG	70818	128.46 ± 22.36	63.73
DBP	mmHG	70564	69.97 ± 14.55	63.86
Saturation	%	68386	94.77 ± 5.04	64.97
Glucose	mg/dL	51713	164.43 ± 76.8	73.51
HCO3	mmol/L	31083	29.42 ± 7.11	84.08
Platelets	10e3/μL	18658	244.54 ± 126.22	90.44
BUN	mg/dL	18473	32.34 ± 26.56	90.54
Neutrophils (#)	10e3/μL	18431	8.12 ± 5.53	90.56
Lymphocytes (#)	10e3/μL	18261	1.25 ± 1.29	90.65
Albumin	g/L	12755	31.46 ± 6.84	93.47
LDH	U/L	11416	844.54 ± 2175.36	94.15

Table S2: List of 13 variables after preprocessing the COVID-19 datasets.

Evaluation Metric	Formula
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$
NRMSE	$\frac{\text{RMSE}}{y_{max}-y_{min}}$
SMAPE	$\frac{1}{n}\Sigma_{i=1}^{n}\frac{\lvert\hat{y}_i - y_i\rvert}{(\hat{y}_i + y_i)/2}$

Table S3: **Evaluation measures of differences between values.** y and $\hat{y}$ refer to the actual and estimated values, respectively. RMSE: Root Mean Squared Error. NRMSE: Normalized Root Mean. SMAPE: Symmetric mean absolute percentage error

Imputation	RMSE	NRMSE	SMAPE
TDI	0.636	0.060	0.858
SoftImpute	0.851	0.079	1.365
Iterative Imputer	0.891	0.084	1.416
MissForest	0.932	0.086	1.058
Mean	1.013	0.095	1.985
Median	1.031	0.097	1.446
KNN	1.124	0.102	1.324

Table S4: **Masking overall performance.** The average value across all variables in MIMIC-III, for difference evaluation.

			AUROC			AUPR		
N	**ML Model**	**Imputation**	**Mean**	**Median**	**SD**	**Mean**	**Median**	**SD**
2,000	**Random Forest**	TDI	**0.814**	**0.825**	0.023	**0.339**	**0.342**	0.069
	CatBoost	TDI	0.813	0.817	0.015	0.313	0.323	0.033
	XGBoost	TDI	0.809	0.81	0.042	0.342	0.355	0.029
	L1 LR	TDI	0.792	0.807	0.033	0.313	0.317	0.05
	2-layer NN	TDI	0.791	0.814	0.051	0.28	0.272	0.048
	L2 LR	TDI	0.791	0.806	0.033	0.316	0.347	0.062
	GRU-D	GRU-D	0.786	0.78	0.015	0.302	0.314	0.082
	SVM (Linear)	TDI	0.786	0.807	0.045	0.322	0.338	0.071
	SVM (RBF)	TDI	0.785	0.772	0.027	0.292	0.314	0.037
	4-layer NN	TDI	0.745	0.756	0.052	0.257	0.277	0.06
10,000	**GRU-D**	GRU-D	**0.817**	**0.818**	0.016	**0.363**	0.359	0.041
	CatBoost	TDI	0.816	0.817	0.012	0.362	**0.365**	0.022
	L1 LR	TDI	0.8	0.801	0.005	0.344	0.359	0.032
	L2 LR	TDI	0.799	0.799	0.005	0.342	0.358	0.034
	SVM (Linear)	TDI	0.798	0.797	0.004	0.338	0.354	0.039
	Random Forest	TDI	0.797	0.797	0.012	0.344	0.346	0.017
	XGBoost	TDI	0.794	0.79	0.01	0.322	0.332	0.025
	SVM (RBF)	TDI	0.762	0.76	0.008	0.323	0.328	0.022
	2-layer NN	TDI	0.748	0.747	0.004	0.283	0.292	0.023
	4-layer NN	TDI	0.688	0.695	0.014	0.251	0.252	0.025

Table S5: **Performance of baseline mortality prediction in MIMIC-III, using 5-fold cross-validation.** For each classifier, only the best performer imputation method is presented. GRU-D does not use external imputation methods. Best measures per sample size are bolded.