Generative Neural Networks for Data Imputation in Longitudinal Epidemiological Studies

Christoph Killing, Kira Elsbernd, Maximilian Wekerle,
Michael Hoelscher, Andrea Rachow, Noemi Castelletti

***Abstract*— Longitudinal epidemiological studies often face challenges with incomplete follow-up and missing data, which can bias results and reduce statistical power. Conventional imputation methods may not adequately capture the complex patterns and dependencies in such multivariate time series data. While more recently developed generative machine learning models offer improved solutions, few methods are available which can handle inconsistently spaced intervals between measurements across long time periods and completely missing time steps, characteristics which are common in real-world studies evaluating long-term health outcomes. This paper introduces a variational autoencoder-based generative neural network designed for imputing partially and fully missing information in irregular time series with extensive missingness. Our approach exploits both correlations between features at a single time step and trends of the same feature over time to reconstruct missing values. Experiments on synthetic data designed to resemble the characteristics of longitudinal epidemiological studies and a case study on a real-world dataset demonstrate the effectiveness of our approach. We show superior performance and parameter stability across varying degrees of missingness and missingness patterns compared to prior work.**

***Index Terms*— Biomedical computing, Data imputation, Generative neural networks, Machine learning**

I. Introduction

LONGITUDINAL epidemiological cohort studies are widely used in health research to track changes in individual health outcomes over time, establish correct timing and sequences of events, and identify causal relationships. These studies often collect health parameters from the same individual repeatedly over time. However, incomplete and interrupted follow-up is an almost universal challenge leading to issues with data completeness. Non-response, attrition, or measurement errors during data collection can result in partial or complete missingness of information at a given time point. The resulting gaps in patient records can introduce bias, reduce statistical power, and ultimately compromise the validity and generalizability of study conclusions [1]. Additionally, irregular follow-up visit schedules due to study design or resource constraints may lead to unevenly spaced measurements in time. Therefore, effective imputation methods capable of handling structured and unstructured missingness while preserving irregular temporal dependencies are needed.

Data imputation, the process of systematically replacing missing, invalid, or inconsistent data with substitute values, has been a critical area of development in both statistical and computational research. Traditional approaches consider individual parameters independently of time step size, for instance, by carrying the last observation forward or taking the mean of the observed values to replace missing observations. However, they are limited in their ability to capture complex temporal dependencies, interparameter correlations, and variability present in medical time series data [2, 3, 4]. Recent advancements in deep learning approaches offer more sophisticated solutions, with variations proposed that rely on both temporal and interparameter correlations for imputation [5]. These models learn underlying data distributions and explicitly account for uncertainty by producing a range of plausible values. However, these networks have often been developed around microscale time series data with regular recording intervals, which are common in technical applications (e.g. [6]). Therefore, the combination of missingness and irregular macroscale recording intervals, between which parameters can change significantly, poses a previously unaddressed challenge to data imputation that requires specialized approaches.

This study presents a novel Variational Autoencoder (VAE)-based generative deep learning approach tailored to the unique requirements of longitudinal epidemiological studies. Our neural network offers robust probabilistic imputation, making it well suited for high-dimensional epidemiological data with irregular follow-up intervals and partially or completely missing time step information. Our main contributions are: 1) a synthetic open-source benchmarking dataset capturing the key characteristic challenges of longitudinal epidemiological data while preserving the ability to quantify imputation quality through the available ground truth information, 2) a novel VAE-based deep neural network for data imputation which addresses the challenges of both irregularly spaced multivariate time series and completely missing time steps,

Submitted for review on October 10th, 2024. Revised June 16th, 2025. This work was supported in part by BMBF-DLR under grant number 01KA2223B and BMBF-DZIF under grant number 01KA2108.

All authors are with the Institute of Infectious Diseases and Tropical Medicine, LMU University Hospital, LMU Munich. (email: {firstname}.{lastname}@med.uni-muenchen.de)

Michael Hoelscher and Noemi Castelletti furthermore are with the Fraunhofer Institute, Immunology, Infection and Pandemic Research, Munich, Germany

Michael Hoelscher and Andrea Rachow furthermore are with the German Centre for Infection Research, Partner Site Munich, Germany and with the Helmholtz Zentrum, German Research Center for Environmental Health, Munich, Germany

and 3) demonstration of the performance of our approach on a challenging real-world dataset evaluating long-term lung health after pulmonary tuberculosis (TB).

In the remainder of this paper, we first present related work in the field of data imputation in Section II, highlighting the unique requirements of irregular time series with fully missing time step information. We then provide a definition of our model in Section III and demonstrate the imputation performance on two datasets, one synthetic and one real-world dataset in Section IV. In Section V, we conclude with the broader translational impact of our findings.

II. Related Work

A. Classical Approaches

Many epidemiological studies use complete case analysis, eliminating the need for data imputation at the cost of losing potentially valuable information and statistical power, especially in datasets with extensive missingness [7]. Classical approaches such as mean, median, or forward imputation, still widely applied today, consider a single recorded parameter from one individual and its evolution over time but can fail to accurately represent overall cohort trends [3]. In contrast, cohort-based approaches such as cohort mean or nearest-neighbor imputation can fail to adequately represent the heterogeneity of individual trajectories over time [4]. Besides approaches which focus on a single parameter over time (univariate time-series), multivariate time series can capture not only how individual parameters change over time but also how they influence each other. Bashir et al. [8] propose an expectation maximization approach to handle missing values in multivariate time series data but rely on further modeling assumptions including that data be missing completely at random, which is not realistic in many real-world applications. Multiple imputation, which pools parameter estimates and variance across several imputed datasets, results in unbiased and efficient estimates but may not perform well with large, high-dimensional datasets with complex characteristics, such as non-linearity, which are increasingly available in health research [9].

B. Deep Learning

With the advancement of deep learning, more complex imputation methods for multivariate time series data have been developed. Deterministic approaches, such as adaptations to Gated Recurrent Units (GRUs) [6, 10] have previously been explored. However, GRUs lack a global perspective on short but irregular time series and are difficult to train due to vanishing gradients. Transformer-based approaches, which can consider all time steps simultaneously, can alleviate these burdens. However, adding the required positional encoding to continuous data remains a significant challenge, limiting the transformer's ability to leverage its key strength: long-range contextual understanding [11]. Generative adversarial networks (GANs), which combine deterministic and probabilistic elements, excel at generating realistic synthetic data but are prone to learning instabilities and require clean training data. The noise inherent in real-world data can make GAN-based imputation unreliable [12].

Autoencoder-based approaches to imputation rely on a deterministic encoding of the data and have been used as a basis for multiple imputation [13]. More robust and commonly used probabilistic approaches include VAEs, which consistently outperform classical approaches [14]. VAEs are trained to encode data into a latent space, typically of lower dimensionality than the original data, where the encoding network generates the parameters of normal distributions that describe the latent space. Samples from these distributions are subsequently decoded in order to reconstruct the original data and dimensionality. The decoding network is therefore inherently designed for imputation. To ensure disentanglement of the latent space, β-VAEs have been proposed. These balance reconstruction loss and the Kullback-Leibler (KL)-divergence of the latent space from a standard normal distribution [15]. Variants such as HI-VAE [16] modify the loss function to account only for observed features. Extending the idea of accounting for missing information to temporal imputation, Fortuin et al. [17] propose a Gaussian process prior VAE (GP-VAE) instead of the standard normal distribution to capture multiscale time dynamics in the latent space. Besides time step-based information, this provides additional temporal context for the imputation. Although their GP-VAE is closest to our proposed method, it struggles with fully missed visits.

C. Time Intervals

Few approaches explicitly address the challenge of imputation in studies with long, uneven time intervals and missing visits. Li and Marlin [18] introduce a continuous convolutional layer to integrate irregular time steps into conventional network architectures, but their method treats irregularly spaced time steps as undersampled and introduces further empty time steps subsequently considered as fully missing. Others [6, 10] have worked around the issue of irregular time intervals by designing a custom GRU cell which considers time since the last observation.

D. Datasets

In the context of deep learning and imputation, one key distinction arises: some approaches are designed to *handle* missing values directly, while others aim to *impute* them. The former includes methods that operate on datasets with missing values, such as those predicting mortality from the PhysioNet dataset [19], without actually filling in the gaps. While these models can effectively make predictions despite missing data, they do not evaluate or improve imputation quality and are therefore outside the scope of this work. In contrast, methods focused on imputing individual missing values face a significant challenge: real-world datasets like PhysioNet lack ground truth for the missing entries. A common workaround is to artificially mask a subset of observed values and assess how well the model reconstructs them. However, in epidemiological studies, where participant recruitment is often limited to the minimum needed for statistical power, this strategy is impractical, as it reduces the already limited amount of usable

data. Instead, imputation performance can be evaluated using domain knowledge, emphasizing the reconstruction of a coherent and clinically plausible representation of an individual patient's health state.

To allow comparison to ground truth, Krishnan et al. [20] proposed *Healing*MNIST, which is based on the handwritten digits from the MNIST dataset, to resemble the properties of a medical time series. Each pixel can be thought of as a single measurement, the collection of which represents a patient's health state through a single frame. The evolution in time is modeled by randomly rotating the frame. Missing measurements are represented by removing a certain share of pixels in each frame. Subsequently, besides the error on individual pixels, the ability of a classifier to accurately distinguish the reconstructed digits has previously been considered as indicative of the reconstruction ability of an imputation algorithm [17, 20].

III. Method

We propose a VAE-based network which evolves around the generation of two latent spaces. Information from time steps which are fully missing can be imputed by considering the overall trajectory of parameters as captured in a global latent space. Partially missing information can be filled by exploiting other measurements from the same time step (local perspective) or by relying on an overall trajectory of a single feature measured over time (global perspective). The overall process, described in this section, is shown in Fig. 1.

A. Problem Statement

We consider a d-dimensional time series $X \in \mathbb{R}^{T\times d}$ of length T. Points $\tau = [\tau_1, ..., \tau_T]$ are unevenly distributed in time, however, $\tau_t < \tau_{t+1}$. At each time step, a d-dimensional observation $x_t = [x_{t,1}, ..., x_{t,d}]$ is made, of which *any* or *all* of the d features may be missing. A boolean missingness mask $m \in \{0,1\}^{T\times d}$ describes these missing data points such that missing data is given by $x^m = x \times m$ and observed data by $x^o = x \times \bar{m}$. Notably, $\sum_d m_t \leq d$ as data of one time step could be missing all together. The imputation problem is then defined as finding the missing values x^m given the observed data points x^o through the means of a generative neural network $p(x|z_c)$ from latent space z_c, shown in Equation (1)

$$p(x|x^o) = \int p(x|z_c)q(z_c|x^o)dz_c. \tag{1}$$

B. Generative and Inference Model

From a known latent space z, a generative network $p_\theta(x|z)$ with parameters θ, can impute observations independently per time step. An encoding $p(z|x)$ is required to construct the latent space from the observed data. The encoding process can consequently be thought of as reconstructing the underlying state given values that were measured or observed. In a medical data framework, the generative process can be interpreted as observing measurable values given a patient's underlying physiological state. However, the true latent features are unknown, making the encoding process intractable. Instead, the

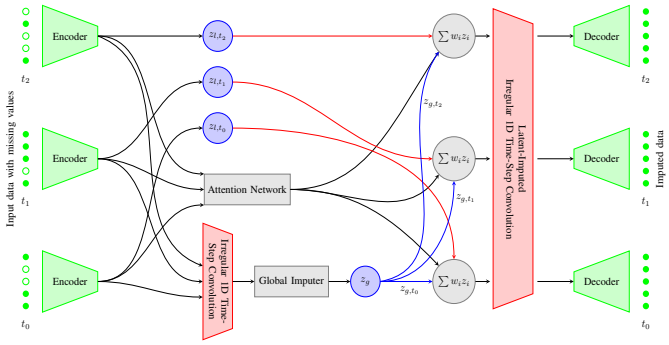


Fig. 1: Overview of our proposed EpiVAE network. Data from time steps $t_{0,1,...}$ is encoded through a fully connected network (with a preceding single convolution layer for MNIST data, not depicted). The resulting local latent representations z_l are (a) used as-is for subsequent mixing (gray circles) with the global latent space z_g, and (b) used to inform an attention network generating the weighting w_i between local and global latent spaces and (c) related to each other through a irregular 1D time-step convolution. The latter informs the global imputation network, generating the global latent space z_g. The contribution of z_g to each latent encoding is determined by the attention network. This combined latent space is fed though the latent imputed irregular 1D time-step convolution before decoding by a fully connected network.

encoding is approximated through $q(z|x)$ which is chosen to be tractable, for instance through a Normal distribution. Minimizing the KL-divergence of $q(z|x)$ from a defined prior distribution $p(z)$, usually also a Normal, then leads to the well-established evidence lower bound. For further details, refer to [15].

To capture local features and allow for a global perspective where data is missing completely at a particular time step, we use a multi-encoder [21] structure in our network. This relies on two inference models focusing on either local or global feature perspectives. First, we use a local inference network $q_\phi(z_l|x^o)$, parameterized by ϕ, which considers each time step independently to encode data into a local latent space z_l. Second, we use a global inference network $q_\psi(z_g|x^o)$, parameterized by ψ, to relate observations across time steps and to overcome the challenge of missed visits by constructing a global latent space z_g. We approximate the true posteriors with Gaussian distributions.

C. Learned Attention

The information contained in the local latent space z_l can result from three possible input modalities: a single time step can be fully observed with all data present, partially observed with some data present, or unobserved with no data present. While in the first case it is beneficial to directly use the local encoding for reconstruction, the latter case requires full reliance on the global latent space z_g. Consequently, in the case of partial observations, a mixture of local and global latent spaces may be preferable.

Therefore, to combine the information contained in the two conditionally independent latent spaces z_l and z_g, we

Algorithm 1 *Epi*MNIST

Given MNIST data and category labels X, C
Sample study time-intervals $t_{1:T} \sim \mathcal{U}(0.5, 2)$
Sample relative frame rotations $r \sim \mathcal{U}(-1, 1)^{J \times C}$

FOR x_i, c in X, C:
  sample disease trajectory $j \sim \mathcal{U}(\mathbb{N} < J)$
  cumulatively rotate x_i by $r_{j,c} \times t$
  apply missingness mask m_i to result

train an attention network. Given single time step missingness information m_t, it produces weights $w \in \mathbb{R}^{2\times d}$ used to combine and reweigh the contributions of the two latent spaces as shown in equation (2)

$$z_c = w_l * z_l + w_g * z_g. \tag{2}$$

To allow for a practical application of our algorithm, we compute the loss only over observed features x^o. However, this leads to a pitfall in (2), where local observations can be exploited to bypass the global pathway all together since z_l contains all x^o. Therefore, during training, we mask certain observations by setting m_t to missing, thereby ensuring gradient-flow through the global network. Furthermore, when an observation is completely missing, the network is not allowed to rely on the local latent space. We implement this by setting the according weights for the local latent space to zero. A more elegant solution could be enforcing a reduced KL-divergence between the two latent spaces, ensuring the comparability of the encoded data. This procedure, however, is only possible in time steps with complete measurements, which we deemed too strong of an assumption for practical applications. We compare this to a hard selection mechanism, which is only relying on the global latent space when observations are completely missing.

D. Irregular Time Step Considerations

To appropriately represent the irregular time structure of the data, time convolutional networks have previously been used to relate neighbouring time steps [17]. We extend this idea to enable our system to learn the uneven spacing of observations by computing the time delta $\Delta t_{i,j} = t_i - t_j$ to all neighbouring observations t_j for each time step t_i before the convolution. Besides the time gap, the sign of the delta also allows the system to identify the relative position to other observations in time. For the computation of relative time step sizes, we normalize the time vector by its mean. We embed the time difference into the feature space by concatenating $\Delta t_{i,j}$ to z_c before performing the convolution.

E. Training

The generator network $p_\theta(x|z)$ consists of the attention mechanism (which balances the influence of local and global features), the time convolution, and a fully connected decoder network. On the inference side, data is encoded through the local encoder network q_ϕ which generates the local Gaussian posterior distribution $\mathcal{N}(\mu_l, \sigma)$ for each time step. On the global path, q_ψ uses the same gradient-free weights as q_ϕ but skips the step of generating a Gaussian distribution. Instead, it uses the gradient-disabled time convolution from the generator network to relate observations. Subsequently, we rely on a fully connected network containing all optimizable parameters ψ of the global network to generate the global Gaussian posterior distribution $\mathcal{N}(\mu_g, \sigma)$.

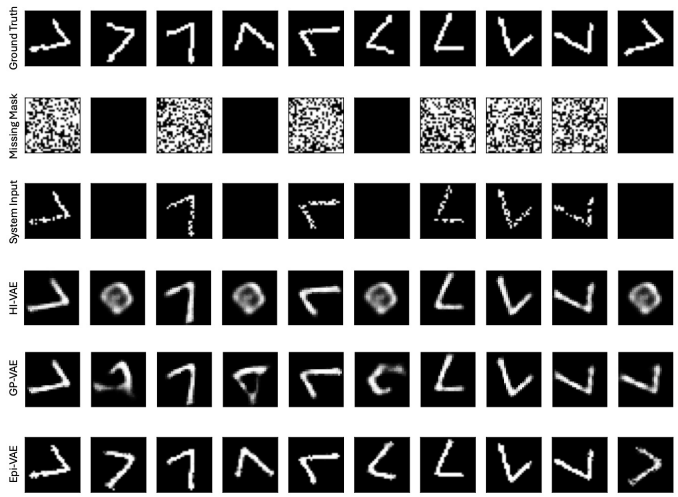


Fig. 2: Visualization of *Epi*MNIST dataset generation process (top three rows) with 40% missingness, both partial and complete, which are represented by columns. Irregular time intervals are represented by varying degrees of rotation between frames (e.g., time steps 1 to 2 and 6 to 7 in the first row). Qualitative imputation results for Hi-VAE, GP-VAE and our Epi-VAE are shown in the bottom three rows. For completely missing time steps, our approach results in the sharpest reconstruction of the handwritten digit. Furthermore, it best captures the underlying time dynamics represented by the random rotation of the digits across columns (e.g., rightmost column).

Following the VAE training paradigm, the parameters p_θ of our generative model and the encoder networks q_ϕ and q_ψ can jointly be trained by maximizing the multi-encoder evidence lower bound

$$\mathcal{L}(\theta, \phi, \psi; x) = \mathbb{E}_{q_\phi(z_l|x)q_\psi(z_g|x)}\left[\log p_\theta(x|z_l, z_g)\right] - \frac{\beta}{2}(\beta_{w_l} \cdot \mathrm{KL}(q_\phi(z_l|x)\|p(z_l)) + \beta_{w_g} \cdot \mathrm{KL}(q_\psi(z_g|x)\|p(z_g))). \tag{3}$$

We evaluate this expression on the observed features. Trade-off parameter β guides the strength of the influence of disentanglement in the latent space [15]. We ensure equal influence of the reconstruction and KL-divergence loss terms over varying degrees of missingness by automatically balancing for the number of observed features through β_w, which represents the ratio of observed features to the dimension of the respective latent space. This decouples the reconstruction loss from the number of observations.

TABLE I: Architecture Details

	*Epi*MNIST	Cohort Study	Benchmark
Encoder	2 x [256]	2 x [128]	2 x [256]
Latent Dimension	256	11	256
Time Convolution Size	3	5	NA
Global Imputer	4 x [256]	4x [11]	NA
Decoder	5 x [256]	3x [128]	3 x [256]
Attention Network	2 x [256]	2 x [11]	NA
Learning rate	1e-3	1e-3	1e-3
Batch Size	256	128	64
Dropout	0.1	0.1	0.0
Beta	0.1	0.2	0.6
Weight Decay	1e-5	1e-5	0.0
Training Epochs	200	2000	20
Runtime [hours]	5	0.2	0.8

Hyperparameters used for training our networks and comparison methods. Benchmark refers to VAE, Hi-VAE and GP-VAE as implemented by [17]. Our implementation uses PyTorch-Lightning 2.2.5. and PyTorch 2.3.1. Training times are reported on a NVIDIA GeForce RTX 4070 Ti GPU.

IV. Experiments

In this section, we show experimental results of our approach on two datasets. First, we quantify imputation performance and compare results to other approaches on a synthetic dataset. Second, we apply our approach to a real epidemiological study following patients through recovery from mycobacterium tuberculosis (TB) infection. We show architecture details in Table I.

A. EpiMNIST

Dataset: In contrast to the random perspective taken by HealingMNIST introduced in Section II-D, changes to an individual patient's health state typically follow a trajectory of recovery or disease development. We therefore propose *Epi*MNIST, a variant with a more structured evolution of health states over time to resemble a multi-year epidemiological cohort study. We construct a time series from a single MNIST digit by rotating it with progressing time and base rotations and missingness on three considerations: 1) observations or study visits are made at defined, unevenly spaced time points; 2) while each patient is unique, some share similar characteristics represented by various versions of the same handwritten digit in MNIST; and 3) several different recovery or disease development journeys are possible for each individual patient. We represent this third consideration by generating a variety of J possible rotations per digit class c. For each digit class, we then randomly assign a trajectory and propagate the image in time according to the given time intervals from the first instance. Additionally, all generated time series start with a random rotation. We apply a missing mask m to each time series which randomly removes a certain share of observations per view and a certain share of views altogether. A representation of this is shown in the first three rows of Fig. 2. The process to generate *Epi*MNIST is described in Algorithm 1. An open-source implementation of this process and our method is available[1].

Results: Table II shows a quantitative comparison of the imputation quality of various approaches on the *Epi*MNIST dataset. The observed mean squared error (MSE) represents the training error, while the imputed MSE represents the deviation of the imputed values from ground truth. To quantify the performance of the reconstruction, we also report the area under the receiving operator characteristics curve (AUROC) for classifying individual-level reconstructions. The last three rows of Fig. 2 depict a qualitative comparison of these approaches. Notably, HI-VAE is only able to reconstruct clouds where data is completely missing while GP-VAE carries forward the last observation when data is missing at the boundary. Our approach is able to infer the particular trajectory with irregular time steps, resulting in the more accurate reconstruction as shown in the last column in Fig. 2 and the AUROC in Table II.

[1]https://github.com/christqoh/epi_imputation

TABLE II: Reconstruction Performance on *Epi*MNIST

Algorithm	MSE observed	MSE imputed	AUROC
VAE [15]	0.0373	0.1120	0.7714
HI-VAE [16]	0.0210	0.1161	0.7824
GP-VAE [17]	0.0153	0.0849	0.8589
Epi-VAE [hard selection]	0.0392	0.0476	0.8952
Epi-VAE [attention]	0.0394	**0.0471**	**0.8969**

Performance of various approaches on *Epi*MNIST with 40% missing values and 40% missed visits. A linear classifier fitted on test reconstructions was used to compute area under receiving operator characteristics. Learned reweighting though the attention network performs only slightly better than hard selection.

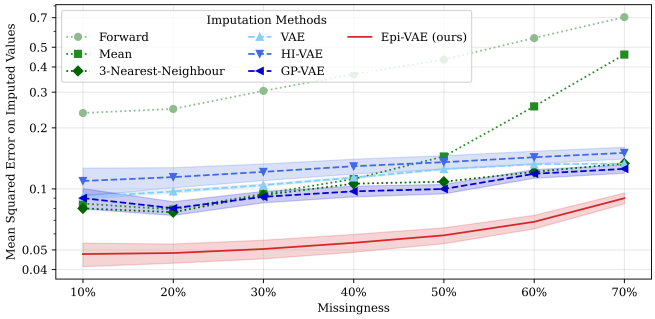


Fig. 3: Comparison of imputation mean squared error and standard deviations (where available) over increasing degrees of random missingness in *Epi*MNIST.

We show the influence of increasing degrees of missingness, which can lead to varying performance of imputation approaches in Figure 3. Additionally, we show the effect of different missingness patterns arising from specific underlying dynamics in the data (e.g., patients withdrawing from the study or constraints in generating a particular measurement) in Table III. Our approach consistently outperforms related methods over both increasing degrees of missingness and different missingness patterns.

B. Longitudinal Tuberculosis Cohort Study

In this section, we present a real-world case study exhibiting the various challenges discussed thus far. It serves to evaluate our proposed approach using clinical domain knowledge due to the absence of ground truth.

Background: TB remains a major global health challenge, with approximately 10 million new cases and 1.6 million

TABLE III: Performance across various Missingness-Patterns and Algorithms

Algorithm / Pattern	Random (a)	Spatial (b)	Temporal (c)	Not at Random (d)
VAE	0.1137 [0.1136, 0.1139]	0.1132 [0.113, 0.1133]	0.1134 [0.1133, 0.1135]	0.1781 [0.178, 0.1782]
HI-VAE	0.1293 [0.1284, 0.1302]	0.1294 [0.1285, 0.1303]	0.1316 [0.1307, 0.1325]	0.1536 [0.1527, 0.1545]
GP-VAE	0.0972 [0.0967, 0.0977]	0.0989 [0.0984, 0.0994]	0.1003 [0.0999, 0.1008]	0.1368 [0.1363, 0.1373]
Epi-VAE	**0.0542** [0.0537, 0.0546]	**0.0528** [0.0524, 0.0533]	**0.0552** [0.0547, 0.0556]	**0.0771** [0.0765, 0.0776]

Performance of several methods across various missingness-patterns. On average, 40% of visits and 40% of observations in the remaining time points are missing. Table reports mean and 99% confidence interval of the mean squared error in the imputed values compared to ground truth. We compare completely at random (a) to spatial (b), representing the fact that if a patient was unable to produce measurement X at time t, they were more likely to be unable to produce the same measurement X at time t+1, (c) temporal increasing with time, representing patients dropping out throughout the study, and (d) missing not at random, where the missingness depends on the underlying data, in our case represented by white pixels being twice as likely to miss compared to black pixels.

deaths worldwide in 2022 [22]. Treatment for active drug-susceptible TB infection typically lasts at least six months and requires regular clinical follow-up and supportive care to monitor response to treatment, manage side effects, and ensure treatment adherence [23]. Additionally, some patients experience residual lung damage which requires ongoing care [24]. Due to the long duration of treatment and recovery periods and the need for consistent monitoring, evaluating post-treatment lung outcomes is often complicated by the previously discussed challenges related to missing data in multi-year longitudinal studies.

Dataset: In the TB dataset used (originating from [25]), patients were recruited at study clinics upon diagnosis with pulmonary TB and followed up for two years. Follow-up times were distributed irregularly throughout the study at 0 and 14 days, and at 2, 4, 6, 9, 12, 18, and 24 months. Baseline characteristics of the cohort as well as the proportion of missing values for key summary statistics are shown in Table IV. In total, 917 time series of at least three recordings per feature were available for analysis. We split our data into train, validation and test set in a 8:1:1 ratio stratified by patient sex, previous TB infection, and HIV status. Subsequently, we train our network under a nine-fold cross-validation scheme and report mean performance of all networks on the test set.

Evaluation: We evaluate our system across three dimensions: 1) alignment of individual-level imputations, 2) overall cohort trends, and 3) robustness to varying degrees of missingness.

Firstly, since medical data is inherently noisy, repeatedly measured parameters can be evaluated with respect to ranges of clinically relevant changes, where out-of-range measures indicate a relevant event. We assume that imputed values between two time points falling within the same range will also lie within that range. Table V shows the share of values fulfilling this condition.

Secondly, besides parameter stability, which focuses on in-

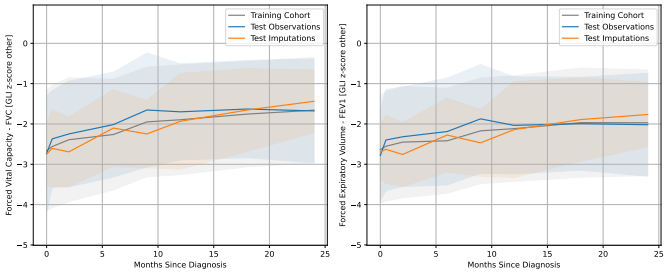
TABLE IV: Epidemiological Study - Summary Statistics

Parameter	Mean	Std. Dev.	Share Missing
HIV Positive [share]	0.3878	-	0.0
Previous TB [share]	0.098	-	0.0
Sex [share female]	0.3344	-	0.0
Height [m]	1.6799	-	0.0
Age at Recruitment [Y]	35.7119	-	0.0
Resting Heart Rate	78.6961	16.6089	0.3252
Blood Pressure Syst.	114.0686	16.3236	0.3335
Blood Pressure Diast.	74.3433	11.3402	0.3262
Forced Vital Capacity [z]	-2.0576	1.4102	0.43
Expiratory Volume 1s [z]	-2.232	1.3203	0.43
Six Minute Walk Test [m]	422.665	76.9861	0.4516
Karnofsky Score	90.2154	7.2623	0.3284
BMI [kg/m2]	20.8818	3.7186	0.3255
St. George's Quest. Score	12.5174	18.1351	0.4161
Forced Vital Capacity [L]	3.0977	0.7782	0.43
Expiratory Volume 1s [L]	2.4139	0.6848	0.43

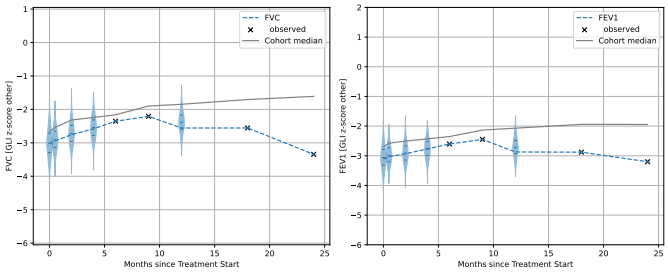
TABLE V: Individual Parameter Imputation Performance

Parameter	Acceptable Range	Imputations Within (attention)	Imputations Within (hard selection)
BMI [kg/m2]	± 2 [26]	**0.8778**	0.8333
Expiratory Volume 1s [L]	± 0.225 [27]	**0.8254**	0.7937
Forced Vital Capacity [L]	± 0.325 [27]	**0.9565**	0.9275
Resting Heart Rate [bpm]	± 10 [28]	0.8298	0.8298
Karnofski Score [Points]	± 10 [29]	0.9868	**1.000**
Six Minute Walk Test [m]	± 54 [30]	**0.9639**	0.9398
Six Minute Walk Test [m]	± 80 [30]	0.9900	0.9900

Relevant medical parameters, associated reference ranges, and share of imputed values within these ranges. Dynamically reweighting latent spaces through attention provides an improvement of up to 5.3% over a hard selection mechanism only relying on the global latent space when observations are completely missing.



(a) Cohort: Imputation mean and distribution over observed and training data.



(b) Individual Example: Black crosses represent observed values. Cohort references are depicted by gray lines. Imputation uncertainty depicted in Violin plots with 25th, 50th, and 75th quantile indicated by horizontal lines.

Fig. 4: Qualitative imputation with respect to cohort trends (a) and a representative individual example (b).

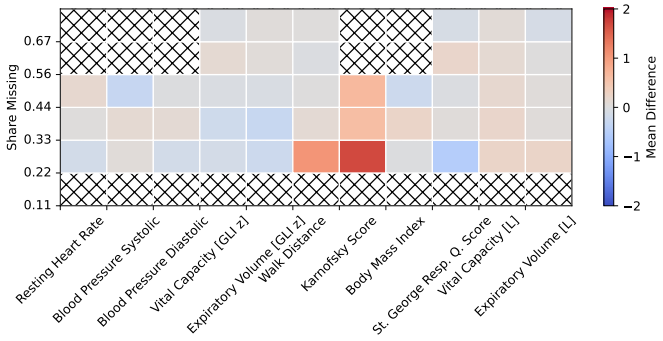


Fig. 5: Visualization of mean difference between imputed and observed values both relative to cohort mean. Values close to zero indicate imputation close to individual patient trajectory. No computation was possible where squares are crossed-out.

dividual trajectories, imputation should capture trends present in the cohort. We illustrate this with the example of spirometry, which was the primary outcome of interest in the study considered. Values were stratified by age, sex, and height according to commonly used reference standards [31] for easier comparability. We show imputed and observed values relative to the overall cohort trend in Fig. 4a and a representative individual example in Fig. 4b. In both instances, the imputation follows the overall cohort trend. Where the individual deviates from the cohort, the imputation captures this as well.

Thirdly, as in the synthetic example, the influence of the share of missing values on the performance of our system is quantified in Fig. 5. The sign of the mean difference indicates whether the imputation tends to over- or underestimate the deviation of an individual from the cohort per observed feature. The system tends to overestimate the Karnofsky score, which is subjectively assigned by a health professional. Overall, however, the deviation of imputed values from the cohort average and observed values for the same individual are similar, suggesting that the imputation aligns well with overall cohort trends independent of the degree of missingness in one feature.

V. Discussion

In this paper, we address the common problem of missing and incomplete data in longitudinal epidemiological studies. We propose an elegant way to convolute several irregularly spaced time steps and combine local and global perspectives of time series data based on the available information at each time step. Our experiments show improved performance compared to other approaches.

Since no synthetic dataset capturing relevant missingness characteristics of longitudinal studies was readily available, we propose the fully parameterizable *Epi*MNIST as a new standard for benchmarking epidemiological imputation algorithms. It focuses on the reconstruction ability where images, which are rotated by an unknown amount, are (up to) fully missing at one time step. Compared to similar methods, our algorithm shows superior imputation performance both qualitatively and quantitatively on this benchmark dataset. Qualitative comparison of our approach to others shows it can impute data accurately, even when an observation is missing completely for a single time step (Fig. 2). We attribute this to our addition of a global latent space and the proposed attention mechanism. Quantitatively, the deviation from ground truth is consistently lower than other approaches across a wide range of degree of missingness (Fig. 3) and various missingness patterns (Table III). Our network reconstructs the digits in our *Epi*MNIST dataset to a quality allowing a classification algorithm to achieve an AUROC of nearly 0.9, outperforming the reconstruction ability of previous methods (Table II). While this suggests strong performance, we further validated our proposed method through a clinical case study.

Applied to a real-world TB dataset, our approach performs well and is capable of handling irregular time intervals between continuous measurements. Imputed values for features expected to be stable over time remain so while values in highly dynamic regions behave according to cohort references. On average, more than 91% of imputed values for all observed features are within the clinically acceptable range (Table V). Furthermore, we adequately capture cohort trends, as visualized in Fig. 2. Imputation is additionally stable over increasing degrees of missingness (Fig. 5).

Having a reliable method to impute missing data in studies evaluating long-term health outcomes during and following TB treatment has several translational advantages. On the individual level, it may help to identify critical periods where patients might be at higher risk for poor outcomes to allow for early intervention or personalized clinical management. On the population level, it can improve insights into the progression of disease and long-term impact on health by reducing bias and increasing reliability of analyses. On the health system level, it allows for a more comprehensive understanding of the burden of TB, providing information for public health planning and resource allocation. Moreover, these advantages can be extended to similar data evaluating long-term health outcomes for other diseases.

Our work has limitations. The global latent space scales quadratically since information from all time steps and input features flows through it concurrently. In the context of epidemiological studies designed to collect a limited number of measurements over time, it is not as much of a concern as in general time series, which consist of more steps. Furthermore, while we address various patterns of missingness, we do not consider noise in the synthetic dataset. As with most synthetic datasets, *Epi*MNIST is a simplified representation of the real-world complexity observed in the health states of individual patients. Additional work on simulating clinically realistic data noise profiles, such as they appear in tests which require active patient participation (e.g., spirometry), can increase the realistic nature of such datasets, thus making them more powerful for validating imputation methods. Generally, imputation is not a replacement for careful study design and conduct and every effort should be made to limit the likelihood of missing and invalid data in the first place. As demonstrated in our real-world example, a purely numeric interpretation is often not sufficient and assessment of results by a medical expert is crucial. Nonetheless, our approach can directly be translated to

many real-world applications and makes a valuable contribution to improve results in longitudinal epidemiological studies with missing data.

REFERENCES

[1] Ibrahim J. G., H. Chu, and M. H. Chen. "Missing data in clinical studies: issues and methods." In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 30.26 (Jan. 2012), pp. 3297–3303.

[2] Jos Twisk and Wieke de Vente. "Attrition in longitudinal studies: How to deal with missing data". In: *Journal of Clinical Epidemiology* 55.4 (2002), pp. 329–337.

[3] John M. Lachin. "Fallacies of last observation carried forward analyses". In: *Clin Trials* 13 (2016).

[4] L. Beretta and A. Santaniello. "Nearest neighbor imputation algorithms: a critical evaluation". In: *BMC Medical Informatics Decision Making* 16 (2016).

[5] Jun Wang et al. *Deep Learning for Multivariate Time Series Imputation: A Survey*. 2024. arXiv: 2402.04059.

[6] Yonghong Luo et al. "Multivariate Time Series Imputation with Generative Adversarial Networks". In: *Neural Information Processing Systems*. 2018.

[7] Bazo-Alvarez JC et al. "Current Practices in Missing Data Handling for Interrupted Time Series Studies Performed on Individual-Level Data: A Scoping Review in Health Research". In: *Journal of Clinical Epidemiology* 13.4 (2021), pp. 603–613.

[8] Faraj A. A. Bashir and Hua-Liang Wei. "Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm". In: *Neurocomputing* 276 (2018), pp. 23–30.

[9] Stef van Buren. *Flexible Imputation of Missing Data*. Chapman & Hall / CRC, 2018.

[10] Wei Cao et al. "BRITS: Bidirectional Recurrent Imputation for Time Series". In: *Neural Information Processing Systems*. 2018.

[11] Parikshit Bansal, Prathamesh Deshpande, and Sunita Sarawagi. "Missing value imputation on multidimensional time series". In: *Proc. VLDB Endow.* 14.11 (July 2021), pp. 2533–2545. ISSN: 2150-8097.

[12] Yonghong Luo et al. "Multivariate Time Series Imputation with Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[13] Ranjit Lall and Thomas Robinson. "The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning". In: *Political Analysis* 30.2 (2022), pp. 179–196.

[14] Ricardo Cardoso Pereira et al. "Reviewing Autoencoders for Missing Data Imputation: Technical Trends, Applications and Outcomes". In: *J. Artif. Intell. Res.* 69 (2020), pp. 1255–1285.

[15] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations*. 2017.

[16] Alfredo Nazábal et al. "Handling incomplete heterogeneous data using VAEs". In: *Pattern Recognition* 107 (2020), p. 107501.

[17] Vincent Fortuin and Dmitry Baranchuk. "GP-VAE: Deep Probabilistic Multivariate Time Series Imputation". In: *PMLR*. Vol. 108. 2020.

[18] Steven Cheng-Xian Li and Benjamin M. Marlin. "Learning from irregularly-sampled time series: a missing data perspective". In: *Proceedings of the 37th International Conference on Machine Learning*. 2020.

[19] A. L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". In: *Circulation* 101.23 (June 2000), e215–e220.

[20] Rahul G. Krishnan, Uri Shalit, and David Sontag. *Deep Kalman Filters*. 2015. arXiv: 1511.05121.

[21] L. Ternes, M. Dane, and S. et al Gross. "A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis". In: *Commun Biol* 60.5 (2022), p. 255.

[22] World Health Organization. *Global Tuberculosis Report*. Online, 2023.

[23] World Health Organization. *WHO consolidated guidelines on tuberculosis: Module 4: Treatment - Drug-susceptible tuberculosis treatment*. Online, 2022.

[24] Taylor J et al. "Residual respiratory disability after successful treatment of pulmonary tuberculosis: a systematic review and meta-analysis." In: *EClinicalMedicine* (May 2023).

[25] Rachow A et al. "TB sequel: incidence, pathogenesis and risk factors of long-term medical and social sequelae of pulmonary TB - a study protocol." In: *BMC Pulm Med.* (Jan. 2019).

[26] Panza E et al. "Changes in body weight and glycemic control in association with COVID-19 Shutdown among 23,000 adults with type 2 diabetes". In: *Acta Diabetol* 60.6 (2023), pp. 787–795.

[27] Herpel LB et al. "Variability of spirometry in chronic obstructive pulmonary disease: results from two clinical trials". In: *Am J Respir Crit Care Med* 173.10 (2006), pp. 1106–1113.

[28] Michel Cucherat. "Quantitative relationship between resting heart rate reduction and magnitude of clinical benefits in post-myocardial infarction: a meta-regression of randomized clinical trials". In: *European Heart Journal* 28.24 (Nov. 2007), pp. 3012–3019.

[29] Mouelhi Y et al. "How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods." In: *Health Qual Life Outcomes* 18.1 (Nov. 2020), p. 136.

[30] Wise RA and Brown CD. "Minimal clinically important differences in the six-minute walk test and the incremental shuttle walking test". In: *COPD* 2.1 (2005), pp. 125–129.

[31] Quanjer PH et al. "Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations." In: *Eur Respir Journal* 40.6 (Dec. 2012), pp. 1324–43.