Benchmarking Missing Data Imputation Methods for Time Series Using Real-World Test Cases

Adedolapo Aishat Toye ATOYE@STEVENS.EDU
Asuman Celik RCELIK1@STEVENS.EDU
Samantha Kleinberg SAMANTHA.KLEINBERG@STEVENS.EDU
Department of Computer Science, Stevens Institute of Technology, USA.

Abstract

Missing data is pervasive in healthcare. Many imputation methods exist to fill in missing values, yet most were evaluated using randomly deleted values rather than the actual mechanisms they were designed to address. We aimed to determine real-world accuracy for missing data imputation with three missing data mechanisms (missing completely at random, MCAR; missing at random, MAR; and not missing at random, NMAR) for state of the art and commonly used imputation methods. Using two time series data targets (continuous glucose monitoring, Loop dataset; heart rate, All of Us dataset) we simulated missingness by masking values for each mechanism, at a range of missingness percentages (5-30%) and tested 12 imputation methods. We evaluated accuracy with multiple metrics including root mean square error (RMSE) and bias. We found that overall, accuracy was significantly better on MCAR than on MAR and NMAR, despite many methods being developed for those mechanisms. Linear interpolation had the lowest RMSE with all mechanisms and for all demographic groups, with low bias. This study shows that current evaluation practices do not provide an accurate picture of real world performance with realistic patterns of missingness. Future research is needed to develop evaluation practices that better capture real-world accuracy, and methods that better address real-world mechanisms.

Data and Code Availability Loop (Lum et al., 2021) is publicly available upon agreement with the dataset authors. The All of Us dataset is available upon registration and agreement with the All of Us Research Program (All of Us Research Program Investigators, 2019). The code is available at https://github.com/health-ai-lab/benchmarking-missing-imputation-methods.

Institutional Review Board (IRB) This study was approved as exempt by the IRB at Stevens Institute of Technology.

1. Introduction

Missing data is common in health research, such as in electronic health records (EHRs) or patient generated data (Wells et al., 2013). Missing values or variables in healthcare may be due to device malfunctions, sensors not measuring or recording data, irregular sampling intervals, a patient's health status, and omissions on surveys due to memory or social desirability (Kwak et al., 2021). For example, continuous glucose monitors (CGM) can fail to record data when a sensor is compressed such as during sleep (Facchinetti et al., 2016). Similarly, a doctor may test a patient's HbA1c more often when it is above reference ranges. In both cases, values are not missing completely at random (MCAR), meaning that missingness is independent of any other variables. Rather, the first case is missing at random (MAR), where the probability of a value being missing depends on other variables (here, sleep), and the second case is not missing at random (NMAR), where the probability of a variable being missing depends on the variable itself (here, HbA1c). Missing data in healthcare is rarely MCAR, and missingness has even been used as a feature to improve classification accuracy (Che et al., 2018).

Strategies for handling missing data depend on the mechanism and task (Sperrin et al., 2020). While MCAR can be ignored without leading to biased inferences, using only complete records reduces study power (Thomas et al., 2022), which is a challenge for machine learning methods that require large datasets. For MAR and NMAR, analyzing only complete records can lead to incorrect inferences, failed inferences (false negatives), and failures when models

are applied to new settings with different patterns of missingness (Zhou et al., 2023; Getzen et al., 2023).

Many methods have been developed for imputing missing values under MAR, NMAR, or combinations of these mechanisms. However, imputation methods are commonly evaluated by randomly deleting a percentage of values and comparing imputed results to the ground truth. As described in methods, we conducted a review of 205 papers on imputation and found that the majority used this evaluation approach (random dropout). A recent review found similar results (Ren et al., 2023). Yet this assumes the mechanism is MCAR, and it is thus unknown whether similar results will be obtained for MAR and NMAR. Intuitively, having an hour total of missing CGM data where every other data point is deleted is less problematic for imputation and downstream tasks compared to having an hour-long window deleted. Previous work showed that larger CGM gaps led to worse results when calculating glucose metrics (Smith et al., 2023). Imputation quality has a significant impact on downstream ML tasks using healthcare data, making understanding accuracy highly important (Shadbahr et al., 2023; Payrovnaziri et al., 2021). These works simulated missingness using random dropout, so may still understate the limitations of current methods. Ultimately, predicting real-world performance from benchmarks depends on how closely the data reflects real-world data scenarios (Morris et al., 2019).

Prior work shows that missing data mechanisms matter, yet there remains a gap in understanding how methods perform under each mechanism. Additionally, limited research has examined how imputation performance varies across subgroups (e.g., by gender), or quantified the bias they may introduce. To address this, we benchmarked 12 single and multiple imputation methods on two types of commonly used time series health data (CGM, physical activity). We examined imputation accuracy 1) across methods using multiple metrics including bias and with gender-stratified data, 2) by missingness percentage, and 3) by mechanism (MCAR, MAR, NMAR). Our results show the need for rigorous evaluation of imputation methods in health settings, highlighting potential bias and performance differences.

2. Related Work

2.1. Missing Data Imputation Methods

Mean and mode imputation (replacing missing instances with a variable's mean or mode) are simple approaches that remain popular in some settings. However, under MAR or NMAR the observed mean and mode may not be representative of the true range of a variable (e.g., a test being done only on sicker patients), and some variables may have a bimodal distribution (e.g., modal height for men and for women). Last observation carried forward (LOCF) can be effective with variables that are recorded only when they change (e.g., insulin pump values being recorded for new settings) but has also been used with variables that change more frequently than they are measured (e.g., bodyweight) (Gadbury et al., 2003).

Other methods have been developed to address each mechanism, with the majority being for MAR. Since MAR missingness depends on other variables, k-nearest neighbors (kNN) imputation identifies neighbors of a missing data point based on observed features and averages their values for the target variable (Pujianto et al., 2019), but cannot be used when all data is missing for a timepoint. Deep learning methods have also been introduced for imputation in healthcare data (Phung et al., 2019; Liu et al., 2023). Multi-directional Recurrent Neural Network (MRNN) is a multiple imputation (MI) approach (multiple values are imputed for each missing point) that outperformed the state of the art when applied to medical datasets including MIMIC-III and UK Biobank (Yoon et al., 2018). However, evaluation was done on the downstream task of prediction and with random dropout. GP-VAE combines variational autoencoders (VAE) with Gaussian processes (GP) to address both MAR and NMAR and was applied to Physionet data (Fortuin et al., 2020) and shown to avoid bias in other healthcare data (Liu et al., 2023). Despite advances in the imputation methods, there still remains a gap in evaluating their performance in real-world health data scenarios, leading to a knowledge gap for machine learning practices in healthcare.

2.2. Imputation Methods Benchmarking

Given the importance of missing data in healthcare, multiple works have benchmarked imputation algorithms (Vahdati et al., 2024), though have mainly done so using random dropout on vital signs (Le et al., 2018), laboratory data (Luo, 2022; Waljee

et al., 2013), genomic data (Shadbahr et al., 2023), and EHR data (Hegde et al., 2019). However, this approach only predicts real world accuracy if data is MCAR. Other works have focused narrowly such as on deep learning based methods (Kazijevs and Samad, 2023), only NMAR (Pereira et al., 2024), or only on multiple imputation (Kontopantelis et al., 2017). Beaulieu-Jones et al. (2018) generated four different mechanisms and evaluated 6 variants of MICE and 6 common methods for imputing missing laboratory values. However, as lab values tend to be infrequent it is an open question whether results will be similar for more high frequency time series data.

Existing benchmarking studies also mainly rely on Root Mean Squared Error (RMSE) to evaluate imputation performance (Jäger et al., 2021; Jadhav et al., 2019; Prakash et al., 2024). However, RMSE is sensitive to outliers and does not capture direction of error, variability, or subgroup disparities. Metrics such as bias, empirical standard error (EmpSE), and coverage probability as proposed by Morris et al. (2019) provide more comprehensive evaluation but have only been used in few studies. For example, Pan and Chen (2023) evaluated three imputation methods under MAR and NMAR, using different metrics including bias and standard error, but only on cross-sectional data. Further research is needed to robustly evaluate more methods on health time series data across all missingness mechanisms and percentages.

3. Methods

3.1. Data

We used two health-related datasets collected in free-living settings. We focused on CGM and physical activity data as both are increasingly used in machine learning healthcare research, yet both face high degrees of missing data when used in the real world.

3.1.1. Loop

The Loop dataset (Lum et al., 2021) was collected during a study evaluating the effectiveness of an automated insulin delivery system in people with Type 1 diabetes (T1D). The dataset includes CGM (every 5 mins measured by Dexcom or Medtronic CGM), activity (Apple watch, activity data from HealthKit), meals (meal time, grams of carbohydrates), insulin (bolus and basal), and other demographic data collected for over 6 months for 558 individuals (adults and children) with T1D. For this study, we used data from 21 participants (7 women, 11 men, 3 not specified) who had sufficient data across the variables (activity, meals) to simulate the three mechanisms. Participants were mean 18.11 $\pm$ 13.84 years of age, and had a mean of 416.71 $\pm$ 101.35 days of CGM data.

3.1.2. All of Us

The All of Us dataset contains data contributed by participants aged 18 and older as part of the All of Us Research Program, an initiative designed to collect health-related data from diverse populations in the United States (All of Us Research Program Investigators, 2019). The data contains electronic health records, survey responses, genomics, and activity data recorded using a Fitbit smartwatch. We focused on the Fitbit activity data, which contains heart rate (HR) and step count recorded at one-minute intervals for over 15,000 participants. For this study, we used data from 40 participants (20 women, 19 men, 1 not specified) who had at least 10,000 HR values (roughly 1 week of data). We used up to 3 months of data for each individual. Participants were mean 48.48 $\pm$ 16.65 years of age, and had a mean of 70.23 $\pm$ 26.98 days of HR data for analysis.

3.2. Missing Mechanisms

We simulated each mechanism (MCAR, MAR, and NMAR) for various target percentages ranging from 5-30% in increments of 5%. For Loop, the target variable was CGM, and for All of Us, it was HR.

MCAR We simulated this mechanism for both datasets by randomly deleting values in the target variable according to each probability (5%, 10%, 15%, 20%, 25%, and 30%). We used a Bernoulli trial to randomly select instances to delete to ensure that the missing data was truly random, with no underlying pattern linking the missing values to any observed or unobserved factors.

MAR This mechanism was simulated by having a variable other than the target trigger missing instances. To replicate realistic situations, values were deleted for a window of time based on the properties of each dataset. For Loop, meals and exercise were used to probabilistically trigger missing values, replicating the situation where sensors fail to transmit data during more active times. After determining a missing value should be created, we then sampled a duration for it from the distribution of previously

observed missing glucose durations (native missingness). We then deleted glucose values for the sampled duration after the meal event. We followed the same process for All of Us, with step count used to trigger missingness, simulating the situation where activity causes a disruption in sensor recording. For Loop, due to the limited number of meal and activity events, we were able to generate up to only 15% missingness while maintaining realistic durations. For All of Us, the full set of percentages (up to 30%) were generated.

NMAR In this mechanism, missingness was triggered by the target variable's value. For Loop, we used CGM values below 70 mg/dL or above 150 mg/dL. These thresholds were selected based on clinical reference values for high and low glucose levels (Kapoor et al., 2020), to create a more challenging cases where clinically meaningful values are absent. We used the same logic for All of Us, with HR thresholds below 60 bpm or above 100 bpm, based on reference ranges for HR (Cheung et al., 2015; Gopinathannair and Olshansky, 2015). The duration was sampled as for MAR. For both datasets this approach was used to introduce missingness percentages of 5%, 10%, 15%, 20%, 25%, and 30%.

3.3. Missing Data Imputation Methods

To determine methods to include in our experiments, we reviewed 205 papers that focused on developing or describing imputation methods. We selected 62 that used time series data and categorized each method based on: missingness mechanism, type of data (e.g., categorical, continuous), and evaluation metrics and approach. From this, we narrowed our focus to continuous-valued time series data and aimed to balance 1) methods addressing each mechanism (MAR, NMAR, combination), 2) reported accuracy, and 3) commonly used methods even if they do not meet other criteria. See Figure 5 (in Appendix) for the flowchart showing the methodology for selecting the imputation methods used in this study, based on existing literature.

This process resulted in 12 methods selected for experiments across four main categories: developed for MAR (MRNN, MICE, kNN, Hot Deck), developed for NMAR (Fourier Transform), developed for MAR and NMAR (GP-VAE, fl-kNN), and commonly used methods (linear interpolation, mean imputation, mode imputation, LOCF, and missForest) as they are still widely used due to their simplicity and computational efficiency. For all methods, we used 5-fold cross validation to evaluate the methods, and report the average across the folds. For the training-based methods (MRNN and GPVAE), we further divided the training subset by individuals into 85% training and 15% validation. Details of the implementation of each approach are described below.

MRNN MRNN uses recurrent neural networks (RNNs) to model missing data and imputes missing values by learning temporal relationships between observed data points (Yoon et al., 2018). We used the implementation provided in the PyPOTS python package with default parameters (Du, 2023).

MICE This is a commonly used statistical method for multiple imputation. It iteratively fills in missing values through chained regression models (Azur et al., 2011). Each missing value is imputed multiple times, creating several complete datasets which are then combined to account for uncertainty of the imputation process. We used IterativeImputer in scikit-learn with default parameters (Pedregosa et al., 2011), which uses Bayesian Ridge Regression for imputation.

kNN This is another commonly used approach that assumes similar instances have similar missing values, and thus averages values of the k-nearest neighbors (Pujianto et al., 2019). We used KNNImputer in scikit-learn (Pedregosa et al., 2011) with default parameters, setting k=5.

Hot deck This is kNN with k=1. Thus, we used KNNImputer with k=1.

Fourier transform This is one of the few methods developed for NMAR. It reconstructs missing values by exploiting regularity and periodicity in the data. We converted the Fourier Transform MATLAB implementation of prior work (Rahman et al., 2015) to Python, so that it could be used in the All of Us workbench.

GP-VAE We used the PyPOTS implementation with default parameters (Du, 2023).

fl-kNN This approach addresses both MAR and NMAR by combining a lagged version of kNN with the Fourier transform. We used the published MATLAB implementation with default parameters (Rahman et al., 2015). This method could not be applied to All of Us data as MATLAB is not available in the workbench.

Linear Interpolation This approach is commonly used for handling missing CGM data (Fonda et al., 2013) and imputes missing values by drawing a straight line between values immediately before and after the missing values assuming a constant rate of change between the points. We implemented this approach using the interpolate function in the pandas library (pandas development team, 2020; McKinney, 2010).

Mean Imputation This is a commonly used baseline method (Acuna et al., 2023) that replaces missing values with the mean of non-missing values. We used SimpleImputer from the scikit-learn library with strategy set to 'mean' (Pedregosa et al., 2011).

Mode Imputation While more often used for categorical variables, this is a common baseline (Jäger et al., 2021) that replaces missing instances by the mode of observed instances. We used SimpleImputer from the scikit-learn library with strategy set to 'most_frequent' (Pedregosa et al., 2011).

LOCF This approach propagates the last known value for a variable until there is a new data point. While it is most applicable to variables with stable values, it is commonly used in health research (Batterham et al., 2013; van Rossum et al., 2023). We used the forward fill method (ffill) with the pandas 'fillna' function (pandas development team, 2020; McKinney, 2010).

missForest This is a non-parametric imputation method that uses random forest models to predict missing values (Stekhoven and Bühlmann, 2012). We used the missForest package in Python (Hindy et al., 2024).

4. Evaluation Metrics

We used multiple metrics to evaluate accuracy of the imputation methods, and provide insight into generalizability of results.

Root Mean Squared Error (RMSE) is commonly used to report the accuracy of missing data imputation methods. We calculated RMSE first across all individuals and then separately for demographic subgroups (i.e., by gender) to assess fairness.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{1}$$

Bias quantifies the systematic direction of error between imputed and actual values (Morris et al., 2019). Since RMSE does not consider direction of errors, this is an important additional metric that indicates if imputed values are systematically over or underestimated.

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{2}$$

Empirical Standard Error (EmpSE) measures the square root of the variance of the predicted imputed values around their mean (Morris et al., 2019). It helps to assess the consistency of the imputation method by indicating how much the imputed values vary across data points.

$$\text{EmpSE} = \sqrt{\text{Var}(\hat{y})} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} \tag{3}$$

where $\bar{\hat{y}}$ is the mean of the predictions.

5. Results

5.1. Comparison of accuracy across methods

Figures 1 and 2 show the mean RMSE values of all imputation methods for all missing mechanisms and percentages for the Loop and All of Us datasets respectively (see Tables 3 and 4 in Appendix for numerical results). On both datasets, linear interpolation had the lowest RMSE for all mechanisms and percentages. We compared each linear interpolation result to the next best method (LOCF) using a t-test. Linear interpolation was significantly better than LOCF for all mechanisms and percentages for All of Us ($p \leq 0.001$) and Loop (MCAR: $p < 0.05$, NMAR and MAR $p < 0.001$), except for MCAR at 5% and 10% ($p = 0.19$ and 0.06 respectively).

Tables 1 and 2 provide the bias values. Linear interpolation had the lowest bias for many mechanisms and percentages on both datasets, suggesting that errors are not systematically positive or negative. For All of Us, linear interpolation had lower bias than LOCF (which was next best in RMSE) for all mechanisms and percentages (all $p < 0.01$) except NMAR at 5% and MCAR at 10% - 30 % (all $p > 0.5$). For loop, linear interpolation again had significantly lower bias than LOCF across all percentages for NMAR and MAR (all $p < 0.005$), but no significant difference was observed for MCAR across all percentages (all $p > 0.5$). While in some cases Fourier transform had

Figure 1: RMSE for all methods, mechanisms, and missing data percentages for the Loop dataset.

Mechanism		MAR				NMAR	MAR+NMAR		Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	Fl-KNN	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	-0.21	-0.31	0.88	24.82	0.04	-2.04	3.53	**0.02**	-0.04	-18.49	0.05	-0.48
	10	-0.19	-0.31	-1.46	23.95	-0.02	-1.64	2.90	**0.01**	0.04	-18.34	**0.01**	-0.49
	15	-2.28	-0.36	3.91	26.52	-0.04	-3.26	1.52	-0.01	-0.14	-22.21	**-0.00**	-0.54
	20	-1.05	-0.36	-0.90	25.84	-0.06	-1.91	2.41	**-0.00**	-0.15	-22.05	**-0.00**	-0.53
	25	-2.12	-0.22	7.65	26.12	-0.06	-1.60	2.60	**-0.00**	**0.00**	-22.44	0.01	-0.42
	30	-2.12	-0.20	-5.40	26.99	-0.05	-3.27	2.19	**0.01**	0.03	-21.86	0.02	-0.44
NMAR	5	-36.92	-35.38	-37.12	-13.43	0.86	-18.5	-13.52	**-0.48**	-37.04	-54.4	3.67	-34.23
	10	-37.91	-35.75	-30.05	-9.60	0.59	-19.31	-15.32	**-0.05**	-37.37	-52.73	3.40	-34.55
	15	-39.89	-36.31	-31.70	-7.09	0.56	-20.56	-13.65	**-0.18**	-38.39	-55.18	3.06	-35.23
	20	-40.26	-36.60	-34.44	-6.32	0.50	-20.12	-14.66	**-0.16**	-38.84	-53.50	3.10	-35.55
	25	-39.49	-37.21	-37.91	-4.94	0.86	-19.14	-15.24	**-0.06**	-39.60	-52.36	2.92	-36.30
	30	-39.39	-36.71	-39.35	-3.54	1.43	-18.07	-11.67	**0.19**	-39.50	-51.12	2.94	-35.98
MAR	5.0	**0.43**	2.06	1.51	22.72	-2.18	-1.92	5.91	2.14	1.55	-17.08	-3.3	1.13
	10.0	3.67	4.58	6.11	23.04	-1.63	**1.08**	5.05	1.70	4.73	-14.38	-3.46	3.11
	15.0	2.22	4.18	1.85	23.02	-2.72	-1.67	5.75	**1.57**	3.44	-17.73	-3.91	2.61

Table 1: The table shows the mean bias/directional error (mg/dL) on the Loop dataset for all the methods. The best result for each row is bolded.

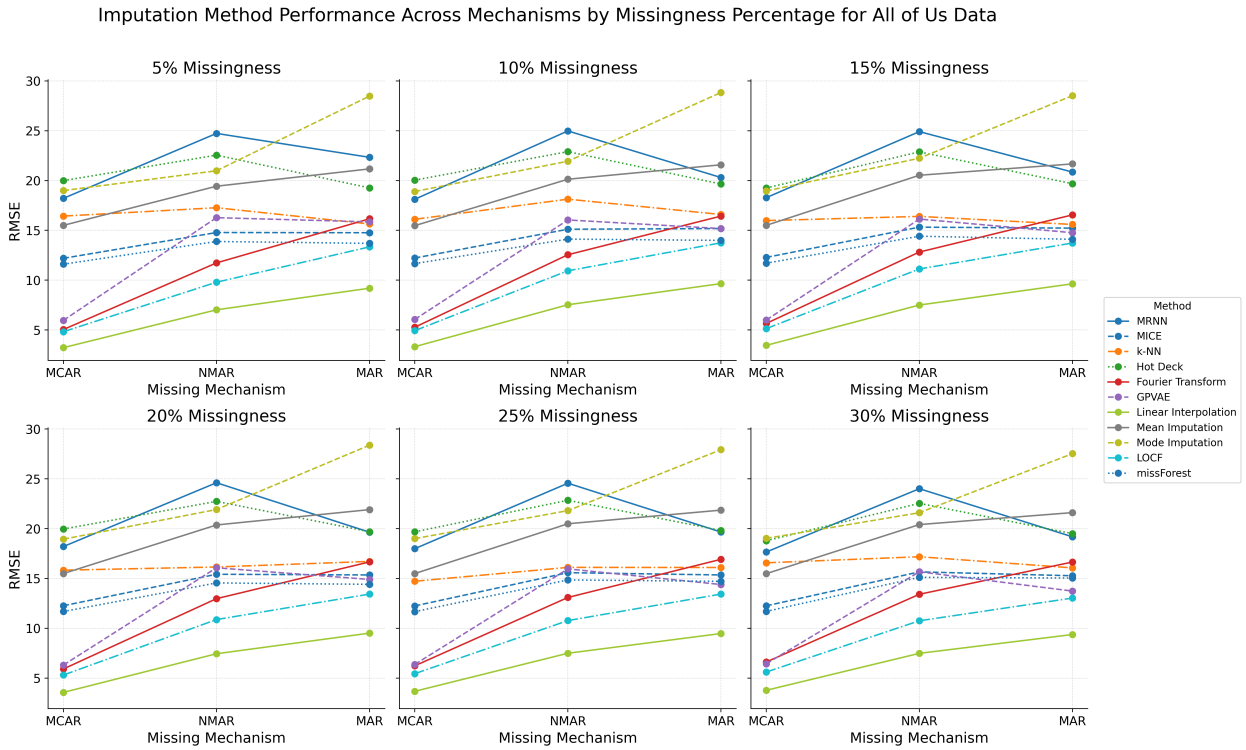

Figure 2: RMSE for all methods, mechanisms, and missing data percentages for the All of Us dataset.

Mechanism		MAR				NMAR	MAR+NMAR	Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	0.27	0.18	0.89	7.48	-0.08	-0.56	**-0.00**	0.15	-9.84	-0.04	0.20
	10	2.09	0.06	-0.03	7.84	**-0.01**	-0.37	**0.01**	0.04	-9.67	0.02	0.11
	15	2.71	0.02	0.97	6.12	-0.03	-0.42	**0.00**	0.03	-9.84	**0.00**	0.08
	20	2.35	0.04	0.40	7.57	-0.05	-0.41	**0.00**	0.04	-9.76	-0.01	0.11
	25	3.06	0.02	1.22	7.36	-0.04	-0.55	**-0.00**	-0.01	-9.87	**-0.00**	0.09
	30	1.96	0.02	4.00	5.82	-0.04	-0.49	**-0.01**	0.02	-9.82	**-0.01**	0.08
NMAR	5	1.81	1.64	2.45	9.03	0.20	0.30	**-0.01**	0.99	-8.50	0.56	1.58
	10	2.30	1.60	3.09	9.19	**-0.14**	-0.23	0.47	0.51	-8.83	1.39	1.53
	15	2.51	1.52	1.59	9.01	**-0.06**	-0.21	0.39	0.36	-8.61	1.27	1.43
	20	3.65	1.63	2.07	8.90	**0.15**	0.22	0.41	0.61	-7.81	1.20	1.50
	25	3.24	1.67	1.73	8.19	**-0.03**	**0.03**	0.32	0.53	-7.40	1.04	1.40
	30	4.03	1.55	1.97	8.04	0.18	**-0.07**	0.27	0.43	-6.63	0.99	1.18
MAR	5	-10.94	-6.10	-3.75	-1.07	-0.75	-7.31	**0.32**	-12.25	-22.46	2.48	-5.01
	10	-7.42	-6.53	-6.91	-1.02	-1.15	-7.14	**0.42**	-13.16	-22.82	2.42	-5.45
	15	-8.09	-6.68	-4.97	-0.53	-0.86	-6.67	**0.36**	-13.37	-22.51	2.15	-5.82
	20	-4.50	-6.99	-6.01	-1.52	-1.05	-7.15	**0.35**	-13.80	-22.43	2.01	-6.53
	25	-3.12	-7.16	-6.03	-0.92	-0.98	-6.64	**0.34**	-13.93	-21.97	1.94	-7.21
	30	1.71	-7.21	-5.39	-0.75	-0.84	-6.07	**0.27**	-13.84	-21.66	1.60	-7.92

Table 2: The table shows the mean bias/directional error (bpm) on the All of Us dataset for all the methods. The best result for each row is bolded.

the lowest bias, this is at the expense of higher RMSE overall. On the other hand, many other methods had large negative bias values for NMAR on Loop, and for MAR on All of Us. The difference in bias across datasets and mechanisms shows the need for testing in a target domain with realistic mechanisms.

Tables 5 and 6 (in Appendix) provide the EmpSE values for both datasets. Fourier Transform had the highest EmpSE for most of the mechanisms and percentages suggesting high variability in the imputed values. Mean and mode imputation had the lowest EmpSE suggesting that there is low variability in the imputed values which is expected given that these methods impute missing values with a constant value (mean or mode). Since EmpSE is based on the imputed values, a low EmpSE does not necessarily indicate high accuracy, but rather consistency.

5.2. Effect of missingness mechanism on imputation accuracy

Many methods performed better under MCAR compared to MAR and NMAR. For All of Us, results on MCAR were significantly better than MAR (all $p < 0.05$) across all methods and percentages, except for MRNN at 20% ($p = 0.06$), k-NN at all percentages ($p > 0.14$), and Hotdeck at all percentages ($p > 0.47$), where differences between MAR and MCAR were not statistically significant. Results on MCAR were also significantly better than those for NMAR for all methods and percentages (all $p < 0.05$) except for k-NN and mode imputation where differences were not statistically significant (all $p > 0.07$). For Loop, results on MCAR were significantly better than for MAR (all $p < 0.002$) for five methods: Fourier Transform, GPVAE, fl-kNN, linear interpolation, and LOCF. Differences for the other methods were not statistically significant (all $p > 0.25$). Results on MCAR were also significantly better than those for NMAR across all methods and percentages (all $p < 0.05$), except for hotdeck (all percentages).

Examining the mechanism each method was developed for, performance was not consistently better in that condition. For the All of Us dataset, MICE and kNN did not have statistically significant differences in performance between MAR and NMAR (all $p > 0.13$). Hotdeck performed better on MAR than NMAR at all percentages (all $p < 0.02$), and MRNN performed better under MAR for 10-30% missing ($p < 0.001$) but the difference was not significant at 5% ($p = 0.29$). For the Loop dataset, hotdeck did not have statistically significant differences in performance between MAR and NMAR ($p > 0.50$), while MICE had better performance at all percentages (all $p < 0.001$). MRNN had better performance on MAR at all percentages, while kNN had better performance on MAR at 10% and 15% missing (all $p < 0.05$) but not for 5% missing ($p = 0.18$). The NMAR method (Fourier Transform) performed significantly better under NMAR than MAR for All of Us ($p < 0.001$), and Loop at 5% ($p = 0.001$). The difference in performance was not significant at 10% and 15% for Loop ($p = 0.12$ and 0.095 respectively).

5.3. Effect of missingness percentage on imputation of accuracy

For many methods, the best performance within each mechanism was observed at 5% missingness. Linear interpolation performed the best across all percentages, while other methods varied in their rankings for each mechanism. For both datasets, hotdeck had the highest RMSE across all percentages for MCAR. For All of Us, MRNN and mode imputation had the highest RMSE for NMAR, and MAR respectively. For Loop, mode imputation and hotdeck had the highest RMSE for NMAR and MAR respectively.

5.4. Accuracy across subgroups

We next examined RMSE by demographics (men, women). Linear interpolation again had the lowest RMSE for all mechanisms and percentages for both groups on both datasets. Figures 3 and 4 show the difference in mean RMSE values between gender subgroups for all methods, mechanisms, and percentages for the Loop and All of Us datasets respectively. (See Tables 7 and 10 in Appendix for numerical results). For All of Us, linear interpolation had significantly higher accuracy than LOCF, for all mechanisms and percentages across both gender groups (all $p < 0.009$). Similarly, for Loop, linear interpolation had significantly lower RMSE for men than LOCF across all percentages for NMAR and MAR (all $p < 0.007$), but no significant difference was observed for MCAR across all percentages (all $p > 0.09$). However, for women, linear interpolation was significantly better than LOCF across all mechanisms and percentages (MCAR: all $p < 0.02$, NMAR: all $p < 0.0001$, and MAR all $p < 0.008$).

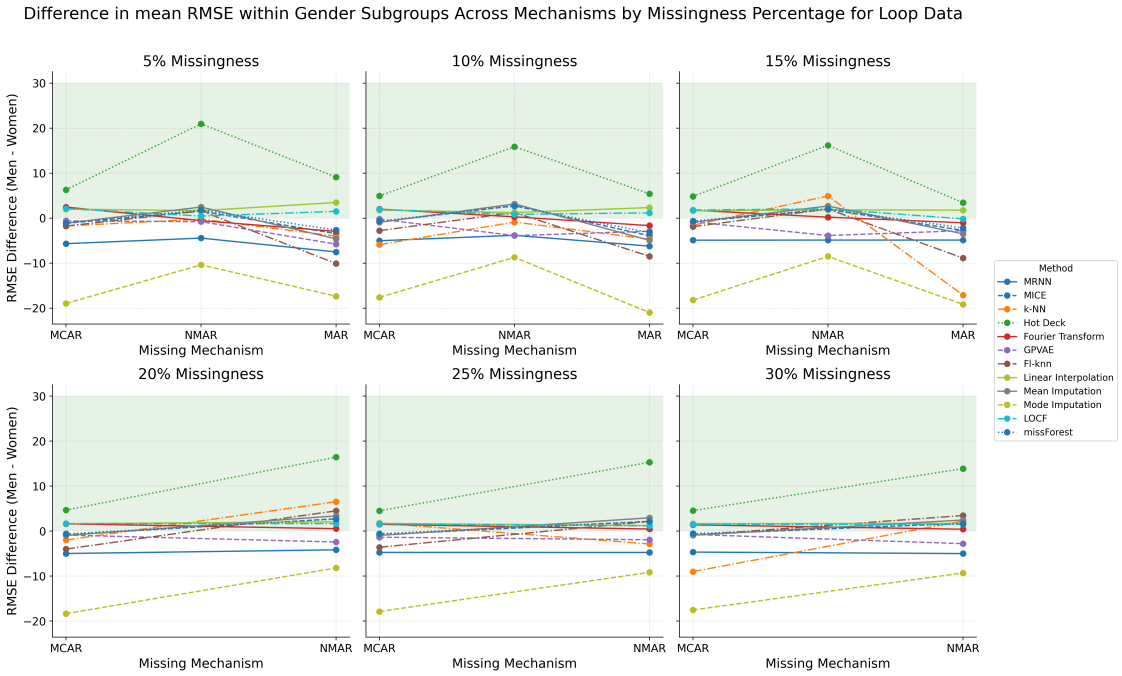

Figure 3: Difference in Root Mean Squared Error (RMSE) by gender for the Loop dataset.

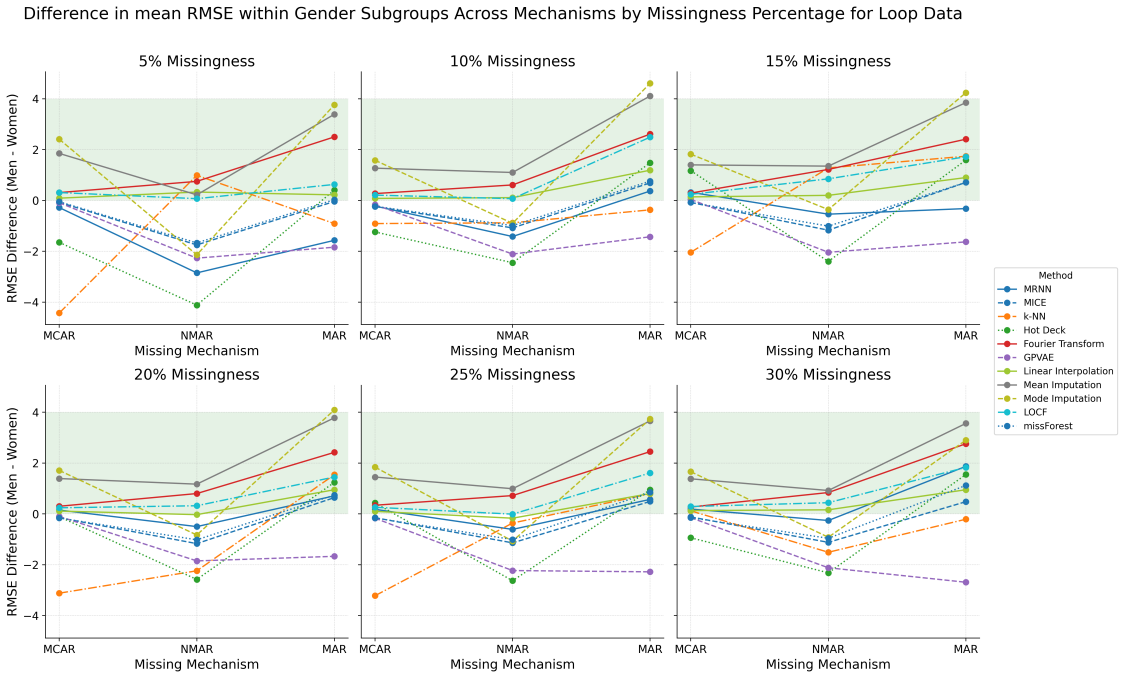

Figure 4: Difference in Root Mean Squared Error (RMSE) by gender for the All of Us dataset.

6. Discussion

Imputation is a key strategy for handling missing data but there has been a lack of evaluation across mechanisms and imputation methods for time series health data. We evaluated 12 methods on two datasets (Loop, All of Us) across varied percentages of missingness and three mechanisms (MCAR, MAR, NMAR). We found that most methods performed worse on MAR and NMAR data compared to MCAR, and missingness percentage had minimal impact on results. Linear interpolation outperformed all other methods for all settings, achieving the lowest RMSE (including across gender subgroups) with low bias.

While most methods have been evaluated using MCAR, we found that results on MCAR did not predict results on MAR and NMAR. Instead, accuracy on MCAR was significantly better than the other settings in many cases. Some methods developed for MAR did perform better on MAR than NMAR and while the same was true of the NMAR method (Fourier transform), its RMSE for NMAR was roughly triple that of MCAR. This suggests an urgent need for new standards in evaluating imputation methods. Other works have reviewed and proposed methods for generating data according to each mechanism (Santos et al., 2019; Cabrera-Sánchez et al., 2024), yet this is still not standard practice during evaluation of new methods or benchmarking studies. Further, while the ranking of methods was highly similar across mechanisms and missingness rates, the absolute error rates in our results varied substantially. Thus, while the best method for a particular data type could be identified with MCAR, researchers must be cautious in extrapolating from reported accuracy if experiments were conducted using MCAR.

Linear interpolation outperformed all other methods on both datasets, for all mechanisms, at all percentages of missingness. This performance may be attributed to the linear trends observed in the data used in this study (CGM and HR data), where the measurements change at a relatively constant rate over short periods of time. Notably, linear interpolation has much lower computational cost than most of the methods tested (with the exception of mean/mode imputation and LOCF). While prior comparisons have generally not included linear interpolation, top performers in other benchmarking studies (MICE, missForest) were in the bottom half of methods in our study. Pereira et al. (2024) found that for NMAR, MICE performed best however they tested few methods (denoising autoencoder, variational autoencoder, kNN, mean/mode imputation, MICE) and did not include linear interpolation. Beaulieu-Jones et al. (2018) obtained similar results for a similar set of methods, but again did not test linear interpolation or any recent deep learning methods. While no previous works have benchmarked all algorithms we evaluate, MICE has also performed well on MCAR tests (Luo, 2022) and missForest had the lowest error in laboratory data (Waljee et al., 2013). Our results now suggest that 1) linear interpolation should routinely be included as a baseline when benchmarking time series imputation methods and 2) current practices of using linear interpolation for imputing missing CGM (Fonda et al., 2013; Midroni et al., 2018; Butt et al., 2023) and HR (Nickerson et al., 2018; van Rossum et al., 2023) data are appropriate strategies. However, future work is needed to determine upper bounds on the gap lengths that can be successfully filled with linear interpolation and what other data types it also performs well on. Other health data types (e.g., laboratory test results) that do not exhibit linear short-term trends may still require other imputation strategies. Additionally, since we focused on accuracy of imputation, future work is needed to understand how these imputed values may impact the results of downstream tasks such as prediction or inference.

Our approach has some limitations which affect the generalizability of results. First, we focused on widely used data types, but it may be that the ranking of methods varies for other types of time series with different properties and distributions. Second, while we used the most pervasive metric for evaluation (RMSE), along with bias and EmpSE, additional metrics such as discrepancy scores as proposed by (Shadbahr et al., 2023), may be considered to further assess how well the imputed data preserves the underlying distribution. Third, while we used default parameters for the deep learning methods, further hyperparameter tuning may improve the performance of these methods. Lastly, we only examined continuous-valued, rather than categorical, data. Future work is needed to understand how missingness mechanisms affect accuracy for each of those data types using additional evaluation metrics.

7. Conclusion

We evaluated the performance of a variety of time series imputation methods on two healthcare data types

(CGM, HR) with three different missing data mechanisms (MCAR, NMAR, and MAR) using metrics such as RMSE, RMSE stratified by gender subgroups, bias, and EmpSE. Our findings show that current practices of evaluating and benchmarking methods using random deletion of values (MCAR) does not predict performance under MAR and NMAR. Further, the best performing method in our tests (linear interpolation) was often not included in prior benchmarking, yet is computationally inexpensive and robust to high degrees of missing data. Our work highlights the need for imputation methods to be evaluated using a variety of metrics on each mechanism.

Acknowledgments

This work was supported by NIH U54TR004279. The source of the Loop data is the Loop Study (sponsored by the Jaeb Center for Health Research and funded by the Helmsley Charitable Trust), but the analyses, content and conclusions presented herein are solely the responsibility of the authors and have not been reviewed or approved by the study sponsor. The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants.

References

Edgar Acuna, Roxana Aparicio, and Velcy Palomino. Analyzing the performance of transformers for the prediction of the blood glucose level considering imputation and smoothing. *Big Data and Cognitive Computing*, 7(1):41, 2023.

All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

M. Azur, E. Stuart, C. Frangakis, and P. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.

Marijka J Batterham, Linda C Tapsell, and Karen E Charlton. Analyzing weight loss intervention studies with missing data: which methods should be used? *Nutrition*, 29(7-8):1024–1029, 2013.

B. Beaulieu-Jones, D. Lavage, J. Snyder, J. Moore, S. Pendergrass, and C. Bauer. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Med Inform*, 6(1):e11, 2018.

Hatim Butt, Ikramullah Khosa, and Muhammad Aksam Iftikhar. Feature transformation for efficient blood glucose prediction in type 1 diabetes mellitus patients. *Diagnostics*, 13(3):340, 2023.

JF. Cabrera-Sánchez, RC. Pereira, PH. Abreu, and EL. Silva-Ramírez. A perspective on the missing at random problem: synthetic generation and benchmark analysis. *IEEE Access*, 2024.

Z Che, S Purushotham, K Cho, D Sontag, and Y Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1): 6085, 2018.

Christopher C Cheung, Alan Martyn, Norman Campbell, Shaun Frost, Kenneth Gilbert, Franklin Michota, Douglas Seal, William Ghali, and Nadia A Khan. Predictors of intraoperative hypotension and bradycardia. *The American journal of medicine*, 128(5):532–538, 2015.

W. Du. PyPOTS: A python toolbox for data mining on partially-observed time series. *arXiv preprint arXiv:2305.18811*, 2023.

A Facchinetti, S Del Favero, G Sparacino, and C Cobelli. Modeling transient disconnections and compression artifacts of continuous glucose sensors. *Diabetes Technology & Therapeutics*, 18(4):264–272, 2016.

SJ. Fonda, DG. Lewis, and RA. Vigersky. Minding the gaps in continuous glucose monitoring: a method to repair gaps to achieve more accurate

glucometrics. *Journal of Diabetes Science and Technology*, 7(1):88–92, 2013.

V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt. Gp-vae: Deep probabilistic time series imputation. In *AISTATS*, 2020.

G. Gadbury, C. Coffey, and D. Allison. Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond locf. *Obesity Reviews*, 4(3):175–184, 2003.

E Getzen, L Ungar, D Mowery, X Jiang, and Q Long. Mining for equitable health: Assessing the impact of missing data in electronic health records. *Journal of Biomedical Informatics*, 139:104269, 2023.

Rakesh Gopinathannair and Brian Olshansky. Management of tachycardia. *F1000prime reports*, 7, 2015.

H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, and A. Acharya. Mice vs ppca: missing data imputation in healthcare. *Informatics in Medicine Unlocked*, 17:100275, 2019.

Yuen Shing Yan Hindy, Mario Villaizan Vallelado, and Sep905. Missforest in python - arguably the best missing values imputation method, 2024. URL https://doi.org/10.5281/zenodo.13368883.

Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933, 2019.

Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. A benchmark for data imputation methods. *Frontiers in big Data*, 4:693674, 2021.

Rajat Kapoor, Lava R Timsina, Nupur Gupta, Harleen Kaur, Arianna J Vidger, Abby M Pollander, Judith Jacobi, Swapnil Khare, and Omar Rahman. Maintaining blood glucose levels in range (70–150 mg/dl) is difficult in covid-19 compared to non-covid-19 icu patients—a retrospective analysis. *Journal of Clinical Medicine*, 9(11):3635, 2020.

M. Kazijevs and MD. Samad. Deep imputation of missing values in time series health data: a review with benchmarking. *Journal of Biomedical Informatics*, 144, 2023.

E. Kontopantelis, IR. White, M. Sperrin, and I. Buchan. Outcome-sensitive multiple imputation: a simulation study. *BMC Medical Research Methodology*, 17:1–13, 2017.

DHA Kwak, X Ma, and S Kim. When does social desirability become a problem? detection and reduction of social desirability bias in information systems research. *Information & Management*, 58 (7):103500, 2021.

TD. Le, R. Beuran, and Y. Tan. Comparison of the most influential missing data imputation algorithms for healthcare. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 247–251, 2018.

M. Liu, S. Li, H. Yuan, MEH. Ong, Y. Ning, F. Xie, et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial Intelligence in Medicine*, 142:102587, 2023.

JW. Lum, RJ. Bailey, V. Barnes-Lomen, D. Naranjo, KK. Hood, RA. Lal, et al. A real-world prospective study of the safety and effectiveness of the loop open source automated insulin delivery system. *Diabetes Technology & Therapeutics*, 23(5):367–375, 2021.

Y. Luo. Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, 23(1):bbab489, 2022.

W. McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010.

Cooper Midroni, Peter J Leimbigler, Gaurav Baruah, Maheedhar Kolla, Alfred J Whitehead, and Yan Fossat. Predicting glycemia in type 1 diabetes patients: experiments with xgboost. In *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data (KDH@IJCAI 2018)*, pages 79–84, 2018.

Tim P Morris, Ian R White, and Michael J Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

Paul Nickerson, Raheleh Baharloo, Anis Davoudi, Azra Bihorac, and Parisa Rashidi. Comparison of gaussian processes methods to linear methods for

imputation of sparse physiological time series. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4106–4109. IEEE, 2018.

Steven Pan and Sixia Chen. Empirical comparison of imputation methods for multivariate missing data in public health. *International Journal of Environmental Research and Public Health*, 20(2):1524, 2023.

The pandas development team. pandas-dev/pandas: Pandas, 2020.

SN Payrovnaziri, A Xing, S Salman, X Liu, J Bian, and Z He. Assessing the impact of imputation on the interpretations of prediction models: a case study on mortality prediction for patients with acute myocardial infarction. In *AMIA Summits on Translational Science Proceedings*, page 465, 2021.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

RC. Pereira, PH. Abreu, PP. Rodrigues, and MA. Figueiredo. Imputation of data missing not at random: artificial generation and benchmark analysis. *Expert Systems with Applications*, 249, 2024.

S. Phung, A. Kumar, and J. Kim. A deep learning technique for imputing missing healthcare data. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6513–6516, 2019.

Preethi Prakash, Kelly Street, Shrikanth Narayanan, Bridget A Fernandez, Yufeng Shen, and Chang Shu. Benchmarking machine learning missing data imputation methods in large-scale mental health survey databases. *Artificial Intelligence in Health*, 2(1):81–92, 2024.

U. Pujianto, AP. Wibawa, and MI. Akbar. K-nearest neighbor (k-nn) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 83–88, 2019.

SA. Rahman, Y. Huang, J. Claassen, N. Heintzman, and S. Kleinberg. Combining fourier and lagged k-nearest neighbor imputation for biomedical time series data. *Journal of Biomedical Informatics*, 58: 198–207, 2015.

W Ren, Z Liu, Y Wu, Z Zhang, S Hong, and H Liu. Moving beyond medical statistics: A systematic review on missing data handling in electronic health records. *Health Data Science*, 2023.

MS. Santos, RC. Pereira, AF. Costa, JP. Soares, J. Santos, and PH. Abreu. Generating synthetic missing data: a review by missing mechanism. *IEEE Access*, 7:11651–11667, 2019.

T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3(1):139, 2023.

GJ Smith, MB Abraham, M de Bock, J Fairchild, B King, GR Ambler, et al. Impact of missing data on the accuracy of glucose metrics from continuous glucose monitoring assessed over a 2-week period. *Diabetes Technology & Therapeutics*, 25(5): 356–362, 2023.

M Sperrin, GP Martin, R Sisk, and N Peek. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology*, 125:183–187, 2020.

DJ. Stekhoven and P. Bühlmann. Missforest—nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

DM Thomas, S Kleinberg, AW Brown, M Crow, ND Bastian, N Reisweber, et al. Machine learning modeling practices to support the principles of ai and ethics in nutrition research. *Nutrition & Diabetes*, 12(1):48, 2022.

Amin Vahdati, Sarah Cotterill, Antonia Marsden, and Evangelos Kontopantelis. Enhancing data integrity in electronic health records: Review of methods for handling missing data. *medRxiv*, pages 2024–05, 2024.

Mathilde C van Rossum, Pedro M Alves da Silva, Ying Wang, Ewout A Kouwenhoven, and Hermie J Hermens. Missing data imputation techniques for wireless continuous vital signs monitoring. *Journal of clinical monitoring and computing*, 37(5):1387–1400, 2023.

AK. Waljee, A. Mukherjee, AG. Singal, Y. Zhang, J. Warren, U. Balis, et al. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8):e002847, 2013.

B Wells, K Chagin, A Nowacki, and M Kattan. Strategies for handling missing data in electronic health record derived data. *eGEMs*, 1(3):1035, 2013.

J. Yoon, WR. Zame, and M. van der Schaar. Estimating missing data in temporal data streams using multidirectional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66 (5):1477–1490, 2018.

Y Zhou, J Shi, R Stein, X Liu, RN Baldassano, CB Forrest, et al. Missing data matter: an empirical evaluation of the impacts of missing ehr data in comparative effectiveness research. *Journal of the American Medical Informatics Association*, 30(7): 1246–1256, 2023.

Appendix A. Additional Results

Tables 11 - 14 show the confidence intervals around the RMSE estimates. Figure 6 shows an example of the imputed values using different methods for the Loop Data for MCAR at 5%.

Appendix B. Additional Data Details

Loop We first simulated missingness for the maximum percentage (30%) for NMAR. This resulted in 224 individuals achieving this scenario. We then simulated missingness for MAR mechanism at 30% within this group of 224 individuals. The highest percentage we were able to achieve for MAR was 15% for 21 individuals. We found that only 4 individuals achieved 20% for MAR and none for 25% and 30%. Since our goal is to evaluate across all the three missingness mechanisms, we proceeded with further analysis using these 21 individuals.

All of Us We set the minimum requirement of 10,000 HR values which is approximately 1 week of data (HR is recorded at minute intervals) to ensure we are only looking at individuals with substantial amounts of data for our experiments. We then extracted up to 3 months of data from 100 subjects. For these 100 subjects, we simulated missingness for the three mechanisms (MCAR, MAR, NMAR) across the six missingness percentages. Out of the 100 subjects, 40 met the highest percentage (30%) for MAR and NMAR. We then processed with the analysis using these 40 subjects.

B.1. Native Missingness

For Loop. the average native missing duration across the data is 21.95 $\pm$ 20.80 minutes. The female subgroup had a higher average native missing duration of 25.73 $\pm$ 23.13 minutes ($p < 0.05$) compared to the male subgroup, which has an average of 22.27 $\pm$ 20.91 minutes. However, no statistical difference was observed in the average percentage of native missingness between the gender subgroups ($p = 0.14$). Similarly, for All of Us dataset, there was no statistical difference in the average percentage of native missingness between subgroups ($p = 0.30$). The female subgroup also had a higher average native missing duration of 4.09 $\pm$ 6.45 minutes ($p < 0.05$) compared to the male subgroup, which has an average of 3.81 $\pm$ 5.62 minutes.

Mechanism		MAR				NMAR	MAR+NMAR		Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	Fl-KNN	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	55.53	50.76	60.46	80.93	9.66	15.49	32.2	5.27	53.06	68.04	8.91	49.49
	10	55.8	50.86	61.55	81.01	10.04	16.38	32.29	4.88	53.22	67.84	8.99	49.55
	15	56.02	51.44	66.42	81.88	10.8	18.22	32.94	5.19	53.89	71.06	9.76	50.20
	20	56.17	51.35	63.94	81.33	12.12	18.35	33.82	5.31	53.83	70.9	10.15	50.09
	25	55.92	51.33	62.6	81.34	13.09	19.28	34.27	5.41	53.81	71.15	10.51	50.07
	30	55.88	51.32	62.22	81.62	14.32	20.07	33.58	5.53	53.79	70.87	10.96	50.06
NMAR	5	67.56	63.92	73.01	83.09	37.93	39.0	45.28	12.55	66.83	92.01	26.74	62.44
	10	70.24	66.05	71.92	85.78	40.31	41.47	47.2	13.73	69.2	91.93	28.31	64.51
	15	71.75	66.6	78.17	84.64	39.56	42.79	45.92	13.19	70.07	93.97	27.52	65.04
	20	71.96	66.8	75.58	84.95	39.95	42.82	46.85	13.48	70.3	92.89	27.47	65.28
	25	71.88	67.25	77.4	84.75	39.55	41.96	45.55	12.93	70.93	92.34	26.68	65.79
	30	72.03	67.12	74.74	84.35	39.65	41.26	45.36	12.62	70.93	91.44	26.16	65.74
MAR	5	56.59	50.89	66.18	79.83	45.4	40.69	47.79	16.92	53.66	66.21	31.16	50.11
	10	56.27	51.99	64.42	79.72	43.41	37.17	45.72	15.68	54.42	67.02	29.33	50.64
	15	57.18	52.88	66.26	79.84	43.76	36.94	46.19	15.46	55.0	70.6	29.81	51.42

Table 3: The table shows the mean RMSE (mg/dL) on the Loop dataset for all the methods.

Mechanism		MAR				NMAR	MAR+NMAR	Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	18.21	12.19	16.43	19.99	5.04	5.96	3.22	15.5	18.99	4.82	11.6
	10	18.12	12.22	16.11	20.04	5.27	6.06	3.32	15.47	18.9	4.93	11.66
	15	18.28	12.28	15.98	19.25	5.65	5.99	3.45	15.5	18.95	5.13	11.70
	20	18.2	12.26	15.83	19.96	5.91	6.30	3.56	15.49	18.94	5.30	11.68
	25	17.99	12.24	14.72	19.68	6.23	6.37	3.67	15.49	19.0	5.44	11.67
	30	17.65	12.25	16.57	18.78	6.61	6.43	3.77	15.48	19.02	5.60	11.69
NMAR	5	24.73	14.78	17.27	22.55	11.73	16.27	7.02	19.43	20.98	9.79	13.88
	10	24.98	15.11	18.13	22.9	12.55	16.04	7.53	20.13	21.94	10.94	14.12
	15	24.91	15.32	16.4	22.88	12.82	16.13	7.5	20.53	22.24	11.12	14.41
	20	24.61	15.42	16.15	22.74	12.97	16.07	7.44	20.37	21.92	10.87	14.55
	25	24.56	15.58	16.11	22.85	13.09	15.96	7.49	20.49	21.81	10.77	14.85
	30	24.01	15.66	17.18	22.54	13.41	15.67	7.48	20.41	21.6	10.75	15.11
MAR	5	22.33	14.76	15.64	19.26	16.16	15.85	9.19	21.17	28.48	13.34	13.69
	10	20.3	15.18	16.59	19.65	16.43	15.18	9.65	21.58	28.84	13.75	13.99
	15	20.85	15.22	15.59	19.68	16.55	14.76	9.63	21.67	28.52	13.73	14.10
	20	19.64	15.35	16.71	19.69	16.66	14.9	9.51	21.91	28.38	13.43	14.41
	25	19.66	15.35	16.10	19.82	16.92	14.39	9.47	21.86	27.93	13.43	14.71
	30	19.18	15.27	16.06	19.5	16.65	13.73	9.37	21.61	27.53	13.03	15.04

Table 4: The table shows the mean RMSE (bpm) on the All of Us dataset for all methods.

Mechanism		MAR				NMAR	MAR+NMAR		Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	Fl-KNN	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	0.61	14.98	40.05	27.57	53.12	41.85	29.1	52.71	0.0	0.0	53.11	18.9
	10	0.6	15.45	41.2	26.35	53.21	41.71	28.96	52.84	0.0	0.0	53.23	19.27
	15	0.38	16.15	41.43	25.97	53.17	41.37	29.23	53.45	0.0	0.0	53.9	19.7
	20	0.44	16.12	40.69	25.98	53.83	41.54	29.3	53.36	0.0	0.0	53.85	19.59
	25	0.22	16.14	37.68	25.98	53.8	41.15	29.4	53.29	0.0	0.0	53.82	19.55
	30	0.3	16.21	40.04	27.1	53.93	40.74	29.2	53.21	0.0	0.0	53.78	19.47
NMAR	5	0.56	15.35	40.5	31.15	55.71	36.35	29.81	52.81	0.0	0.0	54.96	19.74
	10	0.17	15.67	38.36	27.66	58.4	36.91	31.24	55.21	0.0	0.0	58.0	19.7
	15	0.65	16.32	46.86	27.22	58.96	34.57	31.81	55.82	0.0	0.0	58.52	19.91
	20	0.68	16.08	39.91	27.04	58.75	34.17	31.64	55.82	0.0	0.0	58.19	19.45
	25	0.61	15.79	40.56	26.66	58.99	34.28	31.15	56.2	0.0	0.0	58.73	18.9
	30	0.56	15.68	35.55	26.61	59.11	35.52	31.44	56.42	0.0	0.0	58.85	18.47
MAR	5	0.14	17.61	45.19	38.07	52.81	20.68	29.28	48.21	0.0	0.0	52.0	20.71
	10	0.33	19.87	41.34	37.28	52.52	25.17	28.57	49.37	0.0	0.0	53.15	21.83
	15	1.24	21.52	42.5	37.54	53.42	26.78	28.9	50.07	0.0	0.0	53.77	22.6

Table 5: The table shows the mean EmpSE (mg/dL) on the Loop dataset for all the methods.

Mechanism		MAR				NMAR	MAR+NMAR	Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	0.01	9.24	11.72	11.40	15.51	12.16	15.13	0.00	0.00	15.51	9.89
	10	0.21	9.22	11.90	11.22	15.48	12.21	15.07	0.00	0.00	15.46	9.95
	15	0.14	9.15	11.16	11.20	15.53	12.69	15.08	0.00	0.00	15.50	9.89
	20	0.22	9.19	11.25	11.13	15.51	12.40	15.06	0.00	0.00	15.49	9.88
	25	0.31	9.20	10.40	11.19	15.51	12.43	15.04	0.00	0.00	15.48	9.87
	30	0.36	9.14	10.00	11.08	15.51	12.59	15.01	0.00	0.00	15.47	9.80
NMAR	5	0.04	9.63	10.93	11.42	17.67	10.81	16.28	0.00	0.00	17.72	10.20
	10	0.07	10.32	11.52	11.60	18.25	11.45	17.44	0.00	0.00	19.28	10.67
	15	0.21	10.24	11.59	11.66	18.51	11.69	17.62	0.00	0.00	19.48	10.57
	20	0.21	9.83	10.86	11.36	18.50	10.87	17.42	0.00	0.00	19.21	10.03
	25	0.20	9.67	10.74	11.39	18.31	10.67	17.40	0.00	0.00	19.21	9.59
	30	0.40	9.35	10.39	11.21	18.34	10.74	17.20	0.00	0.00	19.04	8.92
MAR	5	0.06	13.76	14.44	16.20	17.17	10.10	15.09	0.00	0.00	16.40	13.75
	10	0.22	14.16	16.43	15.97	17.26	10.75	15.16	0.00	0.00	16.32	13.78
	15	0.22	14.01	14.58	15.81	17.05	10.84	15.00	0.00	0.00	16.18	13.33
	20	0.17	13.94	15.20	15.63	16.97	10.80	15.17	0.00	0.00	16.33	12.54
	25	0.18	13.75	14.68	15.39	16.89	11.13	15.00	0.00	0.00	16.19	11.48
	30	0.00	13.59	14.21	15.15	16.54	11.54	14.77	0.00	0.00	15.92	10.36

Table 6: The table shows the mean EmpSE (bpm) on the All of Us dataset for all the methods.

Mechanism		MAR				NMAR	MAR+NMAR		Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	Fl-KNN	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	53.73	50.51	60.61	88.05	10.53	15.24	31.85	6.01	52.89	62.19	9.6	49.39
	10	54.25	50.65	60.15	86.1	11.05	16.39	31.73	5.86	53.01	62.35	9.96	49.54
	15	53.75	50.52	66.71	86.07	11.69	17.46	32.21	5.84	52.91	62.3	10.26	49.45
	20	53.91	50.46	61.86	85.9	12.51	17.79	32.69	5.93	52.85	62.12	10.62	49.39
	25	53.75	50.45	63.38	85.84	13.41	18.51	33.27	6.04	52.84	62.61	10.95	49.37
	30	53.71	50.44	57.34	85.8	14.49	19.47	33.48	6.15	52.81	62.4	11.37	49.34
NMAR	5	66.33	64.93	74.61	93.92	37.69	38.80	46.46	12.88	68.20	89.98	26.64	63.78
	10	68.69	67.11	72.26	93.43	40.65	39.67	48.03	14.35	70.42	89.81	28.85	65.69
	15	69.05	67.07	79.65	93.14	39.03	40.84	46.52	13.79	70.85	88.99	27.96	65.65
	20	69.51	67.54	77.87	93.58	39.39	40.95	48.42	14.04	71.32	88.05	27.69	66.12
	25	69.30	67.68	74.58	92.97	38.80	40.48	46.22	13.08	71.71	87.02	26.46	66.35
	30	69.18	67.17	75.25	91.88	39.08	39.40	46.47	12.98	71.32	85.46	26.17	65.89
MAR	5	55.44	50.79	67.26	88.44	45.00	39.63	45.03	18.42	53.26	62.43	32.16	50.39
	10	54.60	50.97	63.78	85.08	43.18	36.46	43.31	16.89	53.03	60.74	30.12	49.98
	15	55.58	51.68	59.33	84.26	43.18	35.91	43.17	16.27	53.66	62.06	29.64	50.62

Table 7: The table shows the mean RMSE (mg/dL) for men on the Loop dataset for all methods.

Mechanism		MAR				NMAR	MAR+NMAR		Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	Fl-KNN	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	59.40	51.67	62.35	81.74	8.08	15.85	33.61	4.00	54.00	81.15	7.49	50.48
	10	59.27	51.51	66.04	81.16	9.04	16.61	34.51	4.06	53.95	79.96	8.02	50.18
	15	58.65	51.37	68.30	81.23	9.92	18.28	34.10	4.21	53.83	80.52	8.45	50.09
	20	58.89	51.28	63.84	81.22	10.89	18.67	36.68	4.27	53.82	80.46	8.94	49.98
	25	58.48	51.26	61.59	81.32	11.90	19.87	36.85	4.33	53.80	80.48	9.34	49.93
	30	58.37	51.19	66.34	81.24	13.10	20.22	34.39	4.48	53.72	79.92	9.86	49.87
NMAR	5	70.77	63.39	74.82	72.96	38.24	39.58	44.84	11.22	65.74	100.36	26.18	61.95
	10	72.51	64.35	73.13	77.56	40.36	43.55	46.78	13.12	67.26	98.52	28.02	63.01
	15	73.93	65.12	74.73	76.97	38.82	44.70	44.49	11.93	68.16	97.50	25.91	63.64
	20	73.67	64.79	71.33	77.14	38.85	43.37	43.92	11.98	67.94	96.24	26.03	63.46
	25	74.04	65.52	77.44	77.67	38.34	42.40	44.06	11.84	68.76	96.18	25.28	64.18
	30	74.17	65.40	73.00	77.99	38.66	42.19	43.01	11.24	68.85	94.76	24.66	64.22
MAR	5	62.95	54.24	71.26	79.33	47.94	45.41	55.13	14.94	57.88	79.79	30.65	53.04
	10	60.82	54.90	68.46	79.68	44.84	39.38	51.81	14.54	57.96	81.73	28.94	53.20
	15	60.46	54.41	76.43	80.83	44.21	38.70	52.05	14.50	57.15	81.21	29.81	52.78

Table 8: The table shows the mean RMSE (mg/dL) for women on the Loop dataset for all methods.

Mechanism		MAR				NMAR	MAR+NMAR	Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	18.06	12.10	13.82	19.27	5.24	5.93	3.28	16.49	20.36	5.02	11.52
	10	18.00	12.01	15.43	19.48	5.42	5.98	3.37	16.09	19.72	5.06	11.46
	15	18.42	12.16	14.76	19.91	5.82	6.00	3.53	16.18	19.89	5.27	11.58
	20	18.25	12.10	14.04	19.97	6.09	6.25	3.63	16.17	19.83	5.44	11.52
	25	18.09	12.09	13.01	20.00	6.42	6.28	3.73	16.20	19.95	5.59	11.51
	30	17.70	12.11	16.44	18.36	6.77	6.36	3.85	16.16	19.87	5.77	11.53
NMAR	5	23.22	13.74	17.78	20.42	12.17	15.03	7.20	19.55	20.02	9.86	12.88
	10	24.17	14.40	17.63	21.66	12.89	14.82	7.60	20.62	21.12	11.03	13.46
	15	24.52	14.57	16.80	21.67	13.48	14.96	7.62	21.15	21.73	11.61	13.75
	20	24.32	14.67	14.92	21.44	13.42	15.00	7.44	20.91	21.17	11.08	13.89
	25	24.19	14.86	15.83	21.54	13.49	14.69	7.42	20.94	20.96	10.80	14.20
	30	23.88	14.96	16.14	21.39	13.88	14.47	7.57	20.83	20.82	11.01	14.50
MAR	5	21.50	14.78	15.23	19.60	17.62	14.85	9.40	23.11	30.47	13.82	13.73
	10	20.44	15.52	16.53	20.48	17.90	14.35	10.32	23.80	31.09	15.14	14.36
	15	20.66	15.56	16.57	20.53	17.89	13.83	10.15	23.74	30.56	14.71	14.44
	20	19.95	15.65	17.37	20.40	18.02	13.92	10.06	23.93	30.35	14.28	14.77
	25	20.05	15.56	16.61	20.32	18.31	13.09	9.94	23.81	29.70	14.36	15.12
	30	20.21	15.47	15.77	20.26	18.17	12.21	9.91	23.49	28.88	14.06	15.59

Table 9: The table shows the mean RMSE (bpm) for men subgroup on the All of Us dataset for all the methods.

Mechanism		MAR				NMAR	MAR+NMAR	Common Methods				
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	18.34	12.17	18.25	20.92	4.93	6.02	3.19	14.64	17.95	4.71	11.57
	10	18.22	12.25	16.34	20.72	5.15	6.15	3.29	14.82	18.14	4.85	11.69
	15	18.10	12.24	16.80	18.74	5.53	5.96	3.39	14.78	18.07	5.03	11.66
	20	18.10	12.25	17.16	20.08	5.79	6.35	3.52	14.78	18.12	5.20	11.68
	25	17.91	12.24	16.23	19.57	6.08	6.45	3.63	14.75	18.11	5.33	11.67
	30	17.51	12.23	16.32	19.30	6.50	6.51	3.72	14.78	18.21	5.48	11.68
NMAR	5	26.07	15.49	16.79	24.54	11.42	17.30	6.87	19.33	22.16	9.79	14.55
	10	25.59	15.48	18.51	24.11	12.28	16.93	7.49	19.52	22.02	10.96	14.45
	15	25.06	15.74	15.52	24.07	12.26	17.00	7.42	19.80	22.09	10.76	14.76
	20	24.82	15.84	17.16	24.02	12.62	16.85	7.47	19.74	22.00	10.76	14.91
	25	24.80	16.00	16.19	24.17	12.77	16.92	7.59	19.95	22.03	10.81	15.21
	30	24.14	16.08	17.65	23.71	13.04	16.59	7.41	19.91	21.74	10.57	15.46
MAR	5	23.07	14.82	16.14	19.19	15.12	16.69	9.18	19.72	26.71	13.19	13.70
	10	20.07	14.83	16.90	19.00	15.29	15.78	9.13	19.69	26.48	12.65	13.60
	15	20.98	14.85	14.83	18.94	15.48	15.46	9.25	19.89	26.32	12.99	13.73
	20	19.24	15.01	15.83	19.16	15.60	15.59	9.11	20.15	26.26	12.83	14.03
	25	19.48	15.07	15.82	19.37	15.86	15.37	9.14	20.15	25.97	12.75	14.27
	30	18.34	14.99	15.98	18.71	15.41	14.90	8.96	19.93	25.98	12.23	14.47

Table 10: The table shows the mean RMSE (bpm) for women subgroup on the All of Us dataset for all the methods.

RMSE (95% CI)								
Mechanism		MAR				NMAR	MAR+NMAR	
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE	Fl-KNN
MCAR	5	55.53 (48.10, 62.97)	50.76 (47.24, 54.29)	60.46 (53.09, 67.83)	80.93 (59.48, 102.38)	9.66 (5.27, 14.04)	15.49 (13.36, 17.61)	32.20 (29.19, 35.20)
	10	55.80 (48.85, 62.76)	50.86 (48.66, 53.06)	61.55 (56.72, 66.38)	81.01 (59.89, 102.13)	10.04 (6.75, 13.33)	16.38 (15.44, 17.31)	32.29 (30.12, 34.45)
	15	56.02 (49.04, 63.00)	51.44 (49.36, 53.52)	66.42 (57.03, 75.81)	81.88 (61.26, 102.50)	10.80 (7.61, 13.99)	18.22 (15.58, 20.86)	32.94 (30.65, 35.23)
	20	56.17 (49.34, 63.00)	51.35 (49.20, 53.50)	63.94 (59.03, 68.86)	81.33 (60.49, 102.17)	12.12 (8.99, 15.25)	18.35 (16.26, 20.43)	33.82 (30.86, 36.79)
	25	55.92 (49.00, 62.84)	51.33 (49.22, 53.44)	62.60 (50.69, 74.51)	81.34 (60.42, 102.25)	13.09 (10.03, 16.14)	19.28 (16.68, 21.88)	34.27 (32.07, 36.46)
	30	55.88 (49.04, 62.72)	51.32 (49.25, 53.39)	62.22 (57.24, 67.19)	81.62 (60.53, 102.71)	14.32 (11.28, 17.35)	20.07 (18.01, 22.14)	33.58 (31.80, 35.36)
NMAR	5	67.56 (61.55, 73.58)	63.92 (58.25, 69.58)	73.01 (65.68, 80.35)	83.09 (63.55, 102.63)	37.93 (34.04, 41.81)	39.00 (37.36, 40.63)	45.28 (39.38, 51.19)
	10	70.24 (61.78, 78.70)	66.05 (63.11, 68.99)	71.92 (69.12, 74.73)	85.78 (68.79, 102.77)	40.31 (36.05, 44.57)	41.47 (33.09, 49.85)	47.20 (43.59, 50.82)
	15	71.75 (63.26, 80.23)	66.60 (63.06, 70.15)	78.17 (72.20, 84.15)	84.64 (67.48, 101.81)	39.56 (35.60, 43.52)	42.79 (35.13, 50.45)	45.92 (41.08, 50.75)
	20	71.96 (63.49, 80.42)	66.80 (63.21, 70.38)	75.58 (69.88, 81.29)	84.95 (67.95, 101.96)	39.95 (34.99, 44.92)	42.82 (34.97, 50.66)	46.85 (41.17, 52.53)
	25	71.88 (62.86, 80.90)	67.25 (63.72, 70.78)	77.40 (68.87, 85.92)	84.75 (67.63, 101.87)	39.55 (34.44, 44.65)	41.96 (37.15, 46.77)	45.55 (40.09, 51.01)
	30	72.03 (62.72, 81.34)	67.12 (64.31, 69.93)	74.74 (70.54, 78.94)	84.35 (68.09, 100.61)	39.65 (35.34, 43.95)	41.26 (35.29, 47.22)	45.36 (41.30, 49.42)
MAR	5	56.59 (48.31, 64.87)	50.89 (46.64, 55.15)	66.18 (55.44, 76.92)	79.83 (60.14, 99.53)	45.40 (43.86, 46.94)	40.69 (34.37, 47.00)	47.79 (44.89, 50.69)
	10	56.27 (51.46, 61.09)	51.99 (49.93, 54.04)	64.42 (56.60, 72.24)	79.72 (60.22, 99.23)	43.41 (40.83, 45.99)	37.17 (34.30, 40.04)	45.72 (40.07, 51.38)
	15	57.18 (52.08, 62.27)	52.88 (50.09, 55.68)	66.26 (62.25, 70.27)	79.84 (62.08, 97.60)	43.76 (39.04, 48.47)	36.94 (30.86, 43.01)	46.19 (40.98, 51.40)

Table 11: The table shows the mean RMSE (mg/dL) with confidence intervals on the Loop dataset for the MAR, NMAR, and MAR+NMAR methods.

RMSE (95% CI)							
Mechanism		Common Methods					
Method	%	LI	Mean	Mode	LOCF	miss-Forest	
MCAR	5	5.27 (0.01, 10.54)	53.06 (48.76, 57.35)	68.04 (46.61, 89.47)	8.91 (4.32, 13.51)	49.49 (46.22, 52.76)	
	10	4.88 (0.93, 8.83)	53.22 (49.83, 56.60)	67.84 (47.66, 88.02)	8.99 (5.55, 12.42)	49.55 (47.59, 51.51)	
	15	5.19 (1.34, 9.04)	53.89 (50.72, 57.07)	71.06 (51.28, 90.84)	9.76 (6.17, 13.35)	50.20 (48.18, 52.22)	
	20	5.31 (1.57, 9.06)	53.83 (50.62, 57.05)	70.90 (51.18, 90.63)	10.15 (6.76, 13.55)	50.09 (48.02, 52.16)	
	25	5.41 (1.78, 9.03)	53.81 (50.62, 57.01)	71.15 (51.50, 90.80)	10.51 (7.19, 13.82)	50.07 (48.05, 52.09)	
	30	5.53 (1.95, 9.11)	53.79 (50.65, 56.92)	70.87 (51.12, 90.62)	10.96 (7.69, 14.23)	50.06 (48.06, 52.05)	
NMAR	5	12.55 (8.87, 16.23)	66.83 (60.99, 72.67)	92.01 (81.67, 102.36)	26.74 (23.71, 29.76)	62.44 (56.83, 68.05)	
	10	13.73 (10.68, 16.77)	69.20 (66.86, 71.53)	91.93 (80.78, 103.07)	28.31 (24.56, 32.05)	64.51 (61.30, 67.73)	
	15	13.19 (10.21, 16.17)	70.07 (66.81, 73.33)	93.97 (80.88, 107.07)	27.52 (23.52, 31.52)	65.04 (61.32, 68.75)	
	20	13.48 (9.91, 17.06)	70.30 (66.97, 73.64)	92.89 (79.65, 106.12)	27.47 (23.01, 31.92)	65.28 (61.54, 69.01)	
	25	12.93 (9.45, 16.41)	70.93 (67.54, 74.32)	92.34 (78.51, 106.16)	26.68 (22.01, 31.35)	65.79 (62.19, 69.40)	
	30	12.62 (9.32, 15.92)	70.93 (68.55, 73.30)	91.44 (77.97, 104.91)	26.16 (22.04, 30.28)	65.74 (63.11, 68.36)	
MAR	5	16.92 (12.58, 21.27)	53.66 (47.41, 59.91)	66.21 (49.47, 82.94)	31.16 (28.56, 33.76)	50.11 (46.39, 53.83)	
	10	15.68 (12.75, 18.62)	54.42 (50.90, 57.94)	67.02 (48.54, 85.51)	29.33 (26.86, 31.81)	50.64 (49.04, 52.24)	
	15	15.46 (12.99, 17.93)	55.00 (51.23, 58.76)	70.60 (51.02, 90.17)	29.81 (26.55, 33.07)	51.42 (48.85, 53.99)	

Table 12: The table shows the mean RMSE (mg/dL) with confidence intervals on the Loop dataset for the common methods.

		RMSE (95% CI)					
Mechanism		MAR				NMAR	MAR+NMAR
Method	%	MRNN	MICE	k-NN	Hot Deck	Fourier Transform	GPVAE
MCAR	5	18.21 (16.91, 19.50)	12.19 (11.08, 13.29)	16.43 (11.94, 20.91)	19.99 (18.18, 21.80)	5.04 (4.58, 5.50)	5.96 (5.38, 6.55)
	10	18.12 (17.15, 19.09)	12.22 (11.30, 13.13)	16.11 (11.93, 20.30)	20.04 (17.67, 22.41)	5.27 (4.91, 5.62)	6.06 (5.55, 6.57)
	15	18.28 (17.16, 19.40)	12.28 (11.35, 13.20)	15.98 (14.52, 17.44)	19.25 (16.98, 21.52)	5.65 (5.19, 6.11)	5.99 (5.47, 6.50)
	20	18.20 (17.08, 19.33)	12.26 (11.29, 13.22)	15.83 (13.92, 17.75)	19.96 (18.01, 21.90)	5.91 (5.39, 6.44)	6.30 (5.66, 6.94)
	25	17.99 (17.14, 18.84)	12.24 (11.30, 13.19)	14.72 (12.56, 16.88)	19.68 (17.80, 21.56)	6.23 (5.71, 6.75)	6.37 (5.64, 7.11)
	30	17.65 (16.53, 18.76)	12.25 (11.32, 13.18)	16.57 (13.00, 20.13)	18.78 (16.38, 21.19)	6.61 (5.97, 7.25)	6.43 (5.97, 6.90)
NMAR	5	24.73 (22.09, 27.37)	14.78 (13.10, 16.45)	17.27 (14.32, 20.22)	22.55 (20.94, 24.16)	11.73 (10.39, 13.08)	16.27 (14.46, 18.07)
	10	24.98 (23.30, 26.67)	15.11 (13.84, 16.38)	18.13 (15.43, 20.82)	22.90 (20.47, 25.33)	12.55 (11.25, 13.85)	16.04 (14.56, 17.51)
	15	24.91 (23.11, 26.72)	15.32 (14.06, 16.59)	16.40 (14.87, 17.94)	22.88 (20.57, 25.19)	12.82 (11.22, 14.42)	16.13 (14.69, 17.56)
	20	24.61 (22.85, 26.37)	15.42 (14.12, 16.71)	16.15 (14.55, 17.74)	22.74 (20.27, 25.20)	12.97 (11.42, 14.51)	16.07 (14.43, 17.70)
	25	24.56 (23.22, 25.90)	15.58 (14.24, 16.93)	16.11 (14.44, 17.77)	22.85 (20.67, 25.02)	13.09 (11.80, 14.37)	15.96 (14.63, 17.30)
	30	24.01 (22.50, 25.52)	15.66 (14.29, 17.03)	17.18 (15.40, 18.97)	22.54 (20.27, 24.81)	13.41 (12.15, 14.68)	15.67 (14.50, 16.83)
MAR	5	22.33 (20.08, 24.58)	14.76 (13.16, 16.36)	15.64 (13.75, 17.54)	19.26 (17.13, 21.38)	16.16 (15.50, 16.83)	15.85 (14.66, 17.04)
	10	20.30 (18.90, 21.69)	15.18 (13.84, 16.51)	16.59 (14.66, 18.53)	19.65 (18.92, 20.37)	16.43 (15.06, 17.79)	15.18 (14.29, 16.06)
	15	20.85 (18.67, 23.02)	15.22 (13.89, 16.55)	15.59 (14.17, 17.02)	19.68 (18.72, 20.64)	16.55 (15.47, 17.62)	14.76 (13.74, 15.78)
	20	19.64 (18.17, 21.11)	15.35 (13.93, 16.77)	16.71 (15.36, 18.06)	19.69 (18.77, 20.60)	16.66 (15.70, 17.63)	14.90 (13.73, 16.06)
	25	19.66 (18.03, 21.30)	15.35 (13.96, 16.74)	16.10 (15.13, 17.07)	19.82 (18.52, 21.12)	16.92 (15.85, 17.99)	14.39 (13.21, 15.57)
	30	19.18 (18.08, 20.27)	15.27 (13.82, 16.71)	16.06 (15.42, 16.71)	19.50 (18.42, 20.58)	16.65 (15.64, 17.67)	13.73 (12.95, 14.51)

Table 13: The table shows the mean RMSE (bpm) with confidence intervals on the All of Us dataset for the MAR, NMAR, and MAR+NMAR methods.

		RMSE (95% CI)				
Mechanism		Common Methods				
Method	%	LI	Mean	Mode	LOCF	miss-Forest
MCAR	5	3.22 (2.99, 3.45)	15.50 (13.95, 17.06)	18.99 (16.39, 21.59)	4.82 (4.37, 5.27)	11.60 (10.77, 12.43)
	10	3.32 (3.12, 3.52)	15.47 (14.08, 16.85)	18.90 (16.38, 21.42)	4.93 (4.64, 5.23)	11.66 (10.94, 12.37)
	15	3.45 (3.22, 3.68)	15.50 (14.04, 16.96)	18.95 (16.35, 21.55)	5.13 (4.77, 5.48)	11.70 (10.97, 12.43)
	20	3.56 (3.30, 3.82)	15.49 (13.96, 17.02)	18.94 (16.27, 21.62)	5.30 (4.86, 5.74)	11.68 (10.92, 12.44)
	25	3.67 (3.39, 3.94)	15.49 (14.00, 16.97)	19.00 (16.39, 21.60)	5.44 (5.02, 5.86)	11.67 (10.94, 12.40)
	30	3.77 (3.50, 4.04)	15.48 (13.98, 16.99)	19.02 (16.41, 21.63)	5.60 (5.18, 6.02)	11.69 (10.96, 12.41)
NMAR	5	7.02 (6.27, 7.77)	19.43 (16.31, 22.56)	20.98 (16.82, 25.15)	9.79 (8.45, 11.13)	13.88 (12.44, 15.33)
	10	7.53 (6.64, 8.42)	20.13 (18.12, 22.14)	21.94 (18.61, 25.27)	10.94 (9.75, 12.14)	14.12 (13.09, 15.15)
	15	7.50 (6.65, 8.36)	20.53 (18.21, 22.84)	22.24 (18.58, 25.90)	11.12 (9.74, 12.51)	14.41 (13.43, 15.40)
	20	7.44 (6.61, 8.26)	20.37 (18.29, 22.46)	21.92 (18.65, 25.18)	10.87 (9.78, 11.96)	14.55 (13.49, 15.62)
	25	7.49 (6.80, 8.18)	20.49 (18.44, 22.55)	21.81 (18.61, 25.02)	10.77 (9.84, 11.70)	14.85 (13.65, 16.04)
	30	7.48 (6.83, 8.13)	20.41 (18.39, 22.44)	21.60 (18.48, 24.71)	10.75 (9.89, 11.61)	15.11 (13.84, 16.39)
MAR	5	9.19 (8.26, 10.12)	21.17 (18.34, 23.99)	28.48 (23.22, 33.74)	13.34 (12.30, 14.38)	13.69 (12.45, 14.93)
	10	9.65 (9.18, 10.12)	21.58 (19.26, 23.90)	28.84 (25.21, 32.48)	13.75 (12.81, 14.68)	13.99 (12.83, 15.14)
	15	9.63 (9.15, 10.11)	21.67 (19.42, 23.92)	28.52 (24.93, 32.10)	13.73 (13.01, 14.46)	14.10 (13.03, 15.17)
	20	9.51 (8.99, 10.02)	21.91 (19.65, 24.17)	28.38 (24.84, 31.93)	13.43 (12.69, 14.17)	14.41 (13.28, 15.53)
	25	9.47 (8.79, 10.15)	21.86 (19.59, 24.14)	27.93 (24.50, 31.36)	13.43 (12.48, 14.38)	14.71 (13.53, 15.89)
	30	9.37 (8.89, 9.84)	21.61 (19.28, 23.94)	27.53 (23.77, 31.30)	13.03 (12.30, 13.77)	15.04 (13.71, 16.37)

Table 14: The table shows the mean RMSE (bpm) with confidence intervals on the All of Us dataset for the common methods.

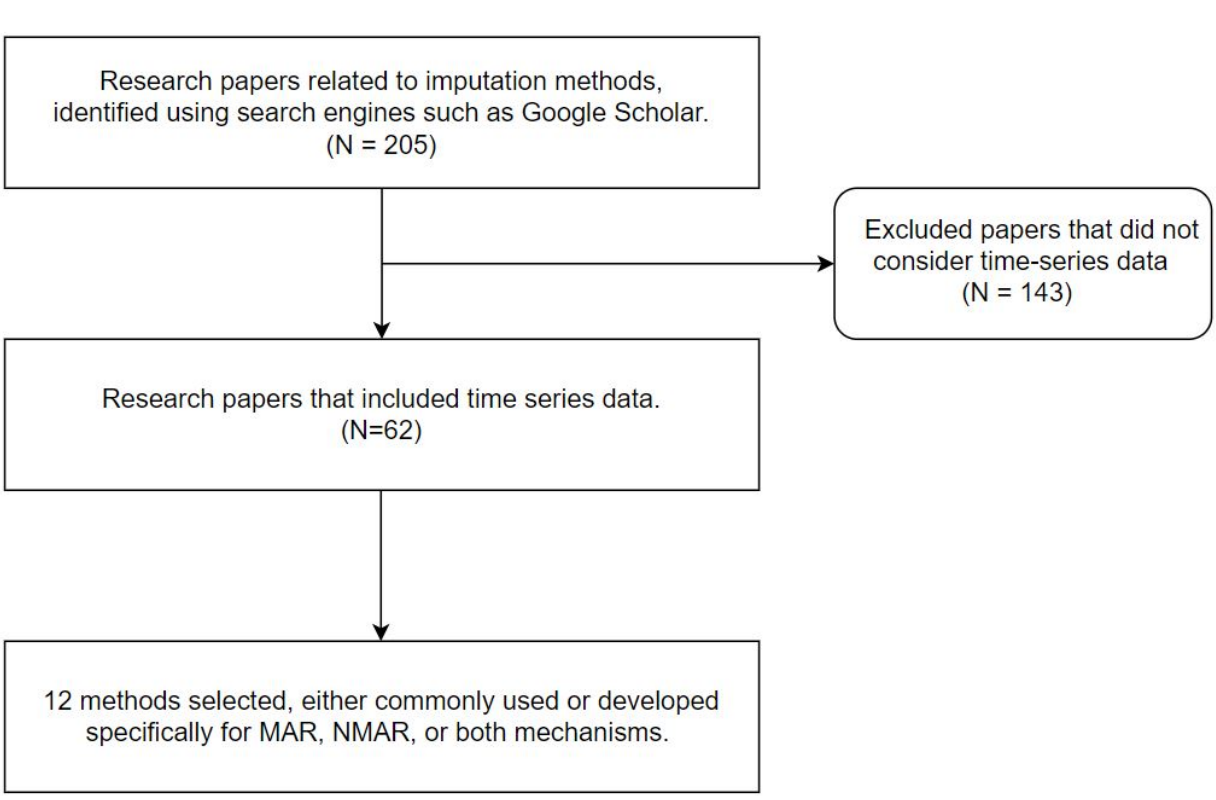

Figure 5: Flowchart showing the methodology of the selection of the methods used in this study.

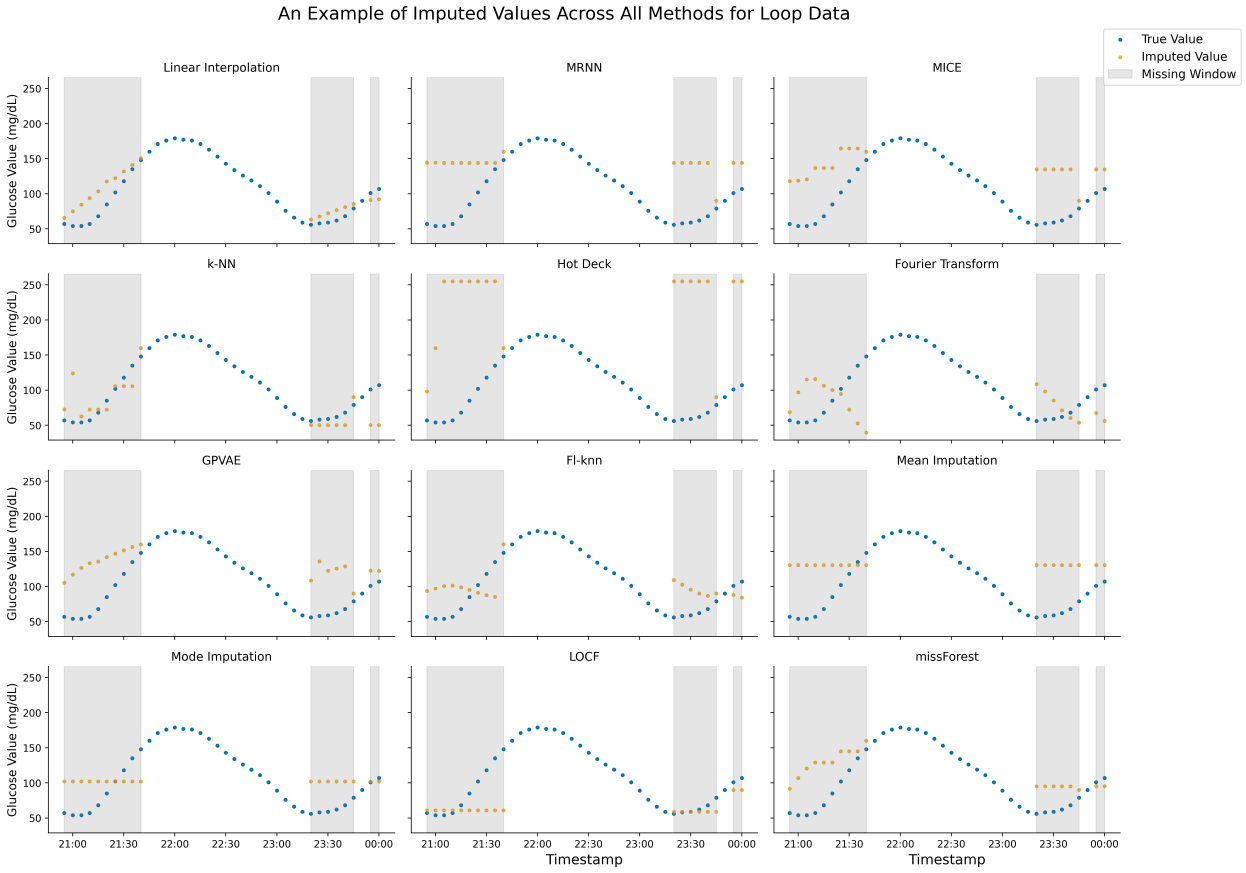

Figure 6: An example of the imputed glucose values using different methods for the Loop Data for MCAR at 5%. The shaded region highlights the missing data window.