Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adaptive Selection of MICE Algorithm Parameters: A Case Study on Pulmonary Function Value Prediction Models

Yeonghui Gang
Daejeon University
Myung-Mo Lee
Daejeon University
Jucheol Moon
California State University Long Beach
Hongjun Kim
hjkim99@dju.kr

Daejeon University

Article

Keywords:

Posted Date: July 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10121301/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Adaptive Selection of MICE Algorithm Parameters: A Case Study on Pulmonary Function Value Prediction Models

Yeonghui Gang[1], Myung-Mo Lee[2], Jucheol Moon[3], and Hongjun Kim[1,*]

[1]Daejeon University, Department of Computer Engineering, Daejeon, 34520, Republic of Korea
[2]Daejeon University, Department of Physical Therapy, Daejeon, 34520, Republic of Korea
[3]California State University Long Beach, Department of Computer Engineering and Computer Science, CA 90840, USA
[*]hjkim99@dju.kr

Abstract

Artificial intelligence-based medical prediction models are often affected by incomplete datasets containing missing values, which can significantly reduce model reliability and predictive performance. Among various imputation approaches, Multiple Imputation by Chained Equations (MICE) is widely used in medical data analysis because it models each variable using the conditional distributions of the remaining variables while reflecting uncertainty in the imputation process. Chronic obstructive pulmonary disease (COPD) is a chronic respiratory disorder characterized by persistent airflow limitation, requiring early diagnosis and continuous monitoring for improved prognosis. In this study, a pulmonary function test dataset collected from 598 subjects over a two-year period was used to investigate the effects of time-series sampling and missing-data imputation on prediction performance. An adaptive selection method was proposed for the initialization strategy and maximum iteration count of the MICE algorithm based on the statistical characteristics of representative pulmonary function indicators, including FEV1, FVC, FEF25–75%, and PEF. Different missing rates were artificially introduced, and the performances of Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) and multivariate iterative regression-based MICE were compared. Experimental results demonstrate that the proposed adaptive MICE strategy maintains stable predictive performance under increasing missing rates and provides reliable predictive information by effectively reflecting multivariate correlation structures.

Introduction

As artificial intelligence models improve in both accuracy and reliability, their application in the medical field continues to expand. However, several challenges remain significant obstacles, including difficulties in securing large datasets, class imbalance, and the prevalence of missing values[1,2]. Medical data is particularly difficult to collect and share on a large scale due to privacy protection regulations and heterogeneity across institutions, equipment, and regions. Electronic medical record data collected in actual clinical environments often contains severe class imbalance, with critical diseases or rare events existing as minority classes. Furthermore, missing values arise from various patterns due to selective test administration, discontinued follow-up observations, and recording omissions. If not properly addressed, these missing values can lead not only to degraded prediction performance but also to biased inference and incorrect clinical decisions[3].

Among various missing-data imputation methods, Multiple Imputation by Chained Equations (MICE) has been widely adopted because it models each variable using the conditional distribution of the remaining variables, appropriately reflects uncertainty in the imputation process, and performs well under the assumption that missing values occur at random[4,5]. MICE employs a fully conditional specification approach, fitting appropriate regression models for each variable in multivariate data containing both

continuous and categorical variables, and then performing iterative imputation in a chained-equation format. Despite its widespread use as the de facto standard tool for multiple imputation in medical and epidemiological research, studies have reported that random forest-based or latent-space-based imputation techniques show superior performance when strong variable interactions or nonlinearity exist, or when the missing mechanism deviates from the Missing-At-Random (MAR) assumption[6]. Therefore, algorithm selection and tuning strategies that account for data characteristics remain important research topics.

In respiratory medicine, these missing-data issues are particularly important in longitudinal prediction tasks for Chronic Obstructive Pulmonary Disease (COPD), which requires early diagnosis and continuous monitoring. COPD is a chronic respiratory disease that causes chronic airflow obstruction due to chronic bronchitis or emphysema. The airway lumen narrows, or lung elasticity decreases, making breathing increasingly difficult in this irreversible chronic respiratory disease. According to the World Health Organization (WHO), COPD ranked as the third leading cause of death worldwide in 2019, accounting for approximately 6% of all deaths[7]. As a disease requiring long-term management, early diagnosis and active management play crucial roles in improving prognosis.

Recent deep learning-based research on predicting COPD severity progression has been actively conducted. However, improper handling of missing data, model uncertainty, and datasets too small relative to the number of predictor variables are identified as the greatest risk factors for bias[8]. The optimal method for handling missing data is not simply to delete cases containing missing data, but rather to mitigate their impact. While various imputation methods have been proposed to replace missing data with appropriate values, no definitive imputation method exists that is suitable for all cohort study datasets[9].

Pulmonary function tests (PFT), primarily used for COPD diagnosis, have the advantages of relatively low cost and repeatability. In particular, indicators such as FVC (Forced Vital Capacity), FEV1 (Forced Expiratory Volume in 1 Second), FEF25-75% (Forced Expiratory Flow at 25-75%), and PEF (Peak Expiratory Flow) are useful for quantitatively evaluating lung function decline. These indicators have close correlations with each other and change nonlinearly over time. In actual clinical settings, patient visit times are irregular, leading to irregular measurement intervals in PFT time-series data. The nonlinearity makes it difficult to preserve clinically meaningful patterns using only univariate imputation methods, making the missing value imputation problem critical when constructing prediction models[10,11].

This paper proposes an adaptive selection method for adjusting the initialization strategy and maximum number of iterations in the MICE algorithm. We analyze the statistical characteristics of representative pulmonary function indicators (FEV1, FVC, FEF25-75%, PEF) used in pulmonary disease diagnosis and validate the proposed MICE algorithm parameter selection method by applying it to actual pulmonary function time-series data. The main contributions can be summarized as follows: First, we interpret pulmonary function test time-series data not as univariate independent time series but as multivariate time-series data with correlation structures between variables, systematically analyzing the clinical significance and statistical characteristics of each pulmonary function indicator. Second, reflecting these characteristics, we present theoretical grounds for initializing MICE imputation techniques and setting iteration counts, applying these to actual pulmonary function data. Third, by comparing MICE and PCHIP (Piecewise Cubic Hermite Interpolating Polynomial) imputation under the same LSTM prediction model structure in environments with different missing rates (5%, 10%, 20%), we quantitatively analyze trends in prediction performance according to imputation methods and their potential applicability.

Methods

This retrospective study used longitudinal pulmonary function test data collected from patients who underwent repeated measurements at Chungnam National University Hospital. The study protocol, including data access, de-identification, and analysis procedures, was reviewed and approved by the Institutional Review Board of Chungnam National University Hospital (Approval No. CNUH IRB 2025-11-008). The Chungnam National University Hospital IRB was selected because the original clinical data used in this study were collected and managed by Chungnam National University Hospital, which served as the data-providing clinical institution for this retrospective study; the authors' affiliated institutions performed secondary data analysis under the approved protocol. All methods were carried out in accordance with the Declaration of Helsinki and relevant guidelines and regulations. Because this study used pre-existing retrospective clinical records and all data were analyzed after de-identification, the requirement for informed consent from the subjects and/or their legal guardians was waived by the Institutional Review Board of Chungnam National University Hospital.

Clinical Meaning and Statistical Properties of PFT Variables

Pulmonary function tests (PFTs) measure time-volume and volume-flow curves during a maneuver from maximal inhalation to maximal exhalation and back to maximal inhalation. This section examines the clinical significance and statistical characteristics of the major PFT measurement indicators. This characterization provides the theoretical background for designing missing-value imputation strategies that reflect variable-specific characteristics.

A. Linearity and Ratio Characteristics of FEV1 and FVC

FEV1 is the volume of air exhaled during the first second after maximal inhalation and serves as a representative indicator of the degree of bronchial constriction. In obstructive lung diseases, including COPD, FEV1 tends to decrease progressively over time, making it useful for predicting disease progression[12]. FVC is the total volume of air exhaled after maximal inhalation and is an indicator related to total lung capacity. It generally exhibits lower variability and more stable characteristics compared to FEV1[13]. Particularly in healthy adult populations, FEV1 and FVC show a strong linear correlation, and prior studies have approximated this relationship using linear regression models[14].

Property 1. Linear relationship between FEV1 and FVC

In healthy adult populations, forced expiratory volume in 1 second (FEV1) and forced vital capacity (FVC) can be approximated by the following linear regression relationship:

$$\mathrm{FEV1} = \alpha \cdot \mathrm{FVC} + \beta + \epsilon \tag{1}$$

, where ϵ is an error term with mean zero.

Rationale. FEV1 and FVC are indicators measured during the same forced expiratory maneuver, derived from the same lung volume-time curve. Therefore, the conditional expectation of FEV1 can be expressed linearly with respect to FVC, establishing $\mathbb{E}[\mathrm{FEV1}|\mathrm{FVC}] = \alpha \cdot \mathrm{FVC} + \beta$. Including observation error yields (1). ■

Property 1 indicates that FEV1 and FVC are not statistically independent, suggesting the limitations of univariate approaches that treat the two variables separately during missing-value imputation.

The ratio indicator widely used clinically in pulmonary function tests is defined as follows:

$$\frac{\text{FEV1}}{\text{FVC}} = \frac{\text{measured FEV1}}{\text{measured FVC}} \tag{2}$$

$$\frac{\text{FVC}}{\text{predicted FVC}} = \frac{\text{measured FVC}}{\text{predicted FVC}} \tag{3}$$

Implication 1. Nonlinearity of the FEV1/FVC Ratio

By Property 1 and (2) and (3), the FEV1/FVC ratio is derived as the following nonlinear relationship[14]:

$$\frac{\text{FEV1}}{\text{FVC}} = \frac{\alpha \cdot \text{FVC} + \beta}{\text{FVC}} = \alpha + \frac{\beta}{\text{FVC}} \tag{4}$$

Rationale. Dividing both sides of (1) by FVC yields $\frac{\text{FEV1}}{\text{FVC}} = \alpha + \frac{\beta}{\text{FVC}} + \frac{\epsilon}{\text{FVC}}$. From an expected value perspective ignoring the error term, (4) holds. ■

Therefore, the FEV1/FVC ratio is a nonlinear indicator dependent on FVC, and simple mean-based approaches may be inappropriate for missing value imputation[14].

B. High Variability Model of FEF25-75%

FEF25-75% represents the average airflow rate between 25% and 75% of FVC and is sensitive to small airway obstruction. This indicator exhibits large inter-individual variability and shows relatively large residual variance even when considering covariates such as age, height, and weight[10,15]. These characteristics indicate that FEF25-75% exhibits higher variability than other pulmonary function indicators.

Property 2. High unexplained variability of FEF25-75%

The mid-expiratory flow indicator FEF25-75% can be conceptualized by the following covariate-based regression equation:

$$\text{FEF25-75\%} = \gamma_0 + \gamma_1 \cdot \text{age} + \gamma_2 \cdot \text{height} + \gamma_3 \cdot \text{weight} + \eta \tag{5}$$

FEF25-75% is defined as the average flow rate in the middle section of FVC and is sensitive to individual small airway function. Therefore, even when it is modeled using regression (5) including age, height, and weight, the variance of the residual term η remains relatively large[17,18].

In practice, FEF25-75% has a coefficient of variation of approximately 20-30%, which is about 2-3 times higher than that of FEV1 or FVC[16,17]. This characteristic suggests the need for imputation techniques that are robust to outliers when handling missing FEF25-75% values.

C.Time-Series Variability of PEF

PEF represents the instantaneous maximum airflow rate reached during maximal-effort exhalation. This value is sensitive to measurement timing and patient effort level, and has been reported to exhibit repetitive variation patterns according to time of day[18]. This within-day variability indicates that PEF is not a static single value but a time-series variable that changes systematically over time.

Assumption 1. Periodic component in PEF over time

Peak expiratory flow (PEF) can be expressed as a time-series variable with the following periodic structure with respect to time t (quarter index):

$$\text{PEF}(t) = \text{PEF}_{\text{baseline}} + A \cdot \sin\left(\frac{2\pi t}{T} + \phi\right) + \xi(t) \tag{6}$$

Rationale. PEF is the instantaneous maximum flow achieved during maximum exhalation and is known to have within-day variability. Modeling the expected value over time as a periodic function yields (6)[19,20]. ■

This highlights the importance of temporal features when imputing missing PEF values.

D.Multivariate Covariance Structure

As discussed above, major pulmonary function indicators possess different statistical characteristics—such as linear dependence, high variability, and periodicity—yet are measured simultaneously during the same forced expiratory process. Therefore, it is reasonable to interpret these variables not as independent univariate data but as multivariate data with mutual dependencies[21].

Property 3. Multivariate covariance structure of pulmonary function indicators

The major pulmonary function indicator vector $\boldsymbol{Y} = [\text{FEV1,FVC,FEF25-75\%,PEF}]^T$ can be expressed with the following variance-correlation structure as a covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \rho_{14}\sigma_1\sigma_4 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_{22}^2 & \rho_{23}\sigma_2\sigma_3 & \rho_{24}\sigma_2\sigma_4 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_{33}^2 & \rho_{34}\sigma_3\sigma_4 \\ \rho_{14}\sigma_1\sigma_4 & \rho_{24}\sigma_2\sigma_4 & \rho_{34}\sigma_3\sigma_4 & \sigma_{44}^2 \end{bmatrix} \quad (7)$$

, where σ_i^2 represents the variance of each pulmonary function indicator, and ρ_{ij} represents the correlation coefficient between indicators i and j. Property 1-2 and Assumption 1 showed that each pulmonary function indicator is measured during the same forced expiratory process and shares different statistical characteristics: linear dependence (FEV1-FVC), high variability (FEF25-75%), and periodicity (PEF). Due to these characteristics, for any two variables $i, j \in$ {FEV1,FVC,FEF,PEF}, correlation coefficient ρ_{ij} can have non-zero values. Accordingly, covariance can be expressed as:

$$\text{Cov}(Y_i, Y_j) = \rho_{ij}\sigma_i\sigma_j \quad (8)$$

Organizing this for all variable pairs yields the covariance matrix in (7). ■

Property 3 justifies that the major pulmonary function indicators are not independent univariate data but multivariate data with an explicit correlation structure. Accordingly, adopting the MICE technique—which can reflect this covariance structure—as a missing-value imputation method is appropriate.

Theoretical Basis for MICE Parameter Settings

Pulmonary function indicators maintain strong inter-variable correlations; however, in actual clinical settings, observation intervals are irregular due to irregular visit cycles and missing values occur frequently. Such missingness alters the statistical structure of the time series and weakens the correlation structure among variables, making it difficult to accurately reflect inter-variable dependencies. This can negatively affect both the learning process and the performance of prediction models. MICE is well suited as a multivariate missing-data handling method to address these issues, and its main hyperparameters—initial imputation strategy and maximum number of iterations—should be set in accordance with the statistical characteristics of pulmonary function variables.

A.Stability of Iterative Imputation Algorithm and Number of Iterations

MICE is a chained-equation-based imputation technique that iteratively estimates missing values of each variable using the conditional distributions of other variables. The stability of this iterative conditional estimation process is closely related to the convergence concepts discussed in Markov Chain Monte Carlo (MCMC) theory.

Background 1. Convergence diagnostics for iterative imputation

In the MICE algorithm, the number of iterations is directly related to the stability of imputation results, and whether the iterative estimation process has stabilized can be evaluated through variance ratio-based indicators. The stability of MCMC-based iterative algorithms can be assessed using the Gelman-Rubin diagnostic indicator. The stability indicator at iteration step k is defined as follows[4,22,23]

$$\hat{R}(k) = \sqrt{\frac{\hat{V}(k)}{W(k)}} \tag{9}$$

, where W is the within-sequence variance and $\hat{V}$ is the total variance including the between-sequence variance. As $\hat{R}(k)$ approaches 1, the iterative estimation process is considered to have stabilized. Since MICE shares the same iterative structure as MCMC-based algorithms, setting the number of iterations is a key factor for ensuring imputation stability.

B.Stability of Iterative Imputation Algorithm and Number of Iterations

The number of iterations required in iterative imputation algorithms is influenced by the covariance structure and convergence characteristics of the data. In particular, the convergence speed of iterative conditional estimation is known to be governed by the minimum eigenvalue of the covariance matrix.

Design Guideline 1. Lower bound for the maximum number of iterations

For multivariate data with covariance structure, the lower bound of the maximum iteration count I_{max} for the iterative imputation algorithm can be expressed as follows[24]:

$$I_{max} \geq max\left\{\left\lceil\frac{log\,(\lambda_{min})}{log\,(\epsilon_{\text{target}})}\right\rceil, 10\right\} \tag{10}$$

Justification. λ_{min} is the minimum eigenvalue of the covariance matrix, representing the direction that converges most slowly in the iterative estimation process. The unconverged ratio after I iterations decreases geometrically, and the minimum number of iterations to reach the target stabilization level ϵ_{target} is derived as in (10)[24].

In practice, a minimum iteration count of 10, as recommended in the literature, is commonly used as the lower bound. ■

In this study, the default iteration count is set to 14 to ensure stable convergence, accounting for the strong collinearity between FEV1 and FVC and the slower convergence associated with the time-series variability of PEF.

C.Adjusting Iteration Count Considering Variable-Specific Variability

Pulmonary function variables differ in their variability, and these differences directly affect the convergence speed in the iterative imputation process.

Heuristic 1. Adjusting iterations according to variable-specific variability

The iteration count I_j for variable j is adjusted proportionally to the coefficient of variation (CV, Coefficient of Variation) of that variable:

$$I_j \propto \frac{\mathrm{CV}_j}{\overline{\mathrm{CV}}} \tag{11}$$

Rationale. FEF25-75% has a coefficient of variation 2-3 times higher than that of FEV1 and FVC. Variables with high variability maintain large variance in their estimated values throughout the iterative imputation process; therefore, applying a uniform iteration count may delay convergence[23,25]. Accordingly, adjusting the iteration count in proportion to the relative variability with respect to the average coefficient of variation is theoretically justified. ■

Therefore, in this study, the iteration count for the high-variability variable FEF25-75% is set to 20.

D.Variable-Specific Setting of Initial Imputation Strategy

The initial imputation strategy in MICE determines the starting point of the iterative imputation process, and any bias introduced at this stage can propagate through all subsequent iterations.

Design Guideline 2. Variable-specific initialization strategy in MICE

According to the statistical characteristics of variable j, the initial imputation strategy S_j is set as follows[28]:

$$S_j = \begin{cases} \text{mean} & \text{if skewness}_j \le \theta \text{ and } \mathrm{CV}_j < \delta \\ \text{median} & \text{if skewness}_j > \theta \text{ or } \mathrm{CV}_j \ge \delta \end{cases} \tag{12}$$

Rationale. For variables that are approximately normally distributed with low variability, mean-based initialization is appropriate because the mean accurately represents the central tendency. Conversely, for variables with high skewness or high variability, median-based initialization provides a more robust starting point, as the mean is sensitive to outliers[25,26]. ■

Accordingly, the mean-based initialization is applied to FEV1 and FVC, while the median-based initialization is applied to FEF25-75% and PEF.

Imputation Processing Procedures, Dataset Construction, and Evaluation Strategy

Given that pulmonary function indicators exhibit strong inter-variable correlations and considerable noise may be introduced during the measurement process, MICE-based missing-value imputation is applied. Random forest is used as the MICE estimator because it can effectively reflect nonlinear relationships and interactions among variables. This configuration enables more stable estimation of the complex dependency structure among pulmonary function indicators.

The overall data processing procedure consists of three main stages: (1) time-series sample window partitioning, (2) missing value imputation and label generation, and (3) training dataset construction and evaluation strategy establishment (Figure 1).

A.Time-Series Sample Window Partitioning

Patient-specific time-series data are partitioned into fixed-length sliding-window units to generate samples suitable for model training. Each window is classified as either a complete sample or a sample containing missing values, based on the presence of missing entries in the raw data before imputation. This classification provides basic information for identifying missing-data patterns and determining the scope of imputation. Additionally, imputation methods are applied to time-series samples containing missing values in order to analyze the performance of the imputation techniques[21].

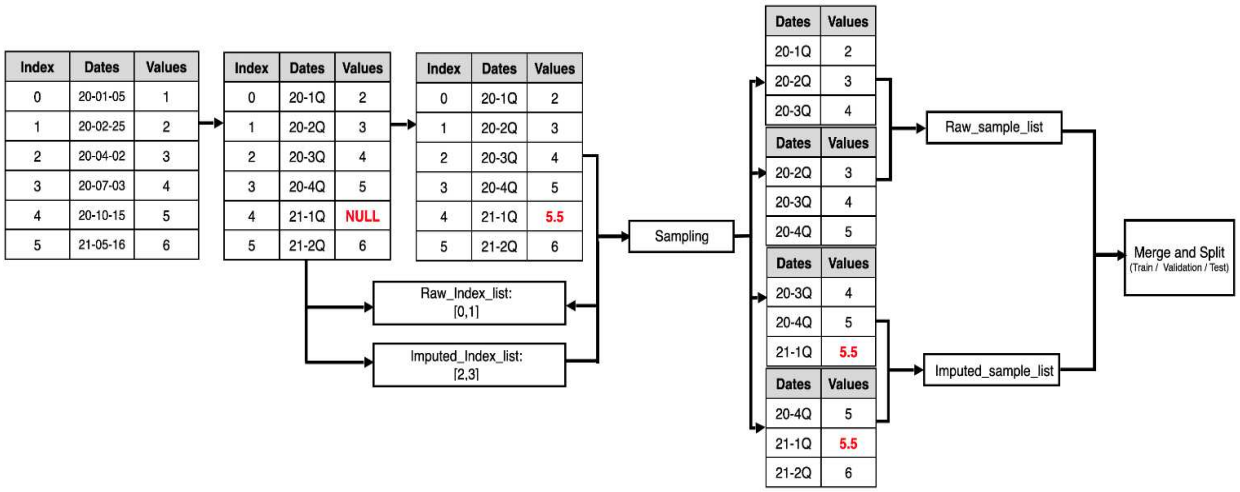


Figure 1. Imputation and sample construction process for PFT time-series data

B.MICE-Based Missing Value Imputation and Label Generation

Two initialization strategies were compared to analyze performance differences according to the MICE initialization method. The mean-only method applies mean-based initialization to all four core pulmonary function indicators (FEV1, FVC, FEF25-75%, and PEF). In contrast, the adaptive method reflects variable-specific statistical characteristics by applying mean-based initialization to FEV1 and FVC, and median-based initialization to FEF25-75% and PEF.

The adaptive initialization strategy is used as the baseline setting because it reflects the variable-specific characteristics analyzed in the Method section. The setting is intended to minimize bias in the initial estimates and ensure convergence stability according to the statistical characteristics of each indicator. In addition, the baseline maximum iteration counts are determined by considering differences in convergence speed across variables. These settings are selected by balancing imputation accuracy and computational efficiency and are used as the reference configuration for analyzing performance

Parameter	Initialization strategy	Max iterations
FEF25-75%	median	20
FEV1	mean	14
FVC	mean	14
PEF	median	14

Table 1. Baseline hyper-parameters configuration for the tuned MICE experiment

differences according to changes in iteration count. Table 1 presents the baseline imputation parameter settings used in this study.

The Result section compares and analyzes the effects of changes in the initialization method and iteration count on prediction performance based on these settings.

After imputation is completed, indicator values such as FVC, FEV1, and FEV1/FVC ratio are recalculated from the imputed variables. Using these pulmonary function indicators, labels corresponding to four COPD categories (Normal, Restrictive, Obstructive, and Mixed) are generated for each time-series sample. These labels serve as reference information reflecting the clinical state of each time-series sample and are used as auxiliary information for subsequent dataset construction and interpretation of the prediction results.

C.Training Dataset Construction and Evaluation Strategy

After imputation, the generated time-series samples are managed separately as original samples and imputed samples. Here, original samples are segments that contained no missing values before imputation, whereas imputed samples refer to segments that originally contained missing values and were subsequently imputed. In the final learning stage, the two sample sets are integrated into a single dataset, randomly shuffled, and split into training, validation, and test sets.

Subsequently, dataset construction and evaluation are performed based on the entire data including imputed samples. This configuration aims to quantitatively analyze the effects of imputation parameter setting changes and missing rate increases on time-series prediction model performance.

LSTM-Based Pulmonary Function Time-Series Prediction Model

An LSTM-based regression model is constructed to predict pulmonary function values at future time points using imputed pulmonary function time-series data. This section defines the time-series data-based prediction problem, describes the structure of the LSTM-based prediction model, and explains the model-training procedure and performance-evaluation metrics.

A.Time-Series Prediction Problem Definition

Pulmonary function test data constitute multivariate time-series data repeatedly measured for each patient over a long period. In this study, the prediction problem is defined such that pulmonary function test data from the past two previous time points are used as input to predict pulmonary function values at the next time point. In other words, from a time-series perspective, data at time points T-1 and T are used as input to predict pulmonary function values at time point T+1. This input-output structure reflects the gradual temporal change patterns of pulmonary function indicators.

The model input consists of multivariate time-series data that include pulmonary function test indicators and the patients' basic clinical information. The major pulmonary function indicators include FVC, FEV1, FEF25-75%, and PEF. These variables are measured during the same forced expiratory process and are mutually correlated while also changing over time. Therefore, the model is designed to jointly utilize multivariate time-series information together to predict pulmonary function values at future time points[27,28].

B.LSTM Structure

LSTM, which is effective for learning temporal dependencies in time-series data, was proposed to overcome the limitations of conventional recurrent neural networks in modeling long-term dependencies and has been widely used in time-series data analysis and prediction tasks[29].

The overall time-series analysis architecture consists of LSTM layers followed by fully connected layers. The LSTM layers extract temporal features from the input time-series data, and the extracted features are transformed into final prediction values through the fully connected layers. Through this process, the model learns the temporal change patterns of pulmonary function indicators and estimates pulmonary function values at the next time point.

The input sequence length of the model is set to 2. This setting is consistent with the problem definition, in which pulmonary function values at the next time point are predicted using data from the two previous time points. This sequence configuration supports stable model training while reflecting the temporal flow of quarterly pulmonary function time-series data.

In addition, the same LSTM-based prediction model structure is maintained across all experiments to compare performance differences according to the imputation technique or imputation parameter settings. In other words, the analysis focuses on the effects of differences in input-data imputation methods on prediction performance rather than on changes in the model structure itself.

C.Model Training Setting and Evaluation Metrics

A regression-based learning approach was applied during model training in accordance with the characteristics of the time-series prediction problem. To evaluate the model's generalization performance more robustly, 5-fold cross-validation (K-fold cross-validation, K=5) was applied. Models were trained independently for each fold, and after training, the final-epoch model for each fold was evaluated on the corresponding test data. Final performance was calculated by averaging the test results across the 5 folds. The Adam optimizer was used for model optimization, with the learning rate set to 0.001 and the batch size set to 32[21].

Mean Squared Error (MSE) is used as the loss function to minimize prediction error during model training:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

, where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the total number of samples.

Additionally, to quantitatively analyze the effect of changes in imputation methods on prediction model performance, regression-based performance metrics MAPE (Mean Absolute Percentage Error), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R^2 (Coefficient of Determination) indicating how much of the variance in actual data the model explains are used:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{14}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (17)$$

, where $\bar{y}$ represents the mean of actual values.

Specifically, MSE is used as the loss function during model training, and RMSE, MAE, MAPE, and coefficient of determination are used to evaluate prediction performance in experimental result analysis. Additionally, prediction performance is compared while maintaining the same LSTM prediction model and varying imputation techniques and imputation parameters.

Results

Experiments were conducted to quantitatively analyze the effects of imputation techniques applied to pulmonary function time-series data on prediction-model performance under missing rates of 5%, 10%, and 20%. MICE, a multivariate correlation-based imputation technique, and PCHIP, a curve-based imputation technique, were compared under the same LSTM-based prediction model and training conditions. Additionally, further experiments were conducted to analyze the effects of the iteration count and initialization method on prediction performance during the MICE imputation process, and the results are reported as average metric values based on 5-fold cross-validation.

Performance Comparison by Imputation Technique

When PCHIP imputation is applied, stepwise age calculation based on January 1 is used for 'age', linear imputation is applied to 'height' and 'weight', and monotone cubic imputation is applied to core pulmonary function indicators. When MICE is applied, the same preprocessing rules are applied to 'age', 'height', and 'weight', but MICE imputation is applied to core pulmonary function indicators.

The performance comparison results by imputation technique across missing rate (Table 2) show that PCHIP provides relatively consistent performance over all missing rate conditions. This indicates that curve-based imputation can effectively reflect the continuous change patterns of pulmonary function time series. Additionally, MICE maintained a low relative error even under the 20% missing rate condition, suggesting that multivariate correlation-based imputation can provide valid input information to prediction models even as missingness increases. Therefore, MICE is an imputation strategy that goes beyond simple value replacement by enhancing the applicability of time-series prediction through its ability to reflect structural relationships among variables.

Missing rate	Method	Configuration	Iterations	MAPE	RMSE	R^2
5%	MICE	adaptive	30	5.859%	0.31153	0.98188
5%	PCHIP	monotone	N/A	5.891%	0.31163	0.98186
10%	MICE	adaptive	30	5.874%	0.31148	0.98181
10%	PCHIP	monotone	N/A	5.876%	0.30919	0.98213
20%	MICE	mean-only	30	5.831%	0.31185	0.98182
20%	PCHIP	monotone	N/A	5.878%	0.31158	0.98184

Table 2. Performance comparison by imputation method and missing rate

Performance Analysis by Iteration Count and Initialization Method

Additional experiments were performed with iteration counts of 14, 20, and 30 to analyze the effect of iteration count on prediction performance during the MICE imputation process. In addition, the mean-only and adaptive methods were compared to evaluate the effect of the initialization method. The mean-only method applies mean-based initialization to all four core pulmonary function indicators, whereas the adaptive method applies mean-based initialization to FEV1 and FVC and median-based initialization to FEF25-75% and PEF, thereby reflecting variable-specific statistical characteristics. This design jointly evaluates the effects of both the initialization strategy and the iteration count on prediction performance.

Under the 5% missing rate condition, the combination of the adaptive method and 30 iterations shows the lowest MAPE. Under the 10% missing rate condition, the same combination also performs best. Under the 20% missing rate condition, the combination of the mean-only method and 30 iterations shows the best performance. These results indicate that the effects of the iteration count and initialization method may vary depending on the missing rate.

The complete 5-fold average results for all 18 configurations are presented Appendix Table A1. Overall, MICE performance did not deteriorate sharply as the missing rate increased, indicating that MICE, by leveraging multivariate correlation structures, can estimate missing values on the basis of inter-variable relationships even in high-missingness conditions. In other words, particularly under high missing rate conditions, approaches that reflect correlation structures among variables, rather than relying only on single variable continuity, can provide meaningful long-term interpretability.

Discussion

The experimental results show that both imputation techniques applied to pulmonary function time-series data maintain relatively consistent prediction performance despite increasing missing rates. PCHIP yields similar performance levels across all missing rate conditions by preserving time-series continuity and monotonicity. By contrast, MICE more directly reflects the multivariate characteristics of pulmonary function data because it utilizes correlation structures among variables.

Particularly, MICE maintain low error levels based on MAPE even as the missing rate increases. The fact that MAPE was 5.831% even in the 20% missing rate environment demonstrates that multivariate correlation-based imputation can provide sufficiently valid information to prediction models even in high-missingness environments. Additionally, the effects of iteration count and initialization method vary according to missing rate, making it difficult to say that one fixed setting is advantageous in all environments. However, both adaptive and mean-only methods show relatively consistent performance when combined with appropriate iteration counts, suggesting that MICE can flexibly respond to various data environments according to setting adjustments.

This study has several limitations. First, the analysis was conducted using a retrospective longitudinal pulmonary function test dataset, and the findings may therefore be influenced by the characteristics of the data source, measurement environment, and follow-up patterns. Second, missing values were artificially introduced at predefined missing rates of 5%, 10%, and 20% to compare imputation strategies under controlled conditions. These experimental missingness settings may not fully reflect real-world missing-data mechanisms, including missing-not-at-random patterns. Third, the same LSTM-based prediction model was used across experiments to isolate the effect of imputation strategies; therefore, the results may not directly generalize to other prediction models or clinical outcomes. Finally, external validation using independent multi-institutional datasets was not performed. Future studies should validate the proposed MICE parameter-selection strategy using larger and more diverse clinical datasets collected from multiple institutions and under real-world missing-data conditions.

Conclusion

This study proposed an adaptive selection method for adjusting the initialization strategy and iteration count of the MICE algorithm and validated its effectiveness for mitigating missing data problems in

pulmonary function time-series data by applying it to an LSTM-based pulmonary function value prediction model. Since pulmonary function indicators have high mutual correlation while variable-specific variation characteristics differ, settings reflecting these characteristics are necessary even in the missing value imputation process. Based on the statistical characteristics of FEV1, FVC, FEF25-75%, and PEF, criteria for setting initialization method and iteration count were established, and prediction performance changes were compared in different missing rate environments.

Results of applying the adaptive selection method for adjusting MICE algorithm initialization strategy and iteration count showed that the MICE-based approach maintained relatively consistent prediction performance even as the missing rate increased. In particular, by reflecting multivariate correlation structures, meaningful prediction information can be provided even in high-missingness environments. Additionally, comparison results with PCHIP under the same prediction model structure showed that MICE demonstrated competitive performance from a relative error perspective and confirmed the possibility of being utilized as an appropriate imputation strategy for long-term follow-up pulmonary function time-series data analysis in that it can jointly utilize structural relationships between variables.

Data availability

The study protocol was approved by the Institutional Review Board (IRB) of Chungnam National University Hospital (Approval No. CNUH IRB 2025-11-008). However, the datasets used in this study are not publicly available because they contain patient-derived clinical information and are subject to privacy and ethical restrictions. Access to the data may be considered only upon reasonable request to the corresponding author and with appropriate approval from the relevant institutional review board.

References

1. M. Liu et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation methods. *Artif. Intell. Med.*, **145**, DOI: https://doi.org/10.1016/j.artmed.2023.102587 (2023).

2. X. Chen et al. Dealing With Missing, Imbalanced, and Sparse Features During the Development of a Prediction Model for Sudden Death Using Emergency Medicine Data: Machine Learning Approach. *JMIR Med Inform.,* **11**, DOI: https://doi.org/10.2196/38590 (2023).

3. M. Ghanem et al. Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review. *Brain Sci.*, **13**, DOI: https://www.doi.org/10.3390/brainsci13121723 (2023)

4. S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R, *J. Stat. Softw*., **45**, DOI: https://www.doi.org/10.18637/jss.v045.i03 (2011).

5. Z. Zhang. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Ann. Transl. Med.*, **4**, DOI: https://www.doi.org/10.3978/j.issn.2305-5839.2015.12.63 (2016).

6. A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am. J. Epidemiol*., **179**, 764-774, DOI: https://www.doi.org/10.1093/aje/kwt312 (2014).

7. World Health Organization. (2024, Aug.). The top 10 causes of death. *WHO Fact Sheets*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

8. L. A. Smith et al. Machine learning and deep learning predictive models for long-term prognosis in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. *The Lancet Digital Health*, **5**, e872-e881, DOI: https://www.doi.org/10.1016/S2589-7500(23)00177-2 (2023).

9. J. Li et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Med. Res. Methodol*., **24**, DOI: https://www.doi.org/10.1186/s12874-024-02173-x (2024).

10. M. R. Simon et al. FEF25-75 and FEV1/FVC in relation to clinical and physiological parameters in Asthmatic Children with Normal FEV1 Values. *J. Allergy Clin. Immunol*., **126**, 527-534, DOI: https://www.doi.org/10.1016/j.jaci.2010.05.016 (2011).

11. C. Sun et al. Time pattern reconstruction for classification of irregularly sampled time series. *Pattern Recognit*., **147**, DOI: https://www.doi.org/10.1016/j.patcog.2023.110075 (2024).

12. A. Balasubramanian et al. Estimating rate of lung function change using clinical data and machine learning models in chronic respiratory disease: Comparison of regression methods. *BMJ Open Respir Res.,* **11**, DOI: https://www.doi.org/10.1136/bmjresp-2023-001896 (2024).

13. P. C. Almeida et al. Longitudinal trends in clinical characteristics and lung function of patients with severe asthma under treatment in Brazil. *BMC Pulm Med.*, **16**, DOI: https://www.doi.org/10.1186/s12890-016-0302-5 (2016).

14. T. Gólczewski, W. Lubiński, and A. Chciałowski. A mathematical reason for FEV1/FVC dependence on age. *Respir. Res*., **13**, DOI: https://www.doi.org/10.1186/1465-9921-13-57 (2012).

15. Y. J. Choi et al. FEF25-75% Values in Patients with Normal Lung Function as a Predictor of Chronic Obstructive Pulmonary Disease. *Int. J. Chron. Obstruct. Pulmon. Dis*., **15**, 1231-1238, DOI: https://www.doi.org/10.2147/COPD.S261732 (2020).

16. C. A. Cox et al. Predicted values for the forced expiratory flow adjusted for forced vital capacity, a descriptive study. *Eur. Respir. J. Open Res.*, **6**, DOI: https://www.doi.org/10.1183/23120541.00426-2020 (2020).

17. S. Stanojevic et al. Reference ranges for spirometry across all ages. *Am. J. Respir. Crit. Care Med*., **177**, 253-260, DOI: https://www.doi.org/10.1164/rccm.200708-1248OC (2008).

18. S. Abut et al. Analysis of Pulmonary Function Test Results by Using Gaussian Mixture Regression Model. *J. Clin. Med. of Kazakhstan*, **18**, DOI: https://www.doi.org/10.23950/jcmk/10919 (2021).

19. H. M. Boezen et al. Distribution of peak expiratory flow variability by age, gender and smoking habits in a random population sample aged 20-70 yrs. *Eur. Respir. J.*, **7**, 1814-1820, DOI: https://www.doi.org/10.1183/09031936.94.07101814 (1994).

20. B. G. Higgins et al. The distribution of peak expiratory flow variability in a population sample. *Am. Rev. Respir. Dis*., **140**, 1368-1372, DOI: https://www.doi.org/10.1164/ajrccm/140.5.1368 (1989).

21. B. Choi, H. Kim, and S. H. Jeon. LSTM-based time series forecasting of pulmonary function test for COPD early diagnosis. *J. Korean Inst. Commun. Inf. Sci*., **49**, 346–355, DOI: https://www.doi.org/10.7840/kics.2024.49.3.346 (2024).

22. E. A. Stuart et al. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *Am. J. Epidemiol*., **169**, 1133-1139, DOI: https://www.doi.org/10.1093/aje/kwp026 (2009).

23. S. van Buuren and K. Groothuis-Oudshoorn. Fully conditional specification (FCS). *Flexible Imputation of Missing Data*, 2nd ed., Boca Raton, FL, USA: CRC Press, chapter 4.5 (2018).

24. S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat*., **7**, 434-455, DOI: https://www.doi.org/10.1080/10618600.1998.10474787 (1998).

25. S. van Buuren. Algorithmic options. *Flexible Imputation of Missing Data*, 2nd ed., Boca Raton, FL, USA: CRC Press, chapter 6.5 (2018)

26. A. Kumar. (2025, Jan.). A better way to handle missing values in your dataset using IterativeImputer. *Towards Data Science*. [Online]. Available: https://towardsdatascience.com/

27. H. Watz et al. Spirometric changes during exacerbations of COPD: a post hoc analysis of the WISDOM trial. *Respir. Res.*, **19**, DOI: https://www.doi.org/10.1186/s12931-018-0944-3 (2018).

28. Y. S. Sim et al. Spirometry and bronchodilator test. *Tuberc. Respir. Dis.*, **80**, 105–112, DOI: https://www.doi.org/10.4046/trd.2017.80.2.105 (2017).

29. S. Gonem, W. Janssens, N. Das, and M. Topalovic. Applications of artificial intelligence and machine learning in respiratory medicine. *Thorax*, **75**, 695–701, DOI: https://www.doi.org/10.1136/thoraxjnl-2020-214556 (2020).

Appendix

Appendix Table A1 from the original article, summarizing 5-fold average performance of all 18 MICE experiments by missing rate, initialization strategy, and iteration count.

Missing rate	Method	Iterations	MSE	MAPE(%)	MAE	RMSE	R^2
5%	adaptive	14	0.00138833	5.870027	0.14351989	0.31184028	0.98185713
5%	adaptive	20	0.00138832	5.884456	0.14216881	0.30526773	0.98255178
5%	adaptive	30	0.00138424	5.858718	0.14330669	0.31153463	0.98188262
5%	mean-only	14	0.00139111	5.890936	0.14314079	0.31139082	0.98188626
5%	mean-only	20	0.00138805	5.871876	0.14290708	0.31119722	0.98185456
5%	mean-only	30	0.00138571	5.871031	0.14255365	0.31030963	0.98197118
10%	adaptive	14	0.00137632	5.867456	0.14257604	0.31001039	0.98200605
10%	adaptive	20	0.00137920	5.859438	0.14336076	0.31186826	0.98180791
10%	adaptive	30	0.00139003	5.874248	0.14311023	0.31147720	0.98180900
10%	mean-only	14	0.00137800	5.854070	0.14356533	0.31194780	0.98181441
10%	mean-only	20	0.00139018	5.885898	0.14275790	0.30854341	0.98219574
10%	mean-only	30	0.00138997	5.884685	0.14279322	0.31071446	0.98188280
20%	adaptive	14	0.00138944	5.883223	0.14307515	0.31135237	0.98184422
20%	adaptive	20	0.00135687	5.788250	0.14354421	0.31184788	0.98182638
20%	adaptive	30	0.00136088	5.847218	0.14344321	0.31187480	0.98180056
20%	mean-only	14	0.00137147	5.847173	0.14207858	0.30640646	0.98245427
20%	mean-only	20	0.00137821	5.884473	0.14242041	0.30828231	0.98224050
20%	mean-only	30	0.00135961	5.830819	0.14351931	0.31185330	0.98182475

Table A1. Five-fold average performance of all 18 MICE experiments by missing rate, initialization strategy, and iteration count

Competing interests (mandatory)

The authors declare no competing interests.

Acknowledgements (optional)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2022-NR072190).

Author contributions

Y.G. conducted the experiments, performed hyperparameter tuning, analyzed the results, and prepared the figures and tables. H.K. conceived the selective application method for MICE hyperparameters, developed its mathematical proof, and supervised the study. M.-M.L. refined the idea of selective MICE hyperparameter application based on clinical and pathological expertise and contributed to the interpretation of the results. J.M. provided expertise in artificial intelligence-based pattern recognition, contributed to the methodological design of the prediction model, and supported the interpretation of model performance and predictive patterns. All authors reviewed, edited, and approved the final manuscript.