Journal of Big Data

METHODOLOGY Open Access

Uncertainty-aware approach for multiple imputation using conventional and machine learning models: a real-world data study

Romen Samuel Wabina[1], Panu Looareesuwan[1*], Suphachoke Sonsilphong[2*], Htun Teza[1], Wanchana Ponthongmak[1], Gareth McKay[3], John Attia[4], Anuchate Pattanateepapon[1], Anupol Panitchote[2] and Ammarin Thakkinstian[1]

*Correspondence:
panu.loo@mahidol.edu;
suphason@kku.ac.th

[1] Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, 270 Rama VI Road, Phaya Thai, Bangkok 10400, Thailand
[2] Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand
[3] Centre for Public Health, School of Medicine, Dentistry, and Biomedical Sciences, Queen's University Belfast, Belfast, Northern Ireland, UK
[4] School of Medicine and Public Health, University of Newcastle, Newcastle, NSW, Australia

Abstract

Missing data poses a significant challenge in clinical real-world studies, often arising from unplanned data collection, misplacement, patient loss to follow-up, and other factors. While multiple imputation by chained equations (MICE) is a widely used method, its sequential nature introduces uncertainty, potentially impacting the prediction model performance. We proposed and evaluated three uncertainty-aware functions (i.e., uncertainty sampling (US), probability of improvement (PI), and expected improvement (EI)) integrated with linear regression (LinearReg), decision tree (DT), random forest (RF), and extreme gradient boosting (XGBoost) using three large datasets: chronic kidney disease (CKD, n = 31,043), hypertension cohort from Ramathibodi Hospital (HT-RAMA, n = 140,047) and Khon Kaen University Hospital (HT-KKU, n = 108,942) with high missing rates. In the CKD cohort, uncertainty-aware models significantly improved performance (evaluated by root mean squared error (RMSE) and mean absolute error (MAE)) over standard MICE, except for XGBoost. LinearReg-EI performed best (RMSE 0.12, MAE 0.36), followed by RF-EI (RMSE 0.22, MAE 0.34), and DT-EI (RMSE 0.21, MAE 0.38). In HT-RAMA, LinearReg-US performed best (RMSE 0.24, MAE 8.15), outperforming RF-US (RMSE 0.92, MAE 8.58) and DT-PI (RMSE 0.96, MAE 8.74). Similarly, in HT-KKU, LinearReg-US performed best (RMSE 0.98, MAE 12.00), followed by RF-PI (RMSE 1.93, MAE 12.90) and DT-US (RMSE 2.10, MAE 12.63). Uncertainty-aware models produced imputed distributions closely resembling the original data, unlike standard MICE. Our findings suggest that incorporating uncertainty functions can improve MICE, particularly for LinearReg, RF and DT. Further research is warranted to validate these findings across diverse clinical settings and model types.

Keywords: Multiple imputation, Uncertainty-aware models, Uncertainty functions, Real-world data, Missing data

Introduction

Missing data presents an unavoidable challenge in clinical research, particularly in real-world longitudinal data studies [1–3]. Data may be missing for various reasons, such as unplanned collection, misplacement, patient loss to follow-up, or a range of other causes. Such missingness represents a critical barrier in clinical research particularly in real-world datasets as improper handling of missing data can lead to biased estimates, decreased statistical power, and less generalizable findings, thereby leading to invalid research conclusions [1]. Various statistical approaches can handle missing data [4–7]. One common method is listwise deletion, which involves removing any data rows that contain missing values. Although this approach is easy and straightforward, it can significantly reduce generalizability, statistical power when many values are missing, and may result in biased parameter estimates when missingness is also related to other factors that affect the outcome [8–10].

Imputation approaches, including single and multiple imputation (MI) techniques, have been used to predict missing data [4–7] based on the assumption that the data are missing at random (MAR). Single imputation typically substitutes missing values with one estimated value such as mean, median, mode, or the last observed value [11]. However, single imputation often leads to bias, underestimation of standard errors, and distortion of data distribution [5, 12]. MI generates predicted values derived from distributions and relationships among other observed features and missing values [13]. Therefore, missing data are replaced with different plausible values for individuals, estimated according to relationships among observed and missing features.

MI typically assumes that data are MAR, i.e., missingness is not associated with the missing value, conditional on the observed data [14]. There are two general approaches, i.e., joint modeling and fully conditional specification, the latter also known as multivariate imputation by chained equations (MICE) [4]. Joint modeling assumes a multivariate normal distribution where imputations are generated from the fitted distribution. However, employing a joint model can present challenges, particularly when data contain high dimensionality with mixed binary and categorical features, which may exhibit non-normal distributions [15, 16]. MICE, on the other hand, imputes missing values using separate univariate conditional distributions for each incomplete variable given all the others, simulating iteratively through each incomplete variable [15]. MICE is composed of three steps: first, it generates/creates multiple m datasets by iteratively imputing missing values using a regression model at multiple iterations; second, the estimation step is then performed to estimate parameters of interest for each m imputed dataset using standard techniques applicable to complete data; finally, these estimates are pooled or combined across all m datasets, providing a comprehensive estimate that accounts for missing data.

Despite its widespread use, consensus on the best framework for MICE across various scenarios remains elusive [17, 18], including model choice, regression algorithm, hyperparameter tuning configurations, and even modeling decisions. The choice of the appropriate model in MICE remains context-dependent and varies according to the type of missing data. By default, MICE uses predictive mean matching for imputing missing values in numeric data while using a logit model for binary/categorical data [19]. However, given the advent of artificial intelligence, several studies have also utilized

machine learning (ML), such as classification and regression trees (CART) [20], decision trees (DT) [21], random forests (RF) [22], and extreme gradient boosting (XGBoost) [23]. These methods have demonstrated promising results in various clinical studies [13, 24]. The performance of different imputation methods can be influenced by various factors such as the amount of missing data [25], its underlying distribution [26], and the complexity of the relationships between observed predictors and missing variables [27]. Existing literature provides few details on hyperparameter tuning and its impact on the performance of imputation methods [28]. Consequently, the selection of hyperparameter values is generally limited to predictive modeling contexts, where choices are guided primarily by prediction performance. Furthermore, the inclusion of outcome variables in the imputation model is controversial as it may lead to increased bias [29] or less robust estimates [24, 30].

While MICE has demonstrated considerable effectiveness and has undergone significant advancements, it remains a challenging problem. This is due to its reliance on using observed features to sequentially predict and impute missing data, which can introduce uncertainty into the predicted values that may not perfectly reflect the true values, potentially affecting the accuracy of subsequent analyses. Once missing values have been imputed, they are used as predictors for sequentially imputing other values; the uncertainty associated with the previous imputation is carried over and can amplify the uncertainty in subsequent imputations. If the prediction model is updated by acquiring the imputed values for subsequent iterations, the performance of the model may be degraded, particularly if the imputed values are a less accurate reflection of the true values.

Incorporation of uncertainty in imputation methods is crucial [31, 32]. Most uncertainty-based imputation studies focus on the MI framework since it was originally constructed to address this problem by generating multiple imputed datasets. However, the original MI only considers the uncertainty within the scope of the missing data itself, typically through the variability introduced by the multiple datasets generated. To the best of our knowledge, only two studies considered imputation uncertainty in the presence of missing data. First, Han and Kang [32] addressed this shortcoming using uncertainty functions (i.e., random selection, confidence, margin, entropy [33], and gini index) in MI frameworks that select samples with low uncertainty. These functions guide the selection of samples for model training to identify which data points have low uncertainty to reduce model degradation [33]. However, these functions are specifically designed for ordinal or nominal data and may be less appropriate for continuous data, especially for real-world data (RWD) in clinical settings where some continuous data are present. In addition, their uncertainty models were created based on the assumption of missing completely at random (MCAR), which may not be applicable to clinical data where MAR is more common. Furthermore, only RF was used within MICE and comparisons with other imputation methods in their study were not considered, which may have been better performance. Second, Tharwat and Schenck [31] extended a previous study by comparing the performance of median imputation, K-Nearest Neighbor, and RF [31] using 18 public datasets with sample sizes ranging from 150 to 3147. Incomplete datasets were created under MCAR with different missingness ratios.

An uncertainty-aware model is a predictive model that not only aims at predictions but also quantifies the uncertainty associated with those predictions. Recently, it has been applied in large language models [34, 35], medical image classification [36], and causal inference [37, 38] to ensure validity of predictions. To quantify uncertainties, ML models incorporate uncertainty functions, also known as acquisition functions, that guide the process for selecting the next data points to evaluate, leveraging the uncertainty estimates from the model [31, 32, 39]. These functions find a balance between sampling points, i.e., where the model is uncertain, and where the model predicts optimal outcomes. This balance is particularly crucial for imputation, as it allows the model to effectively handle missing data by exploring a range of potential values and incorporating the most plausible ones, thereby improving overall data integrity and predictive performance [31, 32]. Known uncertainty functions for handling continuous data include uncertainty sampling (US) [40], probability of improvement (PI) [41], and expected improvement (EI) [42]. The US method aims to improve overall model accuracy by selecting data points where the model is most uncertain. PI prioritizes points with high potential for improvement, while EI balances both potential gains and associated uncertainty.

While the potential of uncertainty-aware functions to enhance imputation methods is significant, their application in MI remains largely unexplored. This research gap presents an opportunity to improve prediction accuracy, particularly in real-world healthcare settings where accurate data is paramount for predicting patient outcomes. Our study addresses this gap by proposing uncertainty-aware methods for prediction models that explicitly account for the inherent uncertainty introduced by imputation, using machine learning (ML), such as classification and regression trees (CART) [20], decision trees (DT) [21], random forests (RF) [22], and extreme gradient boosting (XGBoost) [23].

We propose a novel approach in reducing this uncertainty by strategically querying additional samples based on their estimated uncertainty levels. Our research leverages the widely used MICE framework and incorporates tailored uncertainty functions. To validate our approach, we apply these methods to three large RWD cohorts relating to chronic kidney disease (CKD) and hypertension (HT). The results of these applications are presented and discussed, providing insights into the effectiveness of our proposed uncertainty-aware methods for enhancing imputation in complex healthcare datasets.

Method

Datasets

Two retrospective, longitudinal RWD cohorts were obtained from the data warehouses of the Clinical Epidemiology and Biostatistics (CEB) Department, Faculty of Medicine Ramathibodi Hospital, Mahidol University (RAMA) (see more details on website https://www.rama.mahidol.ac.th/ceb/CEBdatawarehouse/Overview. Accessed 19 February 2025), with additional cohort from collaboration of the Srinagarind Hospital, Khon Kaen University (KKU) (https://www.rama.mahidol.ac.th/ceb/CEBdatawarehouse/multicenter/ht. Accessed 19 February 2025). The RAMA database included 31,043 patients with CKD and 140,047 patients with HT (called HT-RAMA), identified between 1st January 2010 until 31st December 2022, while KKU database contributed 108,942 patients

with HT from 1st January 2015 until 31st December 2023. The CKD-RAMA and HT-RAMA cohorts share several features and include 9.349 overlapping patients (6.7% of HT-RAMA, 30.1% of CKD-RAMA) with both CKD and HT. The CKD cohort was used for internal validation, while both HT-RAMA and HT-KKU cohorts served as external validation datasets to demonstrate the generalizability of the proposed imputation method across different clinical domains and larger datasets. All datasets were approved by the Human Research Ethics Committee of Ramathibodi and Srinagarind Hospitals (COA. No. MURA2024/468 for CKD; COA No. MURA2023/689 for HT-RAMA; and HE681020 for HT-KKU).

The CKD, HT-RAMA, and HT-KKU cohort datasets were independently constructed using eight domains including demographics, physical examination, medical conditions, diagnoses, procedures, laboratory results, medications, and outcomes. The three cohorts were curated through linked hospital numbers and visit dates, and subsequently merged into a single, unified dataset. Laboratory data were also standardized based on Unified Code for Units of Measure (UCUM). For the CKD cohort, we identified CKD patients using International Classification of Diseases 9th and 10th editions (ICD-9 and ICD-10 codes), and eGFR < 60 for more than two consecutive occasions over more than 90 days or more, and patients with glomerulonephritis following confirmation by kidney biopsy. Similarly, anti-hypertensive medication and ICD-10 codes were used to identify patients in HT cohort. Due to the high dimensionality of the data from varying numbers of patient visits, laboratory measures acquired, or medication prescribed at different visits, we aggregated visit data into one-year intervals. To minimize the impact of outliers and prevent the prediction of clinical values that deviate significantly from real-world ranges, truncation was used for all features in both datasets. The range of values for covariates in CKD and HT datasets are listed in Supplementary Table 1.

Baseline characteristics of patients (i.e., demographics, comorbidities, and medications) are described in Supplementary Table 2. The percentage of missing data for each variable in the CKD and HT datasets (HT-RAMA and HT-KKU) ranged from 4.16% to 49.01% in CKD, 15.75–53.32% in HT-RAMA, and 28.20–67.88% in HT-KKU. In the CKD cohort, missingness was minimal for age and serum creatinine (0.34%), while body weight (4.16%), height (6.16%), fasting blood glucose (9.79%), and total cholesterol (10.22%) had relatively lower missing rates. In contrast, the HT datasets exhibited substantially higher missingness rates, particularly in HT-KKU, where height (43.55%) and body weight (42.30%) were frequently missing, compared to HT-RAMA (15.75% and 31.59%, respectively). Lipid profiles were among the most commonly missing variables across all datasets. In CKD, the missing rates for low-density lipoprotein (LDL) and high-density lipoprotein (HDL) were 21.73% and 29.14%, respectively. However, in HT-RAMA, LDL and HDL missingness increased to 48.25% and 53.32%, respectively, while HT-KKU recorded the highest missingness, with LDL absent in 67.33% and HDL in 67.88% of cases, indicating substantial data sparsity for key cardiovascular risk markers. Similarly, triglycerides were missing in 19.63% of cases in CKD, but the rates escalated to 44.72% in HT-RAMA and 67.06% in HT-KKU, further emphasizing the challenge of incomplete lipid data in hypertensive populations. Among the metabolic biomarkers, uric acid (49.01%) and HbA1c (40.72%) exhibited the highest levels of missing data in CKD, reflecting limitations in the availability of glycemic control and kidney function

markers. Additionally, missingness in serum creatinine was substantial in both HT cohorts (28.63% in HT-RAMA and 37.36% in HT-KKU), suggesting challenges in renal function assessment in hypertensive populations. These high levels of missingness, particularly for HbA1c, uric acid, and HDL, pose significant analytical challenges, as nearly half of the data points for these critical clinical indicators are unavailable. Notably, the HT-KKU cohort exhibited the most extensive missingness across key clinical variables, reinforcing the need for robust imputation strategies to enhance data completeness and reliability in subsequent analyses.

Due to variations in data availability and missingness across cohorts, certain features were more prevalent in one dataset than the other. For instance, systolic blood pressure (SBP) and diastolic blood pressure (DBP) were extensively recorded in HT datasets but were unavailable in the CKD dataset since blood pressure may be considered a less significant predictor of CKD progression compared to serum creatinine and uric acid, which were more consistently recorded in the CKD cohort. Conversely, HT datasets had more complete blood pressure data but lacked glucose, HbA1c, and uric acid because these markers are more relevant to metabolic conditions like diabetes or kidney disease. Given these differences, we strategically curated our datasets to ensure methodological consistency and enhance the generalizability of our imputation models. While all eight domains were used to construct both cohorts, we prioritized commonly utilized clinical variables to maintain analytical robustness across CKD and HT populations. To provide a comprehensive overview of comorbidities, medications, and laboratory tests derived from these domains, a detailed descriptive statistics table is presented in Supplementary Table 2.

Amputated data

We applied Van Buuren's guidelines to select an imputation model by comparing the inferences obtained from a fully observed dataset (or complete dataset) to those computed by pooling all MICE estimates from an amputated version of the complete dataset. A complete dataset was created through listwise deletion. Amputation was defined as a process of generating synthetic missing values from the complete data. The missing data patterns necessary for multivariate amputation were specified to create a similar pattern of missing data similar to the original dataset. The overall multivariate amputation procedure is shown in Fig. 1. It begins by determining the missing patterns of the original dataset, where a missing pattern is a specific combination of variables with missing pattern per sample per observation. We identified $k = 296$ missing patterns in the CKD dataset such that [uric acid'] obtained the highest missingness proportion of 32.02%, followed by [HbA1c'] and ['HbA1c', 'uric acid'] with 20.28% and 13.98%, respectively. By specifying the proportion in the multivariate amputation for all different patterns, the amputated dataset had a proportion and number of cases that was similar to the original dataset. However, too many patterns can result in subsets with a few candidates because the amputated data was generated from a complete dataset which represented only 8156 of the total 31,043 patients. Therefore, a threshold of 1% was determined through a trial-and-error approach to balance the trade-off between retaining sufficient candidates in the complete dataset for amputation and minimizing missing patterns. This threshold

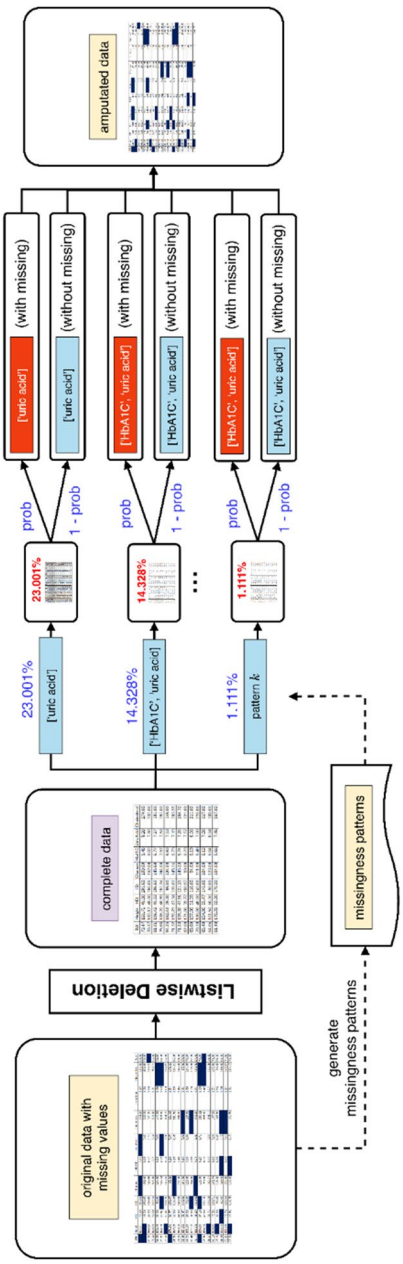

Fig. 1 Schematic overview of the multivariate amputation procedure derived from Schouten et al. (2018)

was chosen to ensure a meaningful reduction in missing patterns, decreasing them from 296 to 16, while preserving enough data for robust analysis.

The complete dataset was then divided into k subsets randomly based upon the selected missing patterns. The subset size depended on the proportion vector represented by the frequency of the specific pattern missing from the complete dataset. The data rows in the subsets were considered a candidate for missingness, based on several factors including the missingness mechanism. Finally, the data rows in the subsets were made missing according to the missing data pattern along with their probability of being missing. The probability (prob) of being missing was calculated from the ratio of the difference between the total number of samples in the original dataset N_{original} and the complete dataset N_{complete} to the total number of records in the original dataset.

$$\text{prob} = \frac{N_{\text{original}} - N_{\text{complete}}}{N_{\text{original}}}$$

This method denotes the probability of the proportion of missing information within the dataset. After the introduction of missing values to each pattern, these subsets were merged to generate an incomplete dataset with missing values in different data rows.

To verify the effectiveness of multivariate amputation, we compared the distribution of each variable in the original with complete and amputated datasets. Data distribution and percentage missingness of the original, complete, and amputated datasets from CKD cohort are presented in Table 1, demonstrating consistency between the complete and amputated dataset across various variables. For instance, BMI and serum creatinine are almost unchanged from a mean of 26.612 and 1.483–26.613 and 1.483, respectively, demonstrating the amputation's ability to preserve the distribution within a clinically relevant range. Similarly, while triglycerides exhibited a slight variation in their distribution between the complete and amputated dataset, it still fell within a range that maintained the original dataset's integrity. We also compared the missing rates between the complete and amputated datasets where missing rates in the amputated dataset should

Table 1 Variable distribution with missing values in original, complete, and amputated datasets from the CKD cohort

Variables	Mean (SD)			Missing Rate (%)	
	Original data	Complete data	Amputated data	Original data	Amputated data
BMI	26.37 (3.85)	26.61 (3.80)	26.61 (3.79)	6.16	5.63
Body Weight	71.26 (12.85)	72.26 (12.239)	72.25 (12.25)	4.16	3.7
Cholesterol	198.64 (52.00)	187.91 (39.93)	187.89 (40.08)	10.22	10.02
Glucose	138.54 (73.29)	145.36 (74.18)	145.03 (73.70)	9.79	9.54
HbA1c	6.87 (1.59)	6.72 (1.44)	6.68 (1.41)	40.72	41.53
HDL	47.07 (14.77)	45.29 (12.77)	45.37 (12.72)	29.14	29.27
Height	159.55 (8.027)	160.16 (7.99)	160.16 (7.99)	6.16	5.63
LDL	107.71 (27.65)	107.14 (27.39)	107.38 (27.33)	21.73	22.08
Serum Creatinine	1.48 (0.91)	1.48 (0.78)	1.48 (0.78)	0.34	0.18
Triglycerides	156.76 (109.12)	159.31 (99.44)	158.90 (98.92)	19.63	19.8
Uric Acid	7.30 (2.15)	7.18 (2.00)	7.19 (1.94)	49.01	50.32

BMI body mass index, *HbA1c* hemoglobin A1c, *HDL* high-density lipoprotein, *LDL* low-density lipoprotein. Cholesterol refers to total cholesterol, and glucose refers to fasting plasma glucose for brevity

closely resemble those in the original dataset. For BMI, body weight, and height, the missing rates in the amputated dataset were slightly lower than those in the original dataset. Lipid profile variables, such as HDL, LDL, total cholesterol, and triglycerides, showed almost identical missing rates between the original and amputated datasets, suggesting a high fidelity in multivariate amputation for these variables. Glucose exhibited missing rates of 9.79% in the original data and 9.54% in the amputated data, while HbA1c had missing rates of 40.72% and 41.53%, respectively. The percentage of missing data for serum creatinine and uric acid between both datasets differed only slightly, with similar mean (SD) values across both datasets. Overall, the multivariate amputation process effectively preserved the missing data patterns, ensuring that the amputated data maintained a distribution like the complete dataset.

Uncertainty-aware MICE

Consider a dataset $\mathbf{X} \in \mathbb{R}^{n \times p}$ containing N patients and p covariates. Let $\mathbf{x}_j = \left(x_{1j}, x_{2j} \cdots, x_{Nj}\right)^T \in \mathbb{R}^N$ denote the j-th covariate, where each entry x_{ij} corresponds to patient i and j-th covariate. For each covariate $\mathbf{x}_j$, the entries are partitioned into observed $\mathbf{x}_j^{\text{obs}}$ and missing $\mathbf{x}_j^{\text{mis}}$ components, such that $\mathbf{x}_j = (\mathbf{x}_j^{\text{obs}}, \mathbf{x}_j^{\text{mis}})$. Here, $\mathbf{x}_j^{\text{obs}}$ represents the subset of x_{ij} values that are observed across i to N patients, while $\mathbf{x}_j^{\text{mis}}$ contains the missing entries. Missingness patterns may vary across all patients; thus, the indices of missing entries differ for each $\mathbf{x}_j$. The MICE framework iteratively imputes missing values by modeling each covariate conditional on the others. For the j-th covariate, let $\mathcal{F}_j \subseteq \{1,2,\cdots,p\} \backslash \{j\}$ denote the set of indices for predictors used to impute $\mathbf{x}_j^{\text{mis}}$. The imputation model for $\mathbf{x}_j^{\text{mis}}$ is formulated as:

$$\widehat{\mathbf{x}}_j^{\text{mis}} = f_j\left(\mathbf{x}_{-j}^{\text{imp}} | \mathbf{x}_j^{\text{obs}}\right),$$

where f_j is a regression function trained on the observed data $\mathbf{x}_j^{\text{obs}}$, and $\mathbf{x}_{-j}^{\text{imp}}$ denotes the current imputed values for the selected covariates, except $\mathbf{x}_j$. The training data for f_j is constructed from patients with observed $\mathbf{x}_j$, expressed as $\mathbf{x}_j^{\text{train}} = \left\{\left(\mathbf{x}_{i,\mathcal{F}_j}^{\text{imp}}, x_{ij}\right) | x_{ij} \in \mathbf{x}_j^{\text{obs}}\right\}$, where $\mathbf{x}_{i,\mathcal{F}_j}^{\text{imp}}$ represents the imputed values of predictors $\mathcal{F}_j$ for the i-th patient. The test data $\mathbf{x}_j^{\text{train}} = \left\{\mathbf{x}_{i,\mathcal{F}_j}^{\text{imp}} | x_{ij} \in \mathbf{x}_j^{\text{mis}}\right\}$, consists of patients with missing $\mathbf{x}_j$.

To initialize the imputation, missing values in $\mathbf{x}_j^{\text{train}}$ are temporarily filled with random draws from a normal distribution parameterized by the observed data:

$$x_{ik}^{\text{init}} \sim \mathcal{N}\left(\mu\left(\mathbf{x}_k^{\text{obs}}\right), \sigma^2\left(\mathbf{x}_k^{\text{obs}}\right)\right), \text{for } x_{ik} \in \mathbf{x}_k^{\text{mis}}$$

ensuring initial imputations reflect inherent uncertainty. This approach follows the original MICE framework, where missing values are firstly estimated from the observed data to enable the imputation process [4, 43]. While this initialization may introduce some bias by centering imputed values around the mean, it serves only as a temporary approximation to ensure algorithmic stability. The iterative refinement process progressively updates these estimates through multiple imputations, progressively aligning them with the conditional expected value while preserving the inherent uncertainty in the data.

The algorithm generates $m = 5$ imputed datasets through $T = 50$ iterations, following the suggestions of van Buuren and Groothuis-Oudshoorn [14, 15, 44, 45]. At iteration t of imputed dataset d, the imputed values $\widehat{\mathbf{x}}_j^{\text{mis}(d,t)}$ are updated as:

$$\widehat{\mathbf{x}}_j^{\text{mis}(d,t)} = f_j^{(d,t)}\left(\mathbf{x}_{-j}^{(d,t)}\right)$$

where $f_j^{(d,t)}$ is the regression model trained on $\mathbf{x}_j^{\text{train}}$ from the previous iteration. The missing rate in the dataset was identified and variables prioritized from the least to the highest missing frequency. Within each iteration, values were standardized to scale the data. We employ randomized search over grid search since it samples from a predefined parameter distribution that significantly reduces computation resources while effectively uncovering optimal hyperparameters. The search space for the model hyperparameters in CKD and HT cohorts is shown in Supplementary Table 5. The best hyperparameters from random search were extracted for subsequent model training. However, to mitigate error propagation, uncertainty functions $\phi(\cdot)$ are applied for $d \geq 2$ to prioritize informative samples. For each patient across all covariates $\mathbf{x}_i^{\text{train}}$, the uncertainty δ_i is quantified, and a subset $\widetilde{\mathbf{x}}_j^{\text{train}(d,t)}$ is selected to refine $f_j^{(d,t)}$. To quantify imputation uncertainty, we define an uncertainty function $\phi_j(\mathbf{x}_i)$ that evaluates the variability of the imputed feature vector $\mathbf{x}_i$ across multiple iterations. This helps prioritize missing values based on their uncertainty, improving the stability of the imputation process. To further reduce reliance on extensive training data, only certain samples are selected for refinement based on an uncertainty function, as not all training samples contribute equally to model training. This ensures that uncertainty samples, which may introduce noise into the model, are avoided. Once f_j is fitted, it is evaluated on $\mathbf{x}_j^{\text{test}}$, and the imputed values $\widehat{\mathbf{x}}_j^{\text{mis}}$ are truncated based on clinically plausible ranges for each variable. Additionally, winsorization is applied by capping values below the 5th percentile and above the 95th percentile, reducing the influence of extreme outliers. This approach minimizes the impact of extreme imputations while ensuring robust estimates.

To prevent overfitting and ensure computational efficiency, early stopping γ was employed like that used in missForest, denoted as γ_{mF}, where the convergence criterion is based on the sum of the squared differences normalized by the squared sum of the current imputed values:

$$\gamma_{\text{mF}} = \sum \frac{\left(\widehat{\mathbf{x}}_j^{\text{mis}(t+1)} - \widehat{\mathbf{x}}_j^{\text{mis}(t)}\right)^2}{\left(\widehat{\mathbf{x}}_j^{\text{mis}(t)}\right)^2}$$

where convergence is reached when the sum of squared differences falls below a predefined threshold. Another convergence criterion was compared; variation threshold (VT) was based on the root mean squared difference normalized by the length of the imputed datasets used, and expressed as:

$$\gamma_{\text{VT}} = \sqrt{\frac{\sum_{j=1}^{p}\left(\widehat{\mathbf{x}}_j^{\text{mis}(t+1)} - \widehat{\mathbf{x}}_j^{\text{mis}(t)}\right)^2}{N_j^{\text{mis}(t)}}}$$

where $N_j^{\text{mis}(t)}$ is the number of missing values in covariate j at iteration t. To monitor convergence, we calculated $\omega_t = |\gamma_{VT}(t) - \gamma_{VT}(t-1)|$, which measures the difference error between iterations. Convergence is achieved if the variation threshold satisfies $\omega_t < \epsilon, \forall i \in \{t-4, t-3, t-2, t-1, t\}$ with $\epsilon = 1 \times 10^{-5}$, ensuring stability. If this criterion holds for five consecutive iterations (i.e., $\{t-4, t-3, t-2, t-1, t\}$), the imputation process for d is terminated. If the convergence is not reached, iteration $t+1$ is further performed using the same training dataset $\mathbf{x}_j^{\text{train}}$ from iteration t. Each imputed dataset d_i is appended into a list for pooling.

After the iterative imputation process is completed for all m datasets, the final imputed values are pooled to obtain a consolidated estimate. The pooled imputation follows Rubin's rules, where the final imputed values for all missing values are computed as

$$\overline{\mathbf{X}}^{\text{pool}} = \frac{1}{m}\sum_{d=1}^{m} \widehat{\mathbf{x}}^{\text{mis}(d)}$$

where $\overline{\mathbf{X}}^{\text{pool}}$ represents the matrix of pooled imputed values across all covariates and all missing entries. The overall algorithm is summarized in Table 2.

The whole analyses were conducted using Python (version 3.9) within a Visual Studio Code integrated development environment. Specifically, LinearReg, DT, and RF were implemented using the scikit-learn library (version 1.3.2), while XGBoost was applied using the XGBoost library (version 1.7.2).

Table 2 Algorithm of Uncertainty-Aware MICE

Algorithm: Uncertainty − Aware MICE

Input: dataset $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p) \in \mathbb{R}^{N \times p}$ with N samples and p covariates
where each variable $\mathbf{x}_j$ has observed $\mathbf{x}_j^{\text{obs}}$ and missing values $\mathbf{x}_j^{\text{mis}}$
Output: imputed dataset $\overline{\mathbf{X}}^{\text{pool}} \in \mathbb{R}^{\text{N} \times p}$
for $d = 1$ to m $= 5$:
  set $\mathbf{X}^{(d)} = X$ *with* initial imputed values for missing data
  for each variable $\mathbf{x}_j$ in $[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p]$:
    for t in 1 to T $= 50$:
      select predictor set $\mathcal{F}_j = (\mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \cdots, \mathbf{x}_{j_f})$
      split data: $\mathbf{x}_j^{\text{train}}$ (cases where $\mathbf{x}_j$ is observed); $\mathbf{x}_j^{\text{test}}$ ($\mathbf{x}_j$ is missing)
      initialize missing values $x_{ik}^{\text{init}} \sim \mathcal{N}\left(\mu(\mathbf{x}_j^{\text{obs}}), \sigma(\mathbf{x}_j^{\text{obs}})\right)$
      if $d \geq 2$:
        calculate uncertainty δ using $\phi_\delta\left(\hat{\mathbf{x}}_i^{\text{train}(d,t)}\right)$
        select samples based on uncertainty $\tilde{\mathbf{x}}_j^{\text{train}(d,t)}$
        train regression model f_j
        if bootstrap $\hat{\mathbf{x}}_j^{\text{mis}_{(d,t)}} = f_j\left(\tilde{\mathbf{x}}_j^{\text{train}(d,t)}\right)$
          else fit regressor f_j on $\mathbf{x}_j^{\text{train}}$
      impute missing values $\hat{\mathbf{x}}_j^{\text{mis}(d,t)} = f_j(\mathbf{x}_j^{\text{test}})$
      winsorize imputed values $\hat{\mathbf{x}}_j^{\text{mis}(d,t)}$
      calculate convergence criteria γ
      stop if $\omega < \epsilon = 1 \times 10^{-5}$ for 5 consecutive iterations **else** $t+1$
  update imputed values $\hat{\mathbf{x}}_j^{\text{mis}(d,t)}$
pool data $\overline{\mathbf{X}}^{\text{pool}} = \frac{1}{m}\sum_{d=1}^{m} \hat{\mathbf{x}}^{\text{mis}(d)}$

Uncertainty functions

Novel query selection methods that consider imputation uncertainty to improve the training process of imputation models were considered. Given a pool of samples with missing values, the imputation uncertainty δ of each sample was quantified by introducing an uncertainty function $\phi(\cdot)$ that aggregates the SD (σ) of the $\mathcal{F}_j$ per missing variable $\mathbf{x}_j^{\text{mis}}$ and selects samples for model training. In this study, three uncertainty functions were proposed which are described below:

a. US: To quantify imputation uncertainty using US, we calculated the variance σ_i^2 of the patient's feature vector $\mathbf{x}_i$ across all imputations m, with degrees of freedom equal to 1, as follows:

$$\sigma_i^2 = \frac{1}{m-1}\sum_{d=1}^{m}\left(x_{ij}^{(d)} - \overline{x}_{ij}\right)^2$$

where $x_{ij}^{(d)}$ is the imputed value of feature j for patient i in the d-th imputation, and $\overline{x}_{ij}$ is the mean imputed value for feature j of patient i across all imputations. Once we calculate the variance for each feature, we computed the total sum of variances across all features for a given set of imputations, which can be represented by the following formula:

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{US}} = \sum_{j=1}^{p}\left(\frac{1}{m-1}\sum_{d=1}^{m}\left(x_{ij}^{(d)} - \overline{x}_{ij}\right)^2\right) = \sum_{j=1}^{p}\left(\sigma_{ij}^2\right)$$

b. PI: While the US prioritizes samples with the least uncertainty, PI selects samples characterized with high uncertainty, underpinned by the assumption that these samples will most likely provide significant improvement in the subsequent iterations. This approach is grounded in the assumption that exploring areas of high uncertainty is likely to yield an improvement in model accuracy and robustness.

Suppose there is $\widehat{x}_{ij}^{(d,t)}$ as the imputed value of variable j of patent i at imputed dataset d, at iteration t. The improvement is defined as $I\left(\widehat{x}_{ij}^{(d,t)}\right) = \max\left(\widehat{x}_{ij}^{(d,t)} - \widehat{x}_{ij}^{(*,t)}, 0\right)$ where $\widehat{x}_{ij}^{(*,t)}$ is the current best imputed value of variable j. Hence, $\widehat{x}_{ij}^{(d,t)} - \widehat{x}_{ij}^{(*,t)}$ is negative if the new imputed value is less than the previous imputed value, which implies a negative improvement. In contrast, if the $\widehat{x}_{ij}^{(d,t)}$ has larger uncertainty, then $\widehat{x}_{ij}^{(d,t)} - \widehat{x}_{ij}^{(*,t)}$ is positive. In this case, $I\left(\widehat{x}_{ij}^{(d,t)}\right)$ provides an indication of model improvement over the current best solution. The probability assigned of $I\left(\widehat{x}_{ij}^{(d,t)}\right) > 0$, i.e., $\widehat{x}_j^{(d,t)}$ having larger improvement than the current best $\widehat{x}_{ij}^{(*,t)}$. Using a Gaussian Process where $\widehat{x}_{ij}^{(d,t)}$ is modeled as a Gaussian distribution, $\widehat{x}_{ij}^{(d,t)}$ is sampled from a normal distribution with mean $\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)$ and variance $\sigma^2\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)$. If $z \sim \mathcal{N}(0,1)$, then $\widehat{x}_{ij}^{(d+1,t)} = \mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) + \sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)z$ is a standard normal distribution. Therefore, the improvement function $I\left(\widehat{x}_{ij}^{(d,t)}\right)$ is rewritten as

$$I\left(\widehat{x}_{ij}^{(d,t)}\right) = \max\left(\widehat{x}_{ij}^{(d,t)} - \widehat{x}_{ij}^{(*,t)}, 0\right)$$

$$I\left(\widehat{x}_{ij}^{(d,t)}\right) = \max\left(\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) + \sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}, 0\right)z \sim \mathcal{N}(0,1)$$

Using the equation above, PI is then derived as:

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{PI}} = 1 - \Theta(z) = \Theta(-z) = \Theta\left(\frac{\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}}{\sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)}\right)$$

where $\Theta(z)$ is the cumulative distribution function (CDF) of z and $z = \frac{\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}}{\sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)}$. The sample with the highest PI is selected for model training.

c. **EI** uses the same assumption as PI where it selects samples with high uncertainty since these samples will most likely provide significant improvement in subsequent iterations. The primary difference compared with PI is that EI considers the magnitude of improvement by calculating the expected value of improvement, denoted as $\mathbb{E}\left[I\left(\widehat{x}_{ij}^{(d,t)}\right)\right]$.

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = \mathbb{E}\left[I\left(\widehat{x}_{ij}^{(d,t)}\right)\right]$$

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = \int_{-\infty}^{\infty} \max\left(\widehat{x}_{ij}^{(d,t)} - \widehat{x}_{ij}^{(*,t)}, 0\right)\Psi(z)dz$$

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = \int_{-\infty}^{\infty} \left[\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) + z\sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}\right]\Psi(z)dz$$

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = \int_{z}^{\infty} \left(\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}\right)\Psi(z)dz + \int_{z}^{\infty} z\sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)\Psi(z)dz$$

$$\begin{aligned}\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = &\int_{z}^{\infty} \left(\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}\right)\Psi(z)dz \\ &+ \int_{z}^{\infty} z\sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\right)dz\end{aligned}$$

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = \left(\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}\right)\int_{z}^{\infty}\Psi(z)dz + \frac{\sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)}{\sqrt{2\pi}}\int_{z}^{\infty} ze^{-\frac{z^2}{2}}dz$$

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = \left(\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}\right)\xi(-\text{z}) + \frac{\sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)}{\sqrt{2\pi}}\left[ze^{-\frac{z^2}{2}}\right]_{z}^{\infty}$$

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = \left(\mu\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}\right)\xi(-\text{z}) + \frac{\sigma\left(\widehat{x}_{ij}^{(1,t)}, \cdots, \widehat{x}_{ij}^{(d,t)}\right)}{\sqrt{2\pi}}\Psi(z)dz$$

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = \left(\mu\left(\widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}\right)\xi(-\text{z}) + \frac{\sigma\left(\widehat{x}_{ij}^{(d,t)}\right)}{\sqrt{2\pi}}\Psi(z)dz$$

where $\Psi(z)$ is the probability density function of the standard normal distribution $\mathcal{N}(0,1)$. Therefore, EI is expressed as:

$$\Phi\left(\mathbf{x}_i^{\text{train}}\right)_{\text{EI}} = \left(\mu\left(\widehat{x}_{ij}^{(d,t)}\right) - \widehat{x}_{ij}^{(*,t)}\right)\xi(-\text{z}) + \frac{\sigma\left(\widehat{x}_{ij}^{(d,t)}\right)}{\sqrt{2\pi}}\Psi(z)dz$$

Evaluation metrics

A series of regression models were incorporated with uncertainty functions during model training. Metrics used for performance evaluation were the root mean squared error (RMSE), mean absolute error (MAE), data distribution, and imputation error as follows:

a) The RMSE represents the quadratic means of the differences between the imputed and observed data. It is one of the most used evaluation metrics. The value of RMSE is always non-negative and a lower value reflects better performance.

$$\text{RMSE} = \sqrt{\left(\frac{\sum_{j=1}^{p}\left(\widehat{\mathbf{x}}_j^{\text{obs}} - \widehat{\mathbf{x}}_j^{\text{pool}}\right)^2}{N}\right)}$$

b) The mean absolute error (MAE) is the mean absolute difference between the actual and the imputed data, and a lower value is preferable to a larger value.

$$\text{MAE} = \frac{1}{N}\sum_{i}^{N}\left|\widehat{\mathbf{x}}_j^{\text{obs}} - \widehat{\mathbf{x}}_j^{\text{pool}}\right|$$

c) Data distribution of each imputed variable was explored by computing the mean and SD based on pooled imputed values from all models tested. Specifically, the mean of the imputed variable across all imputation methods and models provides a measure of central tendency, while the SD indicates the spread or variability of the imputed values.

d) To quantify the imputation calibration for each variable, the imputation calibration coefficient e_j is calculated as the absolute difference between the means of the pooled imputed and complete datasets: $e_j = |\mu_j^{\text{imp}} - \mu_j^{\text{com}}|$. This coefficient measures the alignment between imputed and observed (complete) data distributions, with lower values indicating better alignment and higher imputation accuracy.

Results

MICE was performed with parameters set to create five imputed datasets with a maximum of 50 iterations each. Four models including LinearReg, DT, RF, and XGBoost, were tested both with and without an acquisition function (i.e., US, PI, and EI). The γ_{VT} criterion was employed to determine an appropriate early stopping point if conditions were met.

Model-specific results in CKD

***LinearReg*:** Eleven variables with missing data (0.34–49.01% missing; lowest in serum creatinine and highest in HDL) were imputed. Baseline performance was assessed using

LinearReg models without uncertainty functions on five datasets with 50 iterations for each dataset. LinearReg exhibited an RMSE of 0.168445 and an MAE of 0.46852. The RMSE decreased to 0.14865 when γ_{mF} was implemented but the MAE marginally increased to 0.482385, illustrating a potential trade-off between these error metrics. However, γ_{mF} was imputed in each dataset with only two iterations, raising concerns about overfitting. In contrast, LinearReg with γ_{VT} yielded an RMSE of 0.174566, slightly higher than baseline, but a marginally lower MAE (0.471848) than LinearReg with γ_{mF}. It created five imputations with 18 iterations each, potentially leading to increased model stability compared to γ_{mF}. Therefore, γ_{VT} was chosen as an early stopping criterion for subsequent analyses.

The use of uncertainty functions significantly improved LinearReg models since both RMSE and MAE decreased considerably compared to the baseline model. For instance, LinearReg-US yielded an RMSE of 0.146335 and MAE of 0.441111, while LinearReg-PI and LinearReg-EI achieved even lower RMSEs of 0.131789 and 0.115484, and MAEs of 0.374132 and 0.361053, respectively. Further analysis using bootstrapping with 200 resamples also revealed that LinearReg-US provided a slight improvement (RMSE: 0.110015, MAE: 0.359432), but the impact was more pronounced for LinearReg with RMSE and MAE of 0.124943 and 0.369432 for PI, 0.109231 and 0.346973 for EI, respectively.

***DT*:** The baseline DT without uncertainty functions yielded an RMSE of 0.442594 and MAE of 0.508017. Interestingly, the use of uncertainty functions with bootstrapping of 200 resamples further improved DT performance. The most significant improvement was DT-EI, which yielded the lowest observed RMSE and MAE of 0.213497 and 0.376164, respectively, followed by DT-US and DT-PI with RMSEs of 0.243913and 0.245401 for DT-US and DT-PI, and MAEs of 0.401649 and 0.413497 when compared to baseline DT. However, all LinearReg experiments consistently outperformed their DT counterparts. While integrating uncertainty functions into both LinearReg and DT models reduced error rates, the extent of this reduction varied more significantly among the DT models. Specifically, DT models with US, PI, and EI functions achieved RMSE reductions of 44.89%, 44.55%, and 51.76%, respectively. In contrast, LinearReg models achieved RMSE reductions of 13.13%, 21.76%, and 31.44% for US, PI, and EI functions, respectively, indicating that while both models benefit from uncertainty functions, the impact is more pronounced in DT models.

RF: Baseline RF achieved lower performance compared to both baseline LinearReg and DT models, with an RSME of 1.019152 and MAE of 0.509738. Nonetheless, integration of uncertainty functions with RF demonstrated enhanced model performance, particularly with RF-EI, which achieved the best performance with 0.219435 RMSE and 0.335892 MAE. Additionally, RF-US and RF-PI also achieved lower metrics than the baseline model, though slightly higher than RF-EI, with RMSE of 0.290931 and 0.232560 and MAE of 0.385177 and 0.362912, respectively.

***XGBoost*:** XGBoost models consistently showed the poorest performance across all experiments where the baseline XGBoost produced RMSE and MAE of 1.019152 and 0.509738, which was significantly higher than all baseline models. Further analysis using uncertainty functions failed to make significant improvements, with XGBoost-US achieving the best performance among all XGBoost models with 1.047963 RMSE and 0.442739 MAE.

Table 3 Distribution of variables in complete and imputed datasets of CKD cohort by LinearReg with γ_{VT} early stopping

Variables	Complete Data	Baseline	LinearReg-US	LinearReg-PI	LinearReg-EI
Body Weight	72.26 (12.24)	72.64 (12.84)	72.25 (12.18)	72.25 (12.36)	**72.26 (12.19)**
Cholesterol	187.91 (39.93)	198.72 (51.21)	187.95 (38.88)	187.15 (42.49)	**187.94 (38.88)**
Glucose	145.36 (74.18)	147.49 (75.11)	144.93 (71.21)	143.30 (72.98)	**144.94 (72.95)**
HbA1c	6.72 (1.44)	**6.75 (1.29)**	6.67 (1.11)	7.19 (2.39)	6.67 (1.12)
HDL	45.29 (12.77)	45.06 (12.14)	45.32 (11.20)	45.387 (12.52)	**45.29 (11.95)**
Height	160.16 (7.99)	160.08 (8.16)	160.17 (7.94)	160.15 (8.17)	**160.16 (7.94)**
LDL	107.14 (27.39)	116.27 (36.56)	107.32 (24.96)	**107.05 (29.87)**	107.30 (26.49)
Serum Creatinine	1.48 (0.78)	1.52 (0.86)	1.48 (0.78)	1.48 (0.78)	**1.48 (0.78)**
Triglycerides	159.31 (99.44)	165.99 (102.40)	159.10 (91.18)	159.27 (98.14)	**159.30 (99.23)**
Uric Acid	7.18 (2.00)	7.30 (1.48)	7.19 (1.33)	7.16 (1.43)	**7.19 (1.62)**

The closest distribution with the complete data is highlighted in bold text. Baseline model refers to linear regression (LinearReg) without uncertainty function

US uncertainty sampling, *PI* probability of improvement, *EI* expected improvement, *HbA1c* hemoglobin A1c, *HDL* high-density lipoprotein, *LDL* low-density lipoprotein. Cholesterol refers to total cholesterol, and glucose refers to fasting plasma glucose for brevity

Distribution of the imputed values in CKD dataset

We further explored distributions by comparing the distribution between the imputed values in the amputated and complete datasets. Mean and SDs were estimated for each variable in the complete dataset (serving as the benchmark), and amputated dataset with various uncertainty functions, see Table 3. In this section, we only considered the bootstrapped models with γ_{VT} of the regression algorithms under three different uncertainty functions.

LinearReg: Comparisons of the LinearReg models are described in Table 3. Significant improvements were observed in body weight and height, where LinearReg-EI achieved a 0.001 imputation error against the complete data. Specifically, LinearReg-EI resulted in the mean (SD) of 72.255 (12.186) for weight and 160.162 (7.942) for height, compared to the complete data's benchmark of 72.256 (12.239) and 160.163 (7.986), respectively, across all uncertainty functions (see Table 3). In contrast, baseline LinearReg exhibited the highest deviation of 0.384 kg and 0.083 cm for weight and height, respectively, highlighting the advantage of incorporating uncertainty functions. This trend extended to other variables such as glucose, cholesterol, HDL, triglycerides, uric acid, and serum creatinine, where LinearReg with uncertainty functions consistently outperformed the baseline model. For instance, for glucose levels, LinearReg-EI demonstrated the smallest imputation error at 0.411 mg/dL, followed by LinearReg-US at 0.429 mg/dL, LinearReg-PI at 2.059 mg/dL, and baseline LinearReg at 2.135 mg/dL. Similarly, LinearReg-EI in cholesterol, HDL, and triglycerides also showed minimal deviations of 0.03 mg/dL, 0.002 mg/dL, and 0.004 mg/dL, respectively, with baseline LinearReg consistently generating higher imputation errors across these metrics. Furthermore, despite uric acid showing the highest percentage missing, uncertainty-aware LinearReg models managed to closely approximate the complete data distribution. Notably, LinearReg-EI showed an imputation error of only 0.004 mg/dL, followed by LinearReg-US at 0.007 mg/dL, LinearReg-PI at 0.025 mg/dL, with the baseline LinearReg of 0.119 mg/dL. Regarding LDL, while LinearReg-EI performed favorably in generating a close distribution with

Table 4 Distribution of variables in complete and imputed datasets of CKD cohort by DT with γ_{VT} early stopping

Variables	Complete data	Baseline	DT-US	DT-PI	DT-EI
Body Weight	72.26 (12.24)	72.63 (12.86)	72.25 (12.19)	72.25 (12.23)	**72.25 (12.19)**
Cholesterol	187.91 (39.93)	198.11 (51.63)	188.29 (39.26)	**187.90 (39.29**)	188.05 (39.31)
Glucose	145.36 (74.18)	146.81 (75.74)	**145.36 (74.17)**	144.48 (71.83)	144.94 (71.71)
HbA1c	6.72 (1.44)	6.74 (1.35)	**6.73 (1.41)**	6.66 (1.21)	6.70 (1.22)
HDL	45.29 (12.77)	44.77 (12.31)	45.25 (11.53)	44.94 (11.84)	**45.29 (11.95)**
Height	160.16 (7.99)	160.09 (8.17)	160.17 (7.96)	160.16 (7.97)	**160.16 (7.96)**
LDL	107.14 (27.39)	115.93 (37.00)	**107.23 (26.01)**	107.88 (26.28)	107.93 (26.28)
Serum Creatinine	1.48 (0.78)	1.52 (0.86)	1.48 (0.78)	1.48 (0.78)	**1.48 (0.78)**
Triglycerides	159.31 (99.44)	164.53 (104.54)	159.32 (99.96)	159.29 (99.90)	**159.30 (92.43)**
Uric Acid	7.18 (2.00)	7.28 (1.66)	7.17 (1.46)	7.19 (1.62)	**7.19 (1.49)**

The closest distribution with the complete data is highlighted in bold text. Baseline model refers to decision trees (DT) without uncertainty function

US uncertainty sampling, *PI* probability of improvement, *EI* expected improvement, *HbA1c* hemoglobin A1c, *HDL* high-density lipoprotein, *LDL* low-density lipoprotein. Cholesterol refers to total cholesterol, and glucose refers to fasting plasma glucose for brevity

Table 5 Distribution of variables in complete and imputed datasets of CKD cohort by RF with γ_{VT} early stopping

Variables	Complete data	Baseline	RF-US	RF-PI	RF-EI
Body Weight	72.26 (12.24)	72.64 (12.84)	**72.25 (12.18)**	72.25 (12.18)	72.25 (12.21)
Cholesterol	187.91 (39.93)	198.12 (51.31)	188.92 (39.33)	188.92 (39.25)	**187.67 (39.29)**
Glucose	145.36 (74.18)	147.29 (75.23)	**145.19 (71.55)**	144.75 (71.27)	145.61 (71.28)
HbA1c	6.72 (1.44)	6.75 (1.26)	**6.74 (1.26)**	6.69 (1.17)	6.70 (1.17)
HDL	45.29 (12.77)	44.95 (11.84)	44.98 (11.30)	44.95 (11.22)	**45.15 (11.24)**
Height	160.16 (7.99)	160.09 (8.16)	160.17 (7.95)	160.17 (7.95)	**160.16 (7.97)**
LDL	107.14 (27.39)	116.00 (36.80)	108.49 (26.07)	108.66 (25.88)	**108.08 (25.89)**
Serum Creatinine	1.48 (0.78)	1.52 (0.86)	**1.48 (0.78)**	**1.48 (0.78)**	**1.48 (0.78)**
Triglycerides	159.31 (99.44)	162.91 (103.53)	**159.32 (91.70)**	158.17 (91.06)	157.48 (91.16)
Uric Acid	7.18 (2.00)	7.27 (1.48)	**7.19 (1.38)**	7.19 (1.35)	**7.18 (1.36)**

The closest distribution with the complete data is highlighted in bold text. Baseline model refers to random forests (RF) without uncertainty function

US uncertainty sampling, *PI* probability of improvement, *EI* expected improvement, *HbA1c* hemoglobin A1c, *HDL* high-density lipoprotein, *LDL* low-density lipoprotein. Cholesterol refers to total cholesterol, and glucose refers to fasting plasma glucose for brevity

the complete data, LinearReg-PI achieved the closest match with mean (SD) of 107.054 (29.872) against the benchmark of 107.141 (27.387). Imputation errors from these corresponding models were 0.087, 0.163, 0.175, and 9.129 mg/dL.

DT: Baseline DT yielded the highest imputation errors compared to uncertainty-aware DT models across all variables (see Table 4). This was a significant improvement for HbA1c, in which DT-US achieved the lowest imputation error of 0.005 mg/dL relative to DT-PI and DT-EI with imputation errors of 0.06 and 0.023 mg/dL, respectively. In addition, the DT-US had one-fifth lower imputation error when compared to the baseline LinearReg of 0.026 mg/dL.

Uncertainty-aware models, particularly DT-EI, achieved the closest distribution to the complete dataset for body weight, height, HDL, triglycerides, uric acid, and serum creatinine. However, LinearReg models showed distributions more closely resembling the complete dataset than their DT counterparts. For instance, LinearReg-EI achieved closer distribution for body weight, with a lower imputation error of 0.001 kg, compared to 0.003 kg for DT-EI. Even so, total cholesterol, HbA1c, and glucose provided the lowest imputation error in uncertainty-aware DT model than in LinearReg models. Meanwhile, DT-US achieved the closest distribution for glucose and LDL with an imputation error of 0.006 mg/dL and 0.093 mg/dL. Further detailed comparisons of the DT models are provided in Table 4.

RF: Baseline RF model consistently exhibited the highest imputation errors across all RF models tested (see Table 5). Specifically, variables such as height, cholesterol, HDL, LDL, and uric acid closely approximated the benchmark distribution when using DT-EI. Conversely, for weight, glucose, HbA1c, and triglycerides, DT-US yielded imputed distributions closest to the benchmark. Surprisingly, serum creatinine produced the same distribution of 1.480 mg/dL (0.780) in DT-US, DT-PI, and DT-EI, deviating minimally by 0.003 mg/dL from the complete data distribution.

Imputation errors in RF models were marginally higher compared to LinearReg and DT counterparts despite increased model complexity. For instance, triglycerides had an imputation error of 0.004 mg/dL in LinearReg-EI but increased dramatically to 1.821 mg/dL in RF-EI. Other variables, such as body weight, total cholesterol, HDL, and LDL also followed this trend.

XGBoost: Despite exhibiting higher imputation errors compared to the baseline XGBoost model, uncertainty-aware XGBoost consistently demonstrated superior performance across most variables, with notable exceptions being uric acid and HbA1c (see Table 6). Surprisingly, despite these variables having the highest rates of missing data, the baseline XGBoost model yielded imputed distributions that closely matched the complete dataset more often than the uncertainty-aware XGBoost models. For

Table 6 Distribution of variables in complete and imputed datasets of CKD cohort by XGBoost with γ_{VT} early stopping

Variables	Complete Data	Baseline	XGBoost-US	XGBoost-PI	XGBoost-EI
Body Weight	72.26 (12.24)	72.55 (12.88)	71.90 (12.66)	72.110 (12.304)	**72.13 (12.23)**
Cholesterol	187.91 (39.93)	196.79 (51.76)	181.75 (46.01)	184.758 (41.428)	**185.79 (41.24)**
Glucose	145.36 (74.18)	147.15 (75.43)	141.70 (73.10)	143.804 (71.610)	**144.80 (71.53)**
HbA1c	6.72 (1.44)	**6.67 (1.30)**	5.86 (1.51)	6.294 (1.304)	6.29 (1.24)
HDL	45.29 (12.77)	44.92 (12.85)	42.05 (13.69)	43.523 (12.096)	**44.50 (12.95)**
Height	160.16 (7.99)	160.05 (8.18)	159.79 (8.81)	159.923 (8.378)	**160.16 (8.05)**
LDL	107.14 (27.39)	114.55 (37.21)	96.92 (32.89)	103.342 (27.404)	**105.38 (27.24)**
Serum Creatinine	1.48 (0.78)	1.52 (0.86)	1.49 (0.80)	**1.483 (0.783)**	1.48 (0.80)
Triglycerides	159.31 (99.44)	161.73 (104.67)	149.68 (94.92)	155.24 (91.78)	**158.39 (91.01)**
Uric Acid	7.18 (2.00)	**7.18 (1.57)**	5.87 (1.922)	6.67 (1.54)	6.67 (1.54)

The closest distribution with the complete data is highlighted in bold text. Baseline model refers to XGBoost without uncertainty function

US uncertainty sampling, *PI* probability of improvement, *EI* expected improvement, *HbA1c* hemoglobin A1c, *HDL* high-density lipoprotein, *LDL* low-density lipoprotein. For clarity, cholesterol refers to total cholesterol, and glucose refers to fasting plasma glucose for brevity

instance, uric acid showed imputed mean (SD) values of 7.182 (1.574) mg/dL with XGBoost-PI and XGBoost-EI, closely aligning with the complete data distribution. Similarly, HbA1c exhibited imputed values of 6.670 (1.300) mg/dL with XGBoost-US, indicating a significant deviation from the actual data. Serum creatinine obtained its closest distribution to the complete data using XGBoost-PI. Moreover, it also exhibited minimal variance across all methods, with imputed values consistently hovering around 1.483 (0.783) mg/dL, suggesting stable performance but with slight deviations from the complete data distribution. Detailed comparisons of the XGBoost models are provided in Table 6.

External validation using HT datasets

We validated the proposed imputation methods using two independent HT datasets from RAMA and KKU. Missing data were imputed for nine variables. In HT-RAMA, missing data ranged from 15.75% (height) to 53.32% (HDL); in HT-KKU, the range was 28.2% (SBP/DBP) to 67.88% (HDL). A multivariate amputation procedure was applied, identifying 194 unique missingness patterns in HT-RAMA and 127 patterns in HT-KKU. To ensure the proper replication of both missing rates and the characteristics of the original HT datasets, we selected the top 21 most representative patterns in HT-RAMA and six patterns in HT-KKU for further analysis. Additional details regarding the variable schema used in the imputation process are provided in Supplementary Table 4.

HT-RAMA dataset

The HT-RAMA dataset showed higher rates of missing data than the CKD cohort, led by HDL (53.32%), LDL (48.25%), and triglyceride (44.72%) (see Supplementary Table 2). Demographic variables, such as height and body weight, also had higher missing rates of 15.75% and 31.59%, respectively, versus 6.16% and 4.16% among the CKD cohort. SBP and DBP had identical missing rates of 30.64%. In addition, the missing rate for total cholesterol in the HT dataset was 35.36%, more than three times higher than for CKD cohort (10.22%). Serum creatinine, a critical marker for kidney function, had a missing rate of 28.63% in the HT dataset, in contrast to no missing data in the CKD cohort. The

Table 7 Value distribution of variables with missing values in complete and amputated datasets from HT-RAMA cohort

Variables	Mean (SD)			Missing rate (%)	
	Original data	Complete data	Amputated data	Original data	Amputated data
Height	159.45 (8.75)	159.75 (8.87)	159.82 (8.88)	15.75	13.48
Body Weight	65.71 (15.45)	67.00 (15.82)	67.00 (15.92)	31.59	30.43
SBP	143.67 (22.85)	146.05 (22.36)	145.77 (22.33)	30.64	30.43
DBP	82.91 (11.66)	83.97 (11.88)	84.08 (11.86)	30.64	30.43
Cholesterol	208.67 (51.92)	211.58 (49.32)	211.71 (49.18)	35.36	37.38
HDL	50.16 (13.84)	50.41 (13.75)	50.98 (13.94)	53.32	54.36
LDL	132.81 (42.18)	134.68 (42.44)	134.21 (42.05)	48.25	51.15
Triglycerides	146.88 (93.53)	147.53 (94.01)	145.91 (92.17)	44.72	47.27
Serum Creatinine	0.93 (0.83)	0.88 (0.61)	0.87 (0.6)	28.63	28.99

DBP diastolic blood pressure, *HDL* high-density lipoprotein, *LDL* low-density lipoprotein, *SBP* systolic blood pressure. Cholesterol refers to total cholesterol for brevity

pre- and post-amputation distributions of HT-RAMA data and missing rates for each variable are detailed in Table 7.

HT-KKU dataset

The HT-KKU dataset exhibited high missingness, particularly in HDL (70.34%), LDL (70.34%), and triglycerides (70.34%), with slight increase from the original missing rates (67.88%, 67.33%, and 67.06%, respectively), suggesting that the amputation procedure preserved the missingness structure. Demographic variables such as height (43.55%) and body weight (42.30%) maintained similar missing rates post-amputation (43.4% each), ensuring consistency in data patterns. Similarly, SBP and DBP saw minimal changes (28.20–29.21% for both), while cholesterol (48.41–52.02%) and serum creatinine (37.36–39.03%) showed slight increases. Importantly, the mean values of the amputated data closely align with the original and complete datasets, confirming that the distribution remained stable and consistent despite data amputation. The pre- and post-amputation distributions of HT-KKU data and missing rates for each variable are detailed in Table 8.

Model-specific results in HT cohorts

HT-RAMA

LinearReg: Baseline LinearReg model achieved an RMSE and MAE as high as 11.4946 and 14.5406, respectively. The performance metrics remained the same after applying the γ_{VT} criterion. With early stopping conditions applied, the number of iterations required before stopping varied across the datasets: 29 iterations for the first two datasets, and significantly fewer—7, 6 and 6 iterations respectively—for the last three.

LinearReg-US greatly improved the imputation performance since RMSE and MAE decreased to 0.2427 and 8.1605 at maximum iterations, and 0.2563 and 8.1645 when γ_{VT} was applied. Further smaller improvements were observed with the use of other acquisition functions. LinearReg-PI achieved an RMSE and an MAE of 0.2411 and 8.1457 at maximum iterations without γ_{VT} and 0.2410 and 8.1457 with γ_{VT}. Similarly, LinearReg-EI reached an RMSE of 0.2408 and an MAE of 8.1457 without γ_{VT}, the metrics were 0.2407 and 8.1457 with γ_{VT}, respectively. These uncertainty-aware models performed

Table 8 Value distribution of variables with missing values in complete and amputated datasets from HT-KKU cohort

Variables	Mean (SD)			Missing Rate (%)	
	original data	complete data	amputated data	original data	amputated data
Height	160.46 (8.53)	160.81 (8.55)	160.98 (8.56)	43.55	43.40
Body Weight	63.84 (15.48)	65.66 (15.59)	65.88 (15.91)	42.30	43.40
SBP	135.53 (20.18)	135.50 (18.01)	135.39 (18.01)	28.20	29.21
DBP	77.36 (12.95)	77.54 (12.03)	77.69 (12.05)	28.20	29.21
Cholesterol	184.06 (58.23)	190.74 (53.18)	191.47 (54.21)	48.41	52.02
HDL	51.30 (16.66)	52.12 (16.50)	52.50 (16.81)	67.88	70.34
LDL	124.28 (48.04)	125.15 (47.48)	124.62 (47.15)	67.33	70.34
Triglyceride	150.43 (94.09)	150.97 (95.53)	148.32 (90.49)	67.06	70.34
Serum Creatinine	1.14 (1.25)	1.03 (0.91)	1.03 (0.91)	37.36	39.03

DBP diastolic blood pressure, *HDL* high-density lipoprotein, *LDL* low-density lipoprotein, *SBP* systolic blood pressure. Cholesterol refers to total cholesterol for brevity

Table 9 Model performance in CKD, HT-RAMA, and HT-KKU cohorts with uncertainty functions and γ_{VT} early stopping

Model	Uncertainty-Aware	CKD		HT-RAMA		HT-KKU	
		RMSE	MAE	RMSE	MAE	RMSE	MAE
LinearReg	Baseline	0.1684	0.4685	11.4946	14.5406	26.4892	25.4832
	LinearReg—US	0.1318	0.3741	0.2563	8.1645	**0.9799**	**12.0043**
	LinearReg—PI	0.1100	0.3594	0.241	8.1457	0.9899	12.0245
	LinearReg—EI	**0.1092**	**0.3470**	**0.2407**	**8.1457**	0.9934	12.0246
DT	Baseline	0.4426	0.5080	11.5991	14.9566	26.5210	25.6780
	DT—US	0.3488	0.4159	0.8571	8.6828	**2.0977**	**12.6337**
	DT—PI	0.2439	0.4016	**0.9578**	**8.7398**	2.3035	12.7771
	DT—EI	**0.2135**	**0.3762**	0.9547	8.7477	2.1822	12.6898
RF	Baseline	1.0191	0.5097	11.6277	14.726	26.5356	26.0943
	RF—US	0.2909	0.3852	**0.9173**	**8.5753**	2.0335	12.9682
	RF—PI	0.2326	0.3629	1.2354	8.6382	**1.9315**	**12.9035**
	RF—EI	**0.2194**	**0.3359**	1.4226	8.6578	1.9443	12.9280
XGBoost	Baseline	1.0192	0.5097	19.8503	19.4089		
	XGBoost—US	**1.0480**	**0.4427**	**14.846**	**14.4823**		
	XGBoost—PI	1.6132	0.5330	18.3232	17.6995		
	XGBoost—EI	1.6545	0.5140	18.3345	17.6925		

The root mean squared error (RMSE) and mean absolute error (MAE) are expressed in four decimal points. The best-performing models are indicated in bold text. Baseline refers to models without uncertainty functions

CKD chronic kidney disease, *HT* hypertension, *LinearReg* Linear Regression, *DT* Decision Trees, *RF* Random Forests, *XGBoost* Extreme Gradient Boosting, *US* Uncertainty Sampling, *PI* Probability of Improvement, *EI* Expected Improvement

more poorly by increasing RMSE, which is about 2 times higher when compared to performance in the CKD data, see Table 9.

Despite these observed improvements, the early stopping conditions outlined by γ_{VT} were not met when these acquisition functions were employed. This resulted in all models processing to the maximum of 50 iterations for the last three datasets after completing 29 iterations for the initial two datasets. This pattern was consistent across all regression models tested, including LinearReg, DT, RF, and XGBoost. Consequently, for these models, there was no variation in performance metrics between experiments with maximum iterations and those with the γ_{VT} conditions. The data distribution of the imputed values in HT-RAMA using LinearReg is found in Supplementary Table 7.

***DT*:** The baseline DT model demonstrated slightly higher error rates to the baseline LinearReg model, with an RMSE and an MAE of 11.5991 and 14.9566 versus 11.4946 and 14.5406. The use of acquisition functions improved the performance of DT with DT-US reducing the RMSE and MAE to 0.8571 and 8.6828, while DT-PI achieved at 0.9578 and 8.7398 and DT-EI achieved 0.9547 and 8.7477, respectively. However, all DT experiments consistently yielded lower performances compared to their LinearReg counterparts. Additionally, while the integration of acquisition functions in both models resulted in reduced error rates, the extent of this reduction varied more significantly among DT models. Specifically, while the LinearReg models experienced a substantial decrease in RMSE, achieving a 97.89–97.90% reduction from the baseline model, DT models showed less pronounced improvements: 91.74% in DT-PI, 91.76% in DT-EI, and 92.61% in DT-US. This pattern of relatively smaller reductions in RMSE for machine

learning based models was also observed in other regression models as well. The acquisition functions DT-US, DT-PI, and DT-EI models performed less well, having RMSEs about 2.5 to 4.5 times higher in HT than CKD cohorts, see Table 8. The data distribution of the imputed values in HT-RAMA using DT is reported in Supplementary Table 8.

***RF*:** The baseline RF model had lower performance than both the baseline LinearReg and DT model at 11.6277 RMSE and 14.726 MAE. The incorporation of uncertainty sampling function RF-US dramatically reduced the error rate by 92.11% with RMSE of 0.9173 and MAE of 8.5753. The improvement in performance was lower in other acquisition functions with 89.38% for RF-PI (1.2354 and 8.6382) and 87.8% for RF-EI (1.4226 and 8.6578). These models were poorer, increasing RMSE about 3.15 to 6.48 times higher in the HT cohort than the CKD cohort. The data distribution of the imputed values in HT-RAMA using RF is reported in Supplementary Table 9.

***XGBoost*:** XGBoost models consistently showed the lowest performance across all experiments. They had the highest error rates, with an RMSE of 19.8503 and an MAE of 19.4089. The incorporation of acquisition functions failed to make significant improvements, with XGBoost-US being the best, with RMSE of 14.8460 and MAE of 14.4800. PI and EI functions had minimal reductions of error rate at 7.69% (18.3232 and 17.6995) and 7.64% (18.3445 and 17.6925), respectively. These models performed less well by increasing RMSEs about 11.08–14.17 times higher in the HT compared to the CKD data, see Table 8. The data distribution of the imputed values in HT-RAMA using XGBoost is found in Supplementary Table 10.

HT-KKU

All baseline models in HT-KKU exhibited extremely high RMSE and MAE across all models, with RMSE exceeding 26.48 and MAE above 25.48, reinforcing its poor imputation performance. In contrast, integrating uncertainty-aware functions significantly reduced errors, with LinearReg-US achieving lowest RMSE (0.9799) and MAE (12.0043), followed by RF and DT.

LinearReg: Baseline LinearReg exhibited high imputation errors, with an RMSE of 26.4892 and an MAE of 25.4832. Incorporating uncertainty-aware acquisition functions drastically improved performance, reducing RMSE by approximately 96% and MAE by 53% compared to the baseline. Among these, LinearReg-US achieved the lowest errors, with an RMSE of 0.9799 (96.3% decrease) and an MAE of 12.0043 (52.9% decrease), followed closely by LinearReg-PI (RMSE = 0.9899, MAE = 12.0245) and LinearReg-EI (RMSE = 0.9934, MAE = 12.0246). While the differences between the uncertainty-aware models were minimal, LinearReg-US consistently outperformed the others, demonstrating its superior ability to reconstruct missing data with minimal imputation errors. The data distribution of the imputed values in HT-KKU using LinearReg is found in Supplementary Table 11.

DT: Like LinearReg, baseline DT model exhibited high imputation errors, with an RMSE and MAE as high as 26.5210 and 25.6780, respectively. Incorporating uncertainty-aware functions substantially reduced RMSE by 92.1% (DT-US: 2.0977), 91.3% (DT-PI: 2.3035), and 91.8% (DT-EI: 2.1822) compared to the baseline. Among these, DT-US achieved the best performance, with the lowest RMSE (2.0977) and MAE (12.6337). Despite these improvements, DT models still exhibited higher imputation errors than

LinearReg, with DT-US having an RMSE more than twice that of LinearReg-US (2.0977 vs. 0.9799). Similarly, MAE in DT-US (12.6337) remained 5.2% higher than in LinearReg-US (12.0043), indicating DT was less precise than LinearReg across all uncertainty-aware functions. The data distribution of the imputed values in HT-KKU using DT is found in Supplementary Table 12.

RF: Baseline RF exhibited the highest imputation errors, with an RMSE of 26.5356 and an MAE of 26.0943. After the integration of uncertainty-aware functions, RMSEs significantly reduced by 92.3% (RF-US: 2.0335), 92.7% (RF-PI: 1.9315), and 92.7% (RF-EI: 1.9443) compared to the baseline RF model. Among these, RF-PI demonstrated the best performance with the lowest RMSE (1.9315) and MAE (12.9035), outperforming RF-US and RF-EI. Compared to DT models, RF achieved lower imputation errors, with RF-PI (1.9315 RMSE) improving upon the best DT model, DT-US (2.0977 RMSE), by 7.9%. However, RF models still exhibited higher RMSE than LinearReg, where LinearReg-US (0.9799 RMSE) was 49.3% lower than RF-PI. Similarly, MAE in RF models remained higher than LinearReg, with RF-PI (12.9035) performing 7.5% worse than LinearReg-US (12.0043). Overall, RF-PI provided the most accurate imputation within RF models and outperformed DT across all uncertainty-aware methods. However, RF remained less precise than LinearReg, indicating that RF still benefits from uncertainty-aware functions. The data distribution of the imputed values in HT-KKU using RF is found in Supplementary Table 13.

Discussion

We have introduced a novel method incorporating uncertainty to handle missing values in RWD. Our approach utilizes uncertainty functions (i.e., US, PI, and EI) to provide a more precise estimate by calculating the imputation uncertainty of each sample and aggregating the variances of the imputed values. The US penalizes and prioritizes samples with the least imputation uncertainty. Contrastingly, PI and EI select samples with high uncertainty, which assumes that these samples are most likely to provide significant improvement in subsequent iterations.

To validate our method, we applied it to three large real-world clinical datasets (CKD, HT-RAMA, and HT-KKU), with high rates of missing data, particularly uric acid (49.01%) and HbA1c (40.72%) in the CKD dataset, and HDL and LDL in the HT datasets (HT-RAMA: HDL 53.32%, LDL 48.25%; HT-KKU: HDL 67.84%, LDL 67.47%). Integrating three uncertainty-aware functions (US, PI, EI) with four ML models (LinearReg, DT, RF, XGBoost) significantly improved model performance and yielded more robust imputed values of 11 and 9 variables in CKD and HT data compared to a baseline model without uncertainty functions. Of ML models, LinearReg showed the best overall performance. Patient overlap between the CKD and HT-RAMA datasets (from the same hospital) resulted in overfitting, as indicated by lower RMSE and MAE values in the HT-RAMA dataset compared to the HT-KKU dataset in both baseline models and models with uncertainty functions. In contrast, HT-KKU dataset, which lacks such overlap, exhibited higher error metrics, which indicates a more stringent and realistic assessment of model generalizability.

Our findings highlight the risks associated with using MICE without uncertainty-aware functions, as the imputed distributions for variables deviated significantly from

the true data distribution. This deviation was particularly evident in the HT-KKU dataset, where baseline LinearReg produced RMSE (26.4892) and MAE (25.4832), indicating severe misalignment between the imputed and actual values. Similarly, in the CKD cohort, baseline XGBoost underperformed, producing an RMSE of 1.0192 and MAE of 0.5097, with imputed values substantially deviating from the expected distribution. Such discrepancies raise concerns about the reliability of imputation without uncertainty, as it introduces systematic biases that may distort subsequent analyses. The inability of standard MICE to preserve key statistical properties can lead to inaccurate inferences, particularly in clinical studies where even minor deviations can alter medical decision-making.

LinearReg-EI demonstrated superior performance in the CKD and HT-RAMA datasets, while LinearReg-US achieved the lowest RMSE and MAE in the HT-KKU dataset. In contrast, XGBoost consistently produced the poorest performance across CKD and HT-RAMA datasets. The observed differences in performance between LinearReg versus XGBoost in both datasets may be attributed to the bias-variance trade-off inherent in these models. XGBoost excels in reducing bias through its boosting mechanism; however, this often results in higher variance (uncertainty), especially when the model is not optimally tuned [46, 47]. In addition, the complexity of XGBoost necessitates careful tuning of multiple parameters, including learning rate, maximum depth, and the number of boosting rounds, making it time-consuming and computationally demanding [48]. In this study, XGBoost utilized random search for 25 h in CKD and 13 h in the HT dataset instead of the more exhaustive grid search, which may have contributed to suboptimal performance.

The superior performance of the baseline XGBoost model compared to its uncertainty-aware counterparts and other ML methods warrants further investigation. This unexpected finding may stem from how XGBoost handles residuals during tree construction. Unlike RF, which split nodes based on target purity (e.g., minimizing variance in regression tasks), XGBoost iteratively builds trees to correct residuals (gradients of the loss function) inherently accounting for uncertainty in the imputed values by focusing on prediction errors from prior iterations. Introducing uncertainty-aware functions might disrupt this process, by adding constraints that conflict with the gradient-based updates. Furthermore, XGBoost's complexity requires careful hyperparameter tuning, which is especially challenging under limited computational resources. This makes XGBoost potentially more susceptible to overfitting when uncertainty functions are added. While RF's purity-based splits are better suited to the US, XGBoost's reliance on residuals may make it less responsive to external uncertainty adjustments. This suggests that uncertainty-aware methods may be more compatible with LinearReg, DT, or RF, which do not inherently optimize error correction via residuals. However, further investigation is needed to confirm this hypothesis and explore the response of other boosted-tree models to uncertainty-aware imputation.

Only a small number of studies have incorporated uncertainty in their imputation frameworks [31–33]. However, they used uncertainty functions (such as random sampling, gini index, and entropy) that are only used for classification tasks. Han and Kang [32] originally proposed the US and applied it to RF, artificial neural network, and softmax regression models using 20 benchmark datasets with no missing values [32]. They

created five incomplete versions of each dataset by varying the missing rates between 10%, 20%, 30%, 40%, and 50% based on the MCAR mechanism. Their results demonstrated the effectiveness of incorporating US in improving the performance across all three regression models. In this study, the US also showed superior performance compared to the baseline across all regression models, particularly in the HT-KKU dataset. The US helps prevent inaccurately imputed samples from being selected for training by choosing samples with lowest uncertainty, thereby avoiding any degradation in the performance of the model.

Overall, EI consistently generated the imputed values with the closest distribution to the original dataset across all regression models except XGBoost. One possible reason for this is that EI inherently identifies new points in the imputed distribution that have the potential to offer the most significant improvement over the current best imputed values [42]. Unlike the US, samples with high uncertainty are selected for re-imputation because they represent areas with high potential improvement. This encourages the exploration of new values that might offer better imputation quality [49]. Moreover, while PI effectively identifies areas with a high likelihood of improvement based on lowest uncertainty like EI, it might not balance exploration and exploitation as efficiently as EI [49, 50]. PI tends to focus more on regions of the imputed distribution with known good performance, potentially leading to less exploration of the imputed space. This can result in less diversity in the imputed values and may not always capture the original data distribution as accurately as EI. For example, in the CKD dataset, EI corrected cholesterol distribution more efficiently than US or PI, demonstrating its ability to recover the true data structure. Similarly, in HT-KKU, LinearReg-EI improved RMSE by 96.3% compared to the baseline, showcasing its effectiveness in handling missing data.

Conclusion

In conclusion, applying any method of uncertainty functions in MICE could greatly improve the performance of the ML models compared with the models without uncertainty functions. This study fills a crucial gap by extending the application of uncertainty functions to real-world clinical datasets with a high percentage of missingness. LinearReg-EI performed the best and may be applied in the MICE for predicting missing values for both categorical and continuous data. In addition, our framework shows a great promise in enhancing the robustness of predictive models in clinical settings by selecting samples accounting for their uncertainty to mitigate the degradation in model performance. Our results suggest that uncertainty-aware imputation can be generalized to other clinical domains, making it a valuable tool for improving data integrity and decision-making in healthcare research.

Abbreviations

CKD	Chronic kidney disease
DBP	Diastolic blood pressure
DT	Decision trees
EI	Expected improvement
HbA1c	Hemoglobin A1C
HDL	High-density lipoproteins
HT	Hypertension
KKU	Khon Kaen University
ICD	International classification of diseases

LDL	Low-density lipoproteins
LinearReg	Linear regression
MAE	Mean absolute error
MAR	Missing at random
MCAR	Missing completely at random
MICE	Multiple imputation by chained equations
ML	Machine learning
PI	Probability of improvement
RAMA	Ramathibodi Hospital, Mahidol University
RF	Random Forest
RMSE	Root mean squared error
RWD	Real-world data
SBP	Systolic blood pressure
US	Uncertainty Sampling
VT	Variation threshold
XGBoost	Extreme gradient boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-025-01136-3.

Supplementary Material 1.

Acknowledgements
This study was an important part of Romen Samuel Wabina's training in Doctor of Philosophy (Data Science for Healthcare and Clinical Informatics), the Faculty of Medicine Ramathibodi Hospital, Mahidol University.

Author contributions
R.S.W.: Conceptualization, Data Curation, Research Methodology, Data Analysis, Validation, Writing—Original Draft; P.L.: Conceptualization, Data Curation, Research Methodology, Data Analysis, Validation, Writing— Critically Review & Editing; S.S.: Conceptualization, Data Curation, Research Methodology, Data Analysis, Validation, Writing—Critically Review & Editing; H.T.: Data Curation, Research Methodology, Data Analysis, Validation, Writing—Original Draft; W.P.: Data Curation, Research Methodology, Writing—Review & Editing; G.M.: Writing—Review & Editing, Research Methodology; J.A.: Writing—Review & Editing, Research Methodology; A. Pat.: Data Curation, Research Methodology; A. Pan.: Data Curation, Validation, Writing—Critically Review & Editing; A.T.: Conceptualization, Research Methodology, Data Analysis, Writing—Review & Editing, Supervision, Funding Acquisition.

Funding
Open access funding provided by Mahidol University. This study was funded by the National Research Council of Thailand N42A640323. The study funder had no role in the design or conduct of the study.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
This study was approved by the Institutional Review Boards of the Faculty of Medicine Ramathibodi Hospital, Mahidol University and Srinagarind Hospitals, Khon Kaen University: COA. No. MURA2024/468 for CKD, COA No. MURA2023/689 for HT-RAMA; and HE681020 for HT-KKU. Due to a retrospective study design, the informed consents were waived by the Ethics Committee.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 October 2024 Accepted: 25 March 2025
Published online: 17 April 2025

References

1. Little RJ, D'Agostino R, Cohen M, Dickersin K, Scott E, Farrar J, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367(14):1355–60.
2. Scheffer J. Dealing with missing data. Res Lett Inf Math Sci. 2002;3(1):153–60.
3. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.
4. Enders CK, Mistler SA, Keller BT. Multilevel multiple imputation: a review and evaluation of joint modeling and chained equations imputation. Psychol Methods. 2016;21(2):222.

5. Zhang Z. Missing data imputation: focusing on single imputation. Ann Transl Med. 2016;4(1):9.
6. Murray JS. Multiple imputation: a review of practical and theoretical findings. Statist Sci. 2018;3(2):142–59.
7. Longford NT. Missing data and small-area estimation: modern analytical equipment for the survey statistician. Berlin: Springer Science & Business Media; 2005. p. 3732.
8. McPherson S, Barbosa-Leiker C, Burns GL, Howell D, Roll J. Missing data in substance abuse treatment research: current methods and modern approaches. Exp Clin Psychopharmacol. 2012;20(3):243.
9. Hawthorne G, Hawthorne G, Elliott P. Imputing cross-sectional missing data: comparison of common techniques. Aust NZ J Psychiatry. 2005;39(7):583–90.
10. King G, Honaker J, Joseph A, Scheve K. Listwise deletion is evil: what to do about missing data in political science. Boston: Annual Meeting of the American Political Science Association; 1998.
11. Little RJ, Rubin DB. Single imputation methods. Stat Anal Miss Data. 2002. https://doi.org/10.1002/9781119013563.ch4.
12. Rubin DB. An overview of multiple imputation. Proceeding survey research methods section of the American statistical association. Princeton: Citeseer; 1988.
13. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol. 2015;15:1–14.
14. Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. Can J Cardiol. 2021;37(9):1322–31.
15. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16(3):219–42.
16. Arnold BC, Castillo E, Sarabia JM. Conditionally specified distributions: an introduction. Statist Sci. 2001;16(3):249–74.
17. Luengo J, García S, Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl Inf Syst. 2012;32:77–108.
18. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol. 2001;27(1):85–96.
19. Allison P. Imputation by predictive mean matching: Promise & Peril. Stat Horiz. 2015.
20. Javadi S, Bahrampour A, Saber MM, Garrusi B, Baneshi MR. Evaluation of four multiple imputation methods for handling missing binary outcome data in the presence of an interaction between a dummy and a continuous variable. J Probab Statist. 2021;2021(1):6668822.
21. Loh W-Y, Zhang Q, Zhang W, Zhou P. Missing data, imputation and regression trees. Stat Sin. 2020;30(4):1697–722.
22. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014;179(6):764–74.
23. Aydin ZE, Ozturk ZK. Performance analysis of XGBoost classifier with missing data. Manch J Artif Intell Appl Sci. 2021;2(02):2021.
24. Casiraghi E, Wong R, Hall M, Coleman B, Notaro M, Evans MD, et al. A method for comparing multiple imputation techniques: a case study on the US national COVID cohort collaborative. J Biomed Inf. 2023;139: 104295.
25. Lee JH, Huber JC Jr. Evaluation of multiple imputation with large proportions of missing data: how much is too much? Iran J Public Health. 2021;50(7):1372.
26. Santos MS, Soares JP, Henriques Abreu P, Araújo H, Santos J, editors. Influence of data distribution in missing data imputation. Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21–24, 2017, Proceedings 16; 2017: Springer.
27. Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. Appl Artif Intell. 2019;33(10):913–33.
28. Jäger S, Allhorn A, Biebmann F. A benchmark for data imputation methods. Front Big Dat. 2021;4: 693674.
29. Sisk R, Sperrin M, Peek N, van Smeden M, Martin GP. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: a simulation study. Stat Methods Med Res. 2023;32(8):1461–77.
30. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092–101.
31. Tharwat A, Schenck W. Active learning for handling missing data. IEEE Trans Neural Netw Learn Syst. 2024;36(2):3273–87.
32. Han J, Kang S. Active learning with missing values considering imputation uncertainty. Knowl Based Syst. 2021;224: 107079.
33. Settles B, Craven M, editors. An analysis of active learning strategies for sequence labeling tasks. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008.
34. Yang Y, Li H, Wang Y, Wang Y. Improving the reliability of large language models by leveraging uncertainty-aware in-context learning. ArXiv. 2023. https://doi.org/10.48550/arXiv.2310.04782.
35. Li J, Tang Y, Yang Y. Know the unknown: an uncertainty-sensitive method for LLM instruction tuning. ArXiv. 2024. https://doi.org/10.48550/arXiv.2406.10099.
36. Wu Y, Li X, Zhou Y. Uncertainty-aware representation calibration for semi-supervised medical imaging segmentation. Neurocomputing. 2024;595: 127912.
37. Durso-Finley J, Barile B, Falet J-P, Arnold DL, Pawlowski N, Arbel T. Probabilistic temporal prediction of continuous disease trajectories and treatment effects using neural SDEs. ArXiv. 2024. https://doi.org/10.48550/arXiv.2406.12807.
38. Martín Vicario C, Rodríguez Salas D, Maier A, Hock S, Kuramatsu J, Kallmuenzer B, et al. Uncertainty-aware deep learning for trustworthy prediction of long-term outcome after endovascular thrombectomy. Sci Rep. 2024;14(1):5544.
39. Gal Y, Islam R, Ghahramani Z. Deep bayesian active learning with image data. International Conference on Machine Learning. PMLR. 2017.
40. Nguyen V-L, Shaker MH, Hüllermeier E. How to measure uncertainty in uncertainty sampling for active learning. Mach Learn. 2022;111(1):89–122.
41. Li G, Chen Z, Yang Z, He J. Novel learning functions design based on the probability of improvement criterion and normalization techniques. Appl Math Model. 2022;108:376–91.
42. Qin C, Klabjan D, Russo D. Improving the expected improvement algorithm. Adv Neural Inform Process Syst. 2017;30.

43. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377–99.
44. Van Buuren S, Brand JP, Groothuis-Oudshoorn CG, Rubin DB. Fully conditional specification in multivariate imputation. J Stat Comput Simul. 2006;76(12):1049–64.
45. Van Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. J Stat Softw. 2011;45:1–67.
46. Ramraj S, Uzir N, Sunil R, Banerjee S. Experimenting XGBoost algorithm for prediction and classification of different datasets. Int J Control Theor Appl. 2016;9(40):651–62.
47. Aydin ZE, Ozturk ZK. Performance analysis of XGBoost classifier with missing data. Manch J Artif Intell Appl Sci. 2021;2:2021.
48. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev. 2021;54:1937–67.
49. Zhan D, Xing H. Expected improvement for expensive optimization: a review. J Glob Optim. 2020;78(3):507–44.
50. Noè U, Husmeier D. On a new improvement-based acquisition function for Bayesian optimization. ArXiv. 2018. https://doi.org/10.48550/arXiv.1808.06918.

Publisher's Note