

Original Paper

Methods for Addressing Missingness in Electronic Health Record Data for Clinical Prediction Models: Comparative Evaluation

Jean Digitale^{1,2}, PhD, MPH, BSN; Deborah Franzon³, MD; Mark J Pletcher², MD, MPH; Charles E McCulloch², PhD; Efsthios D Gennatas², MBBS, PhD

¹National Clinician Scholars Program, University of California, San Francisco, San Francisco, CA, United States

²Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States

³Department of Pediatrics, Benioff Children's Hospital, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:

Jean Digitale, PhD, MPH, BSN
Department of Epidemiology and Biostatistics
University of California, San Francisco
550 16th St, 2nd Floor
San Francisco, CA 94158
United States
Phone: 1 (415) 476-2300
Email: jean.digitale@ucsf.edu

Abstract

Background: Missing data are a common challenge in electronic health record (EHR)–based prediction modeling. Traditional imputation methods may not suit prediction or machine learning models, and real-world use requires workflows that are implementable for both model development and real-time prediction.

Objective: We evaluated methods for handling missing data when using EHR data to build clinical prediction models for patients admitted to the pediatric intensive care unit (PICU).

Methods: Using EHR data containing missing values from an academic medical center PICU, we generated a synthetic complete dataset. From this, we created 300 datasets with missing data under varying mechanisms and proportions of missingness for the outcomes of (1) successful extubation (binary) and (2) blood pressure (continuous). We assessed strategies to address missing data including simple methods (eg, last observation carried forward [LOCF]), complex methods (eg, random forest multiple imputation), and native support for missing values in outcome prediction models.

Results: Across 886 patients and 1220 intubation events, 18.2% of original data were missing. LOCF had the lowest imputation error, followed by random forest imputation (average mean squared error [MSE] improvement over mean imputation: 0.41 [range: 0.30, 0.50] and 0.33 [0.21, 0.43], respectively). LOCF generally outperformed other imputation methods across outcome metrics and models (mean improvement: 1.28% [range: –0.07%, 7.2%]). Imputation methods showed more performance variability for the binary outcome (balanced accuracy coefficient of variation: 0.042) than the continuous outcome (mean squared error coefficient of variation: 0.001).

Conclusions: Traditional imputation methods for inferential statistics, such as multiple imputation, may not be optimal for prediction models. The amount of missingness influenced performance more than the missingness mechanism. In datasets with frequent measurements, LOCF and native support for missing values in machine learning models offer reasonable performance for handling missingness at minimal computational cost in predictive analyses.

*JMIR Med Inform*2025;13:e79307; doi: [10.2196/79307](https://doi.org/10.2196/79307)

Keywords: clinical prediction models; imputation; machine learning; missing data; electronic health record

Introduction

Background and Significance

Addressing missing data is necessary for developing a clinical prediction model. Electronic health record (EHR) data are a rich data source but present particular challenges. Missing data may result from lack of documentation or measurement [1]. EHR data are generated via clinical care, with values measured at irregular intervals.

Raw EHR data are often transformed into an analytic dataset by binning variables by time. Missing data arise if a variable is not measured within a time window [2]. Measurement frequency (and resultant missingness) is often linked to how abnormal the value is or is expected to be, such that missingness itself may be informative. Given many algorithms require complete data, a principled approach to address missingness is required.

Techniques for handling missing data for inferential models that seek to describe or causally explain are well established. Missingness is traditionally categorized into three mechanisms [3,4]:

1. Missing completely at random (MCAR)—probability of missingness does not depend on variables in the dataset or depends only on observed values of covariates included in the model; for example, a laboratory technician forgets to record results for a patient, unrelated to any characteristics of that patient or their health [5].
2. Missing at random (MAR)—probability of missingness depends on observed values in the data, including the outcome; for example, height is not recorded for a patient but is related to weight and sex of the patient, which are present in the EHR.
3. Missing not at random (MNAR)—probability of missingness depends on unobserved values; for example, no lactate is measured on a patient because the clinician expects it to be normal.

Bias from MNAR can be intractable for inferential models [4]. Given EHR data are likely MNAR, this could be problematic if also true for clinical prediction models. For inferential models, simple strategies, such as complete case analysis, mean imputation, and last observation carried forward, are known to produce biased results [6,7]. Preferred strategies, such as multiple imputation, incorporate uncertainty into imputed values, thereby accurately characterizing uncertainty in parameter estimates.

Literature on handling missing data in prediction modeling is less developed. Unlike inferential models, which focus on bias and precision in parameter estimates, prediction models prioritize improving predictive accuracy and interpretability [4]. Classic statistical imputation methods may be complex to implement for prediction models [8] or less relevant, particularly as medicine advances toward ever more complex machine learning algorithms [4,9,10]. There is little guidance on best practices to address missing data for clinical prediction models [11]. Methods for handling missing data

are rarely reported, and complete case analysis is the most common approach [11,12]. This may not only result in bias but also risk significant loss of data in high-dimensional EHR datasets [9]. Machine learning is increasingly being used to address missing data, both as imputation models (eg, random forests) [13] and by natively handling missing data in prediction models themselves, bypassing the need for imputation altogether. Tree-based methods [14] are particularly suited for this task [9]. Yet, few studies have compared classic imputation methods with such built-in strategies in EHR data [9,12].

Real-world application of clinical prediction models presents additional challenges. Many risk models currently in practice require complete data or use imputation methods that may be overly simplistic, limiting their usefulness [15,16]. Implementing models prospectively requires data workflows that can be applied in the same way to both retrospective data to build the model and new data for real-time prediction for individuals [4,10,11,15]. There are no established techniques for managing missing data post-model development. Studies assessing methods for handling missing data in prospective applications on individual patients often used datasets containing only a few predictors [16,17]. These findings may not translate to datasets with more variables because some imputation methods may struggle to handle large numbers of correlated features and binary variables are more likely to be perfectly predicted, leading to overfitting. Furthermore, outcome prediction models after addressing missing data were often standard statistical methods such as logistic regression [15,17] or Cox proportional hazard models [16], when different methods for addressing missingness may be preferable for predictive machine learning models.

Objective

We used EHR data from a live use case (predicting extubation readiness of children in the pediatric intensive care unit [PICU]) to generate a synthetic complete dataset to evaluate multiple methods for imputation and their effects on predictive performance. We included methods that learn from training data and apply to new data. As the relative performance of methods varies by type of missingness (MAR, MCAR, and MNAR) and proportion of missing values [18], we varied both in our assessment.

Methods

Study Population

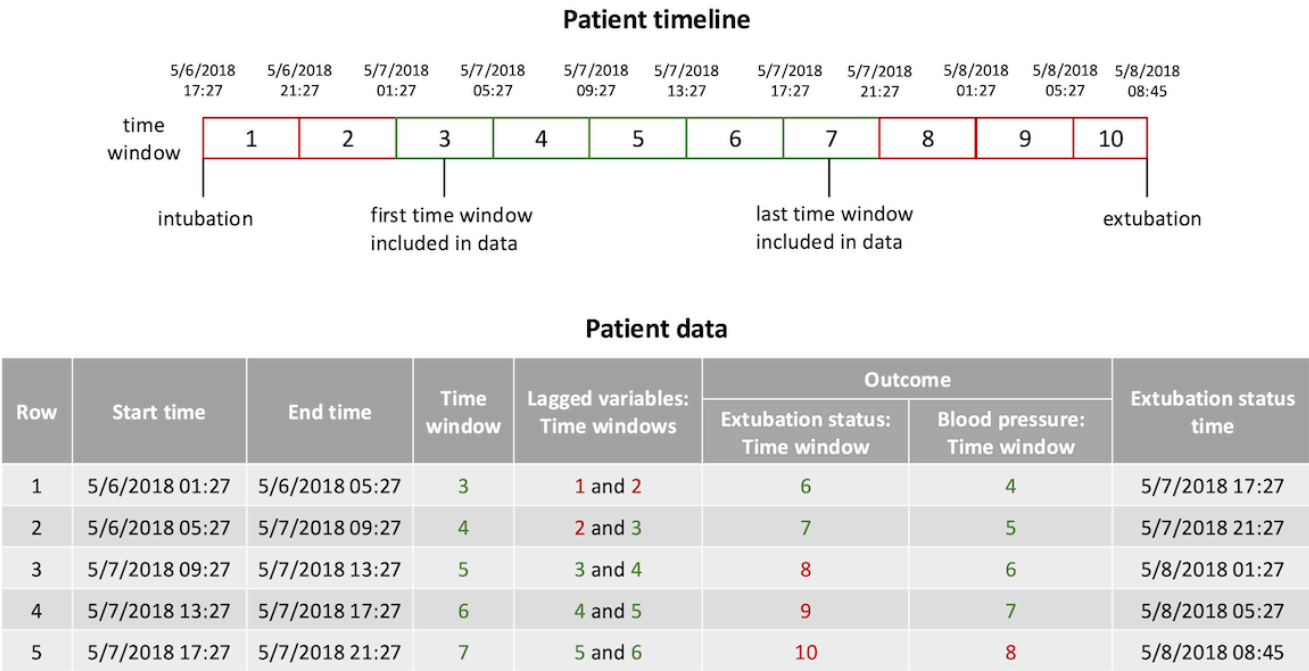
The study population was patients aged >30 days and <18 years old from the PICU at the University of California, San Francisco (UCSF) Benioff Children's Hospital intubated between January 1, 2013, and March 31, 2023. Patient encounters were eligible for the sample if the child was intubated for more than 24 hours. We excluded patients intubated for less than 24 hours as they were likely intubated for surgeries, procedures, or other indications and extubated quickly without complications. Children with repeated intubations were eligible for inclusion for each intubation event.

Data

We used EHR data extracted from the UCSF Clinical Data Warehouse (updated daily from the real-time EHR). Based on expert opinion, peer-reviewed literature, and group consensus of the UCSF Pediatric Critical Care Research Group, we selected a broad range of clinical, physiologic, and laboratory variables in the EHR that could be relevant to extubation readiness including: vital signs, ventilator settings, laboratory values, medications, neurological status, fluid balance, and other patient characteristics (Table S1 in [Multimedia Appendix 1](#)). Raw data were collapsed into 4-hour time windows [1] containing the mean of each numeric or binary

variable and mode of each multi-level categorical variable, resulting in 99 variables. Collapsing variables into time windows made imputation more tractable as it increased the probability of at least 1 nonmissing value per time window and more computationally feasible by decreasing the number of observations on which we needed to impute. In practice, nursing assessments are completed at least every 4 hours (with some assessments in the ICU performed hourly); thus, the 4-hour interval is clinically meaningful. The first time window included in the model ended 12 hours after intubation (Figure 1).

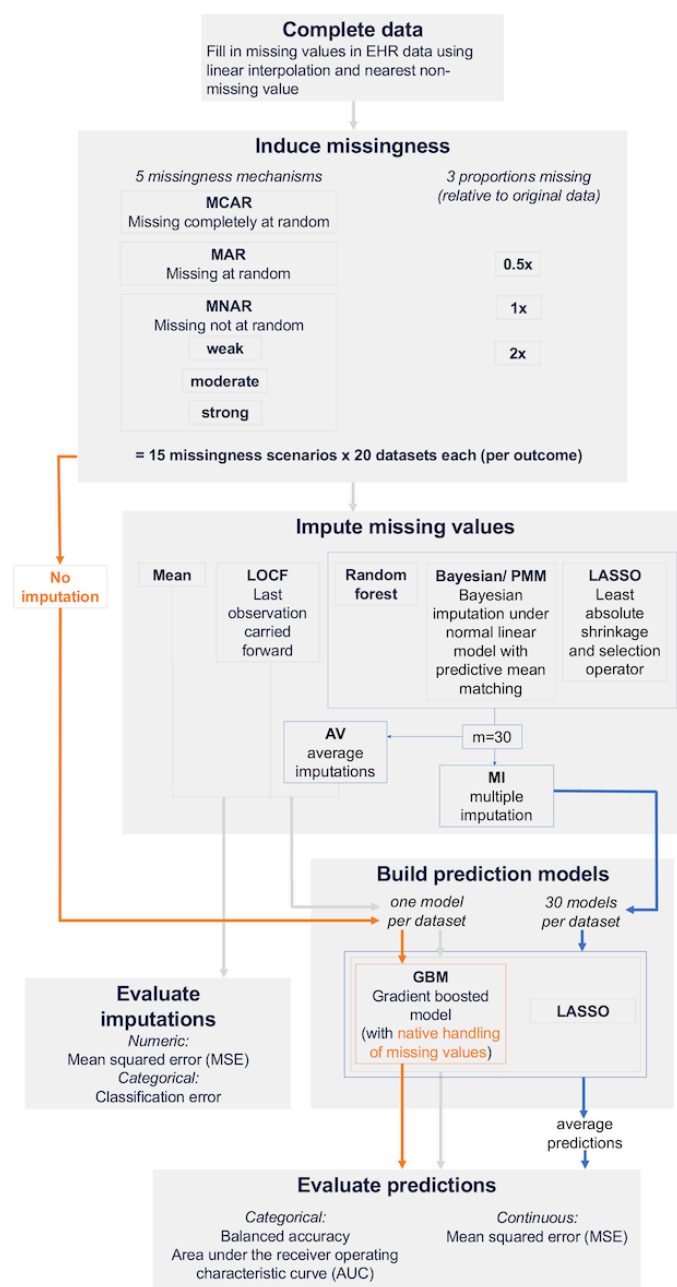
Figure 1. Example patient timeline and resultant data. The example patient timeline depicts 4-hour time windows beginning at intubation and ending at extubation. The table in the figure demonstrates the data resulting from this patient that would be included in the model. This patient was intubated at 17:27 on 5/6/2018 and extubated at 08:45 on 5/8/2018. The first time window included as an observation in the model ended 12 hours after intubation at 05:27 on 5/7/2018. Time-varying data from the prior 2 time windows were included for each observation as lagged variables to capture the trajectory of the patient. The final time window included in the model for this patient is from 17:27 to 21:27 on 5/7/2018. The extubation outcome for each time window was the status 12 hours (or 3 time windows) later. The blood pressure outcome for each time window was the value 4 hours (or 1 time window) later.



The binary outcome of successful extubation was defined as extubation that did not result in reintubation within 48 hours. The extubation outcome assigned to each time window indicated status 12 hours after the end of the time window (Figure 1; [Multimedia Appendix 1](#)), as we aimed to predict successful extubation prior to clinician actions indicating they already decided to extubate a patient. A secondary outcome of age-adjusted systolic blood pressure percentile [19] was added to ascertain whether findings were similar for a continuous outcome. The blood pressure outcome assigned to each time window was the value 4 hours after the end of the time window.

To generate a synthetic dataset with no missing values (Figure 2), we filled in missing numeric values with linear

interpolation between last value observed and next value observed [20]. We filled in all remaining missing values with the nearest nonmissing value. For never observed variables (1.4% of cells), we made reasonable assumptions (eg, we used the standard endotracheal tube [ETT] size formula [age in years/4+4] for pediatrics to fill in missing ETT sizes [21]) and then filled in remaining missingness using the *missRanger* package [22], which implements random forest imputation with predictive mean matching and is optimized for speed and memory efficiency. Analyses were conducted in R version 4.3.2 [23].

Figure 2. Experimental methods flowchart. EHR: electronic health record.

Missingness Mechanisms

We induced missingness in the complete, synthetic dataset to simulate 5 missing data mechanisms:

1. MCAR
2. MAR
3. weak MNAR (twice as much weight given to observed compared with missing variables to determine missingness)
4. moderate MNAR (equal weight given to observed compared with missing variables to determine missingness)
5. strong MNAR (missingness solely based on missing variables)

For each, we created 3 scenarios varying the amount of missingness: approximately 0.5x, 1x, and 2x the percentage

of missing cells in the original data, generating 15 total missingness scenarios. We divided variables into 5 groups (Table S1 in [Multimedia Appendix 1](#)) to create patterns of missing data. The outcome was allowed to influence missingness in MAR and MNAR scenarios ([Multimedia Appendix 1](#)). Using the `ampute` command in the *mice* package [24], we induced missingness in 20 unique datasets (enough to have a distribution of results, but not exceed our computing power) for each of the 15 scenarios, resulting in 300 datasets per outcome. All datasets for both outcomes contained the same predictor variables.

Imputation Methods

To incorporate temporal patterns in the predictors, we added values of time-varying variables from the prior 2 lagged time windows to each row for a total of 265 features. These data

were structured in wide format, such that each row included the current value and its 2 lags as separate variables. This allowed the imputation models to use all 3 time points to inform each other’s missing values. For simplicity, we treated these repeated measurements as distinct variables in imputation models (rather than using a multilevel model). While some studies suggest this approach yields comparable performance [25], others have reported advantages of multilevel models [26]. Data were split into training (75% of intubations) and test (25% of intubations) sets ordered by date, with earlier intubations included in training and later ones in test. We ensured each patient was only in either the training or test set.

To simulate a real-time prediction workflow, imputation models were built using only training data and excluded the outcome [15]. In total, 6 methods to handle missingness were applied to each amputed dataset (Table 1, Multimedia Appendix 1):

1. Mean: Mean imputation is frequently used in practice [27]. We imputed test set values using the unconditional mean (numeric and binary) or mode (factor) in the training set.
2. Last observation carried forward (LOCF): This simple approach is recommended for imputation in time series data where data are available only before the missing value [20] (true in prospective implementation of clinical decision support). We allowed values to be carried forward indefinitely across time windows.
3. Random forest: Traditional statistical imputation relies on parametric assumptions. However, nonparametric methods, such as this, have been shown to outperform established methods (especially in settings with complex interactions and nonlinear relationships) [28]. *Mice* [29] imputes missing values by building a random forest for each variable, identifying observations in the same terminal node, and sampling a donor value from one of these observations.
4. Bayesian imputation under the normal linear model with predictive mean matching (Bayesian/PMM): PMM is a hot deck method where missing values are imputed from cases with observed values matched according to predictions of the imputation model (here, Bayesian imputation under the normal linear model in *mice* [29]). PMM is robust against model misspecification and ensures imputed values are constrained to the range of observed data [7].
5. Least absolute shrinkage and selection operator (LASSO): Regularized models, such as LASSO, are beneficial to handle multicollinearity and prevent overfitting in high-dimensional data. We used *mice* [29] to fit LASSO-penalized regression models on bootstrap samples of observed data and drew imputed values from the resulting distributions.
6. Native support for missing data in prediction model (no imputation required): Some machine learning algorithms can handle missing values directly, without dropping cases or requiring separate imputation. We used gradient boosted trees for our primary prediction model [30]. The LightGBM package [31] allocates missing values to the bins that optimally minimize loss.

Table 1. Methods for handling missing data.

Method	Assumptions	Computational complexity	Limitations	Benefits
Mean imputation	Assumes missing values are similar to the mean	Negligible	- Artificially reduces variance - Disturbs relationships between variables	- Simple to implement - Constant imputation creates patterns that machine learning can exploit
Last observation carried forward	Assumes stability over time	Negligible	- Data remain missing if no prior measurement exists - May not reflect true patient progression - Can introduce bias if trends are not stable over time	- Simple to implement - Reflects how clinicians practice for many variables (assume no changes or re-measure if important)
Random forest	MCAR^a or MAR^b	High computational cost	- Requires significant computational resources - May not perform well in small datasets - Dearth of packages that allow models to impute on new data	- Nonparametric method handles complex interactions and nonlinear relationships well - Works with mixed data types (categorical and continuous)
Bayesian imputation under the normal linear model with predictive	MCAR or MAR	High computational cost	Works best with large samples	Robust against model misspecification

Method	Assumptions	Computational complexity	Limitations	Benefits
mean matching (Bayesian/PMM)	Assumes normality for underlying distribution	High computational cost	Dearth of packages that allow models to impute on new data	Ensures imputed values are within the range of observed data
Least absolute shrinkage and selection operator (LASSO)	MCAR or MAR		- May not capture nonlinear relationships well - Dearth of packages that allow models to impute on new data	- Handles multicollinearity well - Prevents overfitting in high-dimensional data
Native support for missing data in prediction models	- Varies by model - Methods [12] include surrogate splits and allocating missing values to bins that optimally minimize loss	None for imputation	- This capability is not available in all machine learning algorithms - Performance depends on algorithm's internal handling of missing values	No need for explicit imputation

^aMCAR: missing completely at random.
^bMAR: missing at random.

We used available software for imputation to avoid the need to develop custom software and enable our findings to be more readily applicable to practitioners. We did not include deep-learning methods in our experiment as they were impractical with our relatively small sample size. We generated 30 imputations per dataset using the *mice* package [29] for methods 3-5. Models built with the training data were then used to impute on the test set. We tested the imputations 2 ways. *Mice* purposefully incorporates uncertainty into imputations because it is advantageous for inferential analysis. First, we averaged the imputations to get a more stable estimate of each missing value to use in a single outcome model. Second, we implemented multiple imputation by estimating 30 outcome models and averaging the predicted probabilities for a final prediction.

Prediction Model

For the outcome prediction model, we used gradient boosted trees [30]. It is one of the best-performing algorithms in structured data in general and within biomedical datasets [32] and uses all cases in training data, even if they are incomplete. To assess whether imputation method performance was consistent across outcome models, we also compared a LASSO outcome model. However, linear models like LASSO cannot accommodate missing values, so this was only performed on imputed datasets. Outcome models were implemented using the *rtemis* package [33] with *LightGBM* [34] and *glmnet* [35]. We used 5-fold cross-validation in the training set to tune hyperparameters and inverse frequency weighting to upweight the minority class given the data were unbalanced.

Analysis of Imputation Accuracy

It is established that in imputation for statistical inference, focusing on improving accuracy of the imputations at the cost of correctly incorporating true uncertainty leads to biased and invalid results [7]. However, the relationship between

imputation accuracy and prediction model performance is less well-studied. We compared the accuracy of imputations by calculating mean squared error (MSE) for numeric variables and classification error for categorical variables in each dataset and creating box plots. Before calculating MSE, we standardized the variables by dividing by their standard deviations in the complete dataset. To calculate MSE for random forest, Bayesian/PMM, and LASSO, we compared the average of the 30 imputations to the true value. We also assessed whether temporal autocorrelation of each variable was associated with imputation performance across all methods descriptively using scatterplots and quantified using correlation coefficients (Multimedia Appendix 1). In addition, we conducted a sensitivity analysis stratifying imputation error by whether a variable's values had been missing in the original data (and filled in to create the synthetic complete dataset) versus not originally missing, to evaluate whether this initial filling step influenced the apparent performance of the imputation methods.

To assess the accuracy of imputation for different categories of variables, we built linear models for MSE and classification error. Each observation's outcome was the error value for a given variable in a given dataset. Each model included imputation type, missingness type, proportion missing data (0.5x, 1x, and 2x original), and variable group (Multimedia Appendix 1). We included all 3-way and 2-way interactions and completed a backward stepwise elimination procedure based on *P* values (included *P*<.05) to determine the final models.

Analysis of Outcome Model Performance

We assessed the performance of outcome prediction models for extubation with 2 primary metrics: balanced accuracy [36] and area under the receiver operating characteristic curve (AUC) [37]. We present secondary results for sensitivity, specificity, positive predictive value, negative predictive value, and F1 [38]. We assessed the performance of the

outcome prediction models for blood pressure with the primary metric of MSE. Secondary results are presented for mean absolute error, root MSE, and R^2 [39]. We compared these graphically to the performance of a model built with the complete data and calculated the coefficient of variation (CV) to assess variability.

Ethical Considerations

We received ethical approval from the University of California, San Francisco Institutional Review Board (study #17-23751), which granted a waiver of informed consent. No financial incentives were provided to patients. Patient privacy and confidentiality were protected through secure

data storage, restricted access to authorized study personnel, and compliance with institutional and regulatory requirements.

Results

Data

The data contained 886 patients and 1220 intubations, 929 (76.1%) of which ended in successful extubation (Table 2). The median duration of intubation in the PICU was 4.4 (IQR 2.2-8.5) days, leading to 50,187 four-hour time windows in the analytic dataset.

Table 2. Description of sample^a.

Variable	Value
Patients	
Total, n	886
Female patients, n (%)	405 (45.7)
Race or ethnicity, n (%)	
Asian	117 (13.9)
Black	67 (8)
Latinx	340 (40.4)
Other	88 (10.5)
White	229 (27.2)
Intubations	
Total, n	1220
Age at intubation (years), median (IQR)	4.3 (1-12.1)
Outcome, n (%)	
Extubation success	929 (76.1)
Extubation failure	100 (8.2)
Death	96 (7.9)
Tracheostomy	36 (3)
Transfer to another unit	25 (2)
ETT ^b change	34 (2.8)
Duration of intubation (days), median (IQR)	4.4 (2.2-8.5)

^aPatients intubated multiple times during the study period may have multiple intubation events included in the sample. The binary outcome of successful extubation for the model collapsed all other outcome categories. Here, extubation failure is defined as reintubation within 48 hours, death is death before or within 48 hours of extubation, and ETT change is an extubation that was immediately and purposefully replaced by another ETT (eg, to change the size).

^bETT: endotracheal tube.

Missingness for each variable in the original data (collapsed into 4 h time windows) varied from 0% (0/56,287; eg, sex, age) to 77% (43,077/56,287; white blood cells) (Table 3; Table S2 in Multimedia Appendix 1). Overall, 18.2% of cells (1,012,668/5,561,767) were missing (Table

S3 in Multimedia Appendix 1). After simulating missingness, datasets approximating 0.5x, 1x, and 2x missingness of the original data averaged 9.6%, 18.1%, and 35.9% missing cells, respectively.

Table 3. Missingness in original data^a.

Variable	Number of 4-hour time windows missing variable	Percent of 4-hour time windows missing variable
Age at time of intubation (in days)	0	0.0
Medication: total oral morphine equivalents (mg kg ⁻¹)	0	0.0
Sex	0	0.0
Intake or output total milliliter over the prior 12 hours kg ⁻¹	167	0.3
Pulse	420	0.7
Respiratory rate (recorded in vital signs)	1179	2.1
ETT ^b size	6179	11.0
Respiratory pattern: tachypneic	7290	13.0
PEEP ^d	8975	15.9
Exhaled tidal volume kg ⁻¹	10,274	18.3
State behavioral scale	15,888	28.2
Secretion amount (categories: none, scant, small, moderate, large, copious)	16,325	29.0
Glasgow coma scale score	25,751	45.7
pH	25,804	45.8
Upper extremity motor response	28,839	51.2
Cough: present	35,254	62.6
White blood cell count	43,077	76.5

^aExample variables display the range of proportion of missingness in the original data (N=56,287 4-h time windows; these data include more time windows than the final analytic dataset because they were compiled before excluding time windows that were constructed solely for creating lagged variables). Detailed data for all variables are available in Table S2 in [Multimedia Appendix 1](#).

^bETT: endotracheal tube.

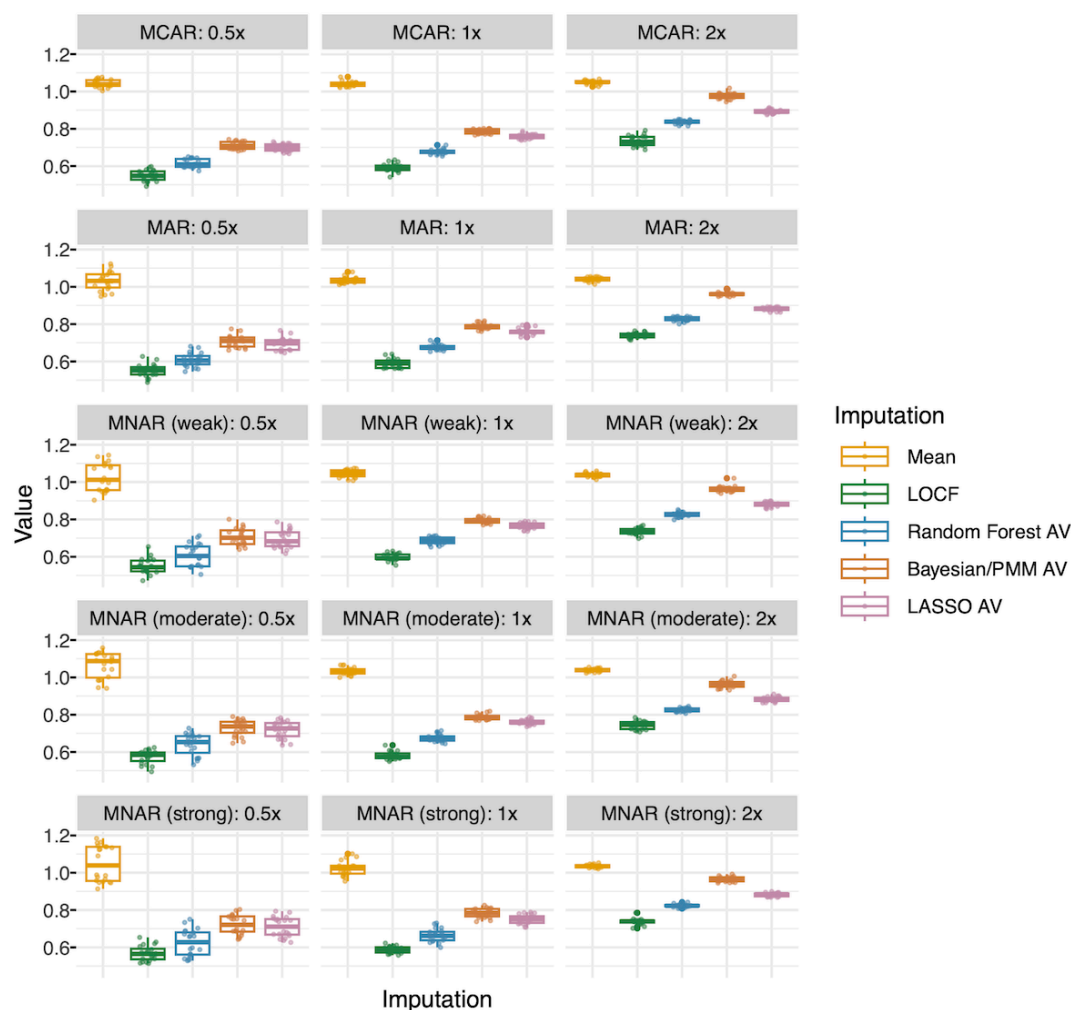
^cPEEP: positive end-expiratory pressure.

Imputation Performance

Across 300 datasets with induced missingness per outcome, the same 176 numeric and 6 categorical variables were

imputed. [Figure 3](#) presents MSE for datasets for the outcome of extubation.

Figure 3. Imputation test performance metrics (extubation): mean squared error. Each point represents mean squared error calculated for 176 numeric variables for 1 of 300 datasets created for the outcome of extubation. There are 20 datasets per missingness scenario and imputation type represented in each box plot. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAR: missing at random; MCAR: missing completely at random; MNAR: missing not at random.



The results for the outcome of blood pressure (which imputed the same variables) were virtually identical ([Multimedia Appendix 2](#)). Performance in the test sets showed that LOCF had the lowest MSE on average in all missingness scenarios for numeric variables (average improvement of MSE compared with mean imputation was 0.41 for the outcome of extubation [range: 0.30, 0.50]). Random forest imputation was consistently second best (0.33 [0.21, 0.43]), followed by LASSO (0.26 [0.15, 0.35]), Bayesian/PMM (0.22 [0.07, 0.34]), and finally, mean imputation (Reference). Performance overall degraded as the proportion of missing data increased, with proportion missing having a greater effect than missingness mechanism. Classification error displayed similar patterns overall for categorical variables ([Multimedia Appendix 3](#)). While LOCF and mean imputation did not generally overfit in the training data compared with the test set, all model-based imputation methods overfit the training data, with random forest doing so the least ([Multimedia Appendices 4 and 5](#)). Temporal autocorrelation was negatively associated with imputation error for all methods except mean imputation, which showed little association ($r=-0.13$; [Multimedia Appendix 6](#)). The association was strongest for

LOCF ($r=-0.92$), followed by random forest ($r=-0.74$). For these, error decreased almost monotonically with increasing autocorrelation, indicating substantially better accuracy for more temporally stable variables.

In a sensitivity analysis, we examined whether the initial filling of originally missing values influenced subsequent comparisons of imputation performance. Overall, MSE was lower for values that were missing in the original data and filled in to create the synthetic complete dataset, especially for LOCF and random forest imputation (2 methods used to fill in the original missingness). Still, both LOCF and random forest continued to achieve the best performance when imputing values not missing in the original data ([Multimedia Appendix 7](#)).

In models for mean squared error, a significant 3-way interaction existed between imputation method, proportion missing, and variable group (Tables S4 and S5 in [Multimedia Appendix 1](#)) due to the fact that (1) all methods except mean imputation degraded with increased missingness and (2) variables with a response of “select all that apply” were more poorly predicted ([Multimedia Appendix 8](#)). There

was no interaction between missingness mechanism and (1) imputation method or (2) variable group.

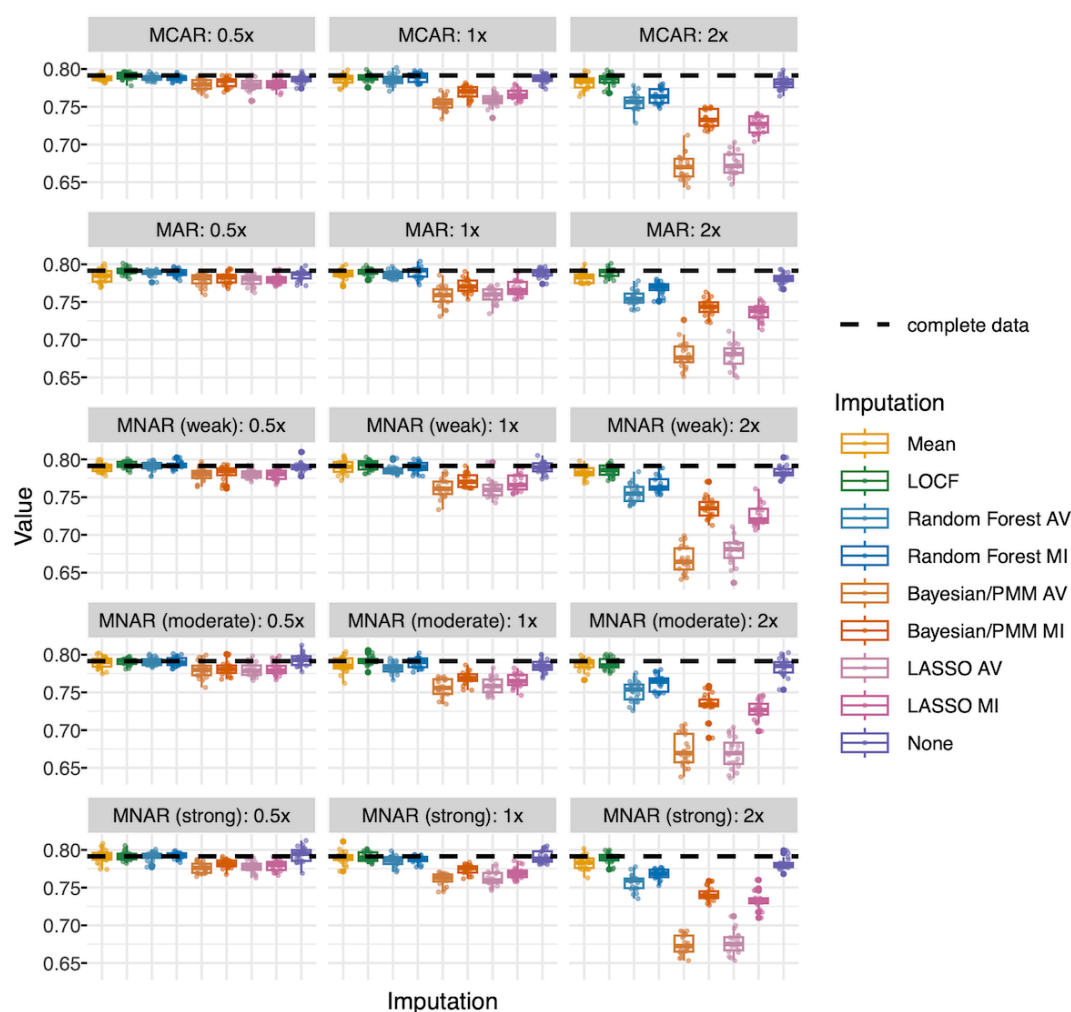
Prediction Model Performance

Gradient Boosted Models

Extubation: Balanced accuracy

For the outcome of extubation using gradient boosted models, balanced accuracy was the highest for LOCF (Figure 4).

Figure 4. Gradient boosted model test performance (extubation): balanced accuracy. Each point represents balanced accuracy for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. Balanced accuracy in the complete dataset is represented by a dashed line. AV, average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAR, missing at random; MCAR: missing completely at random; MI: multiple imputation; MNAR: missing not at random.



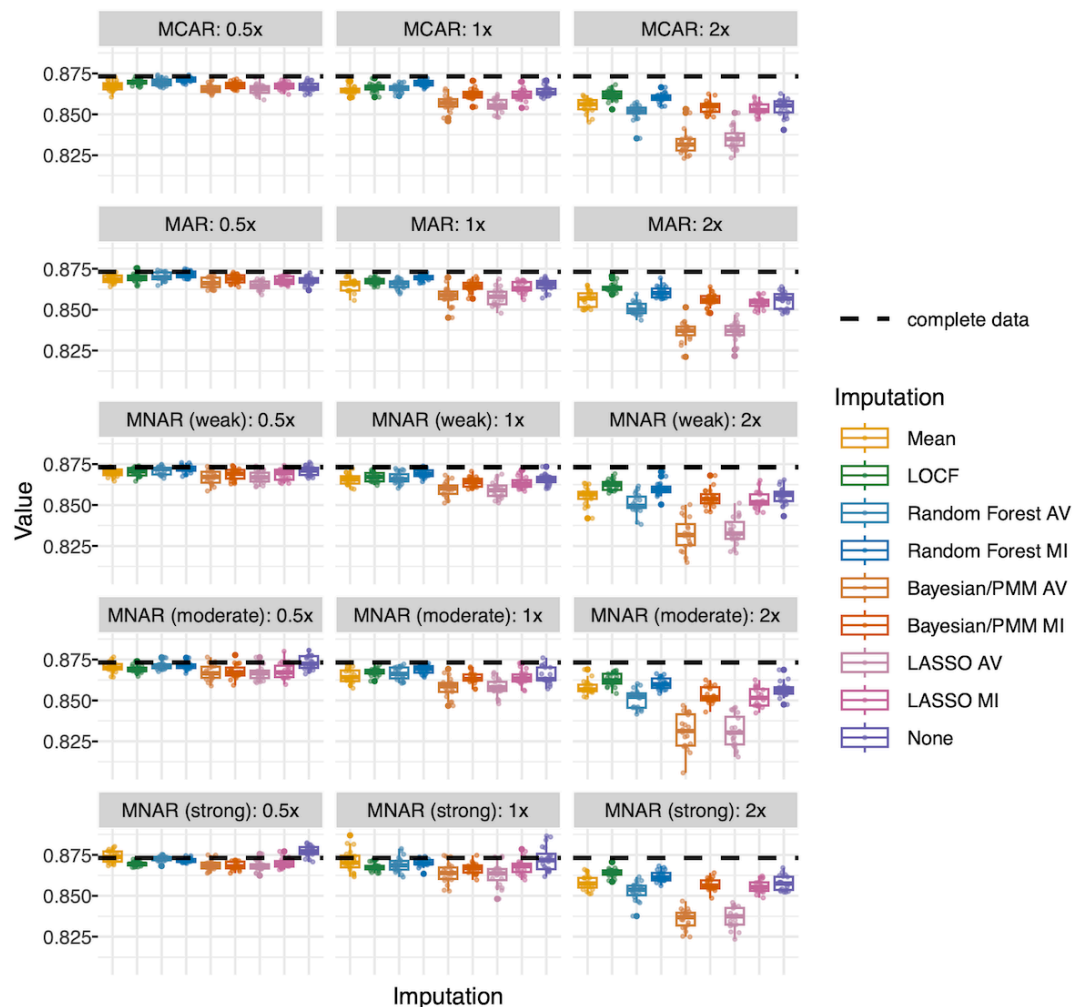
Extubation: AUC

Random forest multiple imputation had the highest AUC for gradient boosted models for 0.5x and 1x missingness, while LOCF did for 2x missingness (Figure 5). Random forest

Mean imputation and no imputation (native support for missing values) performed almost as well. Random forest (both averaged and multiple imputation) also performed well at 0.5x and 1x missingness, but its performance degraded at 2x missingness. The amount of missingness was more influential than the missingness mechanism (Table S6 in Multimedia Appendix 1).

averaged, mean imputation, and no imputation had reasonable performance but degraded at 2x missingness. The amount of missingness continued to have a greater effect on performance than the missingness mechanism.

Figure 5. Gradient boosted model test performance (extubation): AUC. Each point represents AUC for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. AUC in the complete dataset is represented by a dashed line. Other performance outcome metrics (sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV], and F1) are presented in [Multimedia Appendix 9](#). AV, average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAR, missing at random; MCAR: missing completely at random; MI: multiple imputation; MNAR: missing not at random.

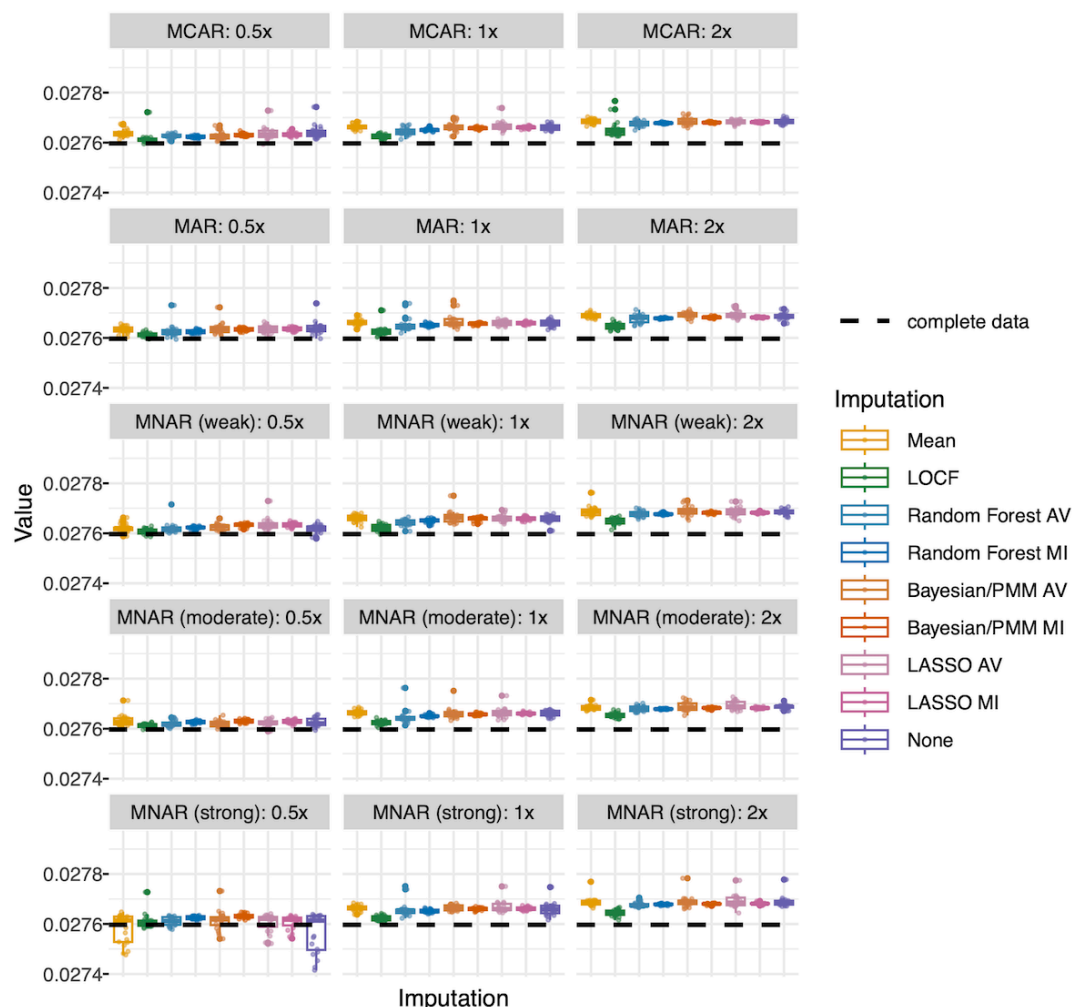


Blood Pressure: MSE

LOCF had the lowest overall MSE ([Figure 6](#)). Random forest (both averaged and multiple imputation) generally had the next lowest MSE. Unlike in extubation models, mean imputation and no imputation did not perform substantially better than other methods. Performance was again more

sensitive to amount of missingness than to missingness mechanism; overall, MSE increased in a stepwise fashion as missingness increased. However, there was less overall variability ($CV=0.001$) between models than for balanced accuracy ($CV=0.042$) and AUC ($CV=0.012$).

Figure 6. Gradient boosted model test performance (blood pressure): mean squared error (MSE). Each point represents MSE for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. MSE in the complete dataset is represented by a dashed line. Other performance outcome metrics (mean absolute error [MAE], root MSE, and R^2) are presented in [Multimedia Appendix 10](#). AV, average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAR, missing at random; MCAR: missing completely at random; MI: multiple imputation; MNAR: missing not at random.



LASSO Models

Extubation: Balanced Accuracy

For the LASSO outcome model, performance and patterns were similar to the gradient boosted model for balanced accuracy, with LOCF demonstrating top performance (Table S7 in [Multimedia Appendix 1](#); [Multimedia Appendix 11](#)). Mean imputation did not perform as well in LASSO models. Interestingly, imputation methods that performed worst—Bayesian/PMM and LASSO—yielded better results in LASSO models than in gradient boosted models, with less variation in performance across methods (CV=0.029).

Extubation: AUC

AUC followed similar patterns to balanced accuracy for LASSO. LOCF had the best performance. Mean imputation again performed worse comparatively, and there was a smaller performance gap between the best and worst imputation methods relative to gradient boosted models

(CV=0.011; Table S7 in [Multimedia Appendix 1](#); [Multimedia Appendix 11](#)).

Blood Pressure: MSE

LOCF had the lowest overall MSE ([Multimedia Appendix 12](#)). MSE again increased as missingness increased. MSE of LASSO models for all imputation methods was higher than gradient boosted models, although LASSO MSE was consistently closer to the MSE for complete data. The range between the best and worst imputation methods was narrower than for gradient boosted models (CV=0.0008).

Overfitting to training data: Both models

Extubation: Balanced Accuracy and AUC

Overfitting to the training data ([Multimedia Appendices 13 and 14](#)) increased as proportion of missingness increased. It was greatest for the worst-performing imputation methods (Bayesian/PMM and LASSO) in both gradient-boosted and

LASSO outcome models. It was higher overall for gradient boosted models than for LASSO outcome models.

Blood Pressure: MSE

Gradient boosted models exhibited more overfitting overall and higher variability in overfitting than LASSO models (Multimedia Appendices 15 and 16).

Discussion

Overview

EHR data present challenges in how to handle missing data when developing clinical prediction models. It is critical to use methods that are transferable to new data when generating real-time predictions for use cases like clinical decision support tools. In a realistic EHR dataset, we compared imputation methods based on (1) imputation accuracy and (2) outcome prediction.

Imputation Performance

LOCF and random forest multiple imputation consistently had the lowest MSE and classification error. LOCF has performed well in health survey and cohort datasets [18,20]. In our data, per-variable LOCF imputation error decreased monotonically with first-order autocorrelation, indicating better performance for more temporally stable variables. Random forest, while not explicitly designed for longitudinal data, was informed by lagged variables in our implementation and therefore also showed some gains for more temporally stable variables, though the association was weaker than for LOCF. By contrast, mean imputation does not account for correlation across repeated measures, and accordingly showed no relationship between autocorrelation and error. In the PICU, where many measurements are recorded frequently and less frequent measurements often indicate presumed stability over time, LOCF may be particularly suitable, while random forest may provide added value when temporal patterns are weaker.

As expected, imputation performance degraded as the proportion of missingness increased. Jäger et al [13] also found imputation performance generally worsened when difficulty (eg, higher missingness fraction and MNAR) increased. We found that missingness proportion affected imputation quality more than missingness mechanism. Although the imputation methods we employed are only theoretically valid in MCAR and MAR settings, performance in MNAR data was fairly similar, possibly due to the availability of repeated measurements over time, which may have captured underlying patterns. This is fortunate, given that EHR data are often likely MNAR [9].

Prediction Model Performance

Many studies do not report imputation performance [9,15,16], focusing instead on prediction performance. In our experiment, LOCF generally yielded the highest prediction performance across outcome types and models. We evaluated whether LOCF's strong performance was influenced by how

we constructed the synthetic, complete dataset. Although it may have been somewhat inflated, sensitivity analyses confirmed that LOCF remained a top performer. Its accuracy was greatest for variables with higher first-order autocorrelation, which may help explain its strong performance in this setting. While criticized in inferential statistics for causing bias and low standard errors [6,40], here it outperformed many multiple imputation methods.

Random forest multiple imputation was the best performing multiple imputation model. Jaeger et al [27] also found it led to the best predictive performance across 12 imputation strategies with 2 different outcome models in a registry dataset. Perez-Lebel et al [9] attributed multiple imputation's improved performance to ensembling (averaging multiple predictors) rather than accurately capturing the distribution of the missing values (the theoretical basis for its use in inferential statistics).

Native support for missing values (no imputation) yielded high balanced accuracy for gradient boosted models and reasonable performance for other metrics. Perez-Lebel et al [9] concluded that it had the best predictive performance in real-world EHR data, with the lowest computational cost. LOCF (which they did not test) also has very little computational cost but may be less broadly applicable.

As with imputation performance, missingness mechanism had less impact on prediction performance than proportion of missingness. Performance degraded substantially as missingness fraction increased, consistent with prior findings [13]. Interestingly, performance was similar across MAR, MCAR, and MNAR scenarios. Some studies suggest explicitly adding indicator variables for missingness to outcome models improves prediction in MNAR settings [9,10,41]. However, others argue against them because they are fragile to operational and practice changes and may not generalize well to other settings [11,12,42]. We therefore did not include such indicators, and performance in MNAR scenarios was comparable to MAR and MCAR scenarios. Machine learning models with native support for missing data may exploit missingness patterns for prediction. If no imputation is employed, missingness patterns should be closely monitored for drift and its effect on predictive performance [42].

Variability between imputation methods differed by model type and performance metric. LASSO outcome models showed less variability than gradient boosted models. In our experiment, imputation method appeared to have a greater impact on the binary outcome than the continuous outcome. Balanced accuracy had the highest variability between methods, followed by AUC; MSE exhibited substantially less variability than either. All imputation methods had lower MSE for gradient boosted models than LASSO models, indicating choice of outcome model mattered more than choice of imputation method.

Relationship Between Imputation and Prediction Performance

The best performing methods for imputation—LOCF and random forest multiple imputation—also performed well

in prediction. Thus, our main results for imputation and prediction performance were largely concordant. Mean imputation had the worst MSE but performed relatively well in gradient boosted models, possibly because constant imputation creates patterns that machine learning can exploit [9,43]. Its performance dropped in LASSO models. Others have noted that more accurate imputation methods do not always yield better predictions [9,27], especially when features are weakly correlated [9].

Overfitting in imputation models reduced imputation accuracy, which subsequently impacted the accuracy of outcome models. Outcome metrics that overfit more (eg, balanced accuracy in gradient boosted models) also showed greater variability between imputation methods in patterns that reflected differences in imputation accuracy.

Imputation and Interpretability

For clinical decision support, it is important to assess how missing data handling affects interpretability. The full promise of artificial intelligence will not be realized if it is not deemed trustworthy and transparent by humans [44,45]. Mean imputation, the most common method of imputation in machine learning clinical prediction models [12], may reduce interpretability for clinicians trying to understand predictions for individual patients, as the mean is not meaningful at the individual level. Native support by algorithms may improve interpretability by relying only on recorded values. LOCF is simple to implement and aligns with clinical reasoning—if a measurement is expected to change and that change is important, it will be remeasured if possible. Finally, if complex methods like random forests (while less interpretable themselves) generate accurate imputations that reflect a biological or clinical relationship between the predictor and outcome, this could lead to more interpretable outcome models that are also more robust over time and across populations [9].

Limitations

The primary limitation of this experiment is its reliance on both a single dataset and a single method for generating the complete dataset on which the analysis is based. However, a key strength is the high fidelity of the data, with detailed, frequent measurements typical of the intensive care unit. This contrasts with EHR data from settings like primary care,

where visits may be months or years apart. Although we predicted both a binary and continuous outcome, our dataset was limited to a single PICU at one academic medical center. Raw data were transformed into a structured format with time windows and summary variables, a common approach for EHR data [2,46]. Thus, our findings may not generalize to other settings or data structures.

We restricted imputation methods in our experiment to those with readily available packages in R that allowed model training on 1 dataset and subsequent imputation in new data. Deep learning methods were not included, nor were novel approaches [47–49] that may outperform tested methods but are more complex to implement for applied practitioners. Performance could also theoretically be improved by combining imputation approaches (eg, a SuperLearner [50]).

Most existing imputation packages do not allow users to save model parameters to apply on new data, limiting the methods available [15]. Some have proposed workarounds, such as stacking data from a new patient with all training data and rerunning multiple imputation models [15,17]. However, this was computationally infeasible in a dataset of our size. Privacy concerns may also prevent access to training data in deployment. Even packages like *mice* that allow imputation on single new cases have limitations—each call refits an iteration of the model, making real-time imputation for new patients infeasible due to speed. The lack of scalable imputation tools remains a barrier to progress to deploying real-time clinical prediction models.

Conclusion

When using EHR data with frequent measurements to build a prediction model, LOCF offers reasonable performance with simple implementation. Native support for missing data in machine learning models, such as gradient boosted trees, is the least computationally intensive approach, with decent performance and potentially broader applicability than LOCF. While multiple imputation is the gold standard for inferential models, it is extremely computationally intensive, may not be optimal for prediction models, and may not be feasible in real time. As clinical prediction models continue to integrate into real-time patient care, addressing missing data appropriately remains essential.

Acknowledgments

ChatGPT was used for assistance in writing R code and to make minor copyedits to the manuscript for conciseness and clarity.

Funding

This work was supported by NIH F31HL156498 (National Heart, Lung, and Blood Institute, Digitale).

Data Availability

The data underlying this article cannot be shared publicly because they contain protected health information.

Authors' Contributions

Conceptualization: CEM (equal), DF (equal), EDG (equal), JD (lead), MJP (equal),

Data curation: DF (supporting), JD (lead)

Formal analysis: JD (lead)

Funding acquisition: CEM (supporting), DF (supporting), EDG (supporting), JD (lead), MJP (supporting)

Methodology: CEM (equal), DF (equal), EDG (equal), JD (lead), MJP (equal)

Resources: JD (lead), MJP (supporting)

Software: EDG (supporting), JD (lead)

Supervision: CEM (equal), DF (equal), EDG (equal), MJP (equal)

Validation: JD (lead)

Visualization: JD (lead)

Writing – original draft: JD (lead)

Writing – review & editing: CEM (supporting), DF (supporting), EDG (supporting), JD (lead), MJP (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary methods and tables.

[\[PDF File \(Adobe File\), 728 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Imputation test performance metrics (blood pressure): mean squared error. Each point represents mean squared error calculated for 176 numeric variables for 1 of 300 datasets created for the outcome of blood pressure. There are 20 datasets per missingness scenario and imputation type represented in each box plot. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward.

[\[PDF File \(Adobe File\), 126 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Imputation test performance metrics (extubation and blood pressure): classification error. Each point represents classification error calculated for 6 categorical variables for 1 dataset (of 300 created for the outcome of extubation and 300 created for the outcome of blood pressure). There are 20 datasets per missingness scenario and imputation type represented in each box plot. The mice implementation of LASSO cannot accommodate multiclass categorical outcomes; thus, we used a simple classification tree. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward.

[\[PDF File \(Adobe File\), 249 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Imputation performance difference between train and test (extubation and blood pressure): mean squared error. Each point represents the difference in mean squared error between training and test sets calculated for 176 numeric variables for 1 dataset (of 300 created for the outcome of extubation and 300 created for the outcome of blood pressure). There are 20 datasets per missingness scenario and imputation type represented in each box plot. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward.

[\[PDF File \(Adobe File\), 257 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Imputation performance difference between train and test (extubation and blood pressure): classification error. Each point represents the difference in mean squared error between training and test sets calculated for 6 categorical variables for 1 dataset (of 300 created for the outcome of extubation and 300 created for the outcome of blood pressure). There are 20 datasets per missingness scenario and imputation type represented in each box plot. The mice implementation of LASSO cannot accommodate multiclass categorical outcomes; thus, we used a simple classification tree. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward.

[\[PDF File \(Adobe File\), 256 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Autocorrelation versus imputation error by method (extubation). Scatterplot of per-variable temporal autocorrelation (AR(1) coefficient, x-axis) versus imputation error (y-axis; mean squared error (MSE) for numeric variables, classification error for categorical) using 4-hour windows. Each point represents 1 predictor; higher persistence corresponds to lower error for last observation carried forward (LOCF).

[\[PDF File \(Adobe File\), 23 KB-Multimedia Appendix 6\]](#)

Multimedia Appendix 7

Imputation test performance metrics by missingness in original data (extubation and blood pressure): mean squared error. Each point represents mean squared error calculated for 176 numeric variables for 1 dataset stratified by whether the value was missing in the original data (of 300 created for the outcome of extubation and 300 created for the outcome of blood pressure). There are 20 datasets per missingness scenario and imputation type represented in each box plot. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward.

[\[PDF File \(Adobe File\), 769 KB-Multimedia Appendix 7\]](#)

Multimedia Appendix 8

Marginal means for interaction between imputation method, proportion missing, and variable group (extubation and blood pressure). We calculated marginal means for interaction from linear models of mean squared error comparing imputed values to complete dataset (for each outcome: 1 observation [n=264,000] per 176 variables per 300 datasets per 5 imputation methods). We calculated marginal means for interaction from linear models of classification error comparing imputed values to complete dataset (for each outcome: 1 observation [n=9000] per 6 variables per 300 datasets per 5 imputation methods) with random intercepts for each of the 1500 datasets. We included all 3-way and 2-way interactions and completed a backward stepwise elimination procedure (included $P<.05$) to determine the final model. Variable group 4, unlike other variable groups, included mostly indicator variables constructed from “select all that apply” responses in the electronic health record (EHR) (see Table S1). Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward.

[\[PDF File \(Adobe File\), 77 KB-Multimedia Appendix 8\]](#)

Multimedia Appendix 9

Gradient-boosted model test performance (extubation): other metrics. Each point represents performance for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. Performance in the complete dataset is represented by a dashed line. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MI: multiple imputation; NPV: negative predictive value; PPV: positive predictive value.

[\[PDF File \(Adobe File\), 1011 KB-Multimedia Appendix 9\]](#)

Multimedia Appendix 10

Gradient boosted model test performance (blood pressure): other metrics. Each point represents performance for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. Performance in the complete dataset is represented by a dashed line. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAE: mean absolute error; MI: multiple imputation; NPV: negative predictive value; PPV: positive predictive value; RMSE: root mean squared error.

[\[PDF File \(Adobe File\), 617 KB-Multimedia Appendix 10\]](#)

Multimedia Appendix 11

LASSO test performance (extubation): all metrics. Each point represents performance for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. Performance in the complete dataset is represented by a dashed line. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAE: mean absolute error; MI: multiple imputation; NPV: negative predictive value; PPV: positive predictive value; RMSE: root mean squared error.

[\[PDF File \(Adobe File\), 1232 KB-Multimedia Appendix 11\]](#)

Multimedia Appendix 12

LASSO test performance (blood pressure): all metrics. Each point represents performance for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. Performance in the complete dataset is represented by a dashed line. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAE: mean absolute error; MI: multiple imputation; MSE: mean squared error; NPV: negative predictive value; PPV: positive predictive value; RMSE: root mean squared error.

[\[PDF File \(Adobe File\), 720 KB-Multimedia Appendix 12\]](#)

Multimedia Appendix 13

Gradient-boosted model difference between train and test (extubation): All performance metrics. Each point represents a difference in performance between train and test sets for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. The difference in the complete dataset is represented by a dashed line. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAE: mean absolute error; MI: multiple imputation; MSE: mean squared error; NPV: negative predictive value; PPV: positive predictive value; RMSE: root mean squared error.

[\[PDF File \(Adobe File\), 1429 KB-Multimedia Appendix 13\]](#)

Multimedia Appendix 14

LASSO model difference between train and test (extubation): all performance metrics. Each point represents a difference in performance between train and test sets for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. The difference in the complete dataset is represented by a dashed line. AUC: area under the receiver operating characteristic curve; AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAE: mean absolute error; MI: multiple imputation; MSE: mean squared error; NPV: negative predictive value; PPV: positive predictive value; RMSE: root mean squared error.

[\[PDF File \(Adobe File\), 1248 KB-Multimedia Appendix 14\]](#)

Multimedia Appendix 15

Gradient boosted model difference between train and test (blood pressure): all performance metrics. Each point represents a difference in performance between train and test sets for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. The difference in the complete dataset is represented by a dashed line. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAE: mean absolute error; MI: multiple imputation; MSE: mean squared error; NPV: negative predictive value; PPV: positive predictive value; RMSE: root mean squared error.

[\[PDF File \(Adobe File\), 820 KB-Multimedia Appendix 15\]](#)

Multimedia Appendix 16

LASSO model difference between train and test (blood pressure): all performance metrics. Each point represents a difference in performance between train and test sets for 1 dataset. There are 20 datasets per missingness scenario and imputation type represented in each box plot. The difference in the complete dataset is represented by a dashed line. AV: average (average of 30 imputations); Bayesian/PMM: Bayesian imputation under the normal linear model with predictive mean matching; LASSO: least absolute shrinkage and selection operator; LOCF: last observation carried forward; MAE: mean absolute error; MI: multiple imputation; MSE: mean squared error; NPV: negative predictive value; PPV: positive predictive value; RMSE: root mean squared error.

[\[PDF File \(Adobe File\), 712 KB-Multimedia Appendix 16\]](#)

References

1. Wells BJ, Chagin KM, Li L, Hu B, Yu C, Kattan MW. Using the landmark method for creating prediction models in large datasets derived from electronic health records. *Health Care Manag Sci*. Mar 2015;18(1):86-92. [doi: [10.1007/s10729-014-9281-3](https://doi.org/10.1007/s10729-014-9281-3)] [Medline: [24752545](https://pubmed.ncbi.nlm.nih.gov/24752545/)]
2. Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *J Am Med Inform Assoc*. Dec 9, 2020;27(12):1921-1934. [doi: [10.1093/jamia/ocaa139](https://doi.org/10.1093/jamia/ocaa139)] [Medline: [33040151](https://pubmed.ncbi.nlm.nih.gov/33040151/)]
3. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. [doi: [10.1093/biomet/63.3.581](https://doi.org/10.1093/biomet/63.3.581)]
4. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol*. Sep 2020;125:183-187. [doi: [10.1016/j.jclinepi.2020.03.028](https://doi.org/10.1016/j.jclinepi.2020.03.028)] [Medline: [32540389](https://pubmed.ncbi.nlm.nih.gov/32540389/)]
5. Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for equitable health: assessing the impact of missing data in electronic health records. *J Biomed Inform*. Mar 2023;139:104269. [doi: [10.1016/j.jbi.2022.104269](https://doi.org/10.1016/j.jbi.2022.104269)] [Medline: [36621750](https://pubmed.ncbi.nlm.nih.gov/36621750/)]
6. Lachin JM. Fallacies of last observation carried forward analyses. *Clin Trials*. Apr 2016;13(2):161-168. [doi: [10.1177/1740774515602688](https://doi.org/10.1177/1740774515602688)]
7. Buuren S. Flexible Imputation of Missing Data. 2nd ed. CRC Press; 2018. URL: <https://stefvanbuuren.name/fimd/sec-pmm.html> [Accessed 2022-11-04]

8. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biom J*. Jul 2015;57(4):614-632. [doi: [10.1002/bimj.201400004](https://doi.org/10.1002/bimj.201400004)] [Medline: [25630926](https://pubmed.ncbi.nlm.nih.gov/25630926/)]
9. Perez-Lebel A, Varoquaux G, Le Morvan M, Josse J, Poline JB. Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*. Apr 15, 2022;11:giac013. [doi: [10.1093/gigascience/giac013](https://doi.org/10.1093/gigascience/giac013)] [Medline: [35426912](https://pubmed.ncbi.nlm.nih.gov/35426912/)]
10. 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*. Aug 2023;32(8):1461-1477. [doi: [10.1177/09622802231165001](https://doi.org/10.1177/09622802231165001)]
11. Tsvetanova A, Sperrin M, Peek N, Buchan I, Hyland S, Martin GP. Missing data was handled inconsistently in UK prediction models: a review of method used. *J Clin Epidemiol*. Dec 2021;140:149-158. [doi: [10.1016/j.jclinepi.2021.09.008](https://doi.org/10.1016/j.jclinepi.2021.09.008)] [Medline: [34520847](https://pubmed.ncbi.nlm.nih.gov/34520847/)]
12. Nijman S, Leeuwenberg AM, Beekers I, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J Clin Epidemiol*. Feb 2022;142:218-229. [doi: [10.1016/j.jclinepi.2021.11.023](https://doi.org/10.1016/j.jclinepi.2021.11.023)] [Medline: [34798287](https://pubmed.ncbi.nlm.nih.gov/34798287/)]
13. Jäger S, Allhorn A, Bießmann F. A benchmark for data imputation methods. *Front Big Data*. 2021;4:693674. [doi: [10.3389/fdata.2021.693674](https://doi.org/10.3389/fdata.2021.693674)] [Medline: [34308343](https://pubmed.ncbi.nlm.nih.gov/34308343/)]
14. Breiman L, Friedman J, Stone C, Olshen R. *Classification and Regression Trees*. Wadsworth & Brooks; 1984. URL: <https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-olshen-charles-stone> [Accessed 2025-11-10] ISBN: 9781315139470
15. Hoogland J, van Barneveld M, Debray TPA, et al. Handling missing predictor values when validating and applying a prediction model to new patients. *Stat Med*. Nov 10, 2020;39(25):3591-3607. [doi: [10.1002/sim.8682](https://doi.org/10.1002/sim.8682)] [Medline: [32687233](https://pubmed.ncbi.nlm.nih.gov/32687233/)]
16. Nijman SWJ, Hoogland J, Groenhof TKJ, et al. On behalf of the UCC-CVRM and UCC-SMART study groups. Real-time imputation of missing predictor values in clinical practice. *Eur Heart J - Digit Health*. Mar 1, 2021;2(1):154-164. [doi: [10.1093/ehjdh/ztaa016](https://doi.org/10.1093/ehjdh/ztaa016)]
17. Janssen KJM, Vergouwe Y, Donders ART, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem*. May 2009;55(5):994-1001. [doi: [10.1373/clinchem.2008.115345](https://doi.org/10.1373/clinchem.2008.115345)] [Medline: [19282357](https://pubmed.ncbi.nlm.nih.gov/19282357/)]
18. Ribeiro C, Freitas AA. A data-driven missing value imputation approach for longitudinal datasets. *Artif Intell Rev*. Dec 2021;54(8):6277-6307. [doi: [10.1007/s10462-021-09963-5](https://doi.org/10.1007/s10462-021-09963-5)]
19. Rosner B, Cook N, Portman R, Daniels S, Falkner B. Determination of blood pressure percentiles in normal-weight children: some methodological issues. *Am J Epidemiol*. Mar 15, 2008;167(6):653-666. [doi: [10.1093/aje/kwm348](https://doi.org/10.1093/aje/kwm348)] [Medline: [18230679](https://pubmed.ncbi.nlm.nih.gov/18230679/)]
20. Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. *J Clin Epidemiol*. Oct 2003;56(10):968-976. [doi: [10.1016/s0895-4356\(03\)00170-7](https://doi.org/10.1016/s0895-4356(03)00170-7)] [Medline: [14568628](https://pubmed.ncbi.nlm.nih.gov/14568628/)]
21. Ahmed RA, Tube B. *Endotracheal Tube*. StatPearls Publishing; 2023. [Medline: [30969569](https://pubmed.ncbi.nlm.nih.gov/30969569/)]
22. Mayer M. *missRanger: fast imputation of missing values*. 2023. URL: <https://CRAN.R-project.org/package=missRanger> [Accessed 2025-10-25]
23. A language and environment for statistical computing. R Core Team. URL: <https://www.R-project.org> [Accessed 2025-11-04]
24. Schouten RM, Lugtig P, Vink G. Generating missing values for simulation purposes: a multivariate amputation procedure. *J Stat Comput Simul*. Oct 13, 2018;88(15):2909-2930. [doi: [10.1080/00949655.2018.1491577](https://doi.org/10.1080/00949655.2018.1491577)]
25. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Med Res Methodol*. Dec 12, 2018;18(1):168. [doi: [10.1186/s12874-018-0615-6](https://doi.org/10.1186/s12874-018-0615-6)] [Medline: [30541455](https://pubmed.ncbi.nlm.nih.gov/30541455/)]
26. Li J, Yan XS, Chaudhary D, et al. Imputation of missing values for electronic health record laboratory data. *NPJ Digit Med*. Oct 11, 2021;4(1):1-14. [doi: [10.1038/s41746-021-00518-0](https://doi.org/10.1038/s41746-021-00518-0)]
27. Jaeger BC, Cantor R, Sthanam V, Xie R, Kirklin JK, Rudraraju R. Improving outcome predictions for patients receiving mechanical circulatory support by optimizing imputation of missing values. *Circ Cardiovasc Qual Outcomes*. Sep 2021;14(9):e007071. [doi: [10.1161/CIRCOUTCOMES.120.007071](https://doi.org/10.1161/CIRCOUTCOMES.120.007071)] [Medline: [34517728](https://pubmed.ncbi.nlm.nih.gov/34517728/)]
28. 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*. Mar 15, 2014;179(6):764-774. [doi: [10.1093/aje/kwt312](https://doi.org/10.1093/aje/kwt312)] [Medline: [24589914](https://pubmed.ncbi.nlm.nih.gov/24589914/)]
29. Buuren SV, Groothuis-Oudshoorn K. *mice: multivariate imputation by chained equations in R*. *J Stat Softw*. 2011;45(3). [doi: [10.18637/jss.v045.i03](https://doi.org/10.18637/jss.v045.i03)]

30. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist.* 2001;29(5):1180-1232. [doi: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451)]
31. Shi Y, Ke G, Soukhavong D, et al. Lightgbm: light gradient boosting machine. 2023. URL: <https://CRAN.R-project.org/package=lightgbm> [Accessed 2025-10-25]
32. Olson RS, Cava WL, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. In: *Biocomput 2018 World Scientific.* 2018:192-203. [doi: [10.1142/9789813235533_0018](https://doi.org/10.1142/9789813235533_0018)] ISBN: 978-981-323-552-6
33. Gennatas ED. Towards precision psychiatry: gray matter development and cognition in adolescence. University of Pennsylvania; 2017. URL: <https://repository.upenn.edu/entities/person/8f68097c-4130-40bd-9e6d-453632bbfe1e> [Accessed 2025-11-04]
34. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst Curran Associates, Inc;* 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html [Accessed 2023-12-22]
35. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1-22. [Medline: [20808728](https://pubmed.ncbi.nlm.nih.gov/20808728/)]
36. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. Presented at: 2010 20th Int Conf Pattern Recognit. 3121-3124; 2010.[doi: [10.1109/ICPR.2010.764](https://doi.org/10.1109/ICPR.2010.764)]
37. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Med Res Methodol.* Dec 12, 2022;22(1):316. [doi: [10.1186/s12874-022-01801-8](https://doi.org/10.1186/s12874-022-01801-8)] [Medline: [36510134](https://pubmed.ncbi.nlm.nih.gov/36510134/)]
38. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep.* Aug 2022;12(1):5979. [doi: [10.1038/s41598-022-09954-8](https://doi.org/10.1038/s41598-022-09954-8)]
39. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. *BMJ.* Sep 3, 2024:e078276. [doi: [10.1136/bmj-2023-078276](https://doi.org/10.1136/bmj-2023-078276)]
40. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med.* Oct 4, 2012;367(14):1355-1360. [doi: [10.1056/NEJMs1203730](https://doi.org/10.1056/NEJMs1203730)] [Medline: [23034025](https://pubmed.ncbi.nlm.nih.gov/23034025/)]
41. Sharafoddini A, Dubin JA, Maslove DM, Lee J. A new insight into missing data in intensive care unit patient profiles: observational study. *JMIR Med Inform.* Jan 8, 2019;7(1):e11605. [doi: [10.2196/11605](https://doi.org/10.2196/11605)] [Medline: [30622091](https://pubmed.ncbi.nlm.nih.gov/30622091/)]
42. van Smeden M, Groenwold RHH, Moons KG. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. *J Clin Epidemiol.* Sep 2020;125:188-190. [doi: [10.1016/j.jclinepi.2020.06.007](https://doi.org/10.1016/j.jclinepi.2020.06.007)] [Medline: [32565213](https://pubmed.ncbi.nlm.nih.gov/32565213/)]
43. Josse J, Prost N, Scornet E, Varoquaux G. On the consistency of supervised learning with missing values. *arXiv.* Preprint posted online on 2020. [doi: [10.48550/arXiv.1902.06931](https://doi.org/10.48550/arXiv.1902.06931)]
44. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc.* Jul 1, 2020;27(7):1173-1185. [doi: [10.1093/jamia/ocaa053](https://doi.org/10.1093/jamia/ocaa053)] [Medline: [32417928](https://pubmed.ncbi.nlm.nih.gov/32417928/)]
45. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA.* Dec 4, 2018;320(21):2199-2200. [doi: [10.1001/jama.2018.17163](https://doi.org/10.1001/jama.2018.17163)] [Medline: [30398550](https://pubmed.ncbi.nlm.nih.gov/30398550/)]
46. Gupta M, Gallamoza B, Cutrona N, Dhakal P, Poulain R, Beheshti R. An extensive data processing pipeline for MIMIC-IV. *Proc Mach Learn Res.* Nov 2022;193:311-325. [Medline: [36686986](https://pubmed.ncbi.nlm.nih.gov/36686986/)]
47. Luo Y. Evaluating the state of the art in missing data imputation for clinical data. *Brief Bioinform.* Jan 17, 2022;23(1):bbab489. [doi: [10.1093/bib/bbab489](https://doi.org/10.1093/bib/bbab489)] [Medline: [34882223](https://pubmed.ncbi.nlm.nih.gov/34882223/)]
48. Kazijevs M, Samad MD. Deep imputation of missing values in time series health data: a review with benchmarking. *J Biomed Inform.* Aug 2023;144:104440. [doi: [10.1016/j.jbi.2023.104440](https://doi.org/10.1016/j.jbi.2023.104440)] [Medline: [37429511](https://pubmed.ncbi.nlm.nih.gov/37429511/)]
49. Liu M, Li S, Yuan H, et al. Handling missing values in healthcare data: a systematic review of deep learning-based imputation techniques. *Artif Intell Med.* Aug 2023;142:102587. [doi: [10.1016/j.artmed.2023.102587](https://doi.org/10.1016/j.artmed.2023.102587)]
50. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol.* 2007;6(1). [doi: [10.2202/1544-6115.1309](https://doi.org/10.2202/1544-6115.1309)]

Abbreviations

AUC: area under the receiver operating characteristic curve
CV: coefficient of variation
EHR: electronic health record
ETT: endotracheal tube
LASSO: least absolute shrinkage and selection operator
LOCF: last observation carried forward

MAR: missing at random
MCAR: missing completely at random
MI: multiple imputation
MNAR: missing not at random
MSE: mean squared error
PICU: pediatric intensive care unit
PMM: predictive mean matching

Edited by Arriel Benis; peer-reviewed by Jaime Speiser, Zoltan Kern; submitted 18.Jun.2025; final revised version received 25.Sep.2025; accepted 13.Oct.2025; published 14.Nov.2025

Please cite as:

Digitale J, Franzon D, Pletcher MJ, McCulloch CE, Gennatas ED

Methods for Addressing Missingness in Electronic Health Record Data for Clinical Prediction Models: Comparative Evaluation

JMIR Med Inform2025;13:e79307

URL: <https://medinform.jmir.org/2025/1/e79307>

doi: [10.2196/79307](https://doi.org/10.2196/79307)

© Jean Digitale, Deborah Franzon, Mark J Pletcher, Charles E McCulloch, Efstathios D Gennatas. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 14.Nov.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.