Integrated Analysis for Electronic Health Records with Structured and Sporadic Missingness

Jianbin Tan[*1], Yan Zhang[†1], Chuan Hong[‡], T. Tony Cai[§], Tianxi Cai[¶2], and Anru R. Zhang[‖2]

Abstract

Objectives: We propose a novel imputation method tailored for Electronic Health Records (EHRs) with structured and sporadic missingness. Such missingness frequently arises in the integration of heterogeneous EHR datasets for downstream clinical applications. By addressing these gaps, our method provides a practical solution for integrated analysis, enhancing data utility and advancing the understanding of population health.

Materials and Methods: We begin by demonstrating structured and sporadic missing mechanisms in the integrated analysis of EHR data. Following this, we introduce a novel imputation framework, MACOMSS, specifically designed to handle structurally and heterogeneously occurring missing data. We establish theoretical guarantees for MACOMSS, ensuring its robustness in preserving the integrity and reliability of integrated analyses. To assess its empirical performance, we conduct extensive simulation studies that replicate the complex missingness patterns observed in real-world EHR systems, complemented by validation using EHR datasets from the Duke University Health System (DUHS).

[*]Department of Biostatistics & Bioinformatics, Duke University, NC, USA.
[†]Department of Biostatistics & Bioinformatics, Duke University, NC, USA.
[‡]Department of Biostatistics & Bioinformatics, Duke University, NC, USA
[§]Department of Statistics & Data Science, University of Pennsylvania, PA, USA
[¶]Department of Biostatistics and Department of Biomedical Informatics, Harvard University.
[‖]Department of Biostatistics & Bioinformatics and Department of Computer Science, Duke University, NC, USA
[1]These authors contributed equally.
[2]Joint corresponding authors.

Results: Simulation studies show that our approach consistently outperforms existing imputation methods. Using datasets from three hospitals within DUHS, MACOMSS achieves the lowest imputation errors for missing data in most cases and provides superior or comparable downstream prediction performance compared to benchmark methods.

Discussion: The proposed method effectively addresses critical missingness patterns that arise in the integrated analysis of EHR datasets, enhancing the robustness and generalizability of clinical predictions.

Conclusions: We provide a theoretically guaranteed and practically meaningful method for imputing structured and sporadic missing data, enabling accurate and reliable integrated analysis across multiple EHR datasets. The proposed approach holds significant potential for advancing research in population health.

Keywords: Clinical Prediction, Electronic Health Records, Heterogeneity, Matrix Completion, Population Health

1 Background and Significance

Electronic Health Records (EHRs) have become a cornerstone of modern healthcare, offering rich, multidimensional data that support clinical decision-making, advance scientific research, and guide health policy development (Evans, 2016; Essén et al., 2018; Beaulieu-Jones et al., 2018; Tayefi et al., 2021; Ahuja et al., 2022; Psychogyios et al., 2023; Tian et al., 2024; Tan et al., 2024b; Li et al., 2024). With the widespread adoption of EHR systems, the volume and diversity of collected data have grown significantly (Hemingway et al., 2018; Johnson et al., 2024), creating great potential for integrated analyses to deepen our understanding of population health. Yet, due to the intricate nature of healthcare delivery and data collection practices, we inevitably encounter challenges related to missing data (Madden et al., 2016; Beaulieu-Jones et al., 2018; Hemingway et al., 2018; Haneuse et al., 2021; Tan et al., 2022; Psychogyios et al., 2023; Luo et al., 2025). These gaps not only hinder the ability to derive accurate insights but also limit data utility, opening new opportunities to develop advanced methods for integrated analysis of EHR data.

When integrating datasets from multiple sources, structured and sporadic missingness are two common missing mechanisms that are often encountered in EHR analysis.

Structured missingness typically results from the fact that different data sources are collected from distinct patient populations. For example, due to resource constraints or disease-specific clinical workflows, not all tests or procedures are performed and recorded uniformly across data sources (Madden et al., 2016; Bower et al., 2017; Haneuse et al., 2021). This often leads to systematic gaps and misalignment when integrating data into a unified structure. In addition, disparities in data accessibility across organizations—driven by factors such as privacy regulations, institutional policies, or differences in technological infrastructure (Beard et al., 2012; Keshta and Odeh, 2021; Tertulino et al., 2024)—may further contribute to such structured gaps in the integrated analysis.

Sporadic missingness, by contrast, often arises from incomplete documentation, technical malfunctions, or missing responses to surveys (Wells et al., 2013; Bower et al., 2017; Haneuse et al., 2021). This type of missingness may occur randomly, with mechanisms that vary not only across datasets but also within a single dataset. The coexistence of structured and sporadic missingness poses significant challenges for EHR data analysis; these difficulties may be further compounded by the heterogeneous nature of data across healthcare systems. These challenges highlight the need for advanced imputation methods capable of accommodating diverse missingness patterns in integrated analyses.

Recently, various imputation methods have been developed to address missing data in EHR datasets, ranging from traditional statistical techniques to advanced machine learning approaches. Classical methods, such as mean imputation (Little and Rubin, 2019) or regression imputation (van Buuren and Groothuis-Oudshoorn, 2011; Morris et al., 2014), are straightforward but often struggle to handle the complexity of missingness in EHR data. Machine learning-based techniques, such as random forests, K-NN, principal component analysis, support vector machines, matrix factorization, and deep learning imputation methods (Wang et al., 2015; Madden et al., 2016; Beaulieu-Jones et al., 2017, 2018; Nazabal et al., 2020; You et al., 2020; Yoon et al., 2020; Li et al., 2021; Aidos and Tomás, 2021; Pathak et al., 2022; Psychogyios et al., 2023; Tan et al., 2024a), have shown promise in imputing missing data by leveraging the richness of available information. However, these approaches usually require that data are missing at random or uniformly distributed across the dataset. This assumption may not be reasonable and is often violated in EHR data (Wells et al., 2013; Beaulieu-Jones et al., 2018; Haneuse et al., 2021; Getzen et al., 2023), making such methods less effective in addressing the distinct mechanisms of structured and sporadic missingness.

2 Objective

In this article, we propose a novel data imputation method tailored for datasets with structured and sporadic missingness, paving the way for integrating multiple EHR datasets for their downstream tasks in clinical analysis. Our goal is to develop a robust and theoretically grounded approach that effectively mitigates both systematic and randomly occurring gaps in integrated analysis. By addressing these complexities, we improve the completeness and accuracy of EHR datasets, supporting more precise clinical insights and population health research. A flowchart illustrating our objective is presented in Figure 1.

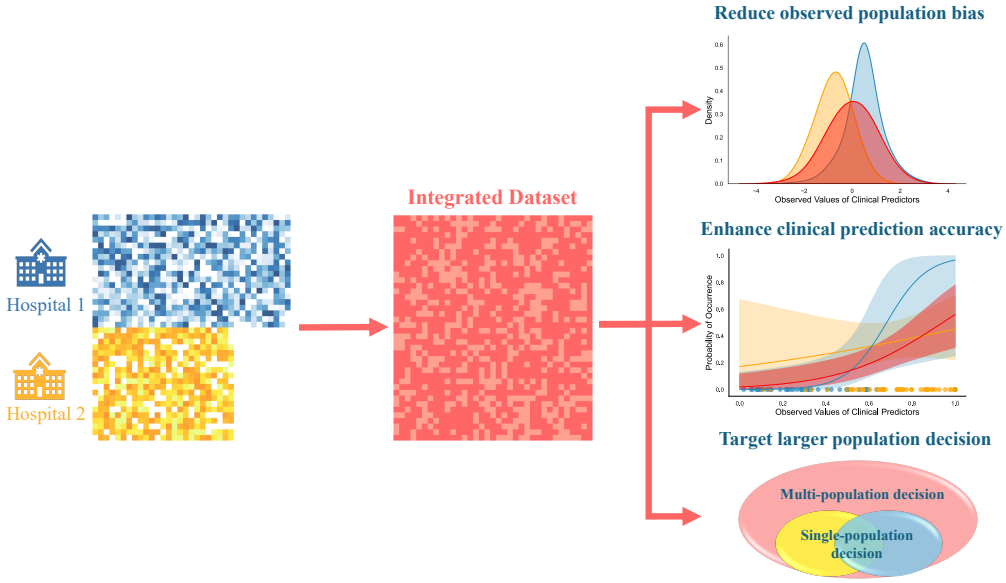

Figure 1: An illustration of the objective of this article.

3 Materials and Methods

In this section, we present a detailed procedure for data imputation of structured and sporadic missingness. We begin with basic notations and definitions that will be used throughout the paper. For any two real numbers a and b, we note $a \wedge b$ and $a \vee b$ as the minimum and the maximum of a and b, respectively. Denote $\|\cdot\|_F$ and $\|\cdot\|$ as the Frobenius and spectral norm of a matrix; their definitions are given in Part A in Supplementary Materials.

3.1 Matrix Completion with Structured and Sporadic Missing Data

We particularly focus on the following setting in this paper. For a high-dimensional low-rank matrix $A \in \mathbb{R}^{p_1 \times p_2}$, we observe its m_1 rows and m_2 columns with noise and possible missing values. Specifically, let $Y = A + Z \in \mathbb{R}^{p_1 \times p_2}$ represent the potential noisy observations without missingness, and $Z \in \mathbb{R}^{p_1 \times p_2}$ denote the noise. To analyze the missingness, we introduce $M \in \mathbb{R}^{p_1 \times p_2}$ as the indicator for observable/missing entries:

$$Y_{ij} \text{ is observed if } M_{ij} = 1; \quad Y_{ij} \text{ is missing if } M_{ij} = 0.$$

Without loss of generality, we permute the structurally missing block to the bottom right corner, and Y and M can be written in the following block form:

$$Y = \begin{array}{c} \begin{matrix} m_2 & p_2 - m_2 \end{matrix} \\ \begin{bmatrix} Y_{(11)} & Y_{(12)} \\ Y_{(21)} & Y_{(22)} \end{bmatrix} \end{array} \begin{array}{c} m_1 \\ p_1 - m_1 \end{array}, \quad M = \begin{array}{c} \begin{matrix} m_2 & p_2 - m_2 \end{matrix} \\ \begin{bmatrix} M_{(11)} & M_{(12)} \\ M_{(21)} & M_{(22)} \end{bmatrix} \end{array} \begin{array}{c} m_1 \\ p_1 - m_1 \end{array}. \tag{1}$$

The block $Y_{(22)}$ is unobserved, so all entries of $M_{(22)}$ are zero. Moreover, entrywise missingness might exist in $Y_{(11)}$, $Y_{(12)}$, and $Y_{(21)}$, meaning that a subset of the entries in $M_{(11)}$, $M_{(12)}$, and $M_{(21)}$ are zero. To model the sporadic missingness, we assume that each entry in the observable rows and columns is missing independently according to a probability matrix Θ. In other words, $M_{ij} \sim \text{Bernoulli}(\Theta_{ij})$ for i, j corresponding to the entries of $M_{(11)}, M_{(12)}, M_{(21)}$.

Here, the rows of Y usually represent patients from multiple EHR systems, while the columns indicate the observed clinical features of the patients. The unobserved block $Y_{(22)}$ may arise when the observed entries in each row are misaligned. This misalignment often occurs in EHR integrated analysis, where the observed clinical features may differ across different sources of datasets (Madden et al., 2016; Bower et al., 2017; Haneuse et al., 2021).

The block-structured missingness in (1) introduces significant challenges in estimating the latent probability matrix Θ. To address this difficulty, we focus on the case of a rank-one probability matrix, i.e., $\text{rank}(\Theta) = 1$. The rank-one missingness model has been widely studied in the matrix completion literature (see, e.g., Keshavan et al. (2010);

Chatterjee (2015); Cho et al. (2017)), which allows the sporadic missingness in Y to occur heterogeneously across patients, clinical features, and data sources. This assumption offers a parsimonious structure and facilitates both algorithmic design and theoretical analysis, but it does not need to be strictly satisfied in practice for the methods to perform well (see Section 3.3.3 for a demonstration).

Our goal is to recover A in an unsupervised manner from the available observations Y_{ij} for those (i, j) pairs where $M_{ij} = 1$, considering both the structured and sporadic missingness mechanisms in the data. Clearly, such a task is impossible in general without further assumptions. Motivated by applications in matrix completion, we assume the original matrix A is low-rank (or nearly low-rank), which enables our method to impute missing values by leveraging low-dimensional shared structure. We postpone more detailed discussions on technical conditions to Part A in Supplementary Materials.

3.2 Algorithm

The proposed algorithm for matrix completion with structured and sporadic missing values consists of five steps and two key components: a spectral method for imputing sporadic missingness and denoising (Steps 1 and 2), and a structured matrix completion method for imputing structured missingness (Steps 3, 4, and 5). The central ideas behind Steps 1 and 2 are mean matching for aligning the observations and parameters, and singular value decomposition for extracting low-rank structures from missing entries and noise. Steps 3, 4, and 5 rely on a simple yet powerful algebraic fact—the Schur complement. Below, we outline the five steps for estimating the noiseless matrix A in detail:

Step 1. (Estimation of Missingness Parameter Θ) Without loss of generality, we permute the structurally missing block to the bottom right corner and assume that Y and M are written in the block structure form (1). In this step, we estimate the missingness parameters, denoted by Θ_{ij}, for all observable rows and columns. Given the assumption that the parameter matrix Θ is rank-one, Lemma 1 in Supplementary Materials indicates that each entry of Θ can be represented as the product of corresponding row and column sums divided by the total sum of the entries in Θ (i.e., $\Theta_{ij} = (\sum_{i'=1}^{p_1} \Theta_{i'j})(\sum_{j'=1}^{p_2} \Theta_{ij'})/(\sum_{i'j'} \Theta_{i'j'})$). Notice that the expectation of M_{ij} is Θ_{ij} (i.e., $\mathbb{E}M_{ij} = \Theta_{ij}$), we estimate Θ_{ij} by computing the corresponding products of the row and column sums of M, normalized by the overall sum of the entries in M.

Step 2. (Imputation of Missing Values) We begin by filling all missing entries of the original data matrix Y with zeros. Subsequently, we normalize each observed entry by dividing it by the corresponding estimated missingness parameter from Step 1 (i.e., $\tilde{Y}_{ij} := Y_{ij}/\hat{\Theta}_{ij}$). This procedure yields a normalized matrix $\tilde{Y}$, which removes biases introduced by missing observations. We then partition the normalized data matrix into four sub-blocks following (1), denoted by $\tilde{Y}_{(11)}, \tilde{Y}_{(12)}, \tilde{Y}_{(21)}$, and $\tilde{Y}_{(22)}$. Notably, all entries of $\tilde{Y}_{(22)}$ are zeros.

Step 3. (Rotation of $\tilde{Y}_{(\bullet 1)}, \tilde{Y}_{(1\bullet)}$) Notice that $\tilde{Y}_{(11)}$, $\tilde{Y}_{(12)}$, and $\tilde{Y}_{(21)}$ are noise observations of the low-rank matrix $A_{(11)}$, $A_{(12)}$, and $A_{(21)}$, we propose to apply the spectral method on $\tilde{Y}$ to extract the underlying low-rank structure. Specifically, we calculate the singular value decomposition for

$$Y^{(\mathrm{col})} = \begin{bmatrix} \frac{(p_1-2m_1)\wedge 0}{p_1} \cdot \tilde{Y}_{(11)} \\ \tilde{Y}_{(21)} \end{bmatrix}, \quad \text{and} \quad Y^{(\mathrm{row})} = \begin{bmatrix} \frac{(p_2-2m_2)\wedge 0}{p_2} \cdot \tilde{Y}_{(11)}, & \tilde{Y}_{(12)} \end{bmatrix}.$$

The resulting singular vectors provide rotation matrices for the observable blocks $\tilde{Y}_{(11)}, \tilde{Y}_{(12)}$, and $\tilde{Y}_{(21)}$, obtaining the rotated matrices $B_{(11)}, B_{(12)}, B_{(21)}$. After the rotations, the significant components hidden in $\tilde{Y}$ are moved to the front ranking rows and columns in $B_{(11)}, B_{(12)}$, and $B_{(21)}$.

Step 4. (Trimming and Rank Determination) In this step we aim at trimming $B_{(11)}, B_{(12)}$, and $B_{(21)}$ to low-rank blocks, where the specific rank $\hat{r}$ needs to be estimated. To do this, we iteratively estimate an appropriate rank for truncation, starting from an initial upper bound and decreasing sequentially. The optimal rank choice satisfies certain stability conditions involving bounds based on the dimensions of observable and missing data, as discussed in Part A in Supplementary Materials.

Step 5. (Assembling and Imputation) In the final step, we treat A as a rank-$\hat{r}$ matrix and reconstruct A by combining the trimmed blocks $B_{(11)}, B_{(12)}, B_{(21)}$ using rotations obtained in Step 3.

By combining the steps described above, our procedure can be implemented as outlined in Algorithm 1 in Supplementary Materials, referred to as Matrix completion with Missing Structurally and Sporadically (MACOMSS). Unlike conventional machine-learning-based imputation methods (Wang et al., 2015; Madden et al., 2016; Beaulieu-Jones et al., 2017, 2018; Li et al., 2021; Pathak et al., 2022; Psychogyios et al., 2023), MACOMSS is tuning-free and straightforward to implement. In addition, the proposed approach effectively

accounts for structured and sporadic missing mechanisms during imputation and incorporates a denoising process for the observed data. In contrast, the existing imputation methods often overlook such complex missing mechanisms and incorporate noise from the observations into imputation processes. This can lead to inaccuracies when the data exhibits structured and sporadic missingness and are observed with relatively high noise.

In practice, we may encounter both continuous and categorical variables simultaneously in matrix completion tasks. For this setting, MACOMSS is applicable to both types of variables based on a sub-Gaussian framework (see Part A of the Supplementary Materials), which encompasses a broad class of continuous or discrete distributions such as Gaussian, uniform, Bernoulli, Binomial, Hypergeometric, and bounded discrete uniform distributions (Vershynin, 2018). Certain unbounded distributions, such as Poisson, may also exhibit sub-Gaussian behavior under appropriate conditions or after a log transformation. Due to this, we always apply a logarithmic transformation to count data before imputation.

Under the sub-Gaussian framework, we analyze the statistical lower and upper bounds of MACOMSS for estimating the latent matrix A from the observed data Y_{ij}s. These results not only highlight the theoretical optimality of MACOMSS in preserving the data integrity.

3.3 Validation

3.3.1 Simulations for Matrix Recovery

We first consider the recovery performance of MACOMSS under various values of m_1 and m_2, which represent the number of observable rows and columns, respectively. Particularly, we fix $p_1 = p_2 = 300$, $r = 3$, and generate $A = UV^\top$, where $U \in \mathbb{R}^{p_1 \times r}$ and $V \in \mathbb{R}^{p_2 \times r}$ are uniformly random orthogonal matrices. We then uniformly randomly select m_1 rows and m_2 columns for observation with missing values under the following settings:

1. We vary m_1 and m_2 among $\{10, 20, \ldots, 100\}$. The missing parameter Θ is constructed as a rank-1 matrix: $\Theta = \alpha\beta^\top$, where $\alpha \in \mathbb{R}^{p_1}$, $\alpha_i \sim 1 - 0.05 \cdot \text{Unif}[0, 1]$, and $\beta \in \mathbb{R}^{p_2}$, $\beta_j \sim 1 - 0.05 \cdot \text{Unif}[0, 1]$. Finally, we corrupt each observation with i.i.d. Gaussian noise as in $Y = A + Z$, and introduce sporadic missingness to Y based on the probability matrix M. Here, the variance of the entries in Z is given as

$\sigma = 0.3 \cdot \|A\|_F / \sqrt{p_1 p_2}$, and M is sampled from Bernoulli distributions with the missing probability being the entries in Θ. Such a setting mimics the real data situation in EHR datasets, where a small portion of sporadic missingness and observational noise typically exist.

2. We consider another setting for investigating the influence of the noise σ and the missingness Θ on the recovery performance. Specifically, we fix $m_1 = m_2 = 50$, and A to be generated in the same way as the previous setting. We let σ vary from $0.2 \cdot \|A\|_F / \sqrt{p_1 p_2}$ to $2 \cdot \|A\|_F / \sqrt{p_1 p_2}$ and set $\Theta = \alpha \cdot \beta^\top$, where $\alpha \in \mathbb{R}^{p_1}, \beta \in \mathbb{R}^{p_2}$, and $\alpha_i, \beta_j \overset{\text{iid}}{\sim} \text{Unif}[1-\eta, 1]$, with η varying from 0 to 0.25. A larger η indicates a higher missing probability.

All experiments in the above settings are repeated for 1,000 times. Based on the observable entries from (Y, M), we apply Algorithm 1 in Supplementary Materials to obtain estimates $\hat{A}$. We utilize the average Frobenius and spectral norm loss of recovery to evaluate the accuracy of $\hat{A}$ for estimating A.

3.3.2 Simulations for Downstream Tasks

Next, we evaluate the performance of MACOMSS for downstream tasks after matrix completion. Specifically, we focus on logistic regressions on the dataset (X, Z), where $X \in \mathbb{R}^{n \times p}$ contains the observations of n samples with p features as predictors, and $Z \in \mathbb{R}^n$ represents the binary responses. Here, n represents the number of patients, p denotes the number of clinical features for patients, Z contains binary outcomes of patients, and X is the predictor matrix, including observed values of clinical features from patients. To mimic the setting in real-world EHR datasets, we suppose that the matrix X presents with both structured and sporadic missing entries and is observed with noise. Our objective is to impute the missing values of X and denoise X, and then use the recovered matrix to predict the response Z.

For data generation, we first sample the matrix X and then generate the response Z using a logistic model; see Section E.2 in Supplementary Materials for details. To introduce sporadic missingness, we generate the observed values of X according to a rank-1 missing probability matrix Θ, similar to those in Section 3.3.1. Specifically, Θ is constructed using $\Theta = \alpha\beta^\top$, where $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$. Each element of α and β is independently drawn from $1 - 0.1 \cdot \text{Unif}[0, 1]$. Given the matrix Θ, we then generate the

label M_{ij} by $M_{ij} \sim \text{Bernoulli}(\Theta_{ij})$, where Θ_{ij} is the (i,j)-th element in Θ. If $M_{ij} = 1$, we then observe the (i,j)-th element of X.

Furthermore, we introduce a structured missing mechanism in the predictor matrix X under the two scenarios:

- In scenario 1, we evaluate different row numbers, n, and assume that the intersection of the final 60% of rows and the last 45 columns is completely missing, while the rest is observable with sporadic missingness.
- In scenario 2, we instead evaluate different column numbers, p, and suppose that the intersection of the final 45 rows and the last 60% of columns is entirely missing, with the remaining entries observable with sporadic missingness.

These two scenarios correspond to cases where the missing block enlarges with the number of rows (patients) or columns (clinical features).

Finally, for each observed entry of X, we add Gaussian noise $\text{Gau}(0, \sigma)$, where σ is set as $\text{SNR} \times \frac{\|X\|_F}{\sqrt{np}}$, with SNR denoting the signal-to-noise ratio. The contaminated and incomplete observations of X, together with the responses Z generated by the noiseless predictor matrix X, are then used for estimation.

We compare MACOMSS with several existing data imputation methods, abbreviated as PMM, BLR, RS, CART, K-NN, VAE, and VAA, where the first five methods are implemented using the `R` package `MICE` (van Buuren and Groothuis-Oudshoorn, 2011). Specifically, Predictive Mean Matching (PMM; Morris et al., 2014) is a regression-based method that provides imputed values by predicting them through regression. Bayesian Linear Regression (BLR; van Buuren and Groothuis-Oudshoorn, 2011) is another regression-based method that incorporates parameter uncertainty using prior distributions. Random Sampling (RS) imputes missing values by randomly selecting from observed values within the same feature. Classification and Regression Trees (CART; Krzywinski and Altman, 2017) use a machine learning technique that partitions the data into subsets and imputes missing values by learning from these partitions. K-Nearest Neighbors (K-NN; Hastie et al., 2001) is another machine learning approach that imputes data using the similarity between data points. In addition, Variational Autoencoders (VAE; Nazabal et al., 2020) are deep generative models that learn a latent representation of the data and reconstruct missing values by sampling from this latent space. Variational neighborhood-aware autoencoders (VAA; Aidos and Tomás, 2021) extend VAE by coupling a latent generative

model with a K-NN, locality-driven imputation step.

We apply the above methods to impute the missing data in X based on its observed entries. In addition, we employ MACOMSS for imputation and denoising of X. Following this, we implement a logistic regression between Z and the imputed matrix of X, utilizing elastic net regularization (Zou and Hastie, 2005) via the R package **glmnet** (Friedman et al., 2010).

To evaluate the matrix recovery performance of each method, we calculate the normalized mean squared errors (NMSE) for the missing entries of X:

$$\text{NMSE}_X = \frac{\|\hat{X}_{\text{miss}} - X_{\text{miss}}\|^2}{\|X_{\text{miss}}\|^2}.$$

Here, X_{miss} and $\hat{X}_{\text{miss}}$ are vectors of the true and estimated values for the missing entries. In addition, we calculate the area under the ROC curve (Huang and Ling, 2005), denoted as AUC, to evaluate predictive performances of classification, where the classification function is estimated from the imputed predictor matrix and the responses Z using logistic regressions. For each setting, we replicate 100 simulations to calculate the NMSE and AUC. In calculating AUC, we also include a complete-data benchmark method where the classification function is estimated from the logistic regression of Z onto the true predictor matrix X.

3.3.3 Experiments in Real Datasets

We evaluate the performance of MACOMSS using EHR data collected from the Duke University Health System (DUHS), accessed through the Duke Clinical Research Datamart (CRDM) (Hurst et al., 2021). The dataset integrates EHR data from three hospitals, each treated as an isolated site: Duke Raleigh Hospital (DRAH) as site 1, Duke Regional Hospital (DRH) as site 2, and Duke University Hospital (DUH) as site 3. Our study focuses on all emergency department (ED) visits and records clinical features of patients across three sites in 2019. To mimic the missing mechanism present in integrated analysis, we first introduce structured and sporadic missingness to the clinical data from different sources under various settings. After that, we apply imputation methods to fill in the missing values and employ the imputed data to predict whether a visit will result in inpatient admission—a classification task on the integrated EHR datasets.

For the classification, the predictors include four categories: (1) Demographics infor-

mation: age and sex; (2) Vital signs: pulse (beats/min), systolic blood pressure (SBP; mm Hg), diastolic blood pressure (DBP; mm Hg), oxygen saturation (SpO_2; %), temperature (°F), respiration (times/min), and acuity level; (3) Comorbidities: indicators of local tumor, metastatic tumor, diabetes with complications, diabetes without complications, and renal disease, for a patient. Comorbidities are defined based on the ICD-10-CM Diagnosis Code; (4) PheCodes: in addition to known comorbidities, all other ICD-10-CM Diagnosis Codes were aggregated into PheCodes to represent more general diagnoses using the ICD-to-PheCode mapping from PheWAS catalog (https://phewascatalog.org/phecodes). We utilized one-digit level PheCodes in the analysis. Given the low incidence rate of the response outcome, we down-sample the data by one-third for cases with outcome 0 at each site. For the final cohort, we include only samples without any initial missingness to facilitate the design of different missing data mechanisms and evaluate imputation performance accurately.

To implement our experiment, we collect 438 predictors from 126,579 visits, coupled with observed indicators of whether a visit will result in inpatient admission as responses. To generate target datasets, we first sample 400 visits and p predictors from the three sources of collected data, forming a $400 \times p$ matrix, where p will vary across different values. For each visit, we always include the predictors related to demographics information, vital signs, and comorbidities, including a total of 13 features. The remaining $(p-13)$ features are randomly sampled from the predictors in PheCodes. Accordingly, we perform a log transformation and standardization on the $400 \times p$ matrix, and divide the 400 visits (including both the predictors and inpatient admission for the visit) into ten folds. We use nine folds as the training set and check the accuracy of classification on the remaining testing fold. We then output the averaged cross-validated classification accuracy for the 400 visits. To ensure the generalizability of our results, we repeat the above process multiple times for other randomly sampled 400 visits and p predictors.

In the training dataset above, we introduce missing values into the $360 \times p$ matrix X using a sporadic missing mechanism. To be more general, we consider a non–rank-one missing matrix to explore more complicated sporadic missing mechanisms. Specifically, the missing probability matrix Θ is constructed as $\Theta = \sum_{m=1}^{4} \lambda_m \alpha_m \beta_m^\top + \varepsilon_\Theta$, where $\alpha_m \in \mathbb{R}^{360}$, $\beta_m \in \mathbb{R}^p$, $\lambda_m = 0.5^{3(m-1)}$, and $\varepsilon_\Theta \in \mathbb{R}^{360 \times p}$ is a matrix containing mean-zero Gaussian noise variables with standard deviation $\frac{1}{20} \left\| \sum_{m=1}^{4} \lambda_m \alpha_m \beta_m^\top \right\| / \sqrt{360p}$. Each element of α_m and β_m is independently drawn from $1 - 0.2 \cdot \mathrm{Unif}[0,1]$. Given the matrix Θ, we then generate the label M_{ij} by $M_{ij} \sim \mathrm{Bernoulli}(\Theta_{ij})$, where Θ_{ij} is the (i,j)-th

element in Θ. If $M_{ij} = 0$, we set the (i, j)-th element of X as missing.

Additionally, we introduce structured missingness by removing observed values in $l\%$ of the rows and columns. The missing rows are randomly sampled from the 360 visits, while the missing columns always include important predictors related to emergency hospital admission, including demographic information, vital signs, and comorbidities (Wallace et al., 2014; Lucke et al., 2018; Brink et al., 2022). The remaining missing columns are randomly sampled from the predictors of PheCodes. We evaluate different values of p and l to examine cases with varying numbers of clinical features and missing proportions.

We apply MACOMSS to impute the incomplete matrix. To evaluate accuracies, we utilize the imputed matrices to perform logistic regressions on their responses, as described in Section 3.3.2. We then output the completion errors, defined by the normalized mean squared errors (NMSEs) between the imputed values and their true observed values, with their corresponding AUC values calculated from the testing dataset, as defined in Section 3.3.2.

4 Results

4.1 Results for Recovery Accuracy

Under the simulation settings 1 - 2 in Section 3.3.1, we present the average Frobenius and spectral norm loss of recovery in Figure 2, where we vary the values of m_1, m_2, η, and σ to examine the theoretical properties and empirical performance of MACOMSS.

By Figures 2, it can be seen that as m_1, m_2 grow or σ, η decrease, i.e., more rows and columns of A are available or the entries of A are observed with a lower noise and fewer sporadic missing values, we achieve better recovery performance using MACOMSS. These results align with the theoretical analysis in Supplementary Materials, demonstrating the stable performance of the proposed algorithm across all values of m_1, m_2, σ, and η.

In Figures 1–2 of the Supplementary Materials, we further examine the performance of MACOMSS for non–low-rank and Poisson count matrices. The results show that our method is effective for both nearly low-rank and Poisson count matrix cases, demonstrating the applicability of MACOMSS.

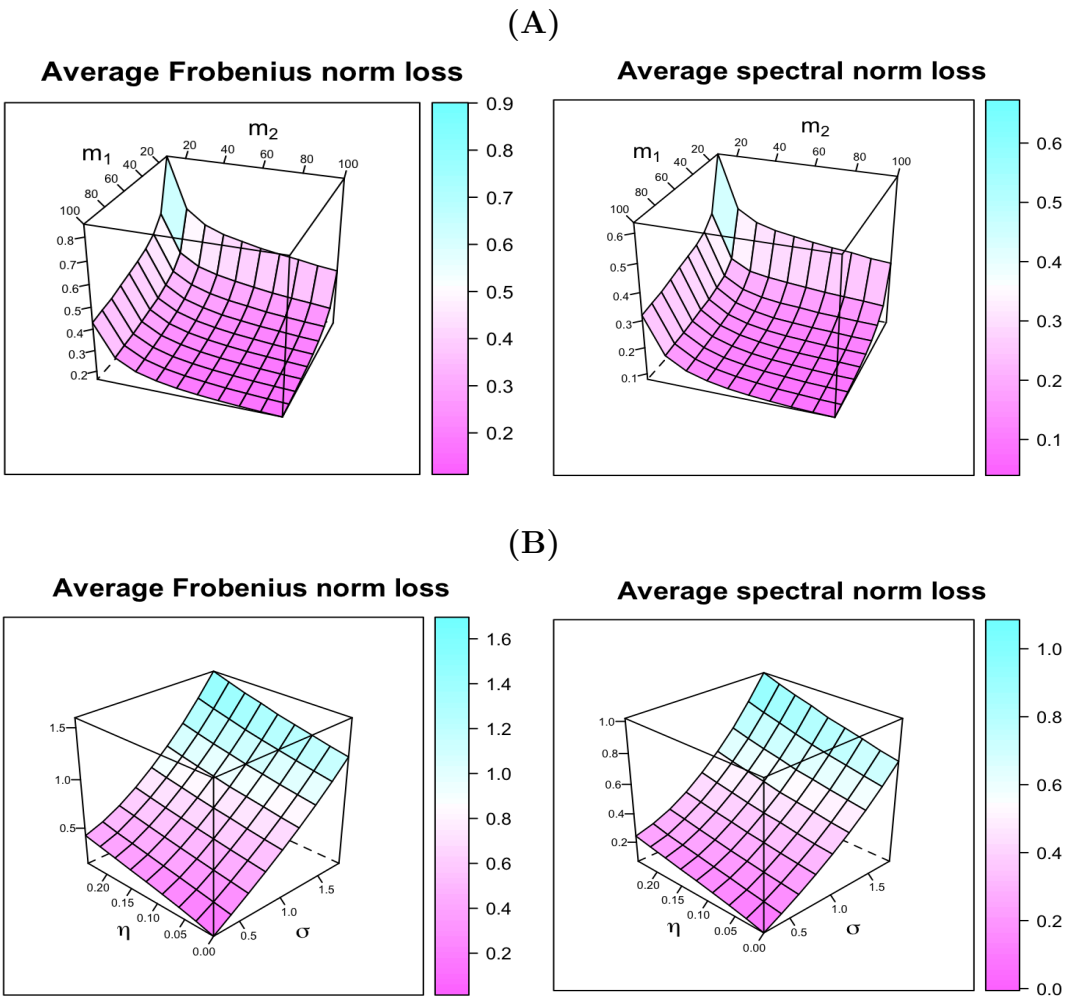

Figure 2: (A): Average Frobenius and spectral norm loss for MACOMSS with varying number of observable rows and columns: $m_1, m_2 \in [10, 100]$. (B): Average Frobenius and spectral norm loss for MACOMSS for varying Θ and σ.

4.2 Results for Downstream Tasks

In this subsection, we examine the performance of MACOMSS for the downstream classification tasks under scenarios 1 - 2 in Section 3.3.2, compared with the existing imputation methods listed in that subsection.

We first consider scenario 1, with the number of rows, n, ranging from 100 to 1000, and the number of columns, p, fixed at 70. The corresponding NMSEs and AUCs are shown in Figure 3. In Figure 3(A), we observe that MACOMSS consistently outperforms

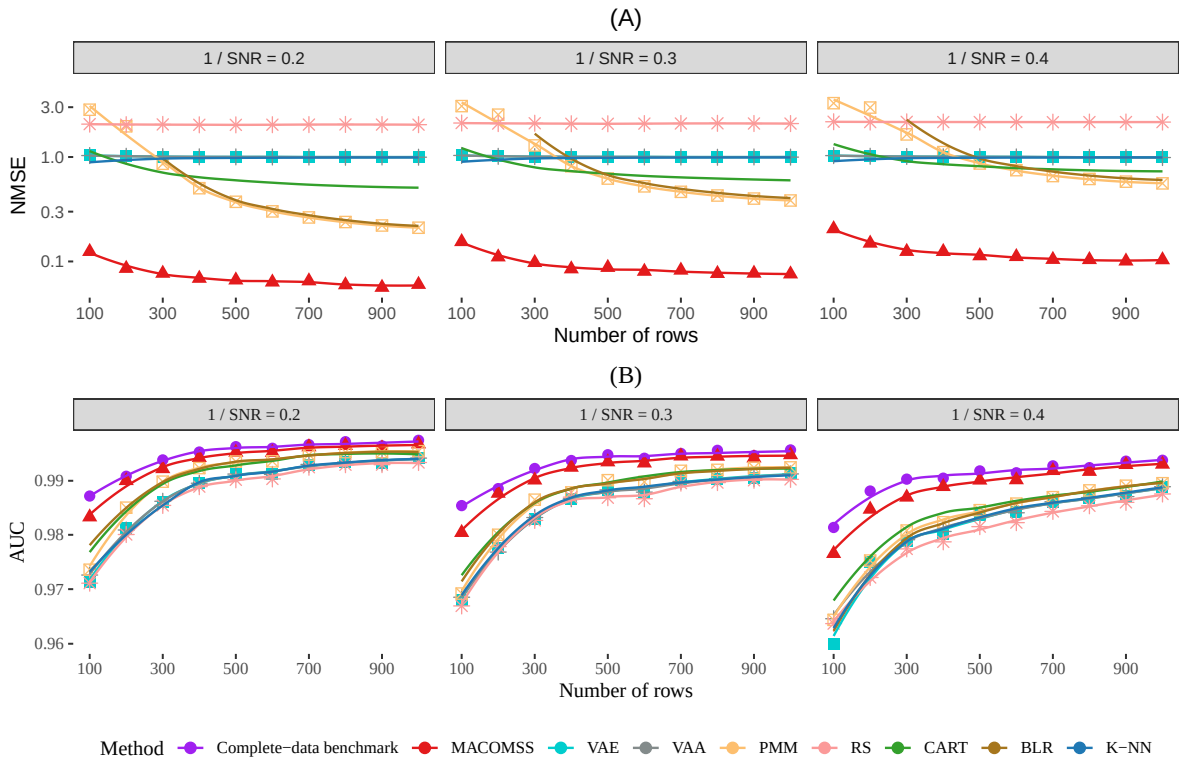

Figure 3: The NMSE (A) and the AUC (for simulated binary outcomes) (B) for scenario 1 for different numbers of rows from different methods.

other methods, achieving smaller NMSEs across all values of n and SNR. Moreover, the performance gap between MACOMSS and the other competing methods widens as SNR decreases, since the latter do not denoise the observed entries. As a result, our method yields more high AUC values, indicating more accurate estimates of regression coefficients and better classification accuracy for the downstream classification task. These advantages become more pronounced as the noise level increases of the data (see Panel (B) of Figure 3).

Moreover, we consider the high-dimensional setting under scenario 2, where the number of rows, n, is fixed at 70, and the number of columns, p, varies from 100 to 200. The corresponding NMSEs and AUCs are presented in Figure 4. In these cases, we similarly observe that MACOMSS outperforms other methods in terms of NMSE and AUC.

In both scenarios 1 and 2, we find that the AUCs of MACOMSS are close to that of

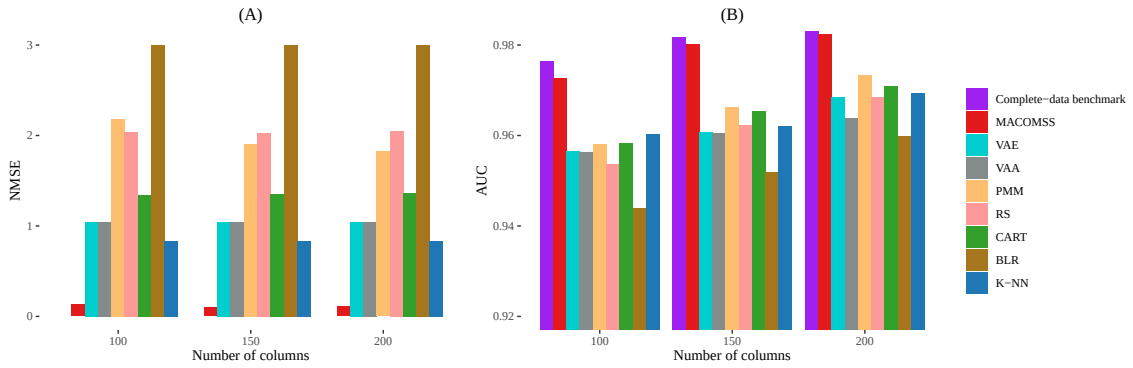

Figure 4: The NMSE (A) and AUC (for simulated binary outcomes) (B) for scenario 2 for different numbers of columns from different methods. We set the SNR to 5 for data generation.

the complete-data case (Panel (B) in Figures 3 - 4), whether we have a large missing block with the increased number of rows (patients) or columns (clinical features). These results indicate that MACOMSS achieves nearly optimal performance in the predictive task using the imputed matrix, owing to its consideration of flexible sporadic and structural missingness mechanisms for data imputation.

4.3 Results for Real EHR Datasets

Before performing analysis, we examine the validity of the low-rank assumption for the EHR dataset in Part E.3 of the Supplementary Materials. Our results show that the data matrix are nearly low-rank, which fulfills the requirement of MACOMSS. Using these EHR datasets, we conduct experiments with different values of p and l under the setting described in Section 3.3.3. In these experiments, we randomly sample 400 patients and p clinical features with a missing block controlled by l, repeated 50 times. We apply MACOMSS to impute these matrices for downstream classification tasks, where the classification accuracy is measured by AUC in testing sets. For comparison, we compute the AUC for logistic regressions using the complete predictor matrix, serving as the complete-data method for classification performance. To examine the impact of block missingness, we also report the AUC obtained by removing clinical features with structural missingness – a conventional approach that disregards structural patterns in downstream tasks.

These AUC values are shown in Figure 5.

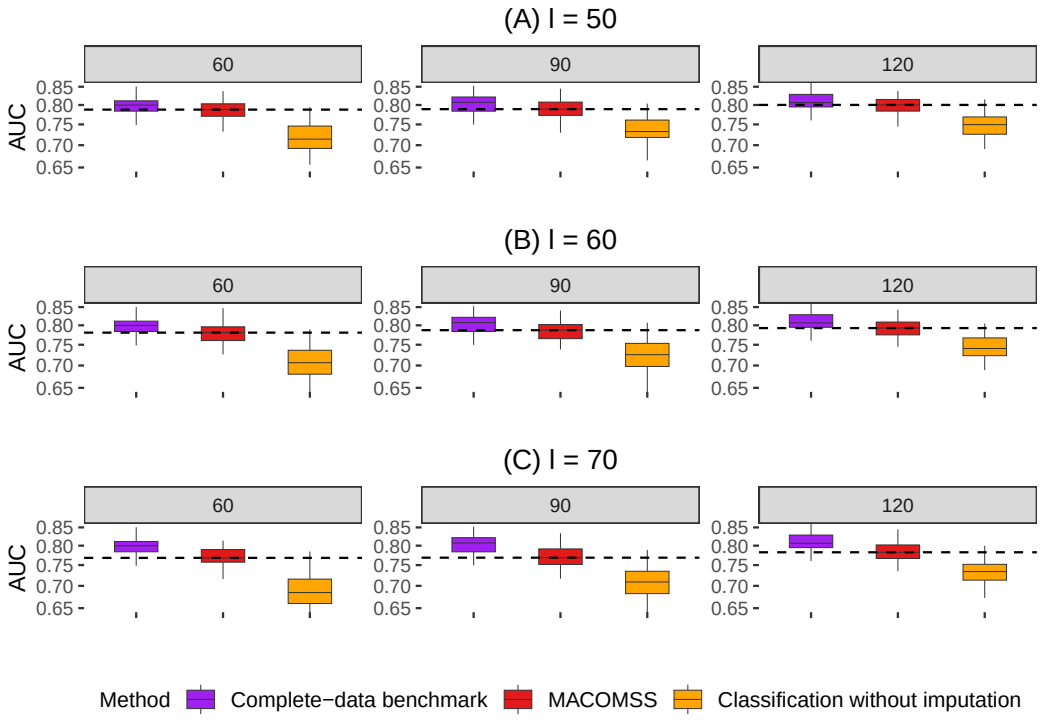

Figure 5: The AUC (for admission prediction) from 50-times experiments with the missing proportions, $l\%$ (main title), ranging from 50 to 70, and the number of features (sub-title), p, ranging from 60 to 120. The dashed horizontal lines indicate the median values of AUC from MACOMSS.

We observe that MACOMSS significantly outperforms classification without imputation, as evidenced by the AUC values in Figure 5. Notably, this recovery achieves a prediction accuracy close to the complete-data method, demonstrating the generalizability of MACOMSS for clinical prediction tasks.

In Figure 6, we illustrate the coefficients from logistic regression models using the above methods, estimated from three randomly selected replication experiments in Figure 5 with $p = 120$ and $l = 50$. We find that MACOMSS identifies clinical features in demographic information, vital signs, and comorbidities—similar to the complete-data benchmark. These features are usually important predictors for emergency hospital admission (Wallace et al., 2014; Lucke et al., 2018; Brink et al., 2022), but they are ignored when no imputation

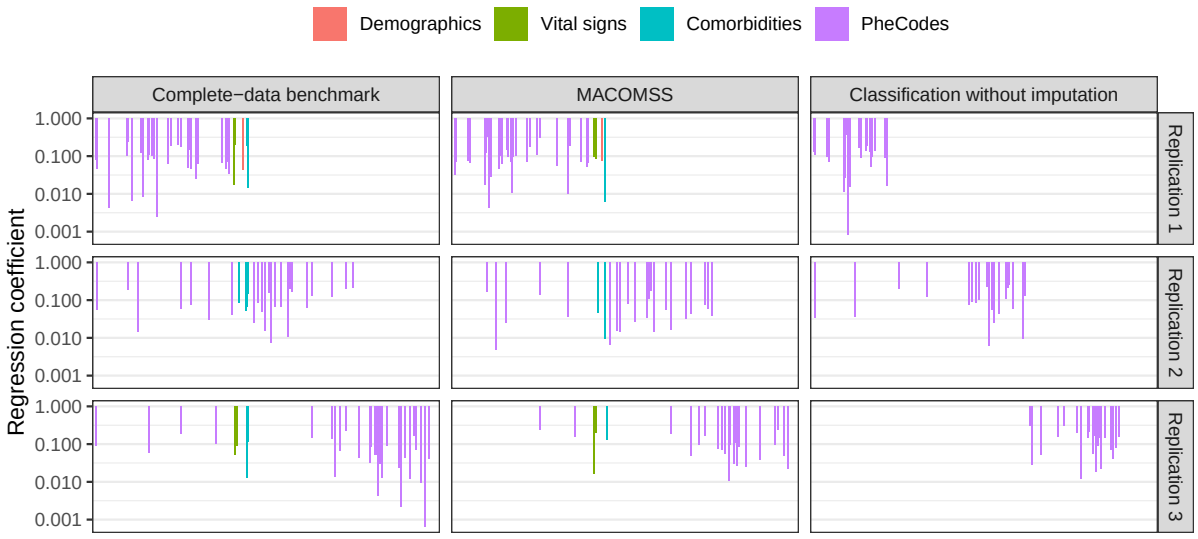

Figure 6: Estimated logistic regression coefficients from three randomly selected replication experiments under the complete-data benchmark, MACOMSS, and classification without imputation.

is performed. These results highlight the necessity of imputing block missingness in the predictor matrix prior to classification, as the imputation may recover important information crucial for predicting patient admission.

We further compare MACOMSS with PMM, RS, CART, BLR, and K-NN listed in Section 3.3.2. We evaluate the imputation accuracy using NMSE, along with the corresponding AUC to assess the performance of downstream predictive tasks. The NMSEs and AUCs of different methods across 50 repeated experiments are presented in Figure 7. We observe that MACOMSS mostly outperforms other methods in terms of NMSE (Panel (A), (C), and (E) in Figure 7). This satisfactory imputation performance of MACOMSS leads to superior prediction accuracy for downstream classification tasks, as evidenced by the AUC results in Panels (B), (D), and (F) of Figure 7.

We also compute the runtime and memory usage benchmarks for the above methods using the DUHS dataset. For dataset of dimension $360 \times p$, with $p \in \{60, 90, 120\}$, we report the averaged runtime and memory consumption for each method in Table 1, evaluated on a server with 2.10 GHz CPU cores and 208 GB of RAM. These results show that MACOMSS achieves favorable computational efficiency compared to other methods.

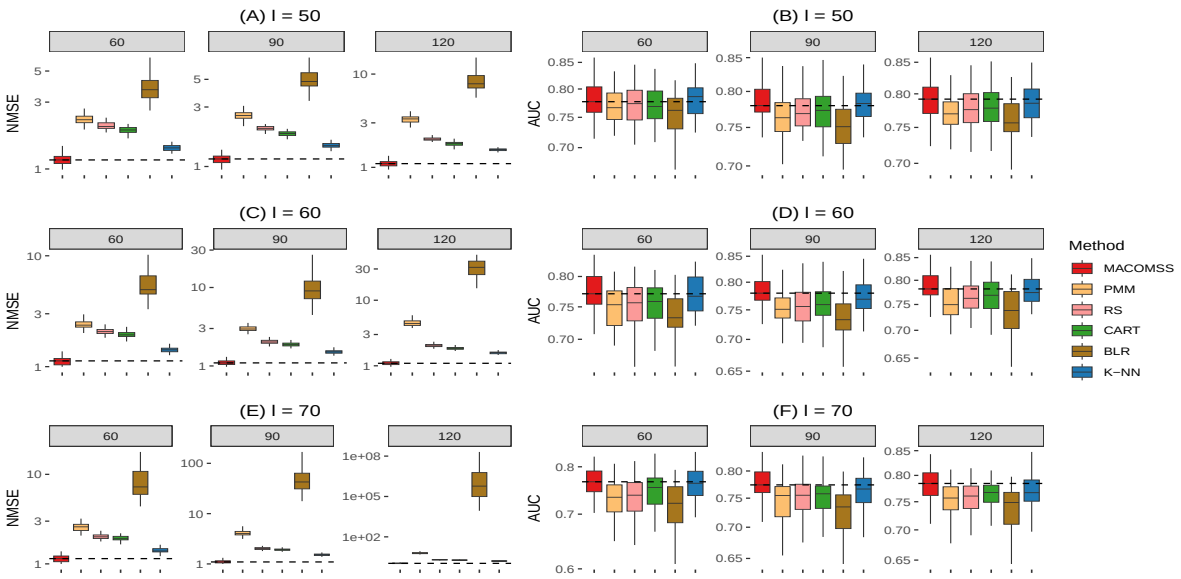

Figure 7: The NMSE ((A), (C), and (E)) and AUC for admission prediction ((B), (D), and (F)) with the missing proportions, $l\%$ (main title), ranging from 50 to 70, and the number of features (sub-title), p, ranging from 60 to 120. The dashed horizontal lines indicate the median values of NMSE or AUC from MACOMSS.

We perform additional experiments to compare the performances between MACOMSS and previous methods by increasing the heterogeneity level in the data. To this end, we randomly divide the patients in the clinical dataset into three groups and generate a new dataset by adding group-specific noise, increasing the population bias and heterogeneity level in the data. Subsequently, we apply MACOMSS and other imputation methods to the new dataset and replicate the previous experiments 50 times. The averaged NMSE under different heterogeneity levels are presented in Figure 8. We observe that the NMSE of MACOMSS becomes smaller compared to that of the other as the heterogeneity level increases. These results suggest that when significant bias and heterogeneity is present in different data sources, existing methods may no longer provide satisfactory imputation for structured and sporadic missingness. Whereas MACOMSS, in general, can achieve superior performance for this setting, which highlights its applicability for integrative analysis of EHR datasets with potential heterogeneity.

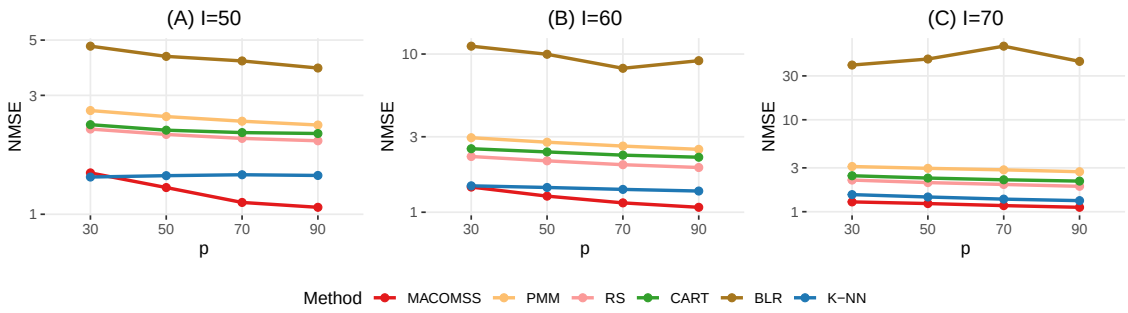

Figure 8: The averaged NMSE is computed over 50 repeated experiments with varying missing proportions $l\%$ and $p = 120$. For each experiment, we randomly partition the patients into three groups and generate a new dataset by adding Gaussian noise to the original dataset according to these groups. For each clinical feature in the k-th group $(k = 1, 2, 3)$, the mean and standard deviation of the added noise are set to $(100 - (k - 2)h)\%$ of the feature's mean and 10% of its standard deviation, respectively. A larger $h\%$ corresponds to a higher level of population bias and heterogeneity.

Table 1: Average runtime (in seconds) and memory usage (in MB) over 50 repeated experiments with $p \in \{60, 90, 120\}$ for different methods on the DUHS dataset.

	MACOMSS	PMM	RS	CART	BLR	K-NN
Time	0.072	4.166	4.173	3.484	7.771	96.707
Memory	17.631	635.871	623.902	464.169	812.366	22859.319

5 Discussion

This article proposes a novel matrix completion method named MACOMSS for imputing datasets with structured and sporadic missingness, effectively addressing key missingness occurring in the integrated analysis of EHR data. Our approach preserves critical information that may be lost due to the intricate healthcare delivery and data collection systems and ensures that downstream clinical and public health analyses remain robust and more generalizable to a larger population. With the theoretical guarantees, simulation validation, and real data analysis, our method demonstrates superiority in data imputation

during integrated analyses, particularly in multi-source EHR studies where traditional imputation methods struggle. Its resilience to data heterogeneity further ensures that population-level health patterns can be studied with greater precision, facilitating better clinical decision-making and more effective public health interventions.

Our work has several limitations that point to directions for future research. First, MACOMSS is designed to treat different data types in a uniform way and does not explicitly accommodate binary or mixed-type variables; in such cases, techniques such as logistic PCA (Lee et al., 2010) or mixed-type factor models (Liu et al., 2023) can be considered either as preprocessing steps or in conjunction with our approach. Second, the current formulation does not exploit temporal smoothness or time-indexed missing patterns that may be present in EHR datasets, which could lead to a loss of estimation efficiency in longitudinal or time series settings. To address this, one could extend MACOMSS with time-based regularization or penalty terms on the singular value decomposition, similar to recent works (Han et al., 2024; Tan et al., 2024b). Third, while we address missing-at-random (MAR)-like missingness, missing-not-at-random (MNAR) remains a challenging scenario for EHR data imputation. In Part E.1 of the Supplementary Materials, we provide a sensitivity analysis to explore robustness under the MNAR mechanism, but further methodological development is needed for principled handling of such patterns.

Overall, MACOMSS inherits challenges common to large-scale, heterogeneous EHR datasets, where diverse data sources, variable quality, and complex missingness patterns can complicate integrated analysis. To apply MACOMSS or other imputation approaches effectively, we recommend a thorough examination of missingness patterns and an assessment of potential population biases prior to EHR analysis, along with applying appropriate transformations for variable types when needed. In parallel, it is essential to validate imputations through downstream predictive performance and evaluations of clinical plausibility, ideally with guidance from domain experts to ensure both statistical rigor and real-world applicability.

6 Conclusion

MACOMSS provides an easy-to-implement and theoretically guaranteed approach for imputing and denoising a matrix integrated from multiple EHR data sources. Our approach effectively leverages the complicated missing mechanisms during imputation, addressing

key challenges where missingness may occur randomly, structurally, and heterogeneously in the data integration processes. These advantages highlight the potential of our method for accurate downstream clinical prediction and precise clinical insights in population health.

Data and Code Availability

The codes to implement MACOMSS are publicly available at https://github.com/Tanjianbin/Macomss.

CRediT Authorship Contribution Statement

Jianbin Tan: Writing – review & editing, Writing – original draft, Software, Investigation, Formal analysis. Yan Zhang: Writing – review & editing, Investigation, Formal analysis, Software, Data curation; **Chuan Hong**: Data curation, Supervision, Writing – review & editing. **T. Tony Cai**: Writing – review & editing, Conceptualization, Supervision. **Tianxi Cai**: Writing – review & editing, Conceptualization, Supervision. **Anru R. Zhang**: Methodology, Writing – review & editing, Writing – original draft, Supervision, Software, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

Funding Information

This work was supported in part by the National Institutes of Health Grant R01HL169347.

References

Ahuja, Y., Zou, Y., Verma, A., Buckeridge, D., and Li, Y. (2022). Mixehr-guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. *Journal of Biomedical Informatics*, 134:104190.

Aidos, H. and Tomás, P. (2021). Neighborhood-aware autoencoder for missing value imputation. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1542–1546. IEEE.

Beard, L., Schein, R., Morra, D., Wilson, K., and Keelan, J. (2012). The challenges in making electronic health records accessible to patients. *Journal of the American Medical Informatics Association*, 19(1):116–120.

Beaulieu-Jones, B. K., Lavage, D. R., Snyder, J. W., Moore, J. H., Pendergrass, S. A., and Bauer, C. R. (2018). Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Medical Informatics*, 6(1):e8960.

Beaulieu-Jones, B. K., Moore, J. H., and CONSORTIUM, P. R. O.-A. A. C. T. (2017). Missing data imputation in the electronic health record using deeply learned autoencoders. In *Pacific symposium on biocomputing 2017*, pages 207–218. World Scientific.

Bower, J. K., Patel, S., Rudy, J. E., and Felix, A. S. (2017). Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. *Current Epidemiology Reports*, 4:346–352.

Brink, A., Alsma, J., van Attekum, L. A., Bramer, W. M., Zietse, R., Lingsma, H., and Schuit, S. C. (2022). Predicting inhospital admission at the emergency department: a systematic review. *Emergency Medicine Journal*, 39(3):191–198.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.

Cho, J., Kim, D., and Rohe, K. (2017). Asymptotic theory for estimating the singular vectors and values of a partially-observed low rank matrix with noise. *Statistica Sinica*, pages 1921–1948.

Essén, A., Scandurra, I., Gerrits, R., Humphrey, G., Johansen, M. A., Kierkegaard, P., Koskinen, J., Liaw, S.-T., Odeh, S., Ross, P., et al. (2018). Patient access to electronic

health records: differences across ten countries. *Health Policy and Technology*, 7(1):44–56.

Evans, R. S. (2016). Electronic health records: then, now, and in the future. *Yearbook of Medical Informatics*, 25(S 01):S48–S61.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

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

Han, R., Shi, P., and Zhang, A. R. (2024). Guaranteed functional tensor singular value decomposition. *Journal of the American Statistical Association*, 119(546):995–1007.

Haneuse, S., Arterburn, D., and Daniels, M. J. (2021). Assessing missing data assumptions in ehr-based studies: a complex and underappreciated task. *JAMA Network Open*, 4(2):e210184–e210184.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., and Botstein, D. (2001). Imputing missing data for gene expression arrays. *Technical Report, Stanford Statistics Department*, 1.

Hemingway, H., Asselbergs, F. W., Danesh, J., Dobson, R., Maniadakis, N., Maggioni, A., Van Thiel, G. J., Cronin, M., Brobert, G., Vardas, P., et al. (2018). Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *European Heart Journal*, 39(16):1481–1495.

Huang, J. and Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310.

Hurst, J. H., Liu, Y., Maxson, P. J., Permar, S. R., Boulware, L. E., and Goldstein, B. A. (2021). Development of an electronic health records datamart to support clinical and population health research. *Journal of Clinical and Translational Science*, 5(1):e13.

Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., and Mark, R. (2024). Mimic-iv (version 3.0). *PhysioNet*, RRID:SCR 007345.

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078.

Keshta, I. and Odeh, A. (2021). Security and privacy of electronic health records: Concerns and challenges. *Egyptian Informatics Journal*, 22(2):177–183.

Krzywinski, M. and Altman, N. (2017). Classification and regression trees. *Nature Methods*, 14(8):757–758.

Lee, S., Huang, J. Z., and Hu, J. (2010). Sparse logistic principal components analysis for binary data. *The Annals of Applied Statistics*, 4(3):1579.

Li, J., Yan, X. S., Chaudhary, D., Avula, V., Mudiganti, S., Husby, H., Shahjouei, S., Afshar, A., Stewart, W. F., Yeasin, M., et al. (2021). Imputation of missing values for electronic health record laboratory data. *NPJ digital medicine*, 4(1):147.

Li, Y., Yang, A. Y., Marelli, A., and Li, Y. (2024). Mixehr-surg: A joint proportional hazard and guided topic model for inferring mortality-associated topics from electronic health records. *Journal of Biomedical Informatics*, 153:104638.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, W., Lin, H., Zheng, S., and Liu, J. (2023). Generalized factor model for ultra-high dimensional correlated variables with mixed types. *Journal of the American Statistical Association*, 118(542):1385–1401.

Lucke, J. A., de Gelder, J., Clarijs, F., Heringhaus, C., de Craen, A. J., Fogteloo, A. J., Blauw, G. J., de Groot, B., and Mooijaart, S. P. (2018). Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years. *Emergency Medicine Journal*, 35(1):18–27.

Luo, F., Tan, J., Zhang, D., Huang, H., and Shen, Y. (2025). Functional clustering for longitudinal associations between social determinants of health and stroke mortality in the us. *The Annals of Applied Statistics*, 19(1):798–820.

Madden, J. M., Lakoma, M. D., Rusinak, D., Lu, C. Y., and Soumerai, S. B. (2016). Missing clinical and behavioral health data in a large electronic health record (EHR) system. *Journal of the American Medical Informatics Association*, 23(6):1143–1149.

Morris, T. P., White, I. R., and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1):75.

Nazabal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. (2020). Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501.

Pathak, A., Batra, S., and Chaudhary, H. (2022). Imputing missing data in electronic health records. In *Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication: MARC 2021*, pages 621–628. Springer.

Psychogyios, K., Ilias, L., Ntanos, C., and Askounis, D. (2023). Missing value imputation methods for electronic health records. *IEEE Access*, 11:21562–21574.

Tan, J., Ge, Y., Martinez, L., Sun, J., Li, C., Westbrook, A., Chen, E., Pan, J., Li, Y., Cheng, W., et al. (2022). Transmission roles of symptomatic and asymptomatic covid-19 cases: a modelling study. *Epidemiology & Infection*, 150:e171.

Tan, J., Liang, D., Guan, Y., and Huang, H. (2024a). Graphical principal component analysis of multivariate functional time series. *Journal of the American Statistical Association*, pages 1–24.

Tan, J., Shi, P., and Zhang, A. R. (2024b). Functional singular value decomposition. *arXiv preprint arXiv:2410.03619*.

Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., and Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549.

Tertulino, R., Antunes, N., and Morais, H. (2024). Privacy in electronic health records: a systematic mapping study. *Journal of Public Health*, 32(3):435–454.

Tian, M., Chen, B., Guo, A., Jiang, S., and Zhang, A. R. (2024). Reliable generation of privacy-preserving synthetic electronic health record time series via diffusion models. *Journal of the American Medical Informatics Association*, 31(11):2529–2539.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Wallace, E., Stuart, E., Vaughan, N., Bennett, K., Fahey, T., and Smith, S. M. (2014). Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. *Medical Care*, 52(8):751–765.

Wang, Y., Chen, R., Ghosh, J., Denny, J. C., Kho, A., Chen, Y., Malin, B. A., and Sun, J. (2015). Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1265–1274.

Wells, B. J., Chagin, K. M., Nowacki, A. S., and Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3).

Yoon, J., Zhang, Y., Jordon, J., and Van der Schaar, M. (2020). Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043.

You, J., Ma, X., Ding, Y., Kochenderfer, M. J., and Leskovec, J. (2020). Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*, 33:19075–19087.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.