scientific reports

Check for updates

OPEN

# Imputation methods for serologic biomarkers in inflammatory bowel disease

Miad Boodaghidizaji, Dermot P. B. McGovern & Dalin Li✉

**Serologic biomarkers have emerged as a powerful tool for the diagnosis of Inflammatory Bowel Disease (IBD) and the differentiation between subgroups of IBD. However, missingness in serologic data can adversely affect the efficacy of any form of statistical or machine learning analysis, leading to biased predictions. This paper provides a thorough comparison of multiple imputation models that can be used for the imputation of serologic data under different missingness scenarios. All major forms of missingness, including Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), were explored in relation to the serologic data. The imputation models used in this study encompass Multiple Imputation (MI) using Chained Equations (MICE), Iterative Imputer (II), and Autoencoders (AE). Across three real IBD cohorts and 2,400 simulated scenarios spanning MCAR/MAR/MNAR and 5–40% missingness, we evaluated imputers on direct accuracy, inferential signal, and predictive utility. No single method is universally optimal: iterative imputers (II-BR/KNN/RF) tend to lead at low–moderate missingness, whereas autoencoder-based (AE/VAE) approaches are more robust as missingness increases; all analyses are performed within-cohort to avoid information leakage.**

**Keywords** Serologic data, Imputation, IBD, MAR, MCAR, MNAR, MICE, Iterative Imputer, Autoencoder

Inflammatory Bowel Disease (IBD), including Crohn's Disease (CD) and Ulcerative Colitis (UC), represents a group of idiopathic disorders characterized by chronic inflammation of different regions of the digestive tract[1]. The precise etiology of IBD remains unclear, but serological markers such as antibodies against the yeast *Saccharomyces cerevisiae* (ASCA), *Escherichia coli* outer membrane porin C (Omp-C), flagellin (cBir1), anti-neutrophil cytoplasmic antibodies (ANCA) and *Pseudomonas fluorescens*–associated sequence I-2 (I2) have emerged as crucial tools for diagnosis, monitoring disease activity, and predicting therapeutic responses[2–4]. Serology in IBD research provides invaluable insights into the immune response dynamics that underpin these disorders, contributing to a more tailored therapeutic approach. Serological antibodies targeting various antigens have proven useful in distinguishing subjects with IBD from those without IBD, as well as in differentiating between individuals with CD and those with UC[5]. Additionally, studies have confirmed the potential of serological markers in predicting IBD even years before diagnosis[6].

Even though a variety of technologies, such as proteomics and multiplex enzyme-linked immunosorbent assay (ELISA), are available to measure serological markers[7], the reliability of serological studies in IBD can be compromised by the presence of missing data. In general, data gaps in clinical studies can result from various sources, including patient non-compliance, logistical challenges in sample collection, and loss to follow-up[8,9]. The presence of missingness, irrespective of the source, may skew the interpretation of the serological landscape in IBD. Traditional methods of managing missing data, such as listwise deletion, are frequently inadequate. These methods can introduce significant biases and distort the analysis, potentially leading to incorrect conclusions about disease mechanisms or efficacy of treatments.

In response to the challenges with missing data, over the years, a myriad of imputation techniques has been developed to address missingness in clinical data, including simple imputation techniques such as mean imputation[10] or more sophisticated techniques such as deep learning models[11]. There is a growing recognition of the potential of machine learning (ML) models to provide sophisticated solutions for imputation that go beyond the limitations of traditional techniques. ML models, such as Random Forest (RF)[12], Neural Networks (NN)[13,14], and Deep Learning (DL)[15] models, utilize complex algorithms to detect patterns in the data, offering a more nuanced approach to handling missing values. These methods can improve the prediction of missing values by learning from the relationships present in the data, potentially better preserving the integrity of serological

F. Widjaja Inflammatory Bowel Disease Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. ✉email: dalin.li@cshs.org

analyses in IBD research. However, their performance and validity depend critically on how and why data are missing.

Given the wide range of available algorithms, it is unclear which imputation technique should be applied specifically to serological data in IBD studies[16]. Consequently, this paper aims to explore and compare innovative ML-based imputation methods applied to serological data in IBD. The focus will be on assessing the performance of different imputation algorithms in a variety of scenarios of missingness and how important imputation can be as a preprocessing step in IBD research when it comes to serologic data. Our goal is practical, task-matched guidance. Because biomedical pipelines pursue two distinct downstream aims—statistical inference and prediction—we evaluate imputers along three axes: direct pointwise accuracy, indirect-inferential performance, and indirect-predictive performance. Finally, we report best methods separately by axis, rather than claiming a universal winner. For ease of reference, a consolidated list of abbreviations used throughout the manuscript is provided in Supplementary Table S0.

## Methods

There are two main approaches to dealing with missing data: ignoring the missing values (complete case analysis, commonly referred to as listwise deletion) and imputing the missing values[17]. Complete case analysis is inefficient due to the loss of sample size and can also introduce bias when the excluded observations differ systematically from the included ones[18]. Thus, complete case analysis is only suitable when the missing data is minimal. To overcome the limitations of removing missing values, one can replace them with plausible values, a process known as imputation[8]. Numerous imputation techniques are available to handle missing data. However, the nature of missingness can strongly influence the capability of different imputers. Hence, it is crucial to first identify the nature of the missing data. In the following sections, we introduce different types of missingness and the imputation techniques available for handling them.

### Missingness types

According to the classification proposed by Rubin[19], missing data can be categorized into three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). MCAR represents a pattern of missing values that is entirely random and unrelated to any variables, whether they are included in the analysis or not.

On the other hand, MAR occurs when the probability of missing data is related to the observed data in the dataset but not to the unobserved data. Finally, MNAR is the scenario in which the missingness is related to the unobserved data rather than the observed data.

In the current study, we systematically simulate all three types of missingness in real-life serological biomarker datasets. Figure S1 illustrates the simulation procedure employed to create these missingness patterns. To generate MCAR, we first compute the total number of entries to be set as missing $N = \lfloor r \cdot n \cdot f \rfloor$ , where $r$ is the target missingness rate and $n$ and $f$ are the numbers of samples and features. We then randomly select $N$ distinct cell positions from the full data matrix and replace those values with Not a Number (NaN). This induces cell-level MCAR missingness, where the missingness rate refers to the proportion of all data entries and not rows.

To generate MAR, a propensity score–based approach is utilized. In brief, non-missing and missing points are treated as 0s and 1s, respectively. We assume each feature (serologic parameter) is a function of other features. Logistic regression is then employed to model the missingness in the feature of interest using other features, which is in turn used to generate the propensity score for missingness. MAR missingness is then generated by sampling missingness indicators according to the fitted propensity scores. To train these logistic regression models, missingness indicators were defined for each feature using the original datasets. Missing values among predictor features were imputed with the median to allow model fitting. The resulting coefficients and intercepts were then applied to the complete ground-truth datasets to generate MAR missingness patterns. The whole process is as follows: let $i = 1,\ldots,n$ index samples and $j = 1,\ldots,f$ index features. For each feature $X_j$, we model the missingness indicator $R_{ij} \in \{0,1\}$ (1 = missing) using a logistic regression that depends only on other features:

$$\text{logit Pr}\,(R_{ij} = 1?X_{i,-j}) = \alpha_{\,j} + X_{i,-j}^{T}\gamma_{\,j}$$

Here, $\alpha_j$ is the intercept, $\gamma_j$ the coefficient vector for the predictors $X_{i,-j}$ (all predictors except $j$), $n$ is the number of samples, and $f$ is the number of features. To fit this model, we create a one-time "baseline" dataset where predictors $X_{i,-j}$ have been single-imputed by their feature-wise medians. We compute propensities $\widehat{\pi}^{\,ij}$ for all $i$, choose a threshold $c_j$ so that exactly $N$ samples (with the largest $\widehat{\pi}^{\,ij}$) satisfy $\widehat{\pi}^{\,ij} \geq c_j$, and set $R_{ij} = 1$ for those samples. The target proportion $r \in \{0.05,0.10,\ldots,0.40\}$ is the desired missingness level per feature. This construction makes $R_{ij}$ depend only on observed covariates (MAR).

To generate MNAR, we assume that most missing values occur at the extremes of the data distribution. We select the majority of indices ($N_{ma}$) from these extremes, defined by the low and high quantiles for the given missing rate. Additionally, a buffer zone between these quantiles is defined to select a minority of indices ($N_{mi}$=$N$- $N_{ma}$), introducing randomness.

These procedures are repeated 100 times for each type of missingness and for missing rates ranging from 0.05 to 0.4, creating 2400 missing scenarios. For each mechanism (MCAR/MAR/MNAR) and rate (5–40%), we induced cell-level missingness on the same fixed subject set (no rows removed); thus, sample size is constant across missingness levels. We emphasize that MNAR mechanisms are fundamentally difficult to handle. This is because different joint models for the data and missingness indicators can induce the same observed-data distribution but imply different completed datasets, so MNAR is not identifiable from the observed data alone without additional assumptions. Our MNAR generator represents one plausible mechanism motivated by clinical intuition but does not cover the general MNAR case. Accordingly, the MNAR analyses should be

viewed as mechanism-specific comparisons among imputers rather than evidence that any method can broadly eliminate MNAR-related bias.

## Data

In this study, we employ three distinct real-world numerical data cohorts to simulate missingness and apply various imputation techniques, as detailed in Table S1. The cohorts include IBD cohorts from the Cedars-Sinai IBD Biobank (MIRIAD, size = 9231), the Risk Stratification and Identification of Immunogenetic and Microbial Markers of Rapid Disease Progression in Children with Crohn's Disease (RISK, size = 431), and the Predictors of Response to Standardized Pediatric Colitis Therapy (PROTECT, size = 1102). Phenotype composition differs across cohorts: RISK contains only CD cases, PROTECT contains only UC, while MIRIAD includes CD, UC, and non-IBD. The original datasets naturally contain missing values. To simulate missingness effectively, we first removed rows with missing values to create complete datasets, which served as ground truth for evaluating imputation performance. However, for MAR simulations, logistic regression models were trained on the original datasets that contained missing values. The fitted coefficients and intercepts were then applied to the complete ground-truth datasets to calculate missingness propensity scores and generate MAR missingness patterns, as outlined in "Missingness types".

Because cohorts differ in phenotype composition and possibly assay characteristics, we summarized and tested baseline feature distributions across cohorts. Supplementary Figure S2 shows cohort-wise boxplots for the five features shared by all cohorts (Anti-I2 shown separately for MIRIAD only). Global differences were assessed with the Kruskal–Wallis (KW) test. P-values were adjusted using the Benjamini–Hochberg (BH) procedure to control the false discovery rate (FDR); BH-adjusted P-values ("q-values") are reported in Table S2. To separate case-mix from cohort effects, we also conducted phenotype-matched pairwise comparisons—MIRIAD-CD vs. RISK-CD and MIRIAD-UC vs. PROTECT-UC—using the Mann–Whitney U (MWU) test with the same BH/FDR adjustment (Supplementary Figure S3, Table S3). Because global differences were statistically significant across cohorts for all shared markers (Kruskal–Wallis test, BH-adjusted q-values shown in Table S2), we adopted a strict within-cohort design—training MAR models, performing imputation, and evaluating performance separately in each cohort—and compared only the resulting performance summaries across cohorts to avoid any information leakage.

Furthermore, by creating a complete dataset, we ensure a robust baseline against which various imputation methods can be evaluated. With this baseline ground truth, we apply direct assessment of imputation performance across all three data types. However, to carry out an indirect assessment of imputation performance, where phenotype information is necessary, we rely exclusively on the MIRIAD dataset, as it is the only dataset with comprehensive clinical phenotype information. In all three cohorts, the input serologic markers including Anti-neutrophil cytoplasmic antibodies (ANCA), anti-Saccharomyces cerevisiae antibodies (ASCA, both IgG and IgA), anti-outer membrane protein C (anti-OmpC), anti-CBir1 Flagellin antibodies (anti-CBir1), and anti-Pseudomonas fluorescens-associated sequence I2 (anti-I2) are selected as input features, as they are highly associated with IBD[20–22].

## Imputation techniques

Data imputation techniques in general are classified as single and multiple imputation techniques[23–25]. Single imputation involves replacing each missing value with a single, specific value, while multiple imputation involves replacing each missing value with several different plausible values, reflecting the uncertainty about what the true value might be. Both methods aim to create a complete dataset that can be analyzed using standard statistical techniques.

In the current study, we focus on multiple imputation by employing different ML models. We use Multiple Imputation by Chained Equations (MICE) imputer[26] in R and the Iterative Imputer (II) of scikit-learn[27] in Python with different common ML models. Additionally, we examine the performance of Autoencoders (AEs) which are based on NNs using PyTorch[28] in Python. MICE imputer utilizes chained equations, involving a series of regression models to impute missing values. For each variable with missing data, a regression model is fitted using the other variables as predictors, iteratively cycling through each variable until convergence is reached. Inspired by the R MICE package, the II in scikit-learn models each feature with missing values as a function of other features in a round-robin fashion. During each iteration, a regression model is trained on the non-missing values, predicting and imputing the missing values until convergence. II can be used as a multiple imputation technique by applying it repeatedly to the same dataset with different random seeds. Both MICE imputer and II can be used with different ML models. Table S4 enlists all the imputers we use in this study.

AEs are a type of NNs designed for unsupervised learning that aim to learn efficient representations of data, typically for dimensionality reduction or feature extraction. They consist of an encoder that maps the input data to a latent space and a decoder that reconstructs the input from this latent representation. Variational Autoencoders (VAEs) extend this concept by introducing a probabilistic approach to the latent space. Instead of mapping inputs to fixed points, VAEs map them to distributions, allowing for the generation of new data samples by sampling from these distributions. This makes VAEs particularly useful for generative tasks and learning more meaningful, continuous latent spaces.

For each method and scenario, we generated $m = 5$ completed datasets to capture imputation variability; while larger $m$ is often recommended in traditional MI settings, $m = 5$ provides a practical balance given the large number of simulated scenarios. MICE has been extensively studied and shown to perform well across a wide range of settings. II (BR/RF/KNN) and AE&VAE produced multiple stochastic draws via resampling/restarts (a pseudo-MI strategy). Further, II (BR) used posterior sampling when available, while VAE used latent sampling. For inferential evaluations we combined estimates across imputations using Rubin's rules. We note in limitations that non-MICE methods do not correspond to draws from a formally specified joint model. Therefore, their

inference results should be interpreted with this caveat, while remaining useful for practice where these methods are widely used. All model abbreviations (e.g., AE, VAE, II (BR/RF/KNN)) are defined at first mention and used consistently throughout the manuscript.

## Assessment techniques

We evaluated each imputation method with one direct accuracy metric and two indirect downstream metrics that reflect inferential and predictive utility. All evaluations are performed within cohort; only the resulting summaries are compared across cohorts to avoid information leakage. For every missingness type and rate, we generated $m = 5$ multiply imputed datasets and repeated the simulation $R = 100$ times. To summarize across imputations and repeats, we report a conservative adjusted value for each metric *Met*:

$$\text{Adjusted}_{\lambda}\ (Met) = \begin{cases} \mu\ (Met) + \lambda\, \sigma\ (Met), & if\ smaller\ is\ better \\ \mu\ (Met) - \lambda\, \sigma\ (Met), & if\ larger\ is\ better \end{cases} \tag{1}$$

where $\mu$ and $\sigma$ are the mean and standard deviation across imputations/repeats and $\lambda = 0.5$ (pre-specified) provides a modest penalty on instability. This quantity is a ranking summary, not a confidence interval, and is not used for hypothesis testing. For inferential analyses, we first combine estimates across imputations using Rubin's rules and then apply the summary above to penalize run-to-run variability.

*Direct assessment*:

For each feature and scenario, we compute the normalized root mean square deviation (NRMSD) between imputed values and the ground-truth values from the complete dataset:

$$\text{NRMSD} = \sqrt{\frac{\sum_{i=1}^{N_i}\left(x_i^{imp} - x_i^{true}\right)^2}{N[x_{max} - x_{min}]^2}}, \tag{2}$$

where $N$ is the number of imputed entries for that feature, and $x_{max}$ and $x_{min}$ are computed within cohort and feature from the ground-truth dataset. We aggregate NRMSD across imputations and repeats using the adjusted rule above (mean + 0.5·SD). For each scenario, the best imputer is the one with the smallest adjusted NRMSD.

*Indirect assessment*:

(A) Inferential signal via adjusted $-\log_{10}(\text{P})$

For each cohort, feature, and phenotype contrast (CD vs. non-IBD, UC vs. non-IBD, CD vs. UC), we fit univariate logistic regression of phenotype on the feature in each of the $m = 5$ imputed datasets. Let $\widehat{\beta}_k$ and $U_k$ denote the coefficient and its variance in imputed dataset $k$ ($k = 1,\ldots,m$). We combine estimates using Rubin's rules:

$$\bar{\widehat{\beta}} = \frac{1}{m}\sum_{k=1}^{m}\widehat{\beta}_k,\ \ W = \frac{1}{m}\sum_{k=1}^{m}U_k,\ \ \ B = \frac{1}{m-1}\sum_{k=1}^{m}\left(\widehat{\beta}_k - \bar{\widehat{\beta}}\right)^2,\ \ T = W + \left(1 + \frac{1}{m}\right) \tag{3}$$

Here, $W$ is the within-imputation variance, $B$ is the between-imputation variance, and $T$ is the total variance. We form $Z = \bar{\beta}\ /\sqrt{T}$ and compute a two-sided P-value (P) from the large-sample normal distribution where Barnard–Rubin t-approximation yields indistinguishable results at our sample sizes. We summarize inferential signal as

$$\text{S} = 1 - \log_{10}(\text{P}) \tag{4}$$

which is a monotone transform of P that preserves the ordering of evidence and remains well-behaved for very small P-values. We then aggregate to an adjusted $-\log_{10}(\text{P})$ using the mean − 0.5·SD rule. For each scenario, we compare the adjusted value to that from the original (complete) data; the best imputer is the one whose adjusted $-\log_{10}(\text{P})$ is closest to the original (smallest absolute difference).

(B) Predictive utility via adjusted area under the receiver operating characteristic curve (AUC)

Within cohort, we train a Gradient Boosting classifier with stratified 5-fold cross-validation using identical folds across imputers and the original data for each contrast (CD vs. non-IBD, UC vs. non-IBD, CD vs. UC). Imputation was performed prior to cross-validation splitting and was fully unsupervised, i.e., phenotype labels were not used during imputation. Therefore, the predictive analysis reflects comparative benchmarking of imputers under a shared preprocessing scheme rather than a deployment-style pipeline with fold-specific imputation. For each run we compute the AUC,

$$\text{AUC} = \int_0^1 \text{TPR}\,(t)\ d\text{FPR}\,(t) = \Pr\{s\left(X^+\right) > s\left(X^-\right)\} \tag{5}$$

i.e., the probability that a randomly chosen positive receives a higher model score ($s$) than a randomly chosen negative (Wilcoxon/Mann–Whitney interpretation). We aggregate to an adjusted AUC using the mean − 0.5·SD

rule. For each scenario, the best imputer is the one whose adjusted AUC is closest to the original (smallest absolute difference).

## Results

Here, we present the results for direct and indirect analyses of the imputation.

### Direct assessment

The adjusted NRMSD values for all imputers across missingness types and percentages are shown in Fig. 1. A lower NRMSD indicates better performance. Visual inspection confirms that iterative imputers (II) and autoencoder-based models (AE, VAE) generally outperform classical MICE variants, particularly as the percentage of missingness increases. Details of the adjusted NRMSD for the best imputers are listed in Table S5. To systematically summarize these patterns, Table 1 reports the top-performing imputers across low (≤ 10%), medium (10–30%), and high (> 30%) missingness levels. At low missingness, II methods such as II (BR), II (KNN), and II (RF) consistently achieve the best accuracy across MAR, MCAR, and MNAR settings, highlighting their ability to leverage observed correlations when relatively few entries are absent. At medium missingness, performance becomes more diverse: AE and VAE frequently emerge as top models, while II (BR) and II (RF) remain strong competitors, suggesting that both deep-learning and ensemble-based iterative approaches can effectively capture nonlinear relationships when data gaps become more substantial. At high missingness, AE clearly emerges as the most robust imputer across nearly all scenarios, with II (KNN) occasionally matching its performance. This indicates that neural-network–based imputers are better suited to preserving structure in heavily incomplete datasets. Overall, no single model dominates in all conditions; however, the II family (particularly BR and RF) excels when missingness is low to moderate, while AE demonstrates superior robustness as missingness grows.

### Statistical analysis assessment

We evaluated imputers based on how closely their adjusted $-\log_{10}(P)$ approximated those obtained from the original data. For each scenario, the model with the smallest deviation from the original $-\log_{10}(P)$ was identified as the best performer. The complete breakdown for CD vs. non-IBD is provided in Table S6, and summary counts of best-performing models across missingness degrees and mechanisms are presented in Table 2. As shown, no single imputer consistently dominated across all scenarios. Instead, IIs, particularly II (KNN), II

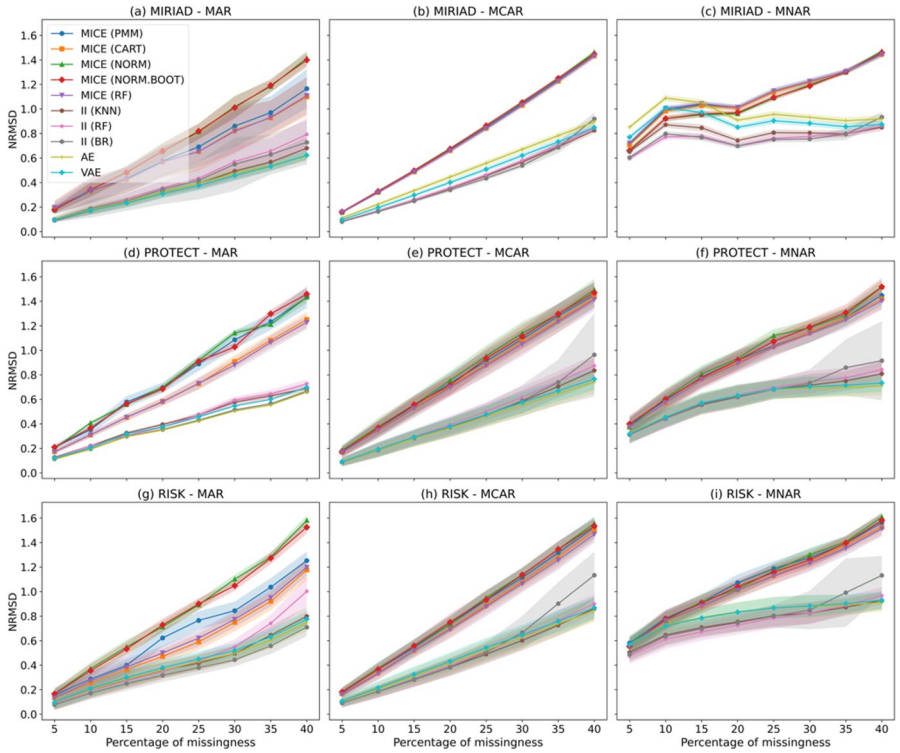


**Fig. 1**. NRMSD as a function of percentage of missingness for the MIRIAD (**a**–**c**), PROTECT (**d**–**f**), and RISK (**g**–**i**) cohorts across different imputers and missingness mechanisms (MAR, MCAR, MNAR). Lower NRMSD indicates better performance. Iterative imputers (II) generally perform best under low missingness, while autoencoder-based models (AE, VAE) become increasingly competitive as missingness rises. At high missingness levels, AE consistently demonstrates the strongest robustness across cohorts and missingness types. Ribbons show ± 1 SD across imputations/repeats. Missingness mechanisms are presented in alphabetical order (MAR, MCAR, MNAR) for consistency across figures and text.

| Missingness degree | Total count | | | | | |
|---|---|---|---|---|---|---|
| | MAR | | MCAR | | MNAR | |
| Low-missingness | II (BR) | 2 | II (BR) | 2 | II (RF) | 6 |
| | II (KNN) | 2 | II (KNN) | 2 | II (KNN) | 0 |
| | AE | 2 | II (RF) | 2 | II (RF) | 0 |
| | II (RF) | 0 | VAE | 0 | VAE | 0 |
| | VAE | 0 | AE | 0 | AE | 0 |
| | M (CART) | 0 | M (CART) | 0 | M (CART) | 0 |
| | M (NORM) | 0 | M (NORM) | 0 | M (NORM) | 0 |
| | M (N. BOOT) | 0 | M (N. BOOT) | 0 | M (N. BOOT) | 0 |
| | M (RF) | 0 | M (RF) | 0 | M (RF) | 0 |
| | M (PMM) | 0 | M (PMM) | 0 | M (PMM) | 0 |
| | COMPLETE | 0 | COMPLETE | 0 | COMPLETE | 0 |
| Top models | II (BR); II (KNN); AE | | II (BR); II (KNN); II (RF) | | II (RF) | |
| Medium-missingness | AE | 4 | II (KNN) | 5 | II (RF) | 4 |
| | II (BR) | 4 | II (BR) | 4 | II (BR) | 4 |
| | VAE | 4 | AE | 3 | II (KNN) | 3 |
| | II (KNN) | 0 | II (RF) | 0 | AE | 1 |
| | II (RF) | 0 | VAE | 0 | VAE | 0 |
| | M (CART) | 0 | M (CART) | 0 | M (CART) | 0 |
| | M (NORM) | 0 | M (NORM) | 0 | M (NORM) | 0 |
| | M (N. BOOT) | 0 | M (N. BOOT) | 0 | M (N. BOOT) | 0 |
| | M (RF) | 0 | M (RF) | 0 | M (RF) | 0 |
| | M (PMM) | 0 | M (PMM) | 0 | M (PMM) | 0 |
| | COMPLETE | 0 | COMPLETE | 0 | COMPLETE | 0 |
| Top models | AE; VAE; II (BR) | | II (KNN) | | II (RF); II (BR) | |
| High-missingness | AE | 3 | II (KNN) | 3 | AE | 3 |
| | II (BR) | 2 | AE | 3 | II (KNN) | 2 |
| | VAE | 1 | II (RF) | 0 | II (RF) | 1 |
| | II (RF) | 0 | II (BR) | 0 | VAE | 0 |
| | II (KNN) | 0 | VAE | 0 | II (BR) | 0 |
| | M (CART) | 0 | M (CART) | 0 | M (CART) | 0 |
| | M (NORM) | 0 | M (NORM) | 0 | M (NORM) | 0 |
| | M (N. BOOT) | 0 | M (N. BOOT) | 0 | M (N. BOOT) | 0 |
| | M (RF) | 0 | M (RF) | 0 | M (RF) | 0 |
| | M (PMM) | 0 | M (PMM) | 0 | M (PMM) | 0 |
| | COMPLETE | 0 | COMPLETE | 0 | COMPLETE | 0 |
| Top models | AE | | II (KNN); AE | | AE | |

**Table 1**. Counts of best-performing imputers (lowest adjusted NRMSD) across three missingness degrees (low, medium, high) and three missingness mechanisms (MAR, MCAR, MNAR), aggregated over all cohorts. For each setting, the top models are listed. Consistent with Fig. 1, IIs dominate under low missingness, a mix of II and AE/VAE perform best under medium missingness, and AE consistently outperforms others under high missingness. Abbreviations: M (PMM, CART, RF, NORM, NORM.BOOT) = R MICE imputer; II = scikit-learn Iterative Imputer with the indicated base model (KNN, RF, BR); AE/VAE = neural network imputers.

(RF), and II (BR), along with AE and VAE, frequently outperformed other methods. Importantly, in most cases, imputation improved agreement with the original $-\log_{10}(P)$ compared to complete-case analysis, underscoring the benefit of imputation and the loss of statistical power when discarding incomplete data.

Closer inspection reveals systematic trends. At low missingness, both II and AE/VAE methods frequently align well with the original $-\log_{10}(P)$ values, suggesting they can recover subtle association signals even with limited data loss. At medium missingness, II (BR) and II (KNN) emerge as the most stable across cohorts, while VAE provides occasional advantages by capturing nonlinear relationships. At high missingness, AE and VAE show greater robustness than traditional methods, reflecting the strength of deep-learning approaches in preserving association structure under more severe data degradation.

The same evaluation criteria were applied to UC vs. non-IBD and CD vs. UC comparisons. Full results for these analyses are provided in Supplementary Tables S7–S10 and corresponding forest plots in Figures S5–S147. Similar patterns emerged: II (KNN), II (BR), and VAE were most effective in UC vs. non-IBD, while II (KNN), II (RF), and II (BR) were most effective in CD vs. UC. Taken together, these findings demonstrate that while

| Missingness degree | Total count | | | | | |
|---|---|---|---|---|---|---|
| | MAR | | MCAR | | MNAR | |
| Low-missingness | II (RF) | 4 | II (KNN) | 4 | AE | 3 |
| | II (KNN) | 4 | II (BR) | 3 | VAE | 3 |
| | VAE | 2 | II (RF) | 3 | II (KNN) | 2 |
| | II (BR) | 1 | AE | 2 | II (BR) | 2 |
| | AE | 1 | VAE | 0 | II (RF) | 1 |
| | M (CART) | 0 | M (CART) | 0 | M (CART) | 1 |
| | M (NORM) | 0 | M (NORM) | 0 | M (NORM) | 0 |
| | M (N. BOOT) | 0 | M (N. BOOT) | 0 | M (N. BOOT) | 0 |
| | M (RF) | 0 | M (RF) | 0 | M (RF) | 0 |
| | M (PMM) | 0 | M (PMM) | 0 | M (PMM) | 0 |
| | COMPLETE | 0 | COMPLETE | 0 | COMPLETE | 0 |
| Top models | II (RF); II (KNN) | | II (KNN) | | VAE; AE | |
| Medium-missingness | II (BR) | 10 | II (KNN) | 12 | II (BR) | 13 |
| | II (KNN) | 7 | II (BR) | 4 | II (KNN) | 5 |
| | VAE | 4 | AE | 4 | AE | 4 |
| | II (RF) | 3 | II (RF) | 4 | II (RF) | 2 |
| | AE | 0 | VAE | 0 | VAE | 0 |
| | M (CART) | 0 | M (CART) | 0 | M (CART) | 0 |
| | M (NORM) | 0 | M (NORM) | 0 | M (NORM) | 0 |
| | M (N. BOOT) | 0 | M (N. BOOT) | 0 | M (N. BOOT) | 0 |
| | M (RF) | 0 | M (RF) | 0 | M (RF) | 0 |
| | M (PMM) | 0 | M (PMM) | 0 | M (PMM) | 0 |
| | COMPLETE | 0 | COMPLETE | 0 | COMPLETE | 0 |
| Top models | II (BR) | | II (KNN) | | II (BR) | |
| High-missingness | II (KNN) | 4 | II (KNN) | 4 | AE | 3 |
| | II (RF) | 4 | II (BR) | 3 | VAE | 3 |
| | VAE | 2 | II (RF) | 3 | II (BR) | 2 |
| | II (BR) | 1 | AE | 2 | II (KNN) | 2 |
| | AE | 1 | VAE | 0 | II (RF) | 1 |
| | M (CART) | 0 | M (CART) | 0 | M (CART) | 1 |
| | M (NORM) | 0 | M (NORM) | 0 | M (NORM) | 0 |
| | M (N. BOOT) | 0 | M (N. BOOT) | 0 | M (N. BOOT) | 0 |
| | M (RF) | 0 | M (RF) | 0 | M (RF) | 0 |
| | M (PMM) | 0 | M (PMM) | 0 | M (PMM) | 0 |
| | COMPLETE | 0 | COMPLETE | 0 | COMPLETE | 0 |
| Top models | II (KNN); II (RF) | | II (KNN) | | VAE; AE | |

**Table 2**. Counts of best-performing imputers (adjusted $-\log_{10}(P)$) across missingness degrees (low, medium, high) and mechanisms (MAR, MCAR, MNAR) for CD vs. non-IBD differentiation. For each setting, the top models are listed. Iterative imputers (II) dominate at medium levels of missingness, while both II and autoencoder-based methods (AE, VAE) frequently appear among the top models at low and high missingness. Abbreviations: M (PMM, CART, RF, NORM, NORM.BOOT) = R MICE imputer; II = scikit-learn Iterative Imputer with the indicated base model (KNN, RF, BR); AE/VAE = neural network imputers.

iterative imputers dominate at moderate levels of missingness, AE and VAE can offer more reliable performance when missingness is extensive. Detailed numerical outputs are available in the supplementary file *statistics.csv*.

To check that inferential precision behaved as expected, we tabulated the Rubin-combined standard errors of the univariate log-odds, $\sqrt{T}$ with $T = W + (1 + 1/m)\,B$ across missingness levels, mechanisms, and phenotype contrasts (Supplementary file *SE_trend_summary.csv)*. Under MCAR and MAR, standard errors generally increased with higher missingness, consistent with information loss. Under MNAR, trends were occasionally non-monotone: misspecification (or increasingly deterministic imputations) can reduce $B$ at high missingness, partially offsetting the rise in $W$, so $\sqrt{T}$ need not increase strictly. Overall, the table shows that MI-based inference loses precision as missingness grows in well-specified settings (MCAR/MAR), with MNAR producing mechanism-dependent departures.

We assessed nominal false-positive control via within-cohort label permutation. For each method and scenario (mechanism $\in$ {MCAR, MAR, MNAR}; missingness 5–40%), we permuted phenotype labels $B_{perm} = 200$ times, fit a univariate logistic model in each of the $m = 5$ imputations, combined estimates using Rubin's rules, and recorded the two-sided P-value; the empirical Type-I error is the proportion of permuted P-values $< \alpha = 0.05$.

Across phenotype contrasts and methods, rejection rates clustered near 0.05: MICE variants were mildly conservative (≈0.02–0.03), whereas iterative imputers and AE/VAE were close to nominal (≈0.04–0.05). Per-scenario values are provided in Supplementary file *type1_summary_all.csv*, and an aggregate view is shown in Supplementary Fig. S4 (bars = rates, error bars = binomial 95% CIs, dashed line = $\alpha = 0.05$). We observed no systematic inflation beyond the $\pm \approx 0.03$ binomial margin expected with $B_{prem} = 200$, indicating that the MI-combined tests are well calibrated under the null.

## Machine learning analysis assessment

We next assessed the impact of imputation on downstream predictive modeling using a GB classifier. Figure 2 shows the average AUC values across different missingness types and percentages, with the black line indicating performance using the original complete datasets. For each scenario, the best model was defined as the imputer whose adjusted AUC most closely matched that of the original data. To complement the graphical results, Table 3 lists the best-performing models and their associated adjusted AUC values, while Supplementary Tables S11–S13 summarize the counts of top imputers across scenarios.

Consistent with the direct and statistical assessments, no single imputer dominated across all settings. Distinct trends emerged by phenotype: in CD vs. non-IBD and CD vs. UC, II (BR) was consistently among the top performers at low and medium levels of missingness, whereas AE became increasingly competitive at high missingness, particularly under MCAR and MNAR. In UC vs. non-IBD, AE generally outperformed other imputers across most missingness types and degrees, while II (RF) provided advantages in MNAR at low missingness. Overall, these findings highlight that iterative imputers (especially II (BR) and II (RF)) are reliable in CD-related comparisons, while AE demonstrates greater robustness in UC-related analyses and at higher missingness levels. We note that AUC reflects predictive utility only and is not a test of inferential calibration, which we assess separately via $-\log_{10}(P)$ and the permutation Type-I error check.

Finally, to consolidate all findings, Table 4 summarizes the best-performing imputers across direct, statistical, and machine learning assessments. Across methods and phenotypes, five models consistently emerged as top performers: II (KNN), II (BR), II (RF), AE, and VAE. These results underscore that while no single method is

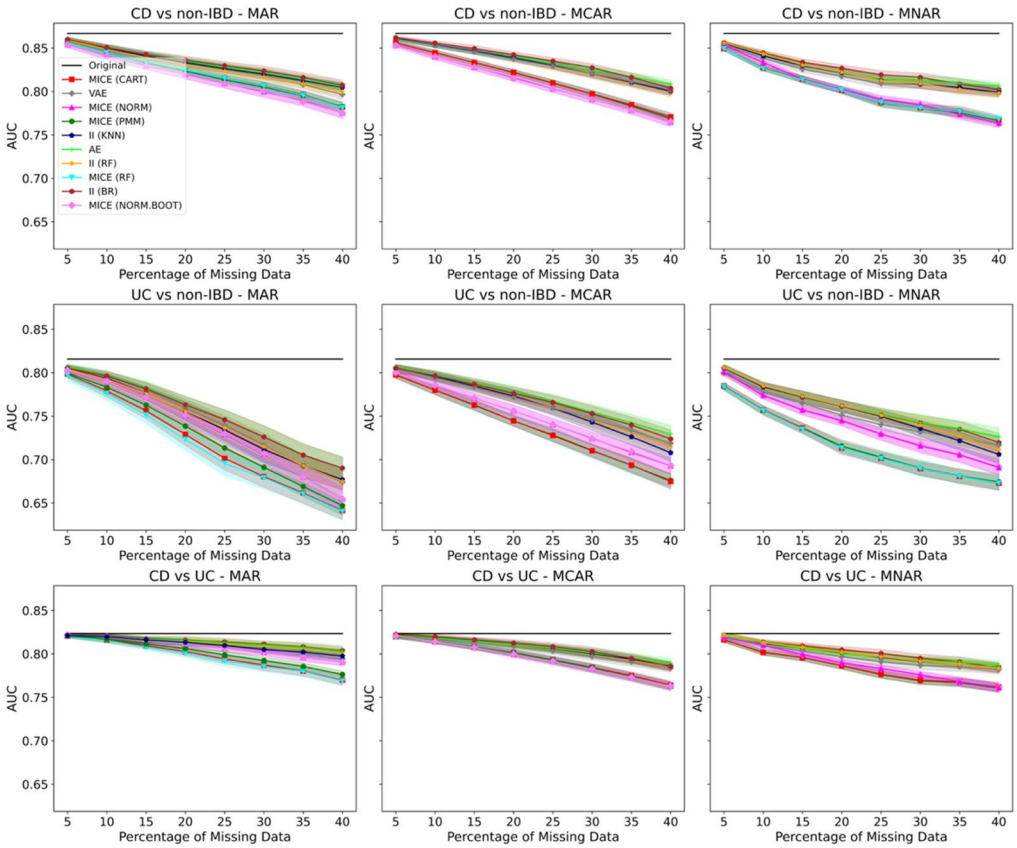


**Fig. 2**. Average AUC as a function of percentage of missingness for different imputers across three classification tasks: CD vs. non-IBD (top row), UC vs. non-IBD (middle row), and CD vs. UC (bottom row), under MAR, MCAR, and MNAR mechanisms. The black horizontal line indicates the performance using the original (complete) data. Iterative imputers (II), particularly II(BR) and II(RF), tend to maintain higher AUCs in CD-related comparisons, whereas AE consistently outperforms others in UC vs. non-IBD at moderate to high missingness. Black horizontal line: original (complete-case) AUC and ribbons show ± 1 SD across imputations/repeats.

| Percentage | MAR | | MCAR | | MNAR | |
|---|---|---|---|---|---|---|
| CD vs. non-IBD | | | | | | |
| 5 | II (BR) | 0.859 | II (BR) | 0.861 | II (BR) | 0.855 |
| 10 | II (BR) | 0.849 | II (BR) | 0.854 | II (BR) | 0.843 |
| 15 | II (BR) | 0.841 | II (BR) | 0.848 | II (BR) | 0.832 |
| 20 | II (BR) | 0.834 | II (BR) | 0.840 | II (BR) | 0.825 |
| 25 | II (BR) | 0.827 | II (BR) | 0.833 | II (BR) | 0.817 |
| 30 | II (BR) | 0.821 | II (BR) | 0.825 | II (BR) | 0.814 |
| 35 | II (BR) | 0.813 | AE | 0.815 | AE | 0.806 |
| 40 | II (BR) | 0.804 | AE | 0.806 | AE | 0.804 |
| Original data | 0.866 | | | | | |
| UC vs. non-IBD | | | | | | |
| 5 | AE | 0.805 | AE | 0.804 | II (RF) | 0.804 |
| 10 | AE | 0.795 | AE | 0.795 | II (RF) | 0.782 |
| 15 | AE | 0.785 | AE | 0.785 | II (RF) | 0.770 |
| 20 | AE | 0.775 | AE | 0.775 | II (RF) | 0.759 |
| 25 | AE | 0.741 | AE | 0.763 | AE | 0.749 |
| 30 | AE | 0.719 | AE | 0.751 | AE | 0.738 |
| 35 | AE | 0.699 | AE | 0.738 | AE | 0.731 |
| 40 | II (BR) | 0.684 | AE | 0.724 | AE | 0.721 |
| Original data | 0.816 | | | | | |
| CD vs. UC | | | | | | |
| 5 | II (BR) | 0.821 | II (BR) | 0.821 | II (RF) | 0.821 |
| 10 | II (BR) | 0.819 | II (BR) | 0.818 | II (BR) | 0.813 |
| 15 | II (BR) | 0.816 | II (BR) | 0.815 | II (BR) | 0.807 |
| 20 | II (BR) | 0.814 | II (BR) | 0.811 | II (BR) | 0.803 |
| 25 | II (BR) | 0.812 | II (BR) | 0.806 | II (BR) | 0.798 |
| 30 | II (BR) | 0.809 | II (BR) | 0.800 | II (BR) | 0.793 |
| 35 | II (BR) | 0.806 | AE | 0.793 | II (BR) | 0.788 |
| 40 | II (BR) | 0.801 | AE | 0.787 | AE | 0.784 |
| Original data | 0.823 | | | | | |

**Table 3**. AUC (adjusted) values for differentiation of CD vs. non-IBD, UC vs. non-IBD, and CD vs. UC using a GB classifier with 5-fold cross-validation under MAR, MCAR, and MNAR mechanisms. Iterative imputers (especially II(BR)) consistently maintain the highest classification performance in CD-related tasks, while AE becomes competitive at higher missingness levels. In UC vs. non-IBD, AE generally outperforms other methods across most missingness degrees, whereas II(RF) shows an advantage in MNAR at low missingness.

universally optimal, iterative imputers tend to excel when missingness is modest, whereas autoencoder-based approaches provide superior resilience when missingness is more extensive.

## Discussion

Missingness in serologic data creates challenges for statistical and ML analyses, potentially leading to biased conclusions if not handled properly. Our study underscores the importance of robust imputation methods in preserving data integrity and utility in serological studies related to IBD. By simulating various missing data scenarios, including MAR, MCAR, and MNAR and evaluating multiple imputation techniques, we aimed to identify the most effective methods. our findings demonstrate that employing imputation methods significantly improves performance in both statistical analyses and prediction tasks compared to cases where no imputation is performed. We also observed that the best-performing imputation models may vary depending on the data size, missing pattern, and assessment approach.

Our findings suggest a practical division of labor. For inference-first workflows that require valid uncertainty (e.g., hypothesis testing), practitioners may prefer MI-style approaches (MICE, II-BR) and aggregate with Rubin's rules. For prediction-first workflows (single completed dataset), II/AE/VAE are often competitive or superior, especially as missingness increases. Reporting results along NRMSD, adjusted $-\log_{10}(P)$, and adjusted AUC thus helps align method choice with the downstream objective.

### Direct assessment

In general, we observe that NRMSD increases with higher percentages of missing data, indicating a decline in imputation accuracy. However, this trend does not hold for certain instances in the MIRIAD dataset with MNAR missingness. This deviation might be attributed to the inherent complexity of the MNAR mechanism, where the probability of missingness in each feature is a function of the feature itself. Additionally, the size of MIRIAD is

| Missingness degree | MAR | MCAR | MNAR |
|---|---|---|---|
| Direct assessment | | | |
| Low-missingness | II (BR); II (KNN); AE | II (BR); II (KNN); II (RF) | II (RF) |
| Medium-missingness | AE; VAE; II (BR) | II (KNN) | II (RF); II (BR) |
| High-missingness | AE | II (KNN); AE | AE |
| Statistical Analysis for CD vs. non-IBD | | | |
| Low-missingness | II (RF); II (KNN) | II (KNN) | II (KNN); II (BR); AE |
| Medium-missingness | II (BR) | II (KNN) | II (BR) |
| High-missingness | II (KNN); II (RF) | II (KNN) | VAE; AE |
| Statistical Analysis for UC vs. non-IBD | | | |
| Low-missingness | II (KNN) | II (BR); VAE | II (BR) |
| Medium-missingness | II (BR) | II (BR) | II (BR) |
| High-missingness | II (KNN) | II (BR) | II (BR); II (KNN) |
| Statistical Analysis for CD vs. UC | | | |
| Low-missingness | II (KNN); II (RF) | II (KNN) | II (BR) |
| Medium-missingness | II (BR) | II (KNN) | II (BR) |
| High-missingness | II (BR) | II (RF) | II (RF) |
| Machine Learning Analysis for CD vs. non-IBD | | | |
| Low-missingness | II (BR) | II (BR) | II (BR) |
| Medium-missingness | II (BR) | II (BR) | II (BR) |
| High-missingness | II (BR) | AE | AE |
| Machine Learning Analysis for UC vs. non-IBD | | | |
| Low-missingness | AE | AE | II (RF) |
| Medium-missingness | AE | AE | AE; II (RF) |
| High-missingness | AE; II (BR) | AE | AE |
| Machine Learning Analysis for CD vs. UC | | | |
| Low-missingness | II (BR) | II (BR) | II (RF); AE |
| Medium-missingness | II (BR) | II (BR) | II (BR) |
| High-missingness | II (BR) | AE | AE; II (BR) |

**Table 4**. Summary of the best-performing imputers across all forms of assessment. The table consolidates results from direct imputation error assessment (NRMSD), statistical analysis using adjusted $-\log_{10}(P)$, and ML classification tasks (adjusted AUC). Iterative imputers (particularly II (BR), II (KNN), and II (RF)) dominate under low to medium levels of missingness, while autoencoder-based methods (AE, VAE) demonstrate stronger robustness at higher missingness, especially in UC-related tasks. This overview highlights that no single imputation strategy is universally optimal; rather, the best method depends on the missingness degree, mechanism, and the type of downstream analysis.

larger than the other cohorts, which might reflect less sensitivity to missingness and explain the deviation in the trend. Indeed, data size can play a key role in determining the best imputer. Specifically, for the MIRIAD dataset, which is the largest cohort in our study, we notice that both IIs and AE-based models are among the best performing cases. Conversely, for the smallest cohort, i.e., RISK, we observe a more noticeable presence of IIs. This suggests that AE-based models can be more effective in scenarios with larger data sizes, highlighting their capability to capture complex patterns in larger datasets. Another important point to consider is that AE-based models demonstrate better performance for MCAR missingness, where it is hard for ML models to learn a non-existent missing pattern. This highlights the capability of AE-based models to impute the data where no explicit relationship exists between the missingness of input features. Complexity and degree of missingness can also play an important role in depicting the right model. For instance, for certain instances with MNAR, RF, BR and AE models outperform other models because they can better capture the dependency of missingness on the input features. Specifically, at higher degrees of missingness, AE outperforms other methods. Furthermore, we did not fit models that explicitly incorporate missingness indicators, so our MNAR analyses serve as a mechanism-specific stress test. Consistent with the inherent difficulty of MNAR, we observed generally larger NRMSD values under MNAR than under MCAR/MAR, and the MNAR results should therefore be interpreted as relative performance within this specific scenario rather than evidence that any method can broadly correct MNAR bias.

Direct assessment provides a method for directly comparing the actual values with the imputed values, offering a clear quantitative measure of accuracy. However, this approach does not assess the overall structure of the imputed data for post-imputation analysis, such as its ability to mimic the original data's statistical behavior or how well the input features collectively predict an outcome. Direct assessment focuses on point-to-point accuracy but overlooks how the imputation impacts the data's holistic properties. Therefore, it is essential to perform indirect assessments to obtain a comprehensive evaluation of the imputation performance. Indirect

assessments help ensure that the imputed data maintains the integrity and predictive power of the original dataset, providing a more complete understanding of the imputation's effectiveness.

### Statistical analysis

As previously mentioned, it is impossible to identify a single optimal imputer, and we observe that the optimal model varies depending on the nature of the missingness and the specific serologic features. For instance, for the ANCA feature in CD vs. non-IBD, we observe a strong presence of VAE and AE, along with a relatively large difference between predicted and original adjusted $-\log_{10}(P)$ values. This lower performance for ANCA may be due to its weaker association with CD compared to other serologic parameters. In general, we observe that II (RF), II (BR) and II (KNN) consistently outperformed others, as evidenced by their leading positions in the direct assessments. Notably, II (KNN) and II (BR) emerged as the most frequently effective imputers. We observe that AE-based imputers perform better when the inherent association is not strong. Overall, in most scenarios, the P-values obtained from the imputed datasets are much closer to the original dataset compared to the complete case datasets. This is significantly highlighted in cases that involve CD, as most serologic parameters have a stronger association with CD compared to UC. Our analysis indicates that choosing the right imputer can yield P-values nearly identical to those of the original dataset.

### Machine learning analysis

Our results indicate that the highest AUC values are observed in cases with no missing data. Additionally, we observe a decline in adjusted AUC values as the percentage of missing data increases. Furthermore, we notice that the decline in the adjusted AUC values for the imputed dataset is more significant for the UC vs. non-IBD as opposed to CD vs. non-IBD and CD vs. UC, which can be attributed to the weaker association of serologic parameters with UC compared to CD. Overall, II (BR) and II (RF) exhibit superior performance at low and medium levels of missingness. Conversely, at high levels of missingness, AE outperforms other methods, underscoring the robustness of AE-based imputers in handling extensive data missingness. Overall, the findings suggest that as the complexity of missingness increases, AE-based models exhibit a stronger capability to capture these intricate patterns. Despite this, the performance differences among the various top imputers remain relatively minor, giving some flexibility when it comes to selecting the imputer, and possibly paving the way for taking into consideration some external parameters, such as computation time, which can play a crucial role in determining the most suitable imputer for specific applications.

Our analysis demonstrated that the top imputation models generally outperform the widely used traditional MICE method. The advantages of these models over traditional MICE are especially notable in scenarios with high missingness where models like II (KNN) and AE excel due to their ability to effectively capture the missing pattern. KNN and RF imputers frequently emerge as top-performing imputers in many studies, which aligns with our findings where II implementations of KNN and RF were among the best performers[29,30]. However, it is important to acknowledge that there is no perfect, universal imputation method. Our findings are consistent with previous studies that highlight the absence of a one-size-fits-all solution for imputation[25]. This underscores the importance of a tailored approach when selecting imputation methods in serologic studies, ensuring that the chosen technique is well-suited to the specific characteristics of the dataset and the research objectives.

It is important to note that although we generated multiple imputations for all methods, the II (RF/KNN) and AE approaches do not correspond to draws from a fully specified joint model. We therefore used multiple stochastic imputations to capture variability and combined estimates using Rubin's rules for consistency across methods. Our simulations used three IBD cohorts and serologic panels; while this strengthens external relevance within IBD, generalization beyond similar antibody panels or clinical contexts should be done with care. Finally, we focused on univariate inference for clarity; multivariable models may alter the relative ordering of methods in specific settings.

## Conclusion

Serologic markers are informative for IBD, yet analyses are sensitive to missing data. Using real serology with simulated MCAR, MAR, and MNAR mechanisms (5–40% missingness), we compared MI approaches with both direct accuracy (NRMSD) and downstream performance (logistic-regression inference and predictive AUC), all evaluated within-cohort. No single imputer was universally best. Performance depended on the missingness mechanism and the analysis goal. At low–moderate missingness under MCAR/MAR, several MICE variants (e.g., PMM/RF/CART) and iterative imputers (e.g., RF/KNN/BR) consistently produced low NRMSD and preserved inferential signal close to the complete-data analysis. Under MNAR, autoencoder approaches (AE/VAE) and tree-based imputers often yielded strong predictive performance, though inferential behavior can vary because of model misspecification. Relative to complete-case analysis, MI retained sample size and yielded more precise, well-calibrated inference. Empirical T from permutation nulls clustered around $\alpha = 0.05$, and Rubin-combined standard errors increased with missingness under MCAR/MAR as expected. Consequently, the level of evidence ($-\log_{10}(P)$) from MI was typically closer to what is observed in the ground-truth data analysis, whereas complete-case analysis tended to attenuate associations by discarding information. Practically, we recommend selecting the imputation strategy to match the primary objective: favor MI methods such as MICE (PMM/RF/CART) or iterative RF/KNN when valid inference is the priority under MCAR/MAR; consider iterative imputers and AE/VAE when the goal is prediction. In all cases, reporting across imputations and mechanisms provides a transparent view of robustness.

## Data availability

The data and the codes used for the imputation and imputation assessment are available at https:/github.com/BoodaghiM/Serologic-Biomarker-Data-Imputation-Applications-in-IBD-Research.

Received: 23 November 2024; Accepted: 20 February 2026
Published online: 26 February 2026

## References

1. Lucas López, R., Grande Burgos, M. J., Gálvez, A. & Pérez Pulido, R. The human gastrointestinal tract and oral microbiota in inflammatory bowel disease: a state of the science review. *APMIS* **125** (1), 3–10 (2017).
2. Kuna, A. T. Serological markers of inflammatory bowel disease. Vol. 23, Biochemia Medica. (2013).
3. Sura, S. P., Ahmed, A., Cheifetz, A. S. & Moss, A. C. Characteristics of inflammatory bowel disease serology in patients with indeterminate colitis. *J. Clin. Gastroenterol.* ;**48**(4). (2014).
4. Lee, W. I., Subramaniam, K., Hawkins, C. A. & Randall, K. L. The significance of ANCA positivity in patients with inflammatory bowel disease. *Pathology* ;**51**(6). (2019).
5. Prideaux, L., De Cruz, P., Ng, S. C. & Kamm, M. A. Serological Antibodies in Inflammatory Bowel Disease: A Systematic Review. *Inflamm. Bowel Dis.* **18** (7), 1340–1355 (2012).
6. Van Schaik, F. D. M. et al. Serological markers predict inflammatory bowel disease years before the diagnosis. *Gut* **62**(5). (2013).
7. Li, X., Conklin, L. & Alex, P. New serological biomarkers of inflammatory bowel disease Vol. 14 *World J. Gastroenterol.*, (2008).
8. Austin, P. C., White, I. R., Lee, D. S. & van Buuren, S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Can. J. Cardiol.* **37** (9), 1322–1331 (2021).
9. Wisniewski, S. R., Leon, A. C., Otto, M. W. & Trivedi, M. H. Prevention of Missing Data in Clinical Research Studies. Vol. 59, *Biol. Psychiatr.*. (2006).
10. Khan, S. I. & Hoque, A. S. M. L. SICE: an improved missing data imputation technique. *J. Big Data*. **7** (1), 37 (2020).
11. Sun, Y., Li, J., Xu, Y., Zhang, T. & Wang, X. Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Syst. Appl.* **227**, 120201 (2023).
12. Tang, F. & Ishwaran, H. Random forest missing data algorithms. *Stat. Anal. Data Min.* ;**10**(6). (2017).
13. Verpoort, P. C., MacDonald, P. & Conduit, G. J. Materials data validation and imputation with an artificial neural network. *Comput. Mater. Sci.* ;147. (2018).
14. Choudhury, S. J. & Pal, N. R. Imputation of missing data with neural networks for classification. *Knowl. Based Syst.* **182**. (2019).
15. Lin, W. C., Tsai, C. F. & Zhong, J. R. Deep learning for missing value imputation of continuous data and the effect of data discretization. *Knowl. Based Syst.* ;**239**. (2022).
16. Gómez-Carracedo, M. P., Andrade, J. M., López-Mahía, P., Muniategui, S. & Prada, D. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometr. Intell. Lab. Syst.* **134**, 23–33 (2014).
17. Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P. & Lee, K. J. Should multiple imputation be the method of choice for handling missing data in randomized trials? *Stat. Methods Med. Res.* **27**(9). (2018).
18. Zhang, Z. Missing data imputation: Focusing on single imputation. *Ann. Transl Med.* **4**(1). (2016).
19. Rubin, D. B. Inference and missing data. *Biometrika* **63**(3). (1976).
20. Pang, Y. et al. Assessment of clinical activity and severity using serum ANCA and ASCA antibodies in patients with ulcerative colitis. *Allergy Asthma Clin. Immunol.* ;**16**(1). (2020).
21. Morgan, N. N. et al. Crohn's Disease Patients Uniquely Contain Inflammatory Responses to Flagellin in a CD4 Effector Memory Subset. *Inflamm. Bowel Dis.* ;**28**(12). (2022).
22. Shome, M. et al. Serological profiling of Crohn's disease and ulcerative colitis patients reveals anti-microbial antibody signatures. *World J. Gastroenterol.* **28**(30). (2022).
23. Faisal, S. & Tutz, G. Multiple imputation using nearest neighbor methods. *Inf. Sci. (N Y).* **570**, 500–516 (2021).
24. Blazek, K., van Zwieten, A., Saglimbene, V. & Teixeira-Pinto, A. A practical guide to multiple imputation of missing data in nephrology. *Kidney Inter.* Vol. 99 (2021).
25. Jadhav, A., Pramod, D. & Ramanathan, K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Appl. Artif. Intell.* **33**(10):913–933. (2019).
26. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**(3). (2011).
27. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**. (2011).
28. Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. (2019).
29. Lin, W. C. & Tsai, C. F. Missing value imputation: a review and analysis of the literature (2006–2017). *Artif. Intell. Rev.* ;**53**(2). (2020).
30. Niass, O., Diongue, A. K. & Touré, A. Analysis of missing data in sero-epidemiological studies. *Afr. J. Appl. Stat.* **2** (1), 29–37 (2015).

## Acknowledgements

This work was supported by the Cedars-Sinai MIRIAD IBD Biobank. The MIRIAD IBD Biobank receives funding from the Widjaja Foundation, Inflammatory Bowel and Immunobiology Research Institute, National Institute of Diabetes and Digestive and Kidney Disease Grants P01DK046763, U01DK062413 and The Leona M. & Harry B. Helmsley Charitable Trust.

## Author contributions

MB conducted the machine learning analysis; MB conducted data preprocessing, collected the data; MB, DL, DM wrote the manuscript; MB, DL, DM helped design the study; DL, DM critically reviewed the manuscript; DL and DM supervised the project; DM provided the funding. All authors reviewed and agreed upon the final manuscript.

## Funding

The study was funded by Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, National Institute of Diabetes and Digestive and Kidney Disease Grants P01DK046763, U01DK062413 and The Leona M. & Harry B. Helmsley Charitable Trust.

## Declarations

### Competing interests
The authors declare no competing interests.

### Ethics approval and consent to participate
All methods were carried out in accordance with relevant guidelines and regulations. All participants signed informed consent forms. All experimental protocols were approved by Cedars-Saini Medical Center (CSMC) Institutional Review Board.

### Consent for publication
All authors have provided their consent for the publication of this manuscript.

### Additional information
**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-41587-z.

**Correspondence** and requests for materials should be addressed to D.L.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.