# Comparing Missing Data Imputation Methods for Patient-Reported Outcomes in Esophageal Cancer Research

Yong Jin Kweon[1], Yousif Salman[2,*], Shayan Dhillon[2,*], Mehrnoush Dehghani[1], Emad A. Mohammed[2], R. Trafford Crump[1,3]

[1]Research Institute of the McGill University Health Centre, Montreal, Canada; [2]Wilfrid Laurier University, Waterloo, Canada; [3]McGill University, Montreal, Canada; [*]These authors contributed equally to this work.

**ABSTRACT**
Missing data is common in patient-reported outcomes (PRO) research, particularly in oncology settings. We evaluated seven methods for handling missing data in esophageal cancer quality of life measurements, namely: Multiple Imputation by Chained Equations (MICE), Variational Autoencoder (VAE), Denoising Autoencoder (DAE), Bayesian Principal Component Analysis (BPCA), a deep autoencoder method with patient-specific embeddings and temporal pattern modeling, SoftImpute (a matrix completion method using iterative soft-thresholded singular value decomposition), and K-Nearest Neighbors (KNN). Using data from McGill University's Esophageal and Gastric Data- and Bio-Bank, we compared these methods across 44 Functional Assessment of Cancer Therapy-Esophageal (FACT-E) quality-of-life variables on execution time, distribution preservation, correlation maintenance, imputation accuracy, and clinical classification performance. Our comprehensive validation framework provides evidence-based recommendations for selecting appropriate imputation methods for esophageal cancer PRO research, which may improve the validity and reliability of research findings in this domain.

**KEYWORDS**
Missing data; Multiple imputation; Machine learning; Patient-reported outcomes; Esophageal cancer; Quality of life

## 1. Introduction

Missing data is common when collecting patient-reported outcomes (PROs), impacting statistical power and potentially biasing results and our interpretation of those results [1, 2]. PROs are collected through standardized survey instruments (i.e., patient-reported outcome measures) with the intention of capturing patients' self-reported symptom burden, functional status, and the impact that these have on their daily activities and quality-of-life. Missing data occurs when values are recorded for some items or subjects but not for all items or subjects in a dataset [3].

The type of missingness - namely, Missing Completely At Random (MCAR), Missing At Random (MAR), or Missing Not At Random (MNAR) - varies based on the underlying reason for non-response [4]. When missingness occurs due to factors such as technical difficulties or survey oversight, it is typically considered MCAR, meaning it is unrelated to the patient's health status or the outcome of interest; for example, a patient might not receive a survey because the clinic was too busy. In contrast, MNAR is characterized by a strong association between missing data and the patient's health, such as absence linked to hospital admission or physical or mental health issues.

In controlled clinical trials, missing PROs data can be mitigated through rigorous study design, protocol integration, and quality assurance, often aided by regulatory

CONTACT Yong Jin Kweon. Email: yong.kweon@mail.mcgill.ca

guidelines and the strategic collection of auxiliary data to support assumptions about MAR mechanisms [5]. However, real-world clinical registries – i.e., databases that capture information about patients' health and healthcare use, often including PROs – lack such stringent controls. These registries face more complex and varied patterns of missingness driven by a wide range of potential causes and fewer dedicated resources for routine data checks.

Real-world clinical registries are more vulnerable to bias and reduced precision if missingness is ignored and statistical analyses only on subjects with complete data [6]. Consequently, robust post-hoc imputation strategies, such as multiple imputation (MI), and a detailed understanding of non-response mechanisms are essential for valid statistical inference in registry analyses [4]. Simple methods like complete case analysis can yield inaccurate or inconsistent clinical conclusions depending on the extent of missing data.

MI is a widely adopted statistical method for addressing missing data that involves imputing multiple plausible values for missing observations [3, 7]. This process creates multiple complete datasets, each with the same size as the original sample, where all missing data have been replaced with plausible values. Statistical analyses are then conducted on each complete dataset, and the results are pooled across samples.

MI generally assumes MAR conditions and linear relationships among variables [8, 9]. Under these MAR assumptions, PRO data present additional complexities including ordinal scales with discrete, bounded responses; potential ceiling or floor effects; and complex non-linear relationships between symptom domains [10, 11]. Recent advances in machine learning offer promising new approaches to imputing missing PROs because they better preserve the complex relationships in PRO data through their ability to model non-linear associations, handle non-parametric distributions, and capture intricate patterns in high-dimensional sparse data [12].

The primary objective of this study is to evaluate traditional statistical methods and more recent machine learning methods for imputing missing PRO data. Our comprehensive benchmark evaluation extends beyond simple imputation accuracy to include computational efficiency, distribution preservation, correlation maintenance, and clinical classification performance. This will provide a holistic assessment of method suitability for different research contexts. The results from this study will support clinical researchers faced with missing PRO data make an informed decision based on their specific research needs, computational constraints, and analytical priorities.

## 2. Methods

This study used data obtained from the McGill University's Division of Thoracic and Upper Gastrointestinal Surgery (i.e., the Division), one of North America's most comprehensive centres for esophageal cancer with over 250 new esophagogastric cancer referrals and 100 esophagectomies performed annually. Since 2006, the Division has maintained the Esophageal and Gastric Data- and Bio-Bank (EGDB Bank), a clinical database that includes longitudinal data collected from patients diagnosed with, and treated for, esophageal cancer. Patients provide informed consent for their data to be collected and used in future research studies. Nearly 100% of newly referred patients are approached for enrolment. This study protocol was approved by the McGill University Health Centre's Centre for Applied Ethics.

### 2.1. Data Sources

Relevant to this study, The Division asks patients to complete the Functional Assessment of Cancer Therapy–Esophageal (FACT-E), a patient-reported outcome measure designed to assess symptom severity and impact in esophageal cancer patients, with established validity and reliability [13]. The FACT-E is comprised of four generic cancer domains (i.e., physical, social/family, emotional, and functional well-being) and one domain focused on symptoms specific to esophageal cancer. There are 44 items, each asking the respondent to rate their level of symptom severity over the past 7 days using a 5-point Likert scale. Responses to the items' Likert scales are scored and aggregated according to the scoring guidelines [13–16].

The Division collects the FACT-E at 11 different time points during the patient's episode of care. Our analysis focused exclusively on one time point: "baseline" (i.e., collected at the patient's first assessment with the thoracic surgeon). Baseline represents the most universal time point that every patient undergoes, providing the largest available sample size and avoiding temporal constraints that could bias certain methods. This cross-sectional focus ensures that specialized longitudinal methods are evaluated on their core imputation capabilities rather than their temporal modeling advantages.

### 2.2. Variables and Preprocessing

The dataset includes numerical and ordinal variables (Table 1). The numerical/ordinal variables (float or int type) included: id (i.e., an anonymous identifier given to each patient) and 44 FACT-E target variables comprising general physical well-being (i.e., variables = GP1-GP7), social/family well-being (i.e., variables = GS1-GS7), emotional well-being (i.e., variables = GE1-GE6), functional well-being (i.e., variables = GF1-GF7), and esophageal cancer-specific subscales (i.e., variables = A_HN1-A_HN5, A_HN7, A_HN10, A_E1-A_E7, A_C6, A_C2, A_ACT11)[1]. Date variable qol_date (i.e., the date that the FACT-E was completed) was retained and transformed into numerical features for the Da Xu et al. [17, 18] deep learning model, which incorporates date-derived features (year, month, day) as part of its architectural design.

**Table 1.** Percentage of missing values for each FACT-E variable

| **GP1** | **GP2** | **GP3** | **GP4** | **GP5** | **GP6** | **GP7** | **GS1** | **GS2** | **GS3** | **GS4** |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.38 | 1.47 | 3.63 | 1.77 | 17.58 | 3.05 | 3.34 | 3.24 | 2.26 | 3.34 | 8.06 |
| **GS5** | **GS6** | **GS7** | **GE1** | **GE2** | **GE3** | **GE4** | **GE5** | **GE6** | **GF1** | **GF2** |
| 6.68 | 10.81 | 43.12 | 1.67 | 4.62 | 8.45 | 3.05 | 3.24 | 2.85 | 4.32 | 7.56 |
| **GF3** | **GF4** | **GF5** | **GF6** | **GF7** | **A_HN1** | **A_HN2** | **A_HN3** | **A_HN4** | **A_HN5** | **A_HN7** |
| 2.36 | 5.60 | 2.26 | 3.63 | 2.75 | 4.22 | 5.60 | 6.78 | 6.68 | 6.09 | 5.99 |
| **A_HN10** | **A_E1** | **A_E2** | **A_E3** | **A_E4** | **A_E5** | **A_E6** | **A_E7** | **A_C6** | **A_C2** | **A_ACT11** |
| 5.60 | 3.14 | 3.05 | 2.65 | 3.14 | 3.63 | 4.52 | 2.95 | 3.05 | 3.14 | 2.75 |

### 2.3. Imputation Methods

We compared Multiple Imputation by Chained Equations (MICE) [19], Variational Autoencoder (VAE) [20], Denoising Autoencoder (DAE) [21], Bayesian Principal Component Analysis (BPCA) [22], a specialized Deep Learning approach based on recent

[1]Variable labels follow the official FACT-E (Functional Assessment of Cancer Therapy-Esophageal) scoring guidelines. Complete variable definitions and scoring instructions are available in the FACT-E administration manual [13] or at www.facit.org.

literature [17, 18], SoftImpute [23, 24], and K-Nearest Neighbors (KNN) [25]. These methods were selected to represent diverse imputation paradigms and address the specific challenges of PRO data. MICE and KNN serve as established baseline methods widely used in clinical research. VAE and DAE represent modern deep learning approaches for complex data reconstruction. BPCA addresses the high-dimensional nature of PRO instruments through dimensionality reduction. The specialized deep learning approach was specifically chosen for its demonstrated effectiveness in clinical patient data with similar longitudinal clinical characteristics, incorporating patient-specific embeddings and temporal patterns that can be applied to cross-sectional baseline data to evaluate whether such patient-specific, longitudinal-based modeling provides benefits even without longitudinal follow-up. SoftImpute was selected for its proven performance in matrix completion tasks with structured missing patterns, which commonly occur in PRO data due to systematic item non-response [23, 24]. This combination ensures comprehensive coverage of statistical, machine learning, and deep learning paradigms for evaluating imputation performance on complex PRO data.

#### *2.3.1. Model Configuration and Comparison Strategy*

To ensure fair comparison across all seven imputation methods, we deliberately avoided extensive hyperparameter tuning and instead used standard, well-established default configurations for each approach. Our primary objective was not to optimize individual methods for marginal performance improvements, but rather to provide a comprehensive benchmarking comparison of popular and relatively recent missing data imputation techniques as they would typically be applied in clinical research settings. This approach reflects real-world usage scenarios where researchers seek reliable, generalizable methods without extensive method-specific optimization.

Specifically, for MICE, we used standard LightGBM parameters with 80 boosting rounds and a maximum depth of 10. VAE employed typical architecture dimensions (64→32→16→8) with standard learning rates (0.001) and batch sizes (64). DAE used conventional five-layer symmetric architectures with standard regularization. BPCA utilized adaptive component selection with established convergence criteria. The Da Xu et al. deep learning approach followed the original paper's architectural specifications. SoftImpute employed adaptive rank selection with standard convergence thresholds. KNN used commonly recommended neighbor counts (n=5) with established distance metrics.

The following subsections detail the architectural design and implementation approach for each of the seven imputation methods evaluated.

#### *2.3.2. MICE (Multiple Imputation by Chained Equations)*

MICE is an iterative statistical method that creates multiple plausible datasets by modeling each incomplete variable as a function of all other variables through chained conditional distributions. The algorithm cycles through variables sequentially, fitting a prediction model for each variable with missing values using all other variables as predictors, then sampling from the posterior predictive distribution to fill missing values. Mathematically, for variable $Y_j$ with missing values, MICE fits: $Y_j^{(t+1)} = f_j(Y_{-j}^{(t)}) + \epsilon_j$, where $Y_{-j}^{(t)}$ represents all other variables at iteration $t$, $f_j$ is the prediction function, and $\epsilon_j$ is the error term.

We used MICE with the `miceforest` package, which uses LightGBM [26] as the base model for handling non-linear relationships. For each target variable, all other columns

in the dataset were used as predictors through a variable schema configuration. The implementation involved 5 imputation iterations, where each iteration cycles through all variables with missing values. Hyperparameters were selected based on computational efficiency and model stability: 80 boosting rounds per model (chosen to balance model complexity with training time), maximum tree depth of 10 (to prevent overfitting while capturing interactions), and 10 parallel threads for computation. The OpenMP environment variable (OMP_NUM_THREADS = 10) was configured at the system level to control the number of CPU threads used by the underlying LightGBM library, optimizing computational performance without thread conflicts.

MICE excels in handling complex relationships between variables through its iterative chained equations approach, making it particularly suitable for datasets with multiple correlated variables like FACT-E subscales [19, 27]. The method provides theoretical guarantees for valid statistical inference when the missing at random assumption is satisfied and naturally handles different variable types within the same framework [28, 29]. The flexibility allows the incorporation of auxiliary variables that may not be of direct interest but improve imputation quality. Additionally, MICE's iterative nature enables it to capture non-linear relationships through the use of ensemble methods like LightGBM, while maintaining computational efficiency through parallelization.

#### *2.3.3. Variational Autoencoder (VAE)*

VAE are generative deep learning models that learn a probabilistic latent representation of data through an encoder-decoder architecture with variational inference, enabling reconstruction of missing values from learned probability distributions. Unlike standard autoencoders that learn deterministic mappings, VAEs learn probabilistic distributions in the latent space, making them particularly suitable for handling uncertainty in missing data imputation.

We implemented a VAE using `PyTorch` [30] with threshold-based feature selection, excluding predictor columns with greater than 50% missingness while ensuring target columns were included regardless of missingness. All columns were converted to numeric format with initial mean imputation and z-score normalization (subtracting mean and dividing by standard deviation) to ensure equal feature contribution. The encoder architecture follows systematic dimensionality reduction with configurable hidden layers (default: 64→32→16), meaning the network progressively reduces input dimension through three hidden layers with 64, 32, and 16 neurons respectively, learning increasingly abstract representations.

The encoder outputs two vectors: a mean vector ($\mu$) and log-variance vector ($\log\sigma^2$), both of dimension 8, representing Gaussian distribution parameters in the latent space. The reparameterization trick enables gradient-based training despite stochastic sampling by computing z= $\mu + \sigma \odot \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$, allowing gradients to flow through the sampling operation during backpropagation. The decoder mirrors the encoder architecture in reverse (16→32→64→original dimension), reconstructing original input from latent representation. We applied layer normalization for CPU efficiency (operating on individual samples rather than batch statistics), LeakyReLU activations [31], and dropout (rate=0.1). The loss function combines masked mean squared error (MSE) reconstruction loss (calculated only on non-missing values) with KL divergence regularization $D_{KL}[q(z|x)||p(z)]$ that maintains similarity between learned latent distribution $q(z|x)$ and standard normal prior $p(z) = \mathcal{N}(0, I)$, encouraging the latent space to follow a known distribution while maintaining sufficient representational ca-

pacity ($\beta$-VAE formulation for disentanglement). Training parameters included Adam optimizer with learning rate 0.001 and a maximum of 100 epochs.

VAE provides superior handling of non-linear relationships and complex data distributions through its deep neural network architecture with multiple hidden layers and non-linear activation functions [20, 32]. The latent space representation captures underlying patterns in PRO data that may not be apparent through linear methods, such as non-additive interactions between symptom domains, threshold effects where small changes in one variable dramatically affect others, and clustering patterns where patients with similar overall health status may have vastly different individual symptom profiles. The generative nature of VAEs allows them to learn robust representations that can handle various missing data patterns, while the KL divergence regularization maintains similarity between the input and generated data distributions, preventing the model from generating unrealistic values that deviate significantly from the observed data patterns [20, 33].

#### *2.3.4. Denoising Autoencoder (DAE)*

DAE are neural networks trained to reconstruct clean data from artificially corrupted inputs through a symmetric encoder-decoder architecture, learning robust feature representations that can handle missing data patterns. The denoising process forces the model to learn meaningful patterns by reconstructing original data from noisy versions, making it particularly effective for imputation tasks where missing values can be viewed as a form of data corruption.

We utilized a `scikit-learn` [34] based DAE using Multi-Layer Perceptron Regressor (MLPRegressor), a feedforward neural network with multiple hidden layers trained via backpropagation that reconstructs clean data from noisy inputs. We used a feature selection approach based on RandomForestRegressor to identify up to 20 important predictors per target column, excluding columns with more than 30% missingness, while ensuring target columns were included. The DAE then reconstructs missing values using these selected columns, leveraging relationships among them to impute the target variables. For example, GE1's missing values are imputed based on patterns learned from a selected set of up to 20 columns, including target columns and other predictors with less than 30% missingness, identified using RandomForestRegressor. While the model leverages relationships among these selected columns, this differs from our VAE implementation, which includes all columns with less than 50% missingness.

The implementation employed initial mean imputation with z-score normalization (StandardScaler). A key aspect of our approach was selective noise addition, applied only to observed values through $X_{noisy} = X_{scaled} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with noise factor $\sigma = 0.1$, which served as a denoising effect thus improving generalization while preserving missing value patterns. The architecture was designed to adapt based on dataset size, featuring a five-layer symmetric structure where the network compresses input through encoder layers (n→n/2→n/4), reaches a bottleneck representation, then expands through decoder layers (n/4→n/2→n) back to original dimension (Fig. 1). The default configuration (44→64→32→16→32→64→44) creates a symmetric hourglass shape that forces the model to learn compressed representations in the middle layer before reconstruction. We used ReLU activation functions with adaptive learning rate adjustment. The loss function combines masked mean squared error reconstruction loss with L2 regularization: Loss = MSE + $\alpha \sum_j w_j^2$, where $w_j$ are the network weights and $\alpha = 0.0001$ is the regularization strength to prevent overfitting [21, 35]. Training involved 200 iterations with adaptive learning rate adjustments. We applied Adam

optimizer and implemented automatic batch size determination where the batch size is set to 'auto' by `scikit-learn`'s MLPRegressor, which automatically selects min(200, n_samples) to balance computational efficiency with gradient stability.

This approach is particularly valuable for its simplicity, computational efficiency, and framework independence while learning robust representations through its denoising mechanism [21, 36]. The noise injection forces the model to learn more generalizable patterns, reducing overfitting on sparse data, which is especially important for clinical datasets with high missingness rates [37, 38]. Its framework independence and simplicity make it highly practical for clinical research settings where computational resources may be limited, while still capturing non-linear relationships within quality-of-life subscales. The approach, which leverages relationships among selected columns (including both target variables and other predictors with less than 30% missingness), is particularly suitable for capturing non-linear relationships within quality-of-life subscales while incorporating relevant external predictors when beneficial [12].

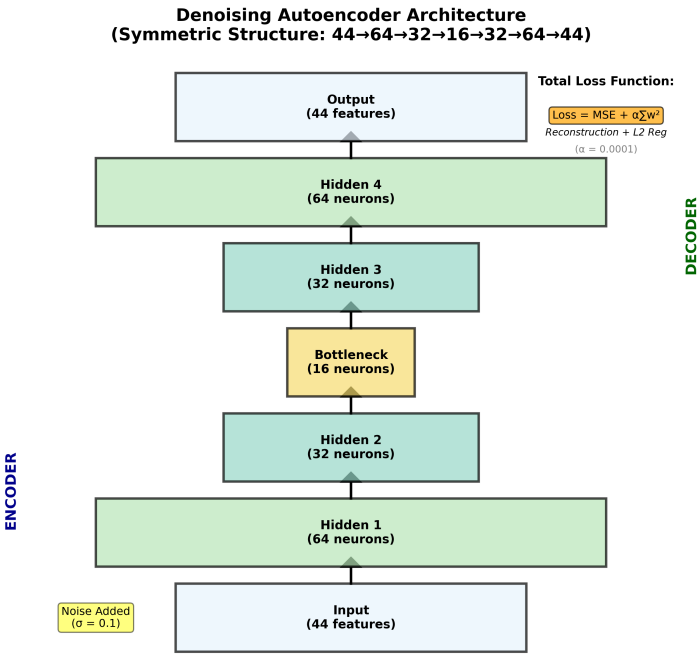


**Figure 1.** Denoising Autoencoder Architecture. The symmetric five-layer structure compresses input through encoder layers (44→64→32→16), reaches a bottleneck representation at 16 neurons, then reconstructs through decoder layers (16→32→64→44). Gaussian noise ($\sigma = 0.1$) is selectively added to observed values during training to improve generalization. The total loss function combines masked mean squared error (MSE) reconstruction loss with L2 regularization ($\alpha = 0.0001$): Loss = MSE + $\alpha \sum w_j^2$, where $w_j$ represents network weights.

### *2.3.5. Bayesian Principle Component Analysis (BPCA)*

BPCA extends traditional principal component analysis (PCA) by treating model parameters as random variables with prior distributions, using variational inference to learn a probabilistic low-dimensional representation that captures uncertainty in missing value estimation. Unlike deterministic PCA, BPCA provides full posterior distributions over parameters, enabling uncertainty quantification for imputed values [22, 39].

We implemented BPCA with variational inference for uncertainty quantification using the `PyTorch` [30] framework. The preprocessing involved initial mean imputation, z-score standardization, and standard PCA for parameter initialization. The model used all columns with less than 50% missingness, including all target columns, for both training and imputation, with all columns converted to numeric format. The architecture implemented a probabilistic PCA formulation where observed data $X \in \mathbb{R}^{n\times d}$ is modelled as $X = ZW^T + \epsilon$, where $Z \in \mathbb{R}^{n\times k}$ are latent variables, $W \in \mathbb{R}^{d\times k}$ is the loading matrix, and $\epsilon$ represents observation noise.

The loading matrix W was modelled with variational parameters: mean $\mu_W \in \mathbb{R}^{d\times k}$ and log-standard deviation $\log \sigma_W \in \mathbb{R}^{d\times k}$ to ensure positivity. For reconstruction uncertainty, we incorporated noise precision $\tau = \exp(\log \tau)$ where $\log \tau$ is a learnable parameter representing the log precision of observation noise. Latent variables $z_i \sim \mathcal{N}(0, I)$ were initialized from standard PCA scores. We used $k = \min(5, d-1)$ principal components to capture main variation patterns while preventing overfitting.

The Bayesian framework incorporated standard normal priors: $W_{ij} \sim \mathcal{N}(0, 1)$ for loadings and $z_i \sim \mathcal{N}(0, I)$ for latent variables. Posterior sampling involved drawing 1,000 samples from the approximate posterior $q(W) = \mathcal{N}(\mu_W, \mathrm{diag}(\sigma_W^2))$ to estimate parameter uncertainty and compute robust imputations. The reparameterization trick enabled gradient-based optimization by rewriting $W \sim q(W)$ as $W = \mu_W + \sigma_W \odot \epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$, allowing gradients to flow through sampling operations during backpropagation.

The loss function combined three components: $\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_z + \mathcal{L}_W$, where $\mathcal{L}_{\text{recon}} = \tau \sum_{i,j\in\text{observed}}(X_{ij} - [ZW^T]_{ij})^2$ is the masked reconstruction error weighted by noise precision, $\mathcal{L}_z = \frac{1}{2}\sum_i \|z_i\|^2$ is the latent variable regularization, and $\mathcal{L}_W = \frac{1}{2}\sum_{ij}(\mu^2_{W_{ij}} + \sigma^2_{W_{ij}} - \log \sigma^2_{W_{ij}} - 1)$ is the KL divergence between the approximate posterior and prior for loadings. Training used Adam optimizer (lr=0.01), ReduceLROnPlateau scheduler for adaptive learning rate adjustment, early stopping (60 epoch patience) to prevent overfitting, batch size min(64, n), and dynamic epochs max(100, min(300, $\lfloor \frac{10000}{n} \rfloor + 30$)) to adapt computational effort to dataset size.

BPCA provides principled uncertainty quantification through its Bayesian framework, offering full posterior distributions for missing values. This is particularly valuable for clinical decision-making, where understanding imputation uncertainty is crucial for risk assessment and treatment planning. The method excels at capturing linear relationships and correlation structures inherent in FACT-E data, making it highly effective for PRO that often exhibit strong inter-scale correlations [22, 40]. Its exceptional computational speed compared to other deep learning methods makes it suitable for rapid prototyping and large-scale applications [39, 41].

#### *2.3.6. Patient-specific and temporal-patterned Deep Learning Autoencoder*

Based on recent Da Xu et al. literature [17, 18], we implemented an autoencoder architecture that combines patient-specific embeddings with temporal pattern modeling originally designed to capture individual trajectories and longitudinal relationships in clinical data, but applied here to cross-sectional baseline data to evaluate whether patient-specific modeling provides benefits even without multiple time points per patient. Our implementation used columns with less than 50% missingness plus all target columns plus patient ID and qol_date. The patient ID enables the model to learn patient-specific representations, while qol_date provides temporal context (such as seasonal effects or cohort differences) that may influence quality-of-life reporting patterns even in cross-sectional data. This feature selection strategy was critical for balancing

model complexity with data availability while ensuring access to patient-specific and temporal information necessary for the approach.

The architecture implements a patient-embedding autoencoder where input data $X \in \mathbb{R}^{n \times d}$ is processed alongside patient identifiers $P \in \{1, 2, ..., k\}^n$ where k = 936 represents the number of unique patients. While 73 patients have multiple baseline entries in our dataset, patient IDs are mapped to dense embeddings $E_p \in \mathbb{R}^{16}$ through an embedding layer to capture patient-specific characteristics, while temporal features (year, month, day) are extracted from qol_date to capture date-related patterns such as seasonal effects or temporal trends in data collection. The autoencoder follows a symmetric five-layer structure (64→32→16→32→64) where the input is compressed through $h^{(l+1)} = f(W^{(l)}h^{(l)} + b^{(l)})$ where $f$ is the ELU activation function [42]. The distinctive decoder concatenates patient embeddings with the latent code: $h^{(0)}_{decoder} = [z; E_p]$ where $z \in \mathbb{R}^{16}$ is the latent representation and [;] denotes concatenation, before reconstruction through subsequent layers. We employed BatchNorm for stability, ELU activations, and dropout (rate=0.3) for regularization.

The loss function combines two components: $\mathcal{L} = \mathcal{L}_{recon} + \lambda \mathcal{L}_{temporal}$, where $\mathcal{L}_{recon} = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} (X_{ij} - \hat{X}_{ij})^2$ is the masked reconstruction loss calculated only on observed values $\Omega$, and $\mathcal{L}_{temporal} = \frac{1}{|S|} \sum_{(i,k)\in S} w_{ik} \|\hat{X}_i - \hat{X}_k\|^2$ is the temporal similarity loss that encourages consistent predictions for similar patients, where $S$ represents patient pairs, $w_{ik}$ are similarity weights based on patient proximity and temporal closeness, and $\lambda$ = 0.4 is the temporal similarity regularization weight. This temporal loss calculates pairwise differences between predictions weighted by patient similarity, encouraging the model to produce similar imputations for patients with comparable characteristics at similar time points. Additional L2 weight decay ($\alpha = 0.0001$) was applied for further regularization.

Training used Adam optimizer (learning rate = 0.001) with ReduceLROnPlateau scheduler (a PyTorch built-in learning rate scheduler that reduces the learning rate by a factor when validation loss plateaus), early stopping (20-epoch patience), 100 default epochs, batch size of 16, and 80/20 train/validation split. For categorical variables, we implemented one-hot encoding for inputs and post-processing to convert predictions back to categories based on maximum probability.

This method uniquely captures patient-specific patterns through personalized embeddings, originally designed for longitudinal clinical data but applied here to cross-sectional baseline data to evaluate whether patient-specific modeling provides imputation benefits even without longitudinal follow-up. The incorporation of temporal information helps identify temporal trends or seasonal effects in data collection, while the patient-specific approach can capture individual differences in PRO reporting patterns that may be clinically relevant. This approach tests whether techniques especially valuable in oncology settings where disease progression and treatment effects follow patient-specific trajectories can also improve imputation quality when applied to single time-point data.

#### *2.3.7. SoftImpute*

SoftImpute is a matrix completion algorithm that iteratively applies soft-thresholded singular value decomposition to find low-rank approximations, alternating between matrix factorization through iterative updates of the singular value decomposition (SVD) decomposition and missing value estimation until convergence. The method is based on nuclear norm minimization, which finds the matrix of lowest rank

that fits the observed data.

We utilized Travis Brady's SoftImpute implementation [43], based on the matrix completion method described in [23, 24], which employs iterative soft-thresholded SVD [44]. The SVD decomposes matrix $X \in \mathbb{R}^{n\times m}$ as $X = UDV^T$, where $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{m\times r}$ are orthogonal matrices containing left and right singular vectors, and $D \in \mathbb{R}^{r\times r}$ is a diagonal matrix of singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \geq 0$.

Our implementation utilized data-driven rank selection with $J = max(2, \lfloor \min(n_{samples}, n_{features})/4 \rfloor)$ components, where the rank is automatically determined based on dataset dimensions rather than requiring manual specification. This adaptive approach ensures computational efficiency by limiting the number of factors while capturing essential data structure—the dominant patterns of variation that explain the majority of variance in the PRO data through the largest singular values.

The algorithm follows these iterative steps: (1) Initialize missing values with column means: $\tilde{X}_{ij} = \bar{X}_j$ for missing entries, and apply z-score standardization to the resulting complete matrix for numerical stability, (2) Compute SVD: $\tilde{X} = UDV^T$, (3) Apply modified soft-thresholding to singular values: $D_{thresholded} = \frac{D}{D+\lambda}$, where $\lambda$ is the regularization parameter, to achieve shrinkage while preserving non-zero singular values. This modification was chosen for improved numerical stability in our dataset, (4) Reconstruct matrix: $\hat{X} = UD_{thresholded}V^T$, (5) Update missing values: $\tilde{X}_{ij} = \hat{X}_{ij}$ for missing entries, (6) Check convergence using the Frobenius norm ratio of the factorized matrices: ratio $= \frac{\|D^2_{old}\|_F + \|D^2\|_F - 2\text{trace}((U^T U_{old} D)(V^T_{old} V D_{old}))}{\max(\|D^2_{old}\|_F, 10^{-9})} < \epsilon$, where $\epsilon = 10^{-5}$, or terminate after a maximum of 100 iterations. The regularization parameter was set to $\lambda = 0$ for the base implementation.

The soft-thresholding operation applies shrinkage to singular values, effectively performing nuclear norm regularization $||X||_* = \sum_i \sigma_i$ (sum of singular values), which encourages low-rank solutions by penalizing model complexity. This is particularly appropriate for PRO data where underlying latent factors (such as overall health status, treatment response, or psychological well-being) may drive correlations across multiple FACT-E subscales, creating natural low-rank structure in the data matrix.

SoftImpute excels in capturing global structure—the overall correlation patterns and relationships that exist across the entire dataset—through matrix factorization, making it particularly effective for datasets with strong correlation patterns like FACT-E subscales. The method is theoretically grounded in convex optimization, providing guaranteed convergence to a global optimum and stable, reproducible results [23, 45]. Its ability to handle high-dimensional sparse data efficiently makes it well-suited for PRO datasets where many variables may be missing simultaneously. The nuclear norm regularization (minimization of the sum of singular values) naturally handles the intrinsic dimensionality of the data, making it robust to overfitting even with high missingness rates. Additionally, the method's computational efficiency and deterministic nature make it highly suitable for clinical research applications requiring reproducible results.

#### *2.3.8. K-Nearest Neighbors (KNN)*

KNN imputation is a non-parametric method that estimates missing values by identifying the k most similar complete observations based on available features and computing weighted averages of their corresponding values. The method operates on the principle of local similarity, where observations with similar feature patterns are expected to have similar values for missing variables.

We implemented KNN imputation [25] using `scikit-learn`'s [34] KNNImputer with adaptive neighbor selection. The method imputes missing values for observation $x_i$ with missing feature $j$ by finding k-nearest neighbors based on available features and computing weighted averages: $\hat{x}_{ij} = \frac{\sum_{l \in N_k(x_i)} w_{il} x_{lj}}{\sum_{l \in N_k(x_i)} w_{il}}$, where $N_k(x_i)$ represents the k-nearest neighbors of $x_i$ and $w_{il}$ are similarity weights.

Our implementation used $n_{neighbors} = 5$ with automatic adjustment based on data sparsity: $k_{adjusted} = \min(k, \max(1, C - 1))$, where $C$ is the number of complete cases available for each variable, ensuring sufficient observations for reliable estimation. The algorithm dynamically adjusts the number of neighbors using $k_{final} = \min(k_{original}, \max(1, \min_{j \in J}(|\{i : x_{ij} \neq \text{NA} \wedge \exists l \neq j : x_{il} \neq \text{NA}\}| - 1)))$, where $J$ represents the set of variables to be imputed, ensuring robust imputation even with sparse data. We used uniform weighting ($w_{il} = 1$) and nan_euclidean distance metric [34] to handle missing values in distance calculations.

The nan_euclidean distance metric is specifically designed for datasets with missing values, computing distances using only dimensions where both observations have valid data: $d(x_i, x_l) = \sqrt{\sum_{j \in O_{il}} (x_{ij} - x_{lj})^2}$, where $O_{il} = \{j : x_{ij} \neq \text{NA} \wedge x_{lj} \neq \text{NA}\}$ represents the set of commonly observed features between observations $i$ and $l$. This makes it particularly suitable for sparse clinical datasets where different patients may have missing values in different variables. The automatic adjustment of $k$ based on data availability ensures that the method remains robust even when few complete cases are available, which is common in cancer research with high attrition rates.

KNN provides intuitive, non-parametric imputation based on similarity between patients, making it particularly interpretable for clinical applications. Clinicians can easily understand that missing values are being estimated based on similar patients, which aligns well with clinical reasoning patterns. The method naturally adapts to local data structures and can capture complex, non-linear relationships without strong distributional assumptions [46, 47], making it robust to various data patterns that may exist with PROs. Its robustness to outliers and ability to work well with small sample sizes make it valuable for clinical datasets where extreme responses may be clinically meaningful rather than erroneous. Additionally, KNN's non-parametric nature means it makes no assumptions about the underlying data distribution, making it particularly suitable for ordinal clinical scales that may not follow normal distributions. The method's ability to preserve local neighborhoods in the feature space can be especially valuable for capturing patient subgroups with similar clinical characteristics or treatment responses.

#### *2.3.9. Computational Environment*

All analyses were conducted on Narval Cluster in the Digital Research Alliance of Canada infrastructure using a single node with 5 CPU cores, 20GB RAM, and 1 GPU (NVIDIA A100-SXM4-40GB) with CUDA 12.2. Python 3.11.5 was used with relevant libraries including scikit-learn, PyTorch, and miceforest.

### *2.4. Evaluation Framework*

To comprehensively evaluate imputation performance, we implemented two complementary assessment approaches specifically designed for high-missingness data, incorporating both continuous accuracy measures and clinically-relevant classification metrics.

### *2.4.1. Direct Dataset Assessment*

Our direct dataset evaluation encompassed multiple dimensions of imputation quality using the original dataset's naturally existing missingness patterns without cross-validation or artificial masking. We analyzed execution time to assess the computational efficiency of each method. Distribution similarity was evaluated using Kolmogorov-Smirnov (KS) tests [48]. This compares observed values against imputed values for each variable using the test statistic $D = \max_x |F_1(x) - F_2(x)|$ to measure the maximum difference between cumulative distribution functions $F_1$ (observed values) and $F_2$ (imputed values). Lower KS statistics and higher p-values indicate better distribution preservation. We conducted correlation preservation analysis by comparing correlation matrices before and after imputation, calculating absolute differences to quantify how well important relationships between variables were maintained. Additionally, we created comprehensive visualizations, including histograms comparing imputed versus observed distributions and heatmaps showing correlation differences across methods.

### *2.4.2. Sparse Cross-Validation Approach*

We developed a specialized cross-validation approach optimized for datasets with moderate missingness rates, introducing artificial missingness by setting non-missing fold values to "NaN" (i.e., missing) for validation. This framework addressed the challenge that standard k-fold cross-validation is inappropriate for sparse data where insufficient complete cases exist. Our approach performed column-wise validation, conducting separate 5-fold cross-validation for each of the 44 target variables. For each column, we identified all non-missing observations and randomly partitioned them into five folds, ensuring a minimum of 5 samples per fold or $n_{observed}/5$ samples, whichever was larger. Columns with fewer than $\max(5, n_{folds} \times 2)$ observations were excluded from validation.

For each imputation method, the imputed values were compared against the original held-out values to assess their respective effectiveness in recovering missing data. We computed Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for each fold and method, along with execution time measurements. Results were averaged across all folds and columns to provide comprehensive performance metrics. This approach enabled robust comparison by ensuring sufficient sample sizes for reliable assessment while maintaining the realistic missing data patterns characteristic of clinical datasets.

### *2.4.3. Clinical Classification Metrics*

Recognizing that FACT-E variables are ordinal scales (0-4), we implemented comprehensive classification evaluation by converting continuous imputed values to discrete clinical categories through rounding to the nearest integer and clipping to the valid range. This approach provides clinically meaningful interpretability as clinicians typically work with discrete FACT-E scores rather than continuous values.

For each imputation method and variable, we calculated multiclass classification metrics including overall accuracy, multiclass Area Under the Curve (AUC) using a one-vs-rest approach, macro-averaged precision and recall, and performance indicators. We computed sensitivity (true positive rate), specificity (true negative rate), positive predictive value (PPV), and negative predictive value (NPV) for each ordinal category, then averaged across categories to provide overall performance measures. These metrics

are particularly relevant for clinical decision-making, where the ability to correctly classify PROs into appropriate severity categories directly impacts treatment planning and patient monitoring.

The evaluation framework generated multiple complementary metrics: quantitative accuracy measures (MAE, RMSE), computational efficiency assessments (execution time), statistical distribution preservation tests (KS test), relationship maintenance analysis (correlation preservation), and classification performance (accuracy, AUC, sensitivity, specificity, PPV, NPV). A ranking system was implemented based on averaged ranks across all metrics, with lower ranks indicating better performance.

## 3. Results

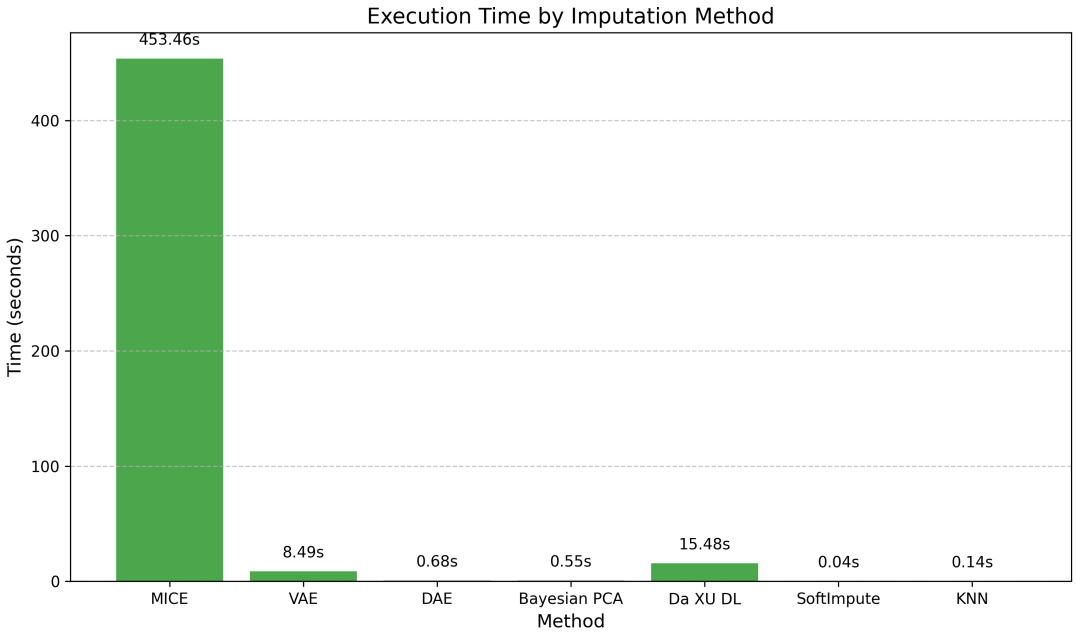


**Figure 2. Execution Time by Imputation Method.** Computational efficiency comparison across seven imputation methods applied to FACT-E quality-of-life data with moderate missingness rates (1.38%-43.12%). MICE required the longest execution time (453.46s), while SoftImpute was fastest (0.04s). The substantial time differences highlight trade-offs between computational efficiency and methodological complexity for clinical research applications.

The analytic dataset comprised 1,018 observations with 46 variables after preprocessing, representing data from 936 unique patients. Multiple baseline entries for some patients occurred due to clinical workflow variations (i.e., repeated assessments during pre-treatment evaluations or administrative scheduling), and all entries were retained given the small number of affected patients (n=73) and to preserve the natural data structure for realistic imputation evaluation. A total of 818 patients were excluded because they had no FACT-E data.

The target variables representing the complete FACT-E instrument (GP1-GP7, GS1-GS7, GE1-GE6, GF1-GF7, A_HN1-A_HN5, A_HN7, A_HN10, A_E1-A_E7, A_C6, A_C2, A_ACT11) exhibited moderate missingness at the row level, with individual

columns ranging from 1.38% (GP1) to 43.12% (GS7)[2] missing values (see Table 1) . Each variable had between 579-1,004 non-missing observations out of 1,018 total rows across the ordinal scale range 0-4, with low uniqueness ratios (0.5-0.9% unique values) characteristic of ordinal clinical scales.

The observed values demonstrated clinically meaningful patterns across the different FACT-E subscales. General physical wellbeing items (GP1-GP7) showed means ranging from 0.56 to 1.68, indicating generally low physical symptom burden in the observed population. Social/family wellbeing items (GS1-GS7) demonstrated higher means (2.11-3.49), suggesting better preserved social functioning. Emotional wellbeing items (GE1-GE6) showed intermediate values (0.67-2.26), while functional wellbeing items (GF1-GF7) demonstrated consistently higher means (2.22-2.57), indicating relatively preserved functional capacity. The esophageal cancer-specific items showed variable patterns, with head/neck items (A_HN1-A_HN10) ranging from 0.56 to 3.36, esophageal-specific items (A_E1-A_E7) ranging from 0.60 to 2.11, and additional concerns items showing intermediate values.

### *3.1. Computational Performance*

Execution times varied dramatically across methods (Fig. 2), reflecting fundamental differences in algorithmic complexity, methodological sophistication, and computational efficiency. KNN exhibited $O(n^2 \times m)$ complexity (0.14 seconds) for pairwise distance calculations across sparse data, with $n$ and $m$ being data-derived parameters. The deep learning approaches (VAE, DAE, and Da Xu et al. deep learning method) all exhibit $O(n \times d \times h \times e)$ complexity typical of neural network training, where $n$ and $d$ are data-derived (sample size and input dimension), while $h$ (hidden layer size) and $e$ (epoch count) are user-tuned hyperparameters. DAE required 0.68 seconds with a scikit-learn implementation, while the Da Xu et al. method required 15.48 seconds due to additional computational burden from patient-specific embedding learning and temporal similarity calculations. VAE completed training at 8.49 seconds, benefiting from GPU acceleration on the NVIDIA A100-SXM4-40GB hardware and efficient batch processing where available.

MICE required the greatest amount of time - 453.46 seconds - due to its $O(k \times m \times n \log n)$ complexity, where $k$ (iterations) and $m$ (variables) are method parameters, $n$ is data-derived sample size, and the $\log n$ term arises from multiple tree construction operations per iteration in the ensemble-based LightGBM implementation requiring sorted data structures for optimal split finding.

By comparison, the matrix factorization approaches demonstrated superior efficiency. SoftImpute required 0.04 seconds with $O(k \times \min(n, m)^2 \times r)$ complexity, where $r$ (rank) and $k$ (iterations) are method parameters, followed by Bayesian PCA at 0.55 seconds with $O(k \times n \times d)$ complexity due to its iterative gradient-based optimization over latent variables and loading matrices, where $k$ is the number of training iterations.

### *3.2. Distribution Preservation*

The KS test (Fig. 3) results revealed substantial differences in distribution preservation capabilities across methods. MICE demonstrated superior performance with the lowest KS statistics across most variables (typically ranging from 0.05-0.35), indicat-

[2]GS7 ("I am satisfied with my sex life") exhibited the highest missingness rate, likely reflecting patient reluctance to respond to intimate personal questions in clinical settings.

ing excellent preservation of the original data characteristics. Many MICE comparisons yielded p-values above 0.05 (Fig. S1), meaning the null hypothesis of identical distributions could not be rejected for numerous variables, demonstrating that imputed values closely matched the observed data distribution.

VAE, DAE, and Bayesian PCA exhibited moderately high KS statistics (typically 0.35-0.7 range) with p-values $< 0.1$ across virtually all variables, indicating significant distributional differences from the observed data. These methods showed consistent patterns of moderate distribution preservation challenges across all FACT-E subscales, suggesting some difficulty with maintaining the discrete ordinal nature of the clinical scales.

The Da Xu et al. deep learning method showed the most extreme distribution preservation problems, with KS statistics frequently reaching 0.85-0.95 across most variables, particularly poor performance on social/family variables, functional and esophageal-specific items. This represents the worst distribution preservation among all evaluated methods, suggesting severe overfitting or inadequate model architecture for these clinical domains.

SoftImpute and KNN demonstrated intermediate to good performance with more consistent results across different variable types. SoftImpute showed KS statistics generally in the 0.2-0.6 range with relatively consistent performance, while KNN demonstrated similar performance (typically 0.25-0.6 range) with less variability across dif-

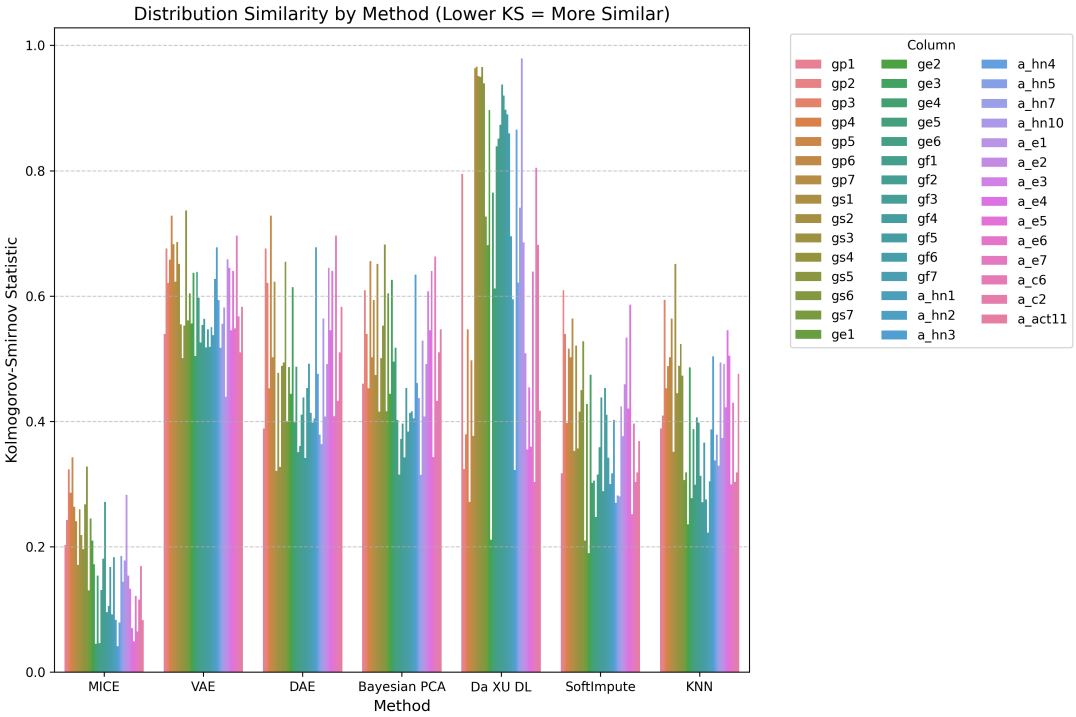


**Figure 3. Distribution Similarity by Imputation Method** Kolmogorov-Smirnov test statistics comparing observed versus imputed value distributions across 44 FACT-E variables (lower values indicate better distribution preservation). MICE demonstrated superior performance with consistently low KS statistics (mostly below 0.35), while Da Xu et al. method [17, 18] showed poor preservation with values frequently reaching 0.85-0.95. VAE, DAE, and Bayesian PCA exhibited moderate performance with most variables in the 0.35-0.7 range, while SoftImpute and KNN showed intermediate to good performance with most variables below 0.6. Results highlight substantial differences in methods' ability to maintain original data distributions.

ferent clinical scales, suggesting both methods provide reasonably robust performance across diverse FACT-E domains.

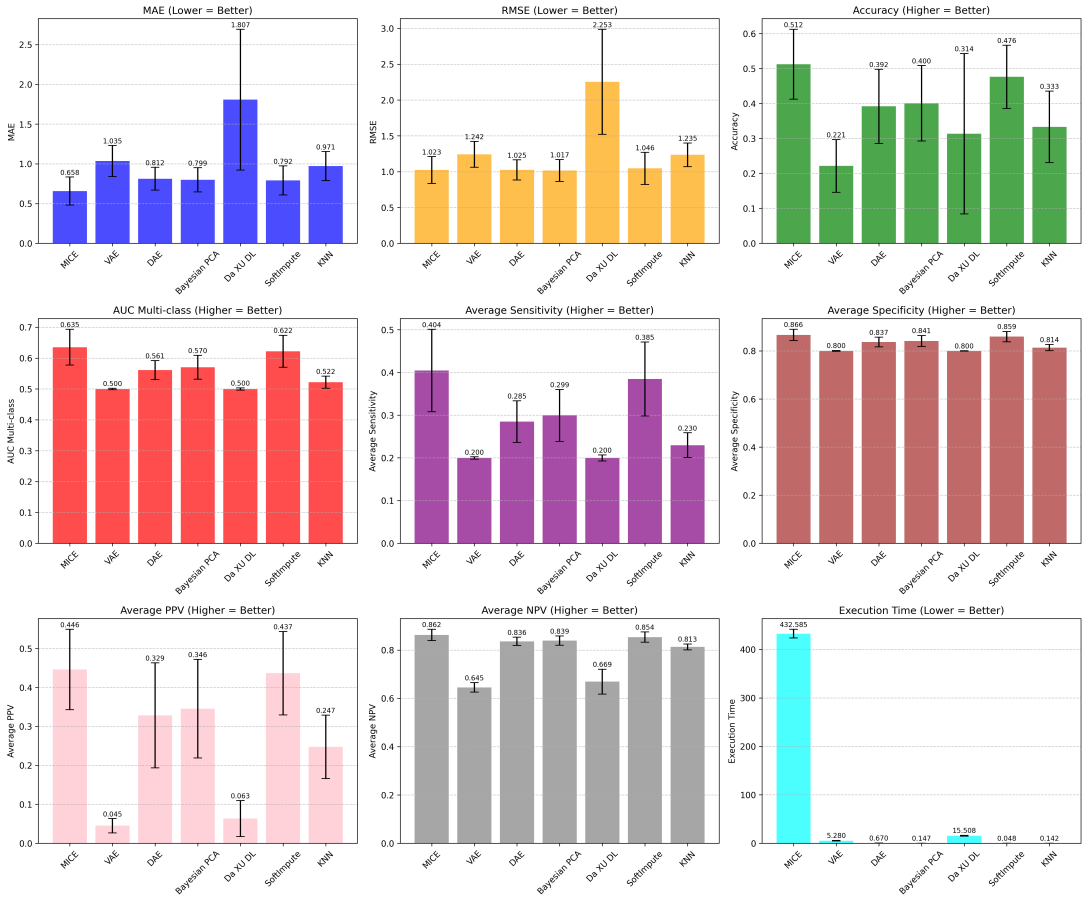


**Figure 4. Comprehensive Performance Evaluation Across Seven Imputation Methods.** Nine performance metrics from sparse cross-validation on 44 FACT-E variables with 1.38-43.12% missingness. MICE achieved the best performance with lowest MAE (0.658), RMSE (1.023), and highest accuracy (0.512) and AUC (0.635). SoftImpute showed competitive performance (accuracy: 0.476) with exceptional efficiency (0.048s). DAE, Bayesian PCA, and KNN demonstrated moderate performance, while Da Xu et al. method consistently exhibited poor accuracy (MAE: 1.807) and classification performance (accuracy: 0.314) despite high computational cost (15.5s). VAE showed inconsistent performance - showing satisfactory results in MAE and RMSE, but not the others. Error bars represent cross-validation standard deviations.

### *3.3. Cross-Validation Performance*

The sparse cross-validation framework successfully evaluated imputation accuracy across varying missingness patterns despite the computational intensity of validating 44 variables with 5-fold cross-validation across 7 methods. The results revealed consistent performance patterns across methods (Fig. 4).

MICE achieved the most stable performance with a MAE of 0.658 and a RMSE of 1.023, demonstrating robust accuracy regardless of the specific clinical domain or missingness pattern. VAE, DAE, Bayesian PCA, SoftImpute, and KNN showed similar moderate performance with MAE values of 1.035, 0.812, 0.799, 0.792, and 0.971, respectively; and RMSE values of 1.242, 1.025, 1.017, 1.046, and 1.235. DAE, Bayesian

PCA and SoftImpute demonstrated competitive accuracy the lowest MAE and RMSE among non-MICE methods, indicating that despite its computational efficiency, it maintains reasonable imputation quality.

The Da Xu et al. deep learning method exhibited the poorest continuous accuracy metrics with MAE of 1.807 and RMSE of 2.253, showing highly variable performance consistent with the distribution preservation findings. This poor performance can be attributed to several methodological factors: the patient embedding layer may have suffered from overfitting due to sparse patient representation, the temporal similarity loss component potentially introduced noise rather than meaningful signal in this cross-sectional study with single time-point data, and the complex architecture with patient-specific parameters may have been inadequately suited for the extreme sparsity patterns in FACT-E data where the model struggled to learn meaningful patient-specific representations from limited observed values.

### *3.4. Clinical Classification Performance*

MICE demonstrated superior classification performance with an accuracy of 0.512, outperforming all other methods. The method achieved balanced sensitivity (0.404) and high specificity (0.866), indicating a strong ability to correctly identify positive and negative cases across the ordinal categories. The multiclass AUC of 0.635 further confirmed MICE's superior discriminative ability. PPV was 0.446, while NPV was 0.862 (Fig. 4).

SoftImpute showed the second-best classification performance with an accuracy of 0.476 and a multiclass AUC of 0.622. DAE and Bayesian PCA showed similar moderate classification performance with accuracy values of 0.392 and 0.400, respectively; and AUC values of 0.561 and 0.570, respectively. While these two methods showed intermediate sensitivity (0.285-0.299), they maintained high specificity (0.837-0.841), indicating reliable identification of patients without specific symptoms. KNN showed moderate performance across most metrics as well, with accuracy of 0.333 and AUC of 0.522.

VAE demonstrated poor classification performance with accuracy of 0.221 and AUC of 0.500, with low sensitivity (0.200). Da Xu et al. method also showed poor performance with an accuracy of 0.314 and AUC of 0.500, with particularly highly variable unstable performance across many metrics, suggesting challenges in correctly identifying patients with specific symptom severities. Both methods show suboptimal PPV and NPV values.

### *3.5. Correlation Preservation Analysis*

The correlation preservation analysis revealed significant differences in how well each method maintained the natural relationships between FACT-E variables (Fig. 5, 6).

All methods showed good correlation preservation, with average correlation difference between 0.0374 and 0.0463 (Fig. 5, 6). To assess robustness of each model under increased sparsity, a critical evaluation since real-world clinical datasets often exhibit varying and potentially higher missingness rates, we conducted an additional correlation preservation analysis using a more challenging sparsity-conditioned dataset with 1,836 observations from 1,754 unique patients by including previously excluded 818 patients, where missingness ranged from 45.3% to 68.5% across FACT-E variables

(Fig. S2, S3).

VAE, DAE, Bayesian PCA, and KNN demonstrated superior correlation preservation with average differences of 0.0375, 0.0377, 0.0386, and 0.0382, respectively, maintaining correlation patterns nearly identical to the original data structure. SoftImpute showed moderate performance with an average difference of 0.0642, representing a

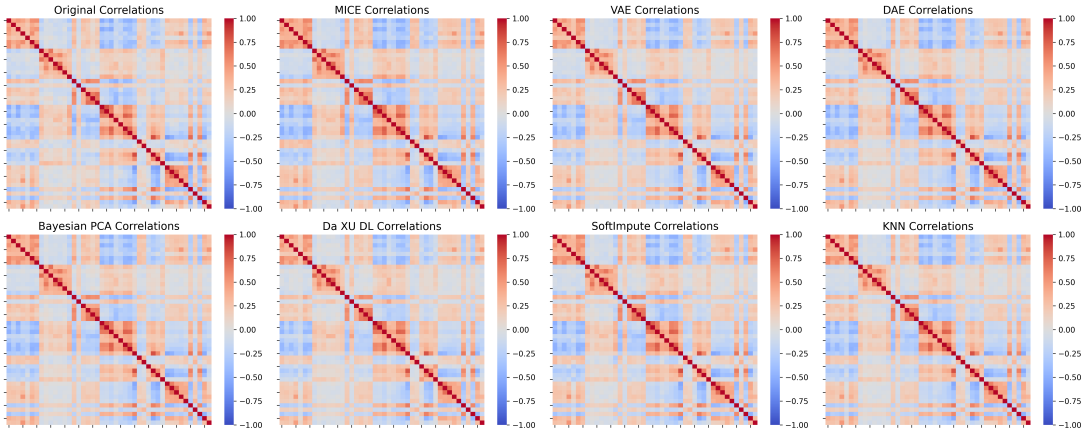


**Figure 5. Correlation Matrix Across Imputed Results and Baseline Correlations in the Original Data.** Heatmaps comparing correlation structures between 44 FACT-E variables for original data (baseline correlations observed in complete cases) and each imputation method. The variables are ordered as: GP1-GP7 (physical well-being), GS1-GS7 (social/family well-being), GE1-GE6 (emotional well-being), GF1-GF7 (functional well-being), A_HN1-A_HN5, A_HN7, A_HN10 (head/neck cancer-specific), A_E1-A_E7 (esophageal cancer-specific), A_C6, A_C2, A_ACT11 (additional cancer concerns). Red indicates positive correlations, blue indicates negative correlations, with intensity representing correlation strength (-1 to +1). All methods maintained correlation patterns reasonably similar to the original structure in the less sparse original dataset with 936 patients.

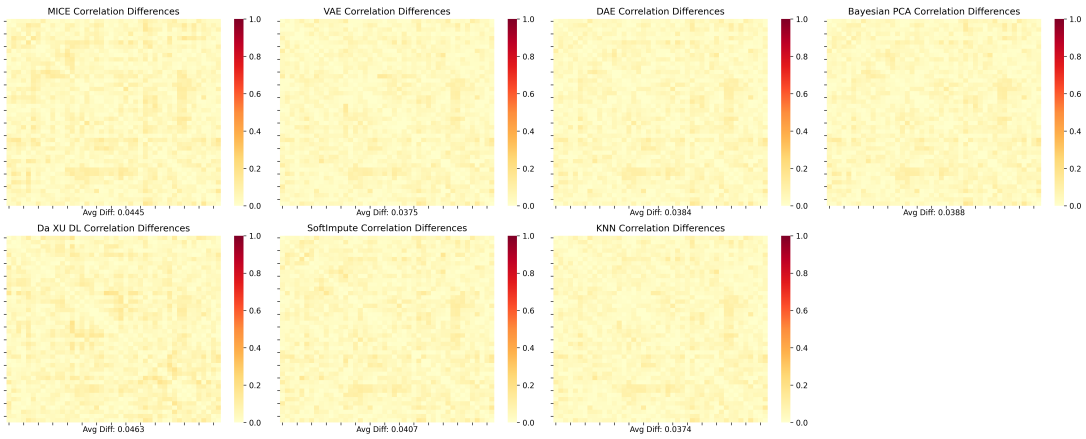


**Figure 6. Absolute Correlation Differences Between Original Data and Each Imputed Matrix.** Absolute differences between original and imputed correlation matrices (lower values indicate better preservation). Variable ordering is the same as Figure 5. VAE (0.0375), DAE (0.0384), Bayesian PCA (0.0388), and KNN (0.0374) achieved exceptional preservation with minimal deviation from original patterns. MICE (0.0445), and Da XU et al. method (0.0463), and SoftImpute (0.0407) showed good preservation. All methods demonstrated reasonable correlation preservation with relatively small differences, indicating that each approach maintains clinically meaningful relationships between FACT-E subscales in the less sparse original dataset with 936 patients.

slight degradation compared to the less sparse original dataset. The correlation matrices after these imputations closely resembled the baseline correlations observed in complete cases, indicating that these methods successfully preserved the established relationships between different FACT-E subscales without introducing artificial associations.

MICE exhibited sensitivity to increased sparsity, with correlation preservation degrading to an average difference of 0.1773. While the method maintained the general structure of correlations between variables, it exhibited notable deviation from the original correlation patterns, particularly in the strength of certain associations. MICE maintains general structure but deviates in specific associations as a consequence of its iterative approach under sparse FACT-E data conditions. However, the overall correlation framework remained clinically recognizable and interpretable.

The Da Xu et al. deep learning method exhibited the poorest correlation preservation among all evaluated methods, with an average correlation difference of 0.3334, showing concerning correlation distortions. The method introduced strong artificial correlations that were not present in the original data, particularly affecting relationships between different FACT-E domains. This pattern suggests potential overfitting to patient-specific patterns rather than learning generalizable quality-of-life structures, which could compromise the clinical validity of imputed values.

**Table 2.** Comprehensive Summary of Imputation Method Performance Across All Evaluation Metrics

| Method | MAE | RMSE | Accuracy | AUC | Sensitivity | Specificity | PPV | NPV | Time (s) | Avg KS Stat | Avg KS p-value | Sparse Correlation Differences | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MICE | 0.658 | 1.023 | 0.512 | 0.635 | 0.404 | 0.866 | 0.446 | 0.862 | 453.457 | 0.171 | 0.388 | 0.177 | 2 |
| SoftImpute | 0.792 | 1.046 | 0.476 | 0.622 | 0.385 | 0.859 | 0.437 | 0.854 | 0.044 | 0.389 | 0.006 | 0.064 | 2.417 |
| Bayesian PCA | 0.799 | 1.017 | 0.4 | 0.57 | 0.299 | 0.841 | 0.346 | 0.839 | 0.548 | 0.492 | 1.31e-4 | 0.039 | 3.333 |
| DAE | 0.812 | 1.025 | 0.392 | 0.561 | 0.285 | 0.837 | 0.329 | 0.836 | 0.679 | 0.488 | 6.35e-4 | 0.038 | 3.833 |
| KNN | 0.971 | 1.235 | 0.333 | 0.522 | 0.23 | 0.814 | 0.247 | 0.813 | 0.135 | 0.404 | 0.006 | 0.038 | 4.167 |
| VAE | 1.035 | 1.242 | 0.221 | 0.5 | 0.2 | 0.8 | 0.045 | 0.645 | 8.486 | 0.591 | 6.42e-06 | 0.038 | 6 |
| Da Xu DL | 1.807 | 2.253 | 0.314 | 0.5 | 0.2 | 0.8 | 0.063 | 0.669 | 15.479 | 0.68 | 0.002 | 0.333 | 6.25 |

*Note: Methods ranked by average rank across all metrics (lower is better). MAE = Mean Absolute Error, RMSE = Root Mean Square Error, AUC = Area Under Curve (multiclass), PPV = Positive Predictive Value, NPV = Negative Predictive Value, KS Stat = Kolmogorov-Smirnov statistic for distribution preservation (lower is better), KS p-value = significance level for distribution similarity test (higher is better). All continuous metrics are rounded to 3 decimal places.*

### *3.6. Ranked Comparison*

Each method was assigned a rank (1=best, 7=worst) for each of the 12 performance metrics (MAE, RMSE, Accuracy, AUC, Sensitivity, Specificity, PPV, NPV, Time, KS Statistic, KS p-value, correlation differences) where lower ranks indicate better performance for MAE, RMSE, time, KS statistic, and correlation differences, while higher ranks indicate better performance for all other metrics. The arithmetic mean was calculated as: Average Rank $= \frac{1}{12}\sum_{i=1}^{12} R_i$, where $R_i$ is the rank for metric $i$. Table 2 presents the ranking results across the complete evaluation framework.

## 4. Discussion

To the best of our knowledge, this study offers the first comprehensive evaluation of missing data imputation methods for PROs in a real-world cancer registry. We evaluated seven approaches to impute missing data from the FACT-E, a common patient-reported outcome measure used for esophageal cancer. Our large registry had moderate missingness rates ranging from 1.38% to 43.12%. Rather than pursuing method-specific optimization for marginal performance gains, we used standard,

generalizable configurations providing practical model selection guidance through performance baselines across 12 evaluation criteria, statistical significance testing for distribution preservation, and computational complexity analysis demonstrating clear performance hierarchies and trade-offs between accuracy, efficiency, and clinical utility. Such guidance is vital for understanding treatment effects, as incomplete PRO data can bias conclusions about treatment efficacy.

The results demonstrate that MICE emerged as the superior method across most evaluation criteria, attributed to its iterative chained equations architecture leveraging LightGBM's gradient boosting to capture complex non-linear relationships while maintaining distributional fidelity via ensemble averaging. MICE achieved exceptional distribution preservation and non-significant differences for multiple variables. This translates into superior clinical classification accuracy, sensitivity, and multiclass AUC, plus the lowest continuous error metrics, compared to the other methods tested. This makes it ideal for clinical applications needing accurate symptom severity categorization. However, MICE showed sensitivity to increased sparsity due to iterative uncertainty propagation amplifying correlation distortions under sparse data conditions. In contrast, VAE, DAE, Bayesian PCA, and KNN maintained robust correlation preservation through their respective methodological strengths: VAE's probabilistic latent space representation, DAE's denoising regularization, Bayesian PCA's probabilistic loading matrix, and KNN's similarity-based local structure preservation. Additionally, MICE's computational demands limit its use in large-scale studies, being about 28 times slower than Da Xu et al. method and over 9,800 times slower than SoftImpute. Matrix factorization methods like SoftImpute demonstrate superior computational scalability for large-scale clinical studies, while iterative ensemble methods face scalability challenges due to multiplicative complexity factors. Deep learning methods showed mixed results: DAE performed reasonably, while the Da XU et al. method exhibited instability with poor overall performance under increased sparsity conditions. SoftImpute balanced efficiency and accuracy via low-rank approximation capturing global correlations, suiting large-scale clinical studies, while KNN provided stable, interpretable results through similarity-based imputation.

Based on the results from this study, we offer several recommendations: MICE for high-stakes applications prioritizing accuracy; SoftImpute for routine clinical use balancing accuracy and efficiency; and Bayesian PCA for research valuing correlation preservation and uncertainty quantification. Deep learning methods for sparse clinical data require careful architectural design: the complex Da XU et al. approach showed poor performance despite high computational cost, while simpler VAE underperformed compared to well-designed DAE, indicating that balanced architectural complexity is essential for effective clinical imputation. The rankings highlights performance hierarchies to guide method selection by research priorities and constraints, serving as a template for future missing data research in PROs to enhance quality-of-life study validity and reliability in cancer care.

Despite comprehensive evaluation across seven imputation methods, the overall performance levels reveal fundamental challenges in PRO data imputation. Even the best-performing method (MICE) achieved only 51.2% classification accuracy, indicating that accurate imputation of PRO remains problematic across all evaluated approaches. The modest performance across all methods likely reflects the complexity of PRO data that extends beyond MAR assumptions. For instance, GS7 ("I am satisfied with my sex life") exhibited disproportionate missingness (43.12%), suggesting MNAR patterns where patients may systematically avoid responding to sensitive questions. This is why the preservation of correlation structure is particularly important for PRO

data, as the established relationships between physical, social, emotional, functional, and disease-specific subscales reflect well-documented clinical patterns in quality-of-life assessment that extend beyond MAR assumptions. Such systematic missingness patterns indicate that traditional imputation approaches may be fundamentally limited for certain PRO domains. Given these limitations, researchers should exercise caution when applying any imputation method to PRO data, consider reporting results both with and without imputation, conduct sensitivity analyses to assess the impact of different imputation approaches on study conclusions, and acknowledge the inherent uncertainty in imputed values when making clinical inferences.

These findings also suggest that synthetic data simulation approaches may have limited applicability to PRO data, given the complex missingness patterns and modest imputation performance observed.

Several limitations should be considered when interpreting these results. Our analysis focused on baseline cross-sectional data from a single-center dataset, potentially limiting generalizability. The framework used artificial missingness under MAR assumptions, but real-world MNAR patterns could affect method performance. Using standard configurations ensured practical applicability but may not reflect optimal performance from extensive hyperparameter tuning.

This comprehensive benchmarking of seven imputation methods for PRO data provides evidence-based performance rankings and practical selection criteria for researchers, while revealing fundamental limitations in PRO imputation accuracy that necessitate cautious application and transparent reporting of methodological choices in cancer quality-of-life research.

### Authors' contributions

Y.J.K. designed and implemented the methodology and analyzed the results. M.D. collected the data. Y.S. and S.D. contributed equally to this work as second authors. E.A.M. and R.T.C. conceived and designed the study and supervised the production of results. Y.J.K., E.A.M. and R.T.C. were involved in writing and revising the manuscript and have approved the final version.

## 5. Data availability

The datasets analyzed in this study are not publicly available due to patient privacy and institutional data sharing restrictions. Requests for data access may be directed to the corresponding author and will be considered on a case-by-case basis following institutional review board approval and establishment of appropriate data use agreements.

## 6. Code availability

The source code and results reproduction are publicly available at our GitHub repository (https://github.com/Crump-Lab/missing_imputation).

**Acknowledgement(s)**

Computations were made on the supercomputer Narval, managed by Calcul Québec and the Digital Research Alliance of Canada (Alliance). The operation of this supercomputer is funded by the Alliance, le ministère de l'Économie, de l'Innovation et de l'Énergie du Québec (MEIE) and le Fonds de recherche du Québec (FRQ).

**Funding**

This research was funded with generous support from the Rossy Cancer Network's Cancer Quality and Innovation program.

**Ethics approval and consent to participate**

This study was approved by the McGill University Health Centre's Centre for Applied Ethics. All patients in the EGDB Bank provided informed consent for their data to be used in research studies.

**References**

[1] Melanie L Bell, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. Handling missing data in rcts; a review of the top medical journals. *BMC medical research methodology*, 14(1):118, 2014.

[2] Joseph M Unger, Riha Vaidya, Dawn L Hershman, Lori M Minasian, and Mark E Fleury. Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation. *JNCI: Journal of the National Cancer Institute*, 111(3):245–255, 2019.

[3] Fiona M Shrive, Heather Stuart, Hude Quan, and William A Ghali. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC medical research methodology*, 6(1):57, 2006.

[4] Lene Kongsgaard Nielsen, Rebecca Mercieca-Bebber, Sören Möller, Louise Redder, Mary Jarden, Christen Lykkegaard Andersen, Henrik Frederiksen, Asta Svirskaite, Trine Silkjær, Morten Saaby Steffensen, et al. Relationship between reasons for intermittent missing patient-reported outcomes data and missing data mechanisms. *Quality of Life Research*, 33(9):2387–2400, 2024.

[5] Olawale F Ayilara, Lisa Zhang, Tolulope T Sajobi, Richard Sawatzky, Eric Bohm, and Lisa M Lix. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and quality of life outcomes*, 17(1):106, 2019.

[6] Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402–406, 2013.

[7] D. Rubin. *Multiple imputation for nonresponse in surveys.* John Wiley Sons, New York, 1987.

[8] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data.* John Wiley & Sons, 2019.

[9] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[10] Bryce B Reeve, Ron D Hays, Jakob B Bjorner, Karon F Cook, Paul K Crane, Jeanne A Teresi, David Thissen, Dennis A Revicki, David J Weiss, Ronald K Hambleton, et al.

Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information system (promis). *Medical care*, 45(5):S22–S31, 2007.

[11] D Cella, JL Beaumont, Z Liu, AT Phan, S Choi, and JC Yao. Neuroendocrine tumor disease state and promis health-related quality of life scores. *Journal of Clinical Oncology*, 28(15_suppl):6118–6118, 2010.

[12] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[13] Gail Darling, David T Eton, Joanne Sulman, Alan G Casson, and David Cella. Validation of the functional assessment of cancer therapy esophageal cancer subscale. *Cancer*, 107(4):854–863, 2006.

[14] Kimberly Webster, David Cella, and Kathleen Yost. The f unctional a ssessment of c hronic i llness t herapy (facit) measurement system: properties, applications, and interpretation. *Health and quality of life outcomes*, 1(1):79, 2003.

[15] Sonya L Eremenco, David Cella, and Benjamin J Arnold. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Evaluation & the health professions*, 28(2):212–232, 2005.

[16] AE Bonomi, DF Cella, EA Hahn, K Bjordal, B Sperner-Unterweger, L Gangeri, B Bergman, J Willems-Groot, P Hanquet, and R Zittoun. Multilingual translation of the functional assessment of cancer therapy (fact) quality of life measurement system. *Quality of Life research*, 5(3):309–320, 1996.

[17] Da Xu, Paul Jen-Hwa Hu, Ting-Shuo Huang, Xiao Fang, and Chih-Chin Hsu. A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management. *Journal of Biomedical Informatics*, 111:103576, 2020.

[18] Da Xu, Jessica Qiuhua Sheng, Paul Jen-Hwa Hu, Ting-Shuo Huang, and Chih-Chin Hsu. A deep learning–based unsupervised method to impute missing values in patient records for improved management of cardiovascular patients. *IEEE Journal of Biomedical and Health Informatics*, 25(6):2260–2272, 2020.

[19] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[20] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[21] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[22] Christopher Bishop. Bayesian pca. *Advances in neural information processing systems*, 11, 1998.

[23] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

[24] Trevor Hastie, Rahul Mazumder, Jason Lee, and Reza Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research*, 16:3367–3402, 2015.

[25] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[26] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[27] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.

[28] Donald B Rubin. Multiple imputation. In *Flexible imputation of missing data, second edition*, pages 29–62. Chapman and Hall/CRC, 2018.

[29] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[31] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.

[32] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[33] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[35] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[36] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[37] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[38] Lovedeep Gondara and Ke Wang. Mida: Multiple imputation using denoising autoencoders. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 260–272. Springer, 2018.

[39] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

[40] David F Cella, David S Tulsky, George Gray, Bernie Sarafian, Elizabeth Linn, Amy Bonomi, Margaret Silberman, Suzanne B Yellen, Patsy Winicour, and Judy Brannon. The functional assessment of cancer therapy scale: development and validation of the general measure. *Journal of clinical oncology*, 11(3):570–579, 1993.

[41] Byanne Malluhi, Nour Basha, Radhia Fezai, Gasim Ibrahim, Hanif A Choudhury, Mohamed Challiwala, Hazem Nounou, Nimir Elbashir, and Mohamed Nounou. Multiscale bayesian pca for robust process modeling of a fischer–tropsch bench scale process. *Chemometrics and Intelligent Laboratory Systems*, 240:104921, 2023.

[42] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[43] Travis Brady. py-soft-impute: https://github.com/travisbrady/py-soft-impute.

[44] Virginia Klema and Alan Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176, 1980.

[45] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[46] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[47] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*, volume 103. Springer, 2013.

[48] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

# Appendices

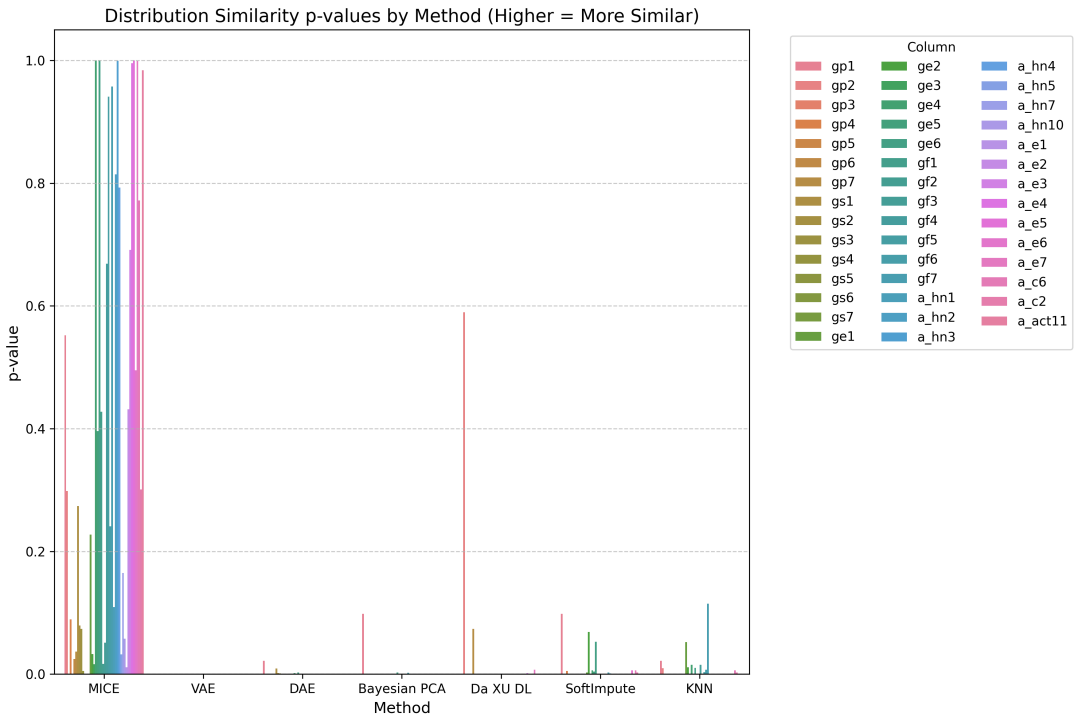


**Figure S1. Distribution Similarity p-values by Imputation Method** Kolmogorov-Smirnov test p-values comparing observed versus imputed value distributions across 44 FACT-E variables (higher values indicate better distribution preservation). MICE demonstrated superior performance with many p-values above 0.05, indicating failure to reject the null hypothesis of identical distributions for multiple variables. Most other methods showed p-values near zero, suggesting significant distributional differences from observed data. Results complement the KS statistics in demonstrating MICE's superior distribution preservation capabilities.

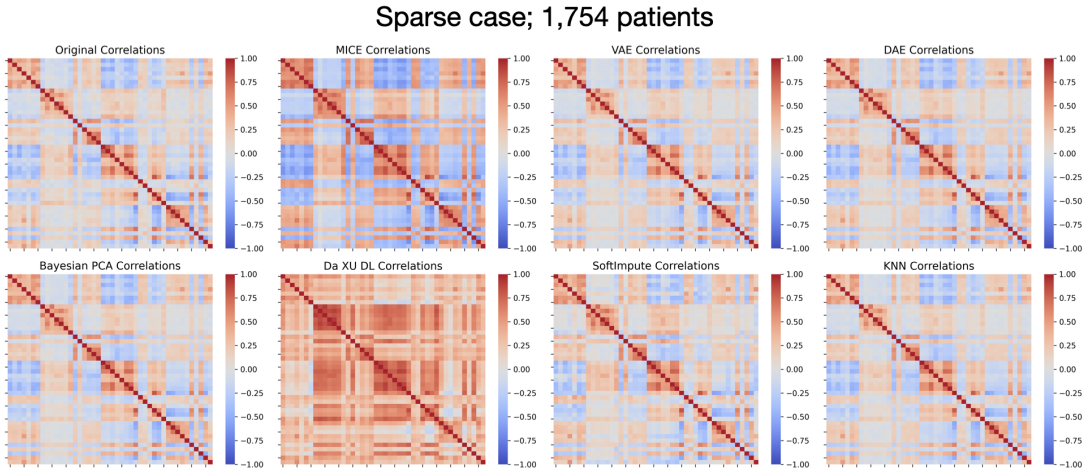


**Figure S2. Correlation Matrix Across Imputed Results and Baseline Correlations in the More Sparse Data.** Heatmaps comparing correlation structures between 44 FACT-E variables for original data (baseline correlations observed in complete cases) and each imputation method under increased missingness conditions (45.3-68.5%). Variable ordering is the same as Figure 5. Red indicates positive correlations, blue indicates negative correlations, with intensity representing correlation strength (-1 to +1). VAE, DAE, Bayesian PCA, and KNN maintained correlation patterns similar to the original structure, while SoftImpute showed moderate preservation. MICE exhibited notable deviations, and Da Xu et al. Da Xu et al. method [17, 18] introduced substantial artificial correlations (excessive red regions) not present in the original data, particularly affecting inter-domain relationships between FACT-E subscales.

**Figure S3. Absolute Correlation Differences Between More Sparse Data and Each Imputed Matrix.** Absolute differences between original and imputed correlation matrices (lower values indicate better preservation) under challenging sparsity conditions. Variable ordering is the same as Figure 5. VAE (0.0375), DAE (0.0377), Bayesian PCA (0.0386), and KNN (0.0382) achieved exceptional preservation with minimal deviation from original patterns. SoftImpute (0.0642) showed moderate preservation, while MICE (0.1773) demonstrated degraded performance. Da Xu et al. method [17, 18] exhibited poor preservation (0.3334) with substantial correlation distortions across all variable pairs, indicating introduction of artificial relationships that compromise clinical interpretability of the quality-of-life structure.