PPMI-Benchmark: A Dual Evaluation Framework for Imputation and Synthetic Data Generation in Longitudinal Parkinson's Disease Research

Conference Paper · July 2025

DOI: 10.5220/0013649700003967

4 authors, including:

Moad Hani
University of Mons
22 PUBLICATIONS **66** CITATIONS
SEE PROFILE

Mahmoudi Saïd
University of Mons
256 PUBLICATIONS **3,442** CITATIONS
SEE PROFILE

PPMI-Benchmark: A Dual Evaluation Framework for Imputation and Synthetic Data Generation in Longitudinal Parkinson's Disease Research

Moad Hani[1][a], Nacim Betrouni[2][b], Saïd Mahmoudi[1][c] and Mohammed Benjelloun[1][d]
[1]*Department of Computer Engineering and Management, University of Mons (UMONS), Belgium*
[2]*Univ. Lille, Inserm, CHU Lille, U1172 – LilNCog – Lille Neuroscience & Cognition, France*

Keywords: Parkinson's Disease, Longitudinal Imputation, Synthetic Data Generation, Clinical Bias Mitigation, HyperImpute, CTGAN, Sliced Wasserstein Distance, PPMI Dataset, Healthcare AI Governance, Multi-Center Reproducibility.

Abstract: Longitudinal datasets like the Parkinson's Progression Markers Initiative (PPMI) face critical challenges from missing data and privacy constraints. This paper introduces PPMI-Benchmark, the first comprehensive framework evaluating 12 imputation methods and 6 synthetic data generation techniques across clinical, demographic, and biomarker variables in Parkinson's disease research. We implement advanced methods including HyperImpute (ensemble optimization), VaDER (variational deep embedding), and conditional tabular GANs (CTGAN), evaluating them through novel metrics integrating sliced Wasserstein distance ($d_{SW} = 0.039 \pm 0.012$), temporal consistency analysis, and clinical validity constraints. Our results demonstrate HyperImpute's superiority in imputation accuracy (MAE=5.16 vs. 5.19–5.57 for baselines), while CTGAN achieves optimal distribution fidelity (SWD=0.039 vs. 0.062–0.146). Crucially, we reveal persistent demographic biases in cognitive scores, with age-related imputation errors increasing by 23% for patients over 70, and propose mitigation strategies. The framework provides actionable guidelines for selecting data completion strategies based on missingness patterns (MCAR/MAR/MNAR), computational constraints, and clinical objectives, advancing reproducibility and fairness in neurodegenerative disease research. Validated on 1,483 PPMI participants, our work addresses emerging needs in healthcare AI governance and synthetic data interoperability for multi-center collaborations.

1 INTRODUCTION

The Parkinson's Progression Markers Initiative (PPMI) dataset has revolutionized neurodegenerative disease research through its comprehensive longitudinal tracking of clinical, imaging, and biomarker data. However, over 42% of variables exhibit missingness rates exceeding 25% by Visit 4 (V04), with critical motor assessments (UPDRS-III) missing in 38% of late-stage patients (Marek et al., 2011). This pervasive missing data presents significant challenges for clinical research, as incomplete records compromise the reliability and validity of downstream analyses, potentially leading to biased conclusions and reduced statistical power.

Traditional approaches to handling missing data in longitudinal Parkinson's disease studies face three fundamental challenges:

1. **Temporal Complexity:** The neurodegenerative progression creates non-linear trajectories that simple imputation methods fail to capture (Postuma et al., 2015). Our analysis shows 38% greater variance in later visit imputations (V06–V09) compared to baseline.
2. **Clinical Plausibility:** Motor (Unified Parkinson's Disease Rating Scale Part III, UPDRS-III) and cognitive (Montreal Cognitive Assessment, MoCA) scores require strict physiological bounds (0–108 and 0–30 respectively) that 22% of baseline methods violated in validation.
3. **Data Heterogeneity:** Multimodal integration of demographic (age, education), clinical (MDS-UPDRS), and biomarker (CSF α-synuclein) variables demands specialized handling of missingness patterns.

[a] https://orcid.org/0000-0003-2342-495X
[b] https://orcid.org/0000-0003-1086-5502
[c] https://orcid.org/0000-0001-8272-9425
[d] https://orcid.org/0000-0002-4020-7327

Hani, M., Betrouni, N., Mahmoudi, S., Benjelloun and M.
PPMI-Benchmark: A Dual Evaluation Framework for Imputation and Synthetic Data Generation in Longitudinal Parkinson's Disease Research.
DOI: 10.5220/0013649700003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 246-259
ISBN: 978-989-758-758-0; ISSN: 2184-285X

Recent methodological advances in both imputation and synthetic data generation offer promising solutions. Ensemble approaches like HyperImpute (Jarrett et al., 2022) dynamically adapt to feature-specific missingness patterns, while deep learning methods such as Variational Deep Embedding with Recurrence (VaDER) leverage temporal dependencies in longitudinal data. Simultaneously, synthetic data generation techniques like Conditional Tabular GANs (CTGAN) (Xu et al., 2019) create privacy-preserving synthetic datasets that maintain statistical properties of the original data without exposing sensitive patient information.

This work builds upon our prior research submitted to the Delta Conference (Hani et al., 2025), which focused on context-aware imputation strategies for longitudinal Parkinson's data. The current paper makes three distinct clinical research contributions:

1. **Synthetic Data Expansion:** First comprehensive evaluation of 6 synthetic generation techniques (including CTGAN and RTVae) specifically adapted for neurodegenerative disease biomarkers, addressing critical privacy challenges in multi-center studies.
2. **Novel Evaluation Metrics:** Integration of sliced Wasserstein distance (d_{SW}) with temporal consistency analysis and clinical validity constraints, enabling multidimensional assessment of data completion methods beyond traditional error metrics.
3. **Bias Quantification Framework:** Systematic measurement of demographic disparities in imputation accuracy across age groups (23% increased MAE for patients $>$70) and education levels, informing equitable PD research practices.

Our methodology extends previous clinical data imputation benchmarks (Luo, 2022) by introducing progression-aware evaluation and biomarker-specific validation protocols.

The remainder of this paper is organized as follows: Section 2 reviews the state-of-the-art in both imputation techniques and synthetic data generation, with a focus on methods applicable to longitudinal clinical data. Section 3 details our methodology for data preparation, implementation of imputation and synthetic data frameworks, and evaluation metrics. Section 4 presents experimental results comparing method performance across different demographic groups and missingness patterns. Section 5 discusses implications for clinical research and highlights key trade-offs between computational efficiency, imputation accuracy, and fairness considerations. Finally, Section 6 concludes with recommendations for researchers working with incomplete longitudinal PD data.

2 STATE OF THE ART

2.1 Missing Data in Longitudinal Studies

Missing data represents a ubiquitous challenge in longitudinal clinical studies, particularly in Parkinson's disease research where patient attrition, incomplete assessments, and varying visit schedules create complex patterns of missingness. The mechanisms underlying missing data significantly impact the appropriate handling strategies and can be categorized into three types according to Rubin's framework (Little and Rubin, 2019). Missing Completely At Random (MCAR) occurs when the probability of missingness is unrelated to any observed or unobserved variables. Missing At Random (MAR) arises when missingness depends only on observed variables, while Missing Not At Random (MNAR) occurs when missingness depends on unobserved factors, including the missing values themselves (Graham, 2009).

In PD longitudinal studies, missingness often follows MNAR patterns, as patients with more severe symptoms may be less likely to complete certain assessments or attend follow-up visits. For instance, Van Buuren (van Buuren, 2018) demonstrated that cognitive decline in PD correlates with higher probabilities of missing data in subsequent cognitive assessments, creating systematic biases that simple imputation methods cannot adequately address. This selective missingness poses significant challenges for researchers, as many statistical methods assume MAR conditions for valid inference.

The consequences of inappropriate handling of missing data extend beyond statistical validity to clinical interpretation and decision-making. Studies by Luo et al. (Luo, 2022) demonstrated that deletion-based approaches can underestimate disease progression rates in PD by selectively removing patients with more rapid decline, while simple imputation methods often distort relationships between variables that are critical for understanding disease mechanisms. These distortions are particularly problematic in precision medicine initiatives that rely on accurate multivariate relationships to identify patient subgroups and personalize treatment approaches.

2.2 Imputation Methods for Longitudinal Data

The landscape of imputation methods spans from traditional statistical approaches to advanced machine learning techniques, each with distinct strengths and limitations for longitudinal clinical data. Traditional cross-sectional methods treat each time point independently, ignoring the temporal structure inherent in longitudinal data.

Mean and median imputation represent the simplest approaches, replacing missing values with central tendency measures of observed data. Despite their computational efficiency, these methods introduce significant statistical issues in longitudinal studies by artificially reducing variance and distorting correlation structures between variables (Donders et al., 2006). This distortion is particularly problematic in PD research, where relationships between motor symptoms, cognitive decline, and biomarkers provide critical insights into disease mechanisms and progression. Additionally, these methods cannot account for individual-specific trajectories, instead imposing average values that may be clinically implausible for specific patients based on their disease stage or demographic characteristics.

K-Nearest Neighbors (KNN) imputation represents an advancement over simple mean/median approaches by identifying similar cases based on distance metrics in feature space. While KNN can capture local patterns and relationships between variables, its performance deteriorates in high-dimensional spaces characteristic of comprehensive clinical datasets (Beretta and Santaniello, 2016). Furthermore, the selection of distance metrics and the number of neighbors (k) significantly influences imputation quality, with suboptimal choices leading to poor performance. In longitudinal studies, standard KNN implementations treat time points independently, failing to exploit temporal dependencies that could improve imputation accuracy.

Multiple Imputation by Chained Equations (MICE) has emerged as a powerful approach for complex clinical datasets by modeling each variable conditionally on others through an iterative process (van Buuren, 2018). MICE creates multiple complete datasets, capturing uncertainty in imputed values through variability across imputations. This statistical framework preserves relationships between variables and produces valid standard errors for downstream analyses. However, MICE typically implements separate models for each variable, potentially missing complex interactions, and its sequential nature can be computationally intensive for large datasets with many variables (White et al., 2011).

Longitudinal methods explicitly incorporate temporal dependencies into the imputation process. Last Observation Carried Forward (LOCF) and Next Observation Carried Backward (NOCB) represent simplistic approaches that propagate observed values to missing time points. While computationally efficient, these methods introduce substantial bias in progressive conditions like PD by failing to account for disease trajectories (Molenberghs and Kenward, 2007). Studies by Engels and Diehr (Engels and Diehr, 2003) demonstrated that LOCF consistently underestimates disease progression rates in neurodegenerative conditions, leading to potentially misleading conclusions about treatment efficacy.

Linear interpolation provides a more sophisticated approach by estimating missing values as weighted averages of adjacent observations. While this method captures linear trends between time points, it struggles with irregular visit schedules and cannot account for non-linear progression patterns common in PD (Diggle et al., 2002). Kalman filtering extends this concept by incorporating state-space modeling to estimate missing values based on system dynamics over time, explicitly modeling both observed and latent variables that drive disease progression (Harvey, 1989). However, this approach requires careful specification of system dynamics and can be computationally intensive.

Linear Mixed Models (LMM) represent a statistically rigorous approach for longitudinal imputation by incorporating both fixed and random effects to model individual-specific trajectories (Laird and Ware, 1982). LMMs accommodate irregularly spaced observations, account for correlation within subjects, and provide valid inference under MAR assumptions. Studies by Verbeke and Molenberghs (Verbeke and Molenberghs, 2000) demonstrated superior performance of LMM-based imputation compared to simpler approaches in longitudinal clinical trials. However, LMMs typically assume linear trajectories and normally distributed errors, which may not capture complex progression patterns in heterogeneous conditions like PD.

Recent advances in machine learning have introduced more flexible approaches to imputation. HyperImpute, developed by Jarrett et al. (Jarrett et al., 2022), uses ensemble optimization to adaptively select imputation strategies for each variable based on cross-validation performance. This approach combines the strengths of multiple methods while mitigating their individual weaknesses, consistently outperforming single-method approaches in heterogeneous clinical datasets. By integrating temporal con-

sistency constraints, HyperImpute preserves plausible progression trajectories essential for PD modeling.

Variational Deep Embedding with Recurrence (VaDER) represents a significant advancement in imputation for longitudinal clinical data. This approach extends traditional variational autoencoders with recurrent neural network components to capture both cross-sectional relationships and temporal dependencies (Fortuin et al., 2020). VaDER learns a latent representation of the data that encodes both individual patient characteristics and disease progression patterns, enabling more accurate imputation of missing values that respect both variable relationships and temporal trends. Comparative studies by Alaa et al. (Alaa et al., 2017) demonstrated that deep generative models like VaDER outperform conventional approaches for variables with complex non-linear relationships and temporal dependencies.

2.3 Synthetic Data Generation Techniques

Synthetic data generation offers a complementary approach to imputation by creating entirely new datasets that maintain statistical properties of the original data while enhancing privacy protection. This capability is particularly valuable for facilitating multi-center research collaborations without exposing sensitive patient information.

Variational Autoencoders (VAEs) represent one of the first deep generative models successfully applied to healthcare data synthesis (Kingma and Welling, 2014). VAEs learn a lower-dimensional latent representation of the data through an encoder network, then generate synthetic samples by sampling from this latent space and transforming through a decoder network. While standard VAEs capture complex distributions and non-linear relationships between variables, they struggle with mixed categorical and continuous features common in clinical datasets and lack mechanisms to incorporate temporal dependencies essential for longitudinal data.

Conditional Tabular GANs (CTGAN), developed by Xu et al. (Xu et al., 2019), address several limitations of traditional generative models for tabular data. CTGAN employs a conditional generator architecture with mode-specific normalization and training-by-sampling to handle the challenges of mixed data types and imbalanced categorical distributions. Clinical applications by Torfi and Fox (Torfi and Fox, 2020) demonstrated that CTGAN preserves statistical relationships critical for disease modeling while maintaining differential privacy guarantees. The conditional nature of CTGAN allows incorporation of temporal dependencies by conditioning generation on previous time points, making it particularly suitable for longitudinal datasets.

Recurrent Temporal Variational Autoencoders (RTVAEs) explicitly model temporal dynamics through recurrent neural network components integrated with VAE architectures (Yingzhen and Mandt, 2018). This approach captures both cross-sectional relationships and longitudinal progression patterns, generating synthetic trajectories that maintain temporal consistency. Studies by Moor et al. (Moor et al., 2020) demonstrated that RTVAEs preserve clinically relevant temporal patterns in synthetic ICU time-series data, enabling more accurate predictive modeling than cross-sectional approaches.

The Generative Adversarial Imputation Network (GAIN), proposed by Yoon et al. (Yoon et al., 2018), represents a hybrid approach that combines principles from both imputation and synthetic data generation. GAIN adapts the GAN framework to the imputation task by treating missing values as masked components that the generator must reconstruct while the discriminator distinguishes between observed and imputed values. This adversarial training process encourages the generator to produce realistic imputations that maintain the joint distribution of the data. However, evaluations by Mattei and Frellsen (Mattei and Frellsen, 2019) revealed limitations in GAIN's ability to capture complex dependencies in heterogeneous clinical datasets.

The Missing Data Importance-Weighted Autoencoder (MIWAE), developed by Mattei and Frellsen (Mattei and Frellsen, 2019), extends VAEs to handle missing data through importance weighting of partial observations. Unlike traditional imputation methods that produce point estimates, MIWAE generates distributions of possible values for missing entries, capturing uncertainty in a statistically principled manner. Comparative evaluations on healthcare datasets demonstrated MIWAE's superior performance in preserving distributional characteristics compared to deterministic approaches, particularly for variables with complex multimodal distributions.

Despite these advances, evaluating synthetic data quality remains challenging. Traditional metrics like precision, recall, and F1-score measure discriminative performance but fail to capture how well the synthetic data preserves the joint distribution of the original dataset (Jordon et al., 2019). Recent work by Choi et al. (Choi et al., 2017) introduced evaluation frameworks specifically designed for healthcare synthetic data, incorporating clinical plausibility constraints and domain-specific utility measures alongside distribution fidelity metrics.

3 METHODOLOGY

3.1 Dataset and Preprocessing

The PPMI dataset contains 4,217 participants across PD, prodromal, and control cohorts; our analysis focuses on the PD (N=891) and prodromal (N=592) groups (total N=1,483), which show the most complex and clinically relevant longitudinal missingness. PD participants are newly diagnosed, untreated, and DAT-deficit confirmed, while prodromal cases are at risk due to clinical features or genetic variants (e.g., *SNCA*, *LRRK2*, *GBA*) (Marek et al., 2011; PPM, ; Marek et al., 2018).

Our analysis focuses on 92 variables spanning multiple domains:

Motor assessments include the Movement Disorder Society Unified Parkinson's Disease Rating Scale Part III (MDS-UPDRS-III, coded as NP3TOT), which evaluates motor function through clinician-administered tests across 33 items covering tremor, rigidity, bradykinesia, and postural stability. Scores range from 0-108, with higher values indicating greater motor impairment (Goetz et al., 2008).

Cognitive evaluations include the Montreal Cognitive Assessment (MoCA, coded as MCATOT), a 30-point screening instrument assessing multiple cognitive domains including executive function, visuospatial abilities, attention, language, and memory. Scores below 26 indicate cognitive impairment, with values progressively decreasing as cognitive decline advances (Nasreddine et al., 2005).

Demographic variables include age, sex, education level, and disease duration, which significantly influence both progression trajectories and assessment scores. Biomarkers encompass cerebrospinal fluid (CSF) measures of α-synuclein, amyloid-β, and tau proteins that reflect underlying pathophysiological processes (Kang et al., 2013).

Our analysis of missingness patterns revealed systematic variations across visits and cohorts (Table 1). Demographic variables showed heterogeneous missingness (0.5-90.5%) compared to clinical assessments like UPDRS scores (16.4-99.7%) and MoCA scores (10.3-100%). These patterns informed our imputation strategy and helped identify potential sources of demographic and temporal bias.

Data preprocessing involved handling outliers through z-score standardization and removal of physiologically implausible values. Categorical variables underwent label encoding, while continuous variables were normalized to ensure comparable scales across measurements. The dataset was split using stratified sampling into training (80%) and validation (20%) cohorts, maintaining the distribution of PD, prodromal groups. We implemented 5-fold cross-validation to ensure reliable performance estimates and reduce overfitting risk.

Table 1: Missing Data Patterns in PPMI Dataset for Selected Variables (PD and Prodromal Cohorts), Across Visits

Feature	Cohort	V02	V04	V06
Age	PD	22.0%	14.3%	19.8%
	Prodromal	25.0%	9.0%	22.0%
UPDRS	PD	16.4%	22.0%	29.2%
	Prodromal	12.8%	34.5%	44.9%
MoCA	PD	0.0%	91.9%	29.0%
	Prodromal	0.0%	100.0%	44.4%

3.2 Imputation Framework

We implemented a comprehensive imputation framework evaluating 12 methods across three categories: cross-sectional, longitudinal, and advanced approaches. The imputation process was formalized as:

$$X_{imp}^{(t)} = f_\theta(X_{obs}^{(t)}, \mathcal{M}^{(t)}, X_{hist}^{(1:t-1)}) \quad (1)$$

where $X_{hist}^{(1:t-1)}$ represents historical data up to visit $t-1$, $\mathcal{M}^{(t)}$ the missingness mask, and f_θ the imputation function with parameters θ.

Cross-sectional methods included mean imputation, median imputation, K-Nearest Neighbors (KNN) with $k = 5$, and Multiple Imputation by Chained Equations (MICE) with 10 iterations. For KNN imputation, we employed Euclidean distance metrics with MinMax scaling to ensure comparable feature contributions to similarity calculations.

Longitudinal methods incorporated temporal dependencies through Last Observation Carried Forward (LOCF), Next Observation Carried Backward (NOCB), linear interpolation, Kalman filtering, and Linear Mixed Models (LMM). The LMM implementation used both fixed effects (for population-level trends) and random effects (for patient-specific trajectories) with the following specification:

$$y_{ij} = X_{ij}\beta + Z_{ij}b_i + \varepsilon_{ij} \quad (2)$$

where y_{ij} represents the observation for subject i at time j, X_{ij} and Z_{ij} are design matrices for fixed and random effects, β represents fixed effect parameters, b_i denotes subject-specific random effects, and ε_{ij} is the error term.

Advanced methods included HyperImpute and Variational Deep Embedding with Recurrence (VaDER). HyperImpute was implemented with 20 iterations of Bayesian hyperparameter optimization

to adaptively select optimal imputation strategies for each variable. The algorithm combines multiple base imputation methods (including KNN, random forests, and deep learning approaches) through stacked generalization, with weights optimized to minimize cross-validation error.

VaDER extends traditional variational autoencoders with recurrent neural network components to capture temporal dependencies in longitudinal data. The architecture employs a β-VAE formulation ($\beta = 0.8$) with GRU layers for temporal modeling, trained to maximize the modified evidence lower bound:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \cdot D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{3}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is the encoder, $p_\theta(\mathbf{x}|\mathbf{z})$ the decoder, and D_{KL} the Kullback-Leibler divergence between the approximate posterior and the prior distribution.

MissForest represents a non-parametric approach to missing data imputation based on random forests. The algorithm iteratively trains a random forest on observed values to predict missing values for each variable, cycling through variables until convergence. This method offers advantages for mixed-type data through its ability to handle continuous and categorical variables simultaneously without requiring distributional assumptions. Additionally, MissForest naturally captures non-linear relationships and interactions between variables, which is particularly valuable for clinical data with complex interdependencies.

3.3 Synthetic Data Generation Framework

Our synthetic data generation framework implemented six methods, with particular emphasis on CTGANs. The CTGAN architecture incorporates mode-specific normalization and conditional generation to handle mixed data types and preserve variable relationships. We enhanced the standard implementation with clinical range constraints through indicator functions in the loss term:

$$\mathcal{L}_{clinical} = \sum_{v=1}^{9} \mathbb{I}_{[0,30]}(\text{MCATOT}_v) + \mathbb{I}_{[0,108]}(\text{NP3TOT}_v) \tag{4}$$

These constraints ensure generated values for cognitive and motor scores remain within physiologically plausible ranges (0-30 for MoCA, 0-108 for UPDRS-III), critical for maintaining clinical validity of the synthetic data.

Recurrent Temporal Variational Autoencoders (RTVAEs) were implemented with bidirectional GRU cells to capture temporal dynamics, with skip connections to preserve information across the encoding-decoding process. The Missing Data Importance-Weighted Autoencoder (MIWAE) extends standard VAEs through importance weighting of partial observations, generating distributions of possible values for missing entries that capture uncertainty in a statistically principled manner.

We also implemented Generative Adversarial Imputation Networks (GAIN), normalizing flows (NFLOW), and autoregressive flow-based models (ARF) to provide comprehensive comparison across generative paradigms. Each method underwent hyperparameter optimization through grid search with 5-fold cross-validation to ensure optimal performance.

3.4 Evaluation Framework

Our evaluation protocol assessed imputation accuracy and synthetic data fidelity across three domains:

Statistical accuracy metrics included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2).

Distribution fidelity was evaluated using sliced Wasserstein distances and Kolmogorov-Smirnov tests.

Clinical validity focused on physiological range preservation for motor (0-108) and cognitive (0-30) scores.

To systematically evaluate demographic bias, we conducted stratified error analyses across three critical dimensions: age groups (< 70 vs. ≥ 70 years), education levels (≤ 12 vs. > 12 years of formal education), and disease duration categories (early-stage < 5 years vs. advanced ≥ 5 years). For each subgroup, we calculated relative error differentials (ΔMAE, ΔRMSE) using the formula:

$$\Delta\text{Metric} = \frac{\text{Metric}_{\text{subgroup}} - \text{Metric}_{\text{reference}}}{\text{Metric}_{\text{reference}}} \tag{5}$$

where positive values indicate bias amplification and negative values denote mitigation. This granular analysis revealed significant disparities in imputation accuracy across demographic strata, particularly for cognitive assessments in older patients with lower educational attainment.

All experiments were conducted on standardized hardware (Intel Xeon CPU @ 2.20GHz, 32GB RAM, NVIDIA Tesla V100 GPU) to ensure comparable benchmarks.

4 RESULTS

4.1 Imputation Performance Comparison

Table 2 presents comprehensive performance metrics for all evaluated imputation methods on the PPMI dataset. HyperImpute achieved the highest R^2 (0.260) and lowest RMSE (7.8934), demonstrating superior performance in explaining variance and minimizing error. Linear Mixed Models followed closely (R^2=0.256), with both methods significantly outperforming simpler approaches that do not account for temporal dependencies.

Table 2: Performance of Imputation Methods on PPMI Dataset.

Method	MAE ↓	RMSE ↓	R^2 ↑
Mean	5.2365	8.0025	0.2408
Median	5.2774	8.0892	0.2467
KNN (k=5)	5.2211	8.0619	0.2523
MICE (n=10)	5.2198	8.0055	0.2501
LOCF	5.3953	8.3133	0.2296
NOCB	5.5719	8.5690	0.1852
Linear Interpolation	5.5618	8.5166	0.1852
Kalman	5.4528	8.2647	0.2282
LMM	**5.1871**	**7.9627**	**0.2561**
HyperImpute (n=20)	**5.1618**	**7.8934**	**0.2600**
VaDER	5.2335	8.0815	0.2460
MissForest	5.3017	8.1254	0.2450

Note: MAE/RMSE rankings may diverge between methods due to error distribution differences. HyperImpute's lower RMSE despite comparable MAE reflects reduced error variance.

Our metric selection balances statistical rigor with clinical interpretability:

- **MAE/RMSE:** Preferred over percentage errors (SMAPE/RMSPE) due to zero-inflation in biomarker measurements (27% of UPDRS-III values = 0) (Armstrong, 1985). Absolute errors provide direct clinical interpretation (e.g., 5-point MAE in MoCA = moderate cognitive stage difference).
- **R^2:** Despite apparent low values (0.26), contextually significant for heterogeneous PD populations - surpasses 0.18-0.24 range from previous PPMI studies (Luo, 2022).
- **Sliced Wasserstein:** Captures distributional fidelity of non-motor symptoms better than KL divergence.

The XGBoost AUC evaluation (Table 5) follows recent synthetic data benchmarks (Jordon et al., 2019), using stratified sampling to handle class imbalance (1:1.5 prodromal:PD ratio, reflecting the actual cohort sizes of 592 prodromal and 891 PD participants).

Among cross-sectional methods, KNN demonstrated the best performance (R^2=0.2523), effectively capturing local patterns in the data. MICE performed similarly (R^2=0.2501) while offering the additional benefit of uncertainty quantification through multiple imputations. Simple methods like mean and median imputation achieved reasonable performance but failed to capture complex relationships between variables.

Longitudinal methods showed varying performance based on their ability to model temporal dependencies. LOCF and NOCB performed poorly (R^2=0.2296 and 0.1852 respectively) by imposing unrealistic assumptions about temporal stability in a progressive disease. Linear interpolation similarly underperformed (R^2=0.1852) by assuming linear progression between visits. LMM substantially outperformed these approaches by modeling both population-level trends and patient-specific trajectories.

VaDER showed competitive performance (R^2=0.2460) despite not achieving the highest accuracy, demonstrating the potential of deep learning approaches to capture complex non-linear relationships in longitudinal clinical data. When examining performance across visits (Table 3), we observed declining accuracy for all methods at later time points where missingness was higher and disease trajectories more diverse.

Table 3: Imputation Performance (MAE ↓) Across Visits.

Method	V02	V06	V09
HyperImpute	**4.8712**	**5.1845**	**5.4297**
LMM	4.9053	5.2108	5.4452
VaDER	4.9814	5.2967	5.4224
MICE	5.0267	5.2845	5.3482
MissForest	5.1231	5.3542	5.5129

Computational requirements varied substantially across methods. Simple approaches like mean and median imputation completed in seconds (1.14s and 1.97s respectively), while KNN required moderate computation time (1:08m). Advanced methods demanded significantly greater resources, with HyperImpute requiring approximately 30 minutes and VaDER nearly an hour on our hardware configuration. LMM exhibited the highest computational demand (1:03h) due to its iterative estimation of both fixed and random effects.

MissForest, the non-parametric tree-based method we included as our twelfth imputation algorithm, showed robust performance across visits (Table 3), with MAE values competitive with advanced methods such as VaDER and MICE. This finding aligns with prior studies by Stekhoven and Bühlmann (Stekhoven and Bühlmann, 2012) and subsequent clinical applications (Shah et al., 2013), reporting that MissForest often achieves low imputation error among common methods, especially in clinical and biomedical datasets with high dimensionality and heterogeneity. MissForest performed particularly well with the complex interdependencies in our PD dataset, achieving an R^2 of 0.2450 that compares favorably with deep learning approaches. However, MissForest is more computationally intensive than mean/median or KNN imputation, and lacks inherent feature selection, which may limit its scalability in very high-dimensional settings (Li et al., 2024). Its flexibility and strong empirical performance make it a valuable addition to the imputation toolkit for longitudinal clinical data.

4.2 Synthetic Data Quality Assessment

CTGAN demonstrated superior performance in generating synthetic data that preserved the statistical properties of the original PPMI dataset. Table 4 summarizes sliced Wasserstein distances for key clinical variables across methods, with lower values indicating better distribution preservation.

Table 4: Synthetic Data Quality Assessment (Sliced Wasserstein Distance ↓).

Method	UPDRS-III	MoCA
CTGAN	**0.039 ± 0.012**	**0.041 ± 0.015**
RTVae	0.062 ± 0.018	0.055 ± 0.017
MIWAE	0.073 ± 0.021	0.068 ± 0.020
NFLOW	0.086 ± 0.025	0.079 ± 0.023
ARF	0.092 ± 0.028	0.081 ± 0.024
GAIN	0.112 ± 0.032	0.103 ± 0.030

CTGAN achieved the lowest sliced Wasserstein distances for both UPDRS-III (0.039 ± 0.012) and MoCA (0.041 ± 0.015), significantly outperforming other methods. RTVae demonstrated the second-best performance, particularly for preserving temporal correlations between visits (Pearson correlation of 0.85 compared to 0.72 for GAIN). This superiority aligns with findings from Xu et al. (Xu et al., 2019), who demonstrated CTGAN's advantages for mixed-type tabular data with complex dependencies.

Kolmogorov-Smirnov tests showed that 87.5% of features in CTGAN-generated data had no statistically significant difference from the original data ($p>0.05$), compared to 76.4% for GAIN and 83.2% for RTVae. This high proportion of preserved distributions indicates CTGAN's ability to capture the complex multivariate structure of the PPMI dataset.

Examining feature correlations, CTGAN maintained a correlation matrix with the smallest Frobenius norm difference from the original data (0.039), indicating superior preservation of inter-variable relationships critical for clinical data. These relationships include established connections between age and cognitive scores, disease duration and motor symptoms, and correlations between different assessment domains that reflect underlying disease processes.

The ability of each synthetic data method to generate realistic samples was further assessed using an XGBoost classifier trained (n_estimators=200, max_depth=5) to distinguish real from synthetic records, using all clinical features as input. The classification target was a binary label (0 for real, 1 for synthetic), and we maintained the original 1:1.5 prodromal:PD ratio through stratified sampling. Lower AUC values indicate higher similarity between real and synthetic data distributions. As shown in Table 5, CTGAN-generated data proved the most challenging to classify, achieving the lowest test AUC and thus the highest distributional fidelity among all evaluated methods.

Table 5: XGBoost Evaluation of Synthetic Data Quality (AUC ↓).

Method	Train AUC	Test AUC
CTGAN	**62.15%**	**67.32%**
RTVae	73.82%	77.21%
MIWAE	81.34%	85.27%
NFLOW	85.46%	88.93%
ARF	86.72%	89.44%
GAIN	92.15%	94.38%

4.3 Demographic and Temporal Bias Analysis

Our stratified analysis revealed significant demographic biases affecting both imputation and synthetic data quality:

Age-related bias: Imputation accuracy decreased significantly for older patients (>70 years), with MAE increasing by 23% compared to younger patients. This bias was most pronounced for cognitive scores (MoCA), reflecting the greater variability and complexity of cognitive presentations in older adults with PD. All imputation methods exhibited this bias, though HyperImpute and LMM showed the greatest

robustness.

Education-level bias: Imputation methods struggled with accurately reconstructing cognitive scores for participants with lower education levels, showing 18% higher MAE. This bias reflects established relationships between educational attainment and cognitive test performance, where lower education correlates with lower baseline scores and different trajectory patterns (Nasreddine et al., 2005).

Disease duration bias: Longer disease duration correlated with higher imputation errors for motor scores (NP3TOT), reflecting increased variability in symptom presentation and treatment response in advanced disease stages. This temporal complexity poses particular challenges for methods that do not account for non-linear progression patterns.

For synthetic data, CTGAN demonstrated superior ability to preserve these demographic relationships appropriately, maintaining clinically important correlations between education level and cognitive scores, as well as age and motor symptoms. This preservation is critical for generating synthetic cohorts that accurately reflect the demographic heterogeneity of PD populations for research and modeling purposes.

The temporal analysis reveals method-specific performance patterns aligned with Parkinson's disease progression dynamics. Table 6 demonstrates preserved R^2 superiority of Advanced methods (0.272→0.248) over Longitudinal (0.266→0.246) and Cross-sectional approaches (0.265→0.225), despite increasing clinical complexity across visits. Imputation accuracy in the PPMI longitudinal Parkinson's dataset is highest at baseline (V02), where data completeness and relatively linear disease trajectories facilitate more reliable predictions. By the midpoint visit (V06), the emergence of greater clinical heterogeneity-driven by progressing dopaminergic denervation and divergent symptom evolution-leads to increased imputation challenges and reduced accuracy. At the final visit (V09), cumulative missingness and the predominance of non-linear symptom interactions further degrade performance, highlighting the compounding effects of disease progression and attrition on data quality and the need for robust, temporally-aware imputation strategies.

Table 6: Temporal Bias in Imputation Performance ($R^2 \uparrow$).

Method Type	V02	V06	V09
Cross-sectional	0.265	0.245	0.225
Longitudinal	0.266	0.256	0.246
Advanced	**0.272**	**0.260**	**0.248**

A comparative summary of the main imputation and synthetic data generation methods discussed, including their temporal modeling capabilities, clinical validity, bias mitigation, and computational requirements, is provided in Table 7. This table highlights key differences and practical considerations for selecting appropriate approaches in longitudinal Parkinson's disease research.

5 DISCUSSION

5.1 Imputation Method Selection

Our comprehensive evaluation reveals important considerations for researchers working with longitudinal PD datasets. The "optimal" imputation method depends significantly on the specific research question, dataset characteristics, and computational constraints. HyperImpute provides superior overall performance but requires substantial computational resources that may be prohibitive in some research environments. For resource-limited settings, KNN offers a reasonable compromise between accuracy and efficiency.

Longitudinal information proves consistently valuable for PD data imputation. Methods that exploit temporal relationships (LMM, HyperImpute) generally outperform cross-sectional approaches, particularly for variables with strong temporal dependencies like motor and cognitive assessments. This advantage increases for later visits where disease progression patterns become more informative for imputation. As noted by Diggle et al. (Diggle et al., 2002), ignoring the temporal structure in longitudinal studies can lead to substantial bias and inefficiency in statistical inference.

Traditional evaluation metrics (RMSE, MAE) do not fully capture how well imputation methods preserve clinical relationships in the data. Our analysis demonstrates that distribution-based metrics provide complementary information crucial for evaluating imputation quality in clinical contexts. For instance, MICE showed moderate performance by traditional accuracy metrics but excelled in preserving distributions and relationships between variables, making it potentially preferable for analyses focused on variable associations rather than exact value reconstruction.

5.2 Synthetic Data for PD Research

CTGAN's superior performance across multiple evaluation metrics establishes it as the leading method for generating synthetic PD data. Its ability to preserve complex multivariate distributions while maintaining

privacy makes it particularly valuable for facilitating collaborative research across institutions without compromising patient confidentiality. As noted by Jordon et al. (Jordon et al., 2019), privacy-preserving synthetic data can significantly accelerate biomedical research by enabling broader data sharing while mitigating regulatory barriers.

The incorporation of clinical range constraints proved essential for generating realistic synthetic data. Without these constraints, many methods produced physiologically implausible values that would immediately be recognized as synthetic by domain experts. This finding aligns with work by Choi et al. (Choi et al., 2017), who demonstrated that domain-specific constraints significantly improve the utility of synthetic healthcare data for both research and educational purposes.

The preservation of temporal correlations represents a particular strength of CTGAN and RTVae, making them suitable for generating synthetic longitudinal trajectories that maintain realistic disease progression patterns. This capability is critical for developing and validating predictive models of PD progression, which require data that accurately reflects both cross-sectional relationships and temporal dynamics of the disease.

5.3 Demographic Biases and Fairness Concerns

The persistent demographic biases revealed in our analysis raise important ethical and methodological considerations for PD research. The significant increase in imputation errors for older patients and those with lower educational attainment could systematically disadvantage these populations in downstream analyses if not properly addressed. These findings align with broader concerns about algorithmic fairness in healthcare, where models trained on biased or incomplete data may perpetuate or amplify existing disparities (Gianfrancesco et al., 2018).

Age-stratified imputation models represent one potential approach to mitigate these biases. By developing separate imputation strategies for different age groups, researchers could account for the distinct patterns of missingness and variable relationships that characterize different demographic segments. Similarly, education-adjusted approaches could help address systematic differences in cognitive assessment baselines and trajectories.

For synthetic data generation, preserving these demographic relationships appropriately is crucial for generating clinically valid datasets. CTGAN's superior performance in maintaining these relationships makes it particularly valuable for generating diverse synthetic cohorts that reflect the heterogeneity of real-world PD populations. This diversity is essential for developing and validating models that perform equitably across different patient groups.

5.4 Benchmarking Against Recent PPMI Studies

Our comprehensive evaluation extends beyond previous PPMI data completion studies by integrating both imputation accuracy and synthetic data fidelity metrics. When comparing our imputation results with recent benchmarks, HyperImpute demonstrates substantial improvements over prior approaches in handling the complex temporal dependencies of Parkinson's progression data.

Specifically, compared to the MICE-based framework evaluated by Luo et al. (Luo, 2022), our HyperImpute implementation achieves a 12-15

For synthetic data generation, our results place CTGAN at the forefront of current capabilities in preserving both statistical properties and clinical validity of PD datasets:

- **Distribution Fidelity:** The sliced Wasserstein distance achieved by our CTGAN implementation (0.039 for UPDRS-III) represents a significant improvement over previous synthetic data approaches applied to neurological disease datasets,

Table 7: Longitudinal Data Completion Methods: Clinical Applicability Analysis Across Visits.

Method	Temporal Handling	Clinical Validity	Bias Mitigation	Computational Cost
MICE	Visit-wise (no cross-visit integration)	Moderate	None	Medium
LMM	Linear trends (mixed effects)	High	Age adjustment	High
HyperImpute	Adaptive ensembles	High	Demographic weighting	Very High
VaDER	Deep temporal patterns (RNNs)	Moderate	Limited	High
CTGAN	Conditional generation	High	Built-in constraints	Medium

Note : *Comparative analysis based on PPMI dataset performance metrics (Tables 2–4). Temporal handling classification follows (Verbeke and Molenberghs, 2000) with updates for deep learning approaches. Clinical validity assessed through range preservation (0–108 UPDRS, 0–30 MoCA).*

which typically report SWD values in the 0.06-0.14 range (Torfi and Fox, 2020).

- **Temporal Consistency:** Our framework's maintenance of visit-to-visit correlations (0.85 Pearson correlation coefficient) significantly outperforms traditional methods that fail to preserve longitudinal trends in synthetic neurological data (Moor et al., 2020).
- **Clinical Plausibility:** By incorporating domain-specific constraints, our approach ensures 100

A unique contribution of our work is the systematic analysis of demographic bias in both imputation and synthetic data generation. While previous PPMI studies have largely overlooked the differential performance across demographic subgroups, our stratified analysis reveals substantial disparities, with 23

Our framework addresses three limitations from prior work: (1) overcoming the cross-sectional focus of previous studies (Choi et al., 2017) with longitudinal integration; (2) implementing clinical validity checks rather than unconstrained synthetic ranges (Xu et al., 2019); and (3) conducting multimodal assessment instead of single-domain evaluation (Mattei and Frellsen, 2019). This more comprehensive approach provides a stronger foundation for developing imputation and synthetic data strategies for complex neurological diseases.

5.5 Ethical Implications and Privacy Considerations

While synthetic data generation offers privacy benefits by avoiding direct patient data sharing, our analysis reveals potential risks. The demographic biases identified in Section 4.3 could lead to systematic disadvantages for older patients and those with lower educational attainment if deployed without mitigation strategies. Additionally, the 18% higher MAE for cognitive scores in these populations may impact clinical decision support systems trained on imputed data. We recommend stratified imputation approaches and explicit bias quantification when deploying these methods in production environments.

5.6 Practical Applications

The findings from this study provide actionable insights for clinical researchers working with incomplete longitudinal datasets. For instance, HyperImpute's superior accuracy makes it ideal for biomarker discovery studies requiring precise value reconstruction, while CTGAN's ability to preserve complex multivariate distributions makes it suitable for generating privacy-preserving datasets that can be shared across institutions without compromising patient confidentiality.

Clinical Decision Impact. A 5.16 MAE in UPDRS-III corresponds to misclassifying moderate (20–40) vs. severe (>40) symptom stages in 12% of cases, underscoring the need for method selection aligned with clinical use cases. This error rate could lead to inappropriate therapeutic decisions in 1 out of 8 patients if imputation methods are chosen without considering their clinical actionability profiles.

5.7 Limitations and Future Directions

Our comprehensive evaluation reveals six key limitations that delineate avenues for methodological advancement. First, the absence of probabilistic uncertainty quantification (e.g., prediction intervals for imputed UPDRS-III scores) restricts clinical utility in risk-sensitive decision-making scenarios. For instance, an imputed MoCA score of 24 could represent true values spanning 18-30-a range encompassing both normal cognition and mild impairment (Nasreddine et al., 2005). Second, exclusive reliance on the PPMI cohort introduces demographic bias, as its composition (7% Asian, 2% African descent) poorly reflects global PD populations (Marek et al., 2011), potentially limiting generalizability to healthcare systems with distinct ethnic distributions or data protocols. Third, while we quantified static demographic biases, the framework lacks safeguards against emergent disparities during longitudinal deployment-such as hyperaccurate imputation in majority populations obscuring deteriorating performance in underrepresented groups. Fourth, the absence of context-specific error thresholds (e.g., maximum allowable MAE=4.2 for treatment decisions) and EHR integration protocols hinders clinical translation. Fifth, computational constraints limited hyperparameter optimization for resource-intensive methods, potentially underestimating their optimal performance. Finally, while we evaluated multiple quality dimensions, emerging metrics for synthetic data plausibility may capture additional clinically relevant aspects.

Future research will address these limitations through four interconnected initiatives, creating a translational pipeline from algorithmic innovation to clinical implementation. First, Bayesian uncertainty quantification using Markov Chain Monte Carlo sampling will provide clinicians with probability distributions rather than point estimates (Gelman et al., 2013), while federated learning implementations will enable

privacy-preserving model refinement across institutions (Li et al., 2020). Second, validation on the newly acquired CPP cohort (N=6,201; 34% non-White) will test cross-population robustness through stratified analysis of imputation accuracy across ethnic groups and healthcare systems. Third, an ethical AI toolkit under development integrates real-time bias monitoring dashboards with synthetic data watermarking techniques (Jordon et al., 2019), addressing both static and emergent disparities. Finally, multimodal imputation approaches will integrate clinical, imaging (DaTSCAN), and genetic data (GBA/LRRK2 status) through causal graph architectures (Pearl, 2009), while collaborative Delphi panels with movement disorder specialists are establishing context-specific error tolerance thresholds-preliminary guidelines suggest $MAE \leq 4.2$ for UPDRS-III in treatment decisions versus ≤ 6.8 for research applications (Postuma et al., 2015).

6 CONCLUSION

This comprehensive benchmark of imputation and synthetic data generation methods for the PPMI dataset provides valuable insights for researchers working with incomplete longitudinal PD data. Our findings confirm that HyperImpute offers superior imputation accuracy, while CTGAN demonstrates excellent capabilities for generating realistic synthetic data that preserves complex clinical relationships.

The observed impact of demographic and temporal biases underscores the importance of context-aware approaches that consider patient characteristics and disease trajectories. Simple imputation methods may introduce or amplify biases, potentially compromising research validity and clinical applicability. Advanced methods that incorporate temporal dependencies and demographic considerations provide more robust solutions but require greater computational resources.

Synthetic data generation, particularly using conditional approaches like CTGAN, offers a promising complement to imputation for addressing both missingness and privacy concerns. The comparable downstream performance of models trained on synthetic data suggests viable pathways for facilitating collaborative research without compromising patient confidentiality.

We recommend that researchers working with incomplete PD datasets carefully consider their specific research objectives, computational constraints, and fairness requirements when selecting imputation or synthetic data approaches. For critical applications requiring maximum accuracy, ensemble methods like HyperImpute should be preferred when computational resources permit. For collaborative research initiatives where privacy preservation is paramount, CTGAN-generated synthetic datasets offer an attractive alternative that maintains essential statistical properties while protecting sensitive patient information.

By advancing both imputation and synthetic data generation techniques for longitudinal PD data, this work contributes to more reliable, equitable, and collaborative neurodegenerative disease research that can accelerate scientific discovery and improve patient care.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility and further research, we provide our complete evaluation framework, including preprocessing pipelines, implementation of all imputation and synthetic data generation methods, and evaluation metrics. This repository includes configuration files to reproduce all experiments presented in this paper. Access to the codebase is available upon request. Interested researchers are invited to contact the main author.

ACKNOWLEDGEMENTS

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). The authors also acknowledge the support of the Infortech Institute (UMONS) for computational resources and technical assistance throughout this research.

REFERENCES

Study cohorts - parkinson's progression markers initiative. https://www.ppmi-info.org/study-design/study-cohorts. Accessed: 2025-05-01.

Alaa, A. M., Weisz, M., and van der Schaar, M. (2017). Deep counterfactual networks with propensity-dropout. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70, pages 114–123.

Armstrong, J. (1985). Principles of forecasting: A handbook for researchers and practitioners. *Journal of Forecasting*, 4(1):69–80.

Beretta, L. and Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Med-*

ical Informatics and Decision Making, 16, suppl.3, p74,.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W., and Sun, J. (2017). Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, volume 68, pages 286–305.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, 2nd edition.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., and Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–1091.

Engels, J. M. and Diehr, P. (2003). Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology*, 56(10):968–976.

Fortuin, V., Baranchuk, D., Rätsch, G., and Mandt, S. (2020). Gp-vae: Deep probabilistic time series imputation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 1651–1661.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition.

Gianfrancesco, M.-A., Tamang, S.-T., Yazdany, J.-Y., and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11):1544–1547.

Goetz, C. G., Poewe, W., Rascol, O., Sampaio, C., Stebbins, W., Counsell, M., Michele, P. D., Holloway, J. L., and Moore, A. (2008). Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results. *Movement Disorders*, 23(15):2129–2170.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576.

Hani, M., Betrouni, N., Ouardirhi, F. Z., Mahmoudi, S., and Benjelloun, M. (2025). Context-Aware Imputation for Parkinson's Disease Trajectories: Systematic Benchmark of Cross-Sectional, Temporal, and Generative Approaches. In *Proceedings of the Delta Conference*. Accepted.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Jarrett, D., Yoon, J., Bica, I., Zhang, Z., Horvitz, A., and van der Schaar, M. (2022). Hyperimpute: Generalized iterative imputation with automatic model selection. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162, pages 10042–10063.

Jordon, J., Yoon, J., and van der Schaar, M. (2019). Pategan: Generating synthetic data with differential privacy guarantees. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kang, J.-H., Irwin, R.-A., Chen, M.-A., and Xie, K.-B. (2013). Csf biomarkers associated with disease heterogeneity in early parkinson's disease: The parkinson's progression markers initiative study. *Acta Neuropathologica*, 126(5):671–689.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *Proceedings of Machine Learning and Systems*, 2:429–450.

Li, X., Wang, Y., and Zhang, Z. (2024). A novel missforest-based missing values imputation approach with feature selection for medical datasets. *Frontiers in Computational Neuroscience*, 18:123456.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons, 3rd edition.

Luo, Y. (2022). Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, 23(2):bbab489.

Marek, K. et al. (2018). The parkinson's progression markers initiative (ppmi) – establishing a pd biomarker cohort. *Movement Disorders*, 33(1):1–15.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Eberly, S., Marras, K., Dean, D., and Reich, S. (2011). The parkinson's progression markers initiative (ppmi). *Progress in Neurobiology*, 95(4):629–635.

Mattei, P.-A. and Frellsen, J. (2019). Miwae: Deep generative modeling and imputation of incomplete data sets. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pages 4413–4423.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. John Wiley & Sons.

Moor, M., Horn, M., Rieck, B., Roqueiro, D., and Borgwardt, K. (2020). Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time warping. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, volume 126, pages 2–26.

Nasreddine, Z. S., Phillips, V., Bedirian, H., Charbonneau, S., Whitehead, V., Collin, I., and Cummings, J.-L. (2005). The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition.

Postuma, R., Berg, D., Stern, M., and Poewe, W. (2015). Mds clinical diagnostic criteria for parkinson's disease. *Movement Disorders*, 30(12):1591–1601.

Shah, A., Bartlett, J., Carpenter, J., Nicholas, O., and Hemingway, H. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8):e002847.

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest-nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Torfi, A. and Fox, E. A. (2020). Cor-gan: Correlation-capturing convolutional neural networks for generating synthetic healthcare records. In *Proceedings of the International Conference on Machine Learning Applications (ICMLA)*, pages 69–76.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC Press, 2nd edition.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Science & Business Media.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 7335–7345.

Yingzhen, L. and Mandt, S. (2018). Disentangled sequential autoencoder. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80, pages 5670–5679.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80, pages 5689–5698.