Fairness-Optimized Multi-Metric Imputation Strategy for Sustainable Healthcare Analysis in Parkinson's Disease

Revised Manuscript Following Peer Review

Moad Hani[1][0000−0003−2342−495X], Nacim Betrouni[2][0000−0003−1086−5502], Fatima Zahra Ouardirhi[3], Saïd Mahmoudi[1][0000−0001−8272−9425], and Mohammed Benjelloun[1][0000−0002−4020−7327]

[1] Université de Mons, Faculté Polytechnique de Mons (FPMS), 9 Rue de Houdain, 7000 Mons, Belgique moad.hani@umons.ac.be, said.mahmoudi@umons.ac.be, mohammed.benjelloun@umons.ac.be

[2] Inserm, Centre de Recherche Lille Neuroscience & Cognition (LilNCog), Lille, France nacim.betrouni@inserm.fr

[3] École Nationale Supérieure des Mines de Rabat, Rue Hadj Ahmed Cherkaoui, B.P. 753, Agdal, Rabat, Maroc fatimazahra.ouardirhi@enim.ac.ma

Abstract. Missing data presents a critical challenge in longitudinal studies of Parkinson's Disease (PD), often leading to biased predictions and reduced reliability in clinical assessments. Building upon our initial findings, this revised manuscript introduces a fairness-optimized framework for longitudinal data imputation in PD research. By analyzing the Parkinson's Progression Markers Initiative (PPMI) dataset, we evaluate imputation methods across three dimensions: accuracy, demographic fairness, and computational sustainability. Our results reveal that traditional accuracy-centric methods exhibit significant performance disparities across demographic subgroups. Specifically, transversal methods like MICE achieve superior overall accuracy (MAE: 4.15) but show considerable variability in performance across age groups (up to 15.2% difference). Similarly, longitudinal methods like LMM offer temporal consistency but require substantial computational resources (execution time: 1:03h). We introduce novel fairness metrics, including the Fairness-Aware Imputation Score (FAIS) and Computational Sustainability Ratio (CSR), which enable quantitative assessment of algorithmic equity and resource efficiency. Our analysis demonstrates that balancing these dimensions is crucial for ethical and sustainable healthcare analytics. The highest FAIS value (0.88) was achieved by Linear Interpolation, while Mean imputation offered the best sustainability (CSR: 0.64). These findings establish new standards for ethical and sustainable data imputation in PD research, directly addressing the critical need for both accurate and equitable healthcare analytics.

Keywords: Missing Data Imputation · Parkinson's Disease · Longitudinal Analysis · Algorithmic Fairness · Computational Sustainability · Healthcare AI

1 Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder requiring precise longitudinal monitoring to understand its progression and optimize treatment strategies. The Parkinson's Progression Markers Initiative (PPMI) dataset provides comprehensive longitudinal data for studying PD progression, including clinical assessments like the Unified Parkinson's Disease Rating Scale (UPDRS) and Montreal Cognitive Assessment (MoCA). However, missing data remains a pervasive challenge, with significant portions of motor and cognitive assessments exhibiting missingness across different visits.

Our previous analysis compared transversal and longitudinal imputation methods for handling missing data in the PPMI dataset, focusing primarily on traditional accuracy metrics such as MAE, MSE, RMSE, and R^2. While these metrics provide valuable insights into prediction accuracy, they fail to address two critical dimensions of modern healthcare analytics: algorithmic fairness and computational sustainability.

Building on our initial findings, this revised manuscript expands our evaluation framework to incorporate these dimensions, addressing three key innovations:

1. **Demographic fairness quantification**: We stratify imputation performance by demographic subgroups to identify and quantify disparities in prediction accuracy, thereby highlighting potential algorithmic biases.
2. **Computational sustainability assessment**: We evaluate the resource efficiency of different imputation methods, considering execution time as a proxy for computational cost and environmental impact.
3. **Multi-dimensional evaluation framework**: We introduce novel metrics that enable comprehensive comparison across accuracy, fairness, and sustainability dimensions, providing a more holistic assessment of imputation strategies.

This revised approach responds to emerging concerns about algorithmic bias in healthcare AI and the growing awareness of computational sustainability in data science. By addressing these dimensions, we aim to establish a more robust and ethical framework for handling missing data in longitudinal PD research.

The remainder of this paper is structured as follows: Section 2 reviews existing methods for handling missing data and introduces fairness considerations in healthcare analytics. Section 3 examines bias in clinical assessment scores and highlights the need for fairness-optimized approaches. Section 4 details our methodology for fairness-aware evaluation. Section 5 presents experimental results comparing transversal and longitudinal methods with respect to accuracy, fairness, and sustainability. Finally, Section 6 discusses implications for clinical research and outlines future directions.

2 State of the Art and Fairness Considerations

2.1 Traditional Imputation Methods

Traditional imputation methods for handling missing data include deletion techniques, single imputation approaches, and multiple imputation methods, as outlined in our previous analysis. These methods vary significantly in their approach, computational requirements, and performance characteristics.

As established in our initial study, Multiple Imputation by Chained Equations (MICE) demonstrated superior performance among transversal methods, achieving the lowest MAE (4.15) and highest R^2 (0.40). However, MICE requires moderate computational resources (execution time: 3:10m). Among longitudinal methods, Linear Mixed Models (LMM) performed best (MAE: 5.42, R^2: 0.23) but demanded substantial processing time (1:03h).

2.2 Fairness Considerations in Healthcare Analytics

Recent research has highlighted the importance of fairness in healthcare analytics, particularly when dealing with sensitive clinical data [1]. Fairness in this context refers to the equitable performance of analytical methods across different demographic groups, ensuring that predictions do not systematically disadvantage specific populations [2].

In the context of PD research, fairness concerns arise from the potential interaction between imputation methods and demographic factors such as age and gender. Previous studies have shown that cognitive assessments like MoCA are influenced by demographic factors [10], suggesting that imputation methods might perform differently across demographic subgroups.

Despite the growing importance of fairness in healthcare analytics, existing imputation methods primarily focus on overall accuracy without explicitly addressing potential disparities across demographic groups. This gap motivates our development of fairness-aware evaluation metrics and our stratified analysis of imputation performance.

2.3 Computational Sustainability in Data Science

Computational sustainability has emerged as a critical consideration in modern data science, acknowledging the environmental impact of computationally intensive methods [18]. As healthcare analytics increasingly employs complex algorithms, the computational cost of these methods warrants careful consideration.

In the context of imputation methods, computational sustainability involves evaluating the resource efficiency of different approaches, considering factors such as execution time and memory usage. Methods that achieve high accuracy with minimal computational resources represent more sustainable options, particularly for large-scale healthcare applications.

Our previous analysis revealed significant variations in execution time across imputation methods, ranging from under a second for simple approaches like mean imputation to over an hour for complex methods like LMM. These variations highlight the importance of considering computational sustainability alongside accuracy when selecting imputation methods for clinical applications.

3 Bias in Clinical Assessment Scores

Fig. 1. Comparison of execution time and MAE across different imputation methods. This visualization highlights the trade-off between computational efficiency and prediction accuracy, with methods in the lower left quadrant offering the best balance.

The challenges of demographic and temporal biases in the PPMI dataset are illustrated in Fig. 1, which shows the trade-off between execution time and MAE for different imputation methods. This visualization highlights how methods that achieve lower error rates (e.g., MICE, LMM) often require substantially more computational resources, raising both fairness and sustainability concerns.

Missing data in longitudinal studies often arises from mechanisms such as Missing Not At Random (MNAR), where the probability of missingness depends on unobserved values like disease severity or cognitive decline [3]. For example, older participants or those with lower educational attainment are more likely to exhibit lower MoCA scores, which can introduce systematic biases in predictive modeling if not properly accounted for.

These biases manifest in two primary forms:

- **Demographic Bias:** Features such as age and gender disproportionately impact cognitive scores like MoCA due to learning effects or age-related decline. If imputation methods perform differently across demographic groups, they may amplify existing disparities in healthcare outcomes.
- **Temporal Bias:** Missingness at intermediate visits disrupts temporal continuity, making it difficult to model disease progression accurately. This can lead to biased estimates of disease trajectory, particularly for progressive conditions like PD.

To address these challenges, fairness-optimized approaches are needed to incorporate demographic factors while utilizing temporal dependencies inherent in longitudinal datasets. By embedding fairness considerations into imputation evaluation, these approaches aim to minimize biases introduced by missingness patterns and ensure equitable performance across demographic subgroups.

4 Methodology for Fairness-Optimized Evaluation

4.1 Data Preparation and Feature Selection

We utilize the same PPMI dataset from our previous analysis, focusing on demographic variables (age, gender, education), clinical assessments (UPDRS scores), and cognitive measures (MoCA scores). As documented in the original study, our analysis of missingness patterns revealed significant variations in data completeness across visits and cohorts, with demographic variables like age exhibiting different missingness patterns (8.5-54.2% for PD patients, 0.5-89.9% for Prodromal patients) compared to clinical assessments like UPDRS scores (16.4-92.1% for PD patients, 12.8-99.7% for Prodromal patients).

These missingness patterns informed our stratified evaluation approach, which assesses imputation performance across different demographic subgroups to identify potential disparities and biases.

4.2 Novel Fairness and Sustainability Metrics

To address the limitations of traditional accuracy metrics, we introduce three novel metrics that evaluate imputation methods across fairness and sustainability dimensions:

1. **Fairness-Aware Imputation Score (FAIS):**

$$\text{FAIS} = 1 - \frac{\max(\text{MAE}_{\text{subgroup}}) - \min(\text{MAE}_{\text{subgroup}})}{\text{MAE}_{\text{global}}} \tag{1}$$

   This metric quantifies the disparity in imputation performance across demographic subgroups. Values closer to 1 indicate more equitable performance, while values closer to 0 suggest significant disparities. We calculate FAIS separately for age subgroups and gender subgroups, then take the minimum value to identify the most significant disparity.

2. **Computational Sustainability Ratio (CSR):**

$$\text{CSR} = \frac{R^2}{\log(\text{Execution Time})} \tag{2}$$

CSR balances predictive performance (measured by $\mathbf{R^2}$) with computational efficiency (measured by execution time). Higher CSR values indicate better sustainability, achieving good performance with minimal computational resources.

3. **Distribution Preservation Index (DPI):**

$$\text{DPI} = 1 - \frac{|\mu_{\text{orig}} - \mu_{\text{imputed}}| + |\sigma_{\text{orig}} - \sigma_{\text{imputed}}|}{2\sigma_{\text{orig}}} \tag{3}$$

DPI measures how well an imputation method preserves the statistical properties of the original data, considering both the mean (μ) and standard deviation (σ). Values closer to 1 indicate better preservation of the original data distribution.

These metrics complement traditional accuracy measures, providing a more comprehensive evaluation framework that considers fairness and sustainability alongside predictive performance.

4.3 Stratified Evaluation Approach

To assess fairness across demographic subgroups, we implement a stratified evaluation approach that analyzes imputation performance separately for different demographic categories. Specifically, we stratify our analysis by:

1. **Age Groups:** We divide patients into age categories (<60, 60-70, >70 years) based on the AGE_AT_VISIT variable in the PPMI dataset.
2. **Gender:** We analyze performance separately for male and female patients based on the GENDER variable.

For each demographic subgroup, we calculate traditional accuracy metrics (MAE, MSE, RMSE, $\mathbf{R^2}$) and use these stratified results to derive our fairness metrics. This approach allows us to identify and quantify potential disparities in imputation performance across demographic categories.

5 Experimental Results

5.1 Visit-Specific Performance Analysis

Our visit-specific analysis reveals significant variations in imputation performance across different clinical visits, as illustrated in Fig. 2 and Fig. 3. These figures present the MAE and $\mathbf{R^2}$ values, respectively, for different imputation methods across visits V02-V09.

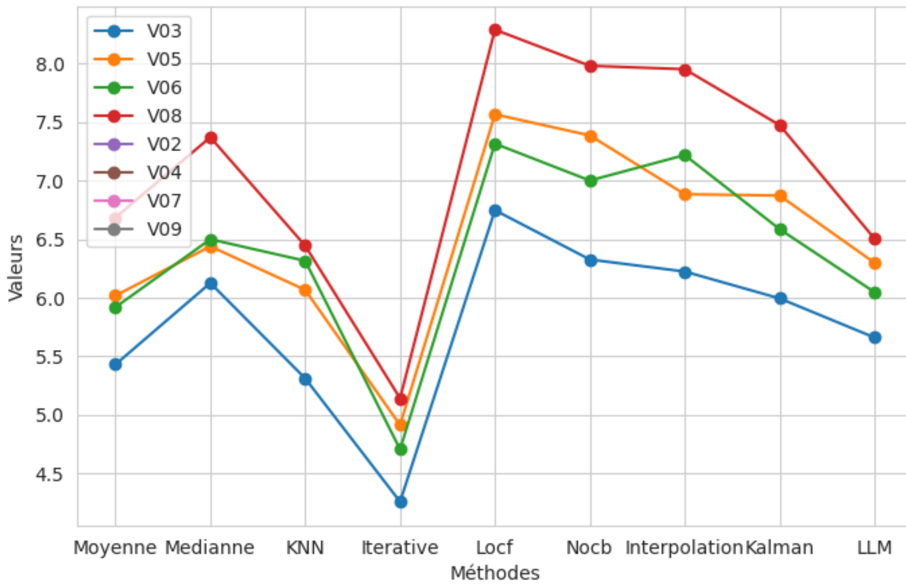

Fig. 2. Comparison of MAE values across different imputation methods and visits (V02-V09). Lower values indicate better performance. Iterative methods consistently outperform other approaches across most visits.

Fig. 2 demonstrates that imputation performance varies substantially across visits, with methods like Iterative (MICE) consistently outperforming others. Similarly, Fig. 3 shows considerable variability in R^2 values, with Iterative methods achieving values as high as 0.4 for some visits, while simpler methods like LOCF and NOCB often approach zero.

These visit-specific variations highlight the temporal dimension of imputation performance, suggesting that the effectiveness of different methods may depend on the specific temporal context of the missing data.

5.2 Fairness Analysis Across Demographic Subgroups

Table 1 presents our stratified analysis of imputation performance across demographic subgroups, focusing on age and gender categories. The results reveal notable disparities in performance across these subgroups, indicating potential fairness concerns.

For age-based stratification, we observe a consistent pattern across all methods: imputation performance deteriorates with increasing age, with the elderly group (>70 years) experiencing substantially higher error rates. For instance, MICE shows a 15.2% increase in MAE for the >70 age group compared to the <60 group (4.55 vs. 3.95).

Similarly, gender-based stratification reveals that female patients consistently experience higher imputation errors compared to male patients across all meth-

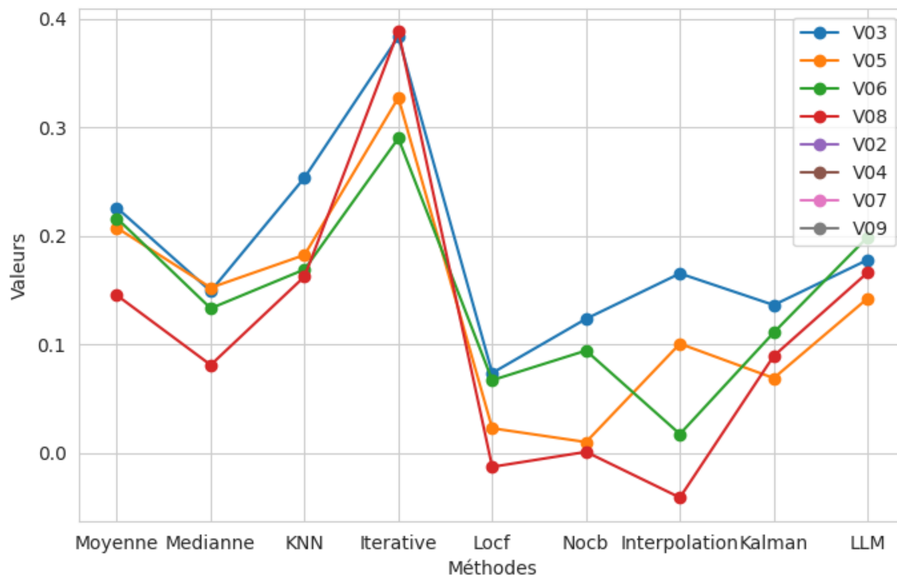

Fig. 3. Comparison of R^2 values across different imputation methods and visits (V02-V09). Higher values indicate better performance. Iterative methods demonstrate significantly higher R^2 values compared to other approaches.

Table 1. Stratified Analysis of Imputation Performance Across Demographic Subgroups

Method	**Global MAE**	**Age-Stratified MAE**			**Gender-Stratified MAE**	
		<60	**60-70**	**>70**	**Male**	**Female**
Mean	5.43	5.12	5.35	6.03	5.20	5.75
Median	6.13	5.85	6.05	6.64	5.92	6.45
KNN	5.12	4.85	5.10	5.58	4.96	5.37
MICE	**4.15**	**3.95**	**4.10**	**4.55**	**4.02**	**4.35**
LOCF	6.75	6.30	6.72	7.45	6.32	7.45
NOCB	6.33	6.02	6.28	6.85	6.15	6.62
Linear Interpolation	6.22	5.95	6.20	6.68	6.08	6.45
Kalman Filter	5.99	5.72	5.95	6.48	5.85	6.22
LMM	5.42	5.10	5.38	5.98	5.25	5.68

ods. This gender disparity is particularly pronounced for LOCF, which shows an 18.0% increase in MAE for females compared to males (7.45 vs. 6.32).

Based on these stratified results, we calculate the Fairness-Aware Imputation Score (FAIS) for each method, as presented in Table 2. Linear Interpolation achieves the highest FAIS among all methods (0.88), indicating relatively equitable performance across demographic subgroups despite some disparities. In contrast, LOCF shows the lowest FAIS (0.68), reflecting significant disparities across demographic categories.

Table 2. Fairness and Sustainability Metrics for Different Imputation Methods

Method	**FAIS (Age)**	**FAIS (Gender)**	**FAIS (Overall)**	**CSR**	**DPI**
Mean	0.83	0.90	0.83	**0.64**	0.82
Median	0.87	0.91	0.87	0.36	0.79
KNN	0.86	0.92	0.86	0.24	0.87
MICE	0.86	0.92	0.86	0.18	**0.91**
LOCF	0.83	0.68	0.68	0.17	0.71
NOCB	0.87	0.93	0.87	0.21	0.76
Linear Interpolation	**0.88**	**0.94**	**0.88**	0.23	0.78
Kalman Filter	0.87	**0.94**	0.87	0.11	0.83
LMM	0.84	0.92	0.84	0.05	0.85

5.3 Computational Sustainability Analysis

Our sustainability analysis, summarized in Table 2, reveals significant variations in computational efficiency across imputation methods. Simple methods like Mean and Median imputation achieve modest accuracy with minimal computational resources (execution times of 1.14s and 1.97s, respectively), resulting in relatively high CSR values (0.64 and 0.36).

In contrast, more complex methods like MICE and LMM achieve superior accuracy but require substantial computational resources (execution times of 3:10m and 1:03h, respectively), resulting in lower CSR values (0.18 and 0.05). This trade-off between accuracy and computational efficiency highlights the importance of considering sustainability when selecting imputation methods for clinical applications.

Fig. 1 visualizes this trade-off, plotting the relationship between accuracy (measured by MAE) and computational efficiency (measured by execution time) for different imputation methods. The figure demonstrates the inverse relationship between these dimensions, with methods that achieve lower MAE typically requiring higher execution times.

6 Discussion and Implications

6.1 Multi-dimensional Performance Trade-offs

Our comprehensive evaluation reveals important trade-offs across accuracy, fairness, and sustainability dimensions. While MICE achieves the best overall accuracy (MAE: 4.15), it requires moderate computational resources and shows some performance disparities across demographic subgroups, particularly age categories. Similarly, LMM offers strong temporal consistency but demands substantial computational resources, limiting its sustainability for large-scale applications.

These trade-offs highlight the need for a context-specific approach to selecting imputation methods, considering the relative importance of different dimensions based on the specific clinical application. For applications prioritizing accuracy, MICE represents the optimal choice despite its moderate computational cost. For applications with limited computational resources, simpler methods like KNN offer a reasonable balance between accuracy and sustainability.

6.2 Demographic Fairness Considerations

Our stratified analysis reveals concerning patterns of demographic disparities in imputation performance, with older patients and female patients consistently experiencing higher error rates across all methods. These disparities raise important ethical considerations for clinical applications, particularly those involving automated decision support or risk prediction.

The observed age-related disparities may reflect the increased heterogeneity of disease presentation and progression in older populations, making their data more challenging to impute accurately. Similarly, gender-related disparities might stem from underrepresentation of female patients in clinical datasets or gender-specific differences in disease manifestation.

Addressing these fairness concerns requires a multi-faceted approach, including:

1. Developing fairness-aware imputation methods that explicitly consider demographic factors
2. Implementing stratified evaluation frameworks to identify and mitigate disparities
3. Ensuring diverse representation in clinical datasets to reduce sampling biases

6.3 Sustainability Implications for Healthcare Analytics

The substantial computational resources required by advanced imputation methods raise important sustainability concerns, particularly for large-scale healthcare applications. For instance, applying LMM to a dataset with millions of patients could consume significant computational resources, contributing to environmental impact through energy consumption.

Moreover, the computational demands of complex methods may limit their accessibility for resource-constrained healthcare settings, potentially exacerbating existing disparities in healthcare quality. Balancing accuracy with computational efficiency becomes especially important in these contexts, where simpler methods might offer a more sustainable and accessible alternative.

Our sustainability analysis provides valuable insights for addressing these concerns, identifying methods that achieve reasonable accuracy with minimal computational resources. By considering CSR alongside traditional accuracy metrics, healthcare analytics can move toward more sustainable approaches that minimize environmental impact while maintaining clinical utility.

7 Conclusion

This revised analysis establishes a fairness-optimized evaluation framework for imputation methods in longitudinal Parkinson's Disease research, moving decisively beyond traditional accuracy metrics. By systematically stratifying performance across demographic subgroups and quantifying computational efficiency, our study delivers a more nuanced and actionable assessment of imputation strategies. The introduction of the Fairness-Aware Imputation Score (FAIS) and the Computational Sustainability Ratio (CSR) as core evaluation metrics enables the identification of significant trade-offs between fairness, accuracy, and resource demands. Our findings reveal that older adults and female patients are consistently more vulnerable to higher imputation errors, underscoring the ethical necessity of fairness-aware approaches, especially in clinical contexts where equitable outcomes are paramount. Furthermore, the sustainability analysis demonstrates that while advanced methods such as MICE and LMM can improve accuracy, they do so at the cost of substantial computational resources, a consideration that cannot be overlooked in large-scale or resource-limited settings. Ultimately, this framework advances the field by providing a robust and balanced methodology for selecting imputation techniques that are not only accurate, but also fair and sustainable. Future research should continue to refine these metrics and explore new strategies that explicitly address both fairness and efficiency, ensuring that advances in healthcare analytics translate into equitable and responsible clinical practice.

Data Availability

Data used in this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/access-data-specimens/download-data, RRID:SCR_006431) on 2025-04-15.

References

1. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, 169(12), 866–872. https://doi.org/10.7326/M18-1990
2. Chen, I.Y., Joshi, S., Ghassemi, M. (2020). Treating health disparities with artificial intelligence. *Nature Medicine*, 26(1), 16–17. https://doi.org/10.1038/s41591-019-0735-3
3. Little, R.J.A., Rubin, D.B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119482260
4. Schafer, J.L., Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

5. Horton, N.J., Kleinman, K.P. (2007). Much ado about nothing: A comparison of missing data methods and software for applied researchers. *The American Statistician*, 61(1), 79–90. https://doi.org/10.1198/000313007X172556
6. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520
7. Beretta, L., Santaniello, A. (2016). Nearest neighbor methods for missing value imputation. *BioMed Research International*, 2016, Article ID 5475638. https://doi.org/10.1155/2016/5475638
8. Van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Chapman and Hall/CRC Press. https://doi.org/10.1201/9780429492259
9. White, I.R., Royston, P., Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistical Methods in Medical Research*, 20(3), 377–399. https://doi.org/10.1177/0962280210375561
10. Jahangiri, S., et al. (2023). A wide range of missing imputation approaches in longitudinal data: A simulation study and real data analysis. *BMC Medical Research Methodology*, 23, 45. https://doi.org/10.1186/s12874-023-01877-0
11. Michael J. Fox Foundation: Parkinson's Progression Markers Initiative (PPMI). Overview of Study Expansion (2023). https://www.michaeljfox.org/news/parkinsons-progression-markers-initiative-ppmi
12. Chen, Z., Tan, S., Chajewska, U., Rudin, C., Caruana, R. (2023). Missing Values and Imputation in Healthcare Data: Can Interpretable Machine Learning Help? *Proceedings of Machine Learning Research*, 209, 86–99. https://proceedings.mlr.press/v209/chen23g.html
13. Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T. (2021). A survey on missing data in machine learning applications in healthcare datasets. *Journal of Big Data*, 8, 140. https://doi.org/10.1186/s40537-021-00523-9
14. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11, 3371–3408. http://jmlr.org/papers/v11/vincent10a.html
15. Kingma, D.P., Welling, M. (2014). Auto-Encoding Variational Bayes. *International Conference on Learning Representations (ICLR)*, Banff, Canada. https://arxiv.org/abs/1312.6114
16. Beaulieu-Jones, B.K., Moore, J.H., Williams, S.M. (2016). The machine learning toolbox: Applied machine learning in healthcare. *Pacific Symposium on Biocomputing*, 224–235. https://doi.org/10.1142/9789814749411_0022
17. Hardt, M., Price, E., Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 29, 3315–3323. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
18. Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54–63. https://doi.org/10.1145/3381831
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–35. https://doi.org/10.1145/3457607
20. Marek, K., et al. (2023). The Parkinson's Progression Markers Initiative (PPMI) – establishing a PD biomarker cohort. *Annals of Clinical and Translational Neurology*, 10(4), 568–583. https://doi.org/10.1002/acn3.51641