

Evaluating Missing Data Imputation Methods for Clinical Trial Analysis

Jiajun Wen, Xiamen Amoytop Biotech Co., LTD

ABSTRACT

Missing data is a common challenge in clinical trial analysis and may affect the reliability of efficacy and safety conclusions. Established approaches such as K-nearest neighbors, multiple imputation by chained equations, and other multiple imputation methods are widely used, but their performance can vary when clinical data include nonlinear relationships, high-dimensional covariates, or complex missingness mechanisms. This paper presents a practical comparison of missing data imputation methods under simulated MCAR, MAR, and MNAR settings using public clinical datasets. The evaluated methods include KNN, MICE, standard Generative Adversarial Imputation Network, warm-start GAIN using pre-filled values, and a pseudo-label conditional GAIN approach that incorporates latent subgroup information through clustering. Root mean square error is used as the primary evaluation metric, supported by visual checks of post-imputation data distributions. The results indicate that warm-start pre-filling can improve the stability and accuracy of standard GAIN in simpler missingness scenarios. In more complex MAR and MNAR settings, the pseudo-label conditional GAIN approach shows improved imputation accuracy compared with traditional multiple imputation and standard GAIN across the tested datasets. This work provides a practical evaluation framework for statisticians, clinical programmers, and data scientists who need to compare established and emerging imputation methods for clinical trial analysis.

Keywords: Missing Data, Clinical Trials, Multiple Imputation, MAR, MNAR, GAIN, PC-GAIN, Machine Learning

INTRODUCTION

Clinical trials rely on complete and credible data to support conclusions about efficacy, safety, and benefit-risk. In practice, missing data can arise from missed visits, early discontinuation, measurement failure, selective testing, remote data collection challenges, or patient dropout. Missing values reduce the usable information in a study and can also introduce bias when the missingness process is related to observed or unobserved patient characteristics (Wood, White, and Thompson 2004; Sterne et al. 2009).

The statistical literature usually describes three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). This distinction is not merely theoretical. PharmaSUG papers on missing outcome data, MNAR sensitivity analysis, and control-based pattern imputation emphasize that the method selected for imputation depends on the assumed missingness mechanism and on the clinical interpretation of the missing pattern (Li 2019; Qian and Lu 2023; Feng and Liu 2023).

Traditional methods such as KNN, MICE, and multiple imputation remain important because they are transparent and familiar in clinical statistics. At the same time, clinical datasets often contain nonlinear relationships, heterogeneous patient subgroups, and high-dimensional covariate patterns. These properties motivate a practical evaluation of whether generative adversarial models can complement established methods rather than replace them. This paper focuses on GAIN and PC-GAIN as machine-learning approaches for imputation, with direct comparison to traditional baselines.

WHY MISSING DATA ASSUMPTIONS MATTER IN CLINICAL TRIALS

A missing data method is only meaningful when its assumptions are aligned with the trial question and the data collection process. Under MCAR, the observed records can be treated as a random subset of the complete records. Under MAR, the probability of missingness can be explained by observed information. Under MNAR, the probability of missingness may depend on the missing values themselves, which makes sensitivity analysis and transparent assumptions especially important (Rubin 1976; Little and Rubin 2019).

Clinical examples rarely look like textbook MCAR. A patient may miss a follow-up assessment because of adverse events, lack of efficacy, disease progression, travel burden, or site-level operational disruption. In longitudinal studies, monotone and intermittent missing patterns can coexist. PharmaSUG work on PROC MI, MNAR statements, tipping point analysis, and ADaM datasets shows that missing data handling is both a statistical issue and a programming traceability issue (Smith and Kosten 2017; Li 2019; Zhao, Li, and Shi 2025).

In the ICH E9(R1) estimand framework, missing data handling should be aligned with the clinical question, the intercurrent event strategy, and the planned sensitivity analyses (International Council for Harmonisation 2019). For this reason, an imputation method should not be evaluated only as a technical data-processing step; it should also be judged by whether its assumptions are compatible with the target estimand and the way missingness may arise during trial conduct.

For that reason, this paper evaluates methods across simulated MCAR, MAR, and MNAR scenarios. The purpose is not to claim that a single method is universally best. The practical question is narrower: under which missingness conditions does warm-start GAIN improve over standard GAIN, and when does PC-GAIN provide useful additional value by incorporating latent subgroup structure?

TRADITIONAL AND GAN-BASED IMPUTATION METHODS

TRADITIONAL BASELINES

Mean imputation, KNN, and MICE were used as practical baselines. Mean imputation is simple but can distort variance and correlation structure. KNN uses local similarity across observed features, which can work well when nearby patients in feature space are clinically similar. MICE fits a sequence of conditional models and is widely used because it can handle multiple variables with missingness while reflecting uncertainty across imputed datasets (van Buuren and Groothuis-Oudshoorn 2011).

These methods are useful comparators because they represent different assumptions about the data. Mean imputation has little structure, KNN relies on local geometry, and MICE relies on a sequence of specified conditional models. A machine-learning method should be judged against these baselines, not only against an uninformative fill value.

GAIN AND WARM-START GAIN

GAIN adapts the generative adversarial network framework to missing data. The generator proposes values for missing entries, while the discriminator attempts to distinguish observed entries from imputed entries. A hint mechanism gives the discriminator partial information about the mask so that the generator must learn more than marginal feature distributions (Yoon, Jordon, and van der Schaar 2018).

In a standard implementation, missing entries may be initialized with zero or random noise before adversarial training. This cold-start strategy is easy to implement but may give the generator a poor starting point when the actual feature distribution is far from zero. In this study, KNN and MICE pre-filling were evaluated as warm-start strategies. The working hypothesis was that a better initial approximation would guide the generator toward a more accurate solution.

Figure 1 illustrates the warm-start version used in this evaluation. Compared with standard GAIN, the prior model supplies a pre-filled matrix from methods such as KNN before generator training. The observed matrix and mask matrix still preserve the original missingness pattern, while the MSE and cross-entropy losses guide reconstruction and adversarial learning.

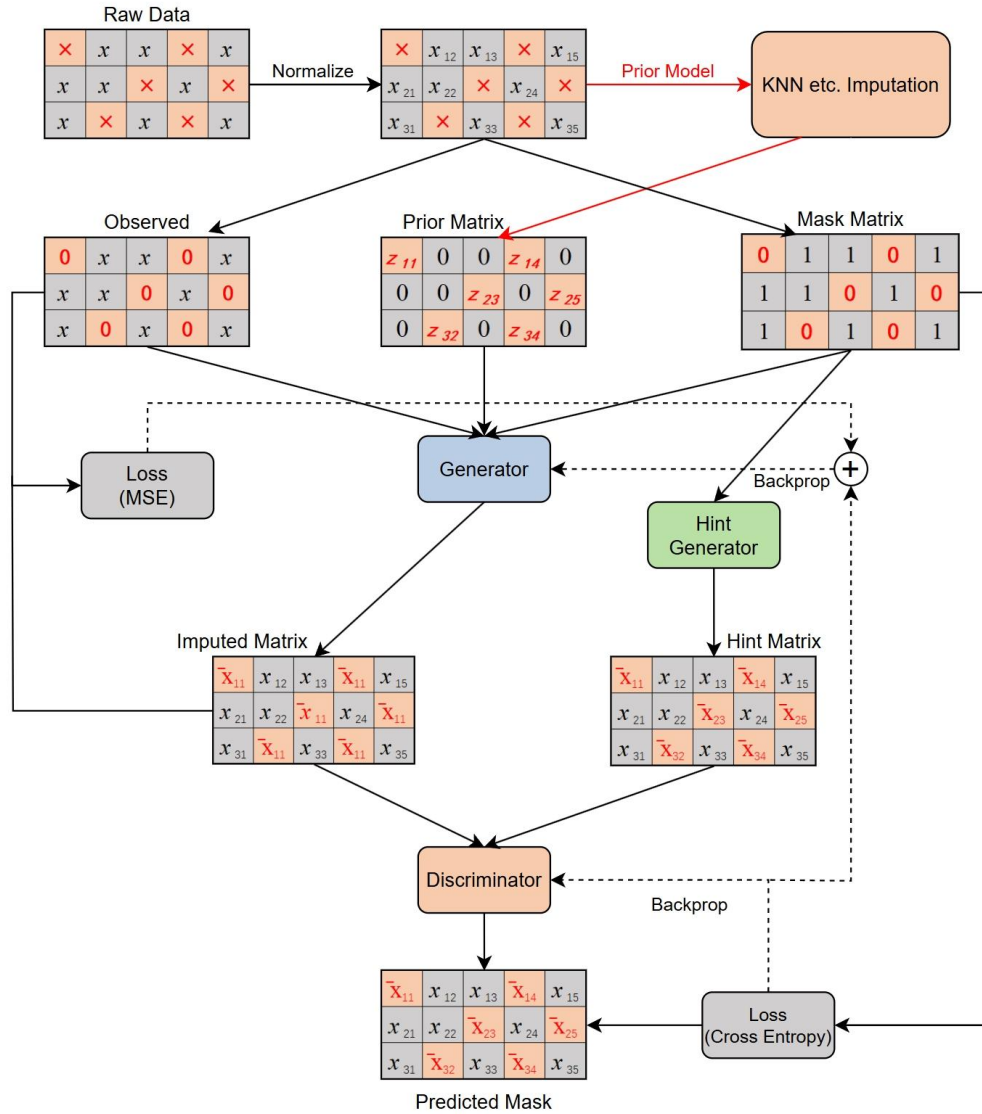


Figure 1. Warm-Start GAIN Imputation Workflow

PC-GAIN

PC-GAIN extends GAIN by using pseudo labels derived from unsupervised clustering. The model first learns latent structure in the partially observed data, assigns pseudo labels, and then uses those labels as conditional information during imputation. This design is related to the broader idea that patient subgroups can carry clinically meaningful structure. Cluster analysis has also been discussed in PharmaSUG as a way to discover natural groupings across related variables (Wittle and Stackhouse 2019).

The intuition is especially relevant under MAR and MNAR settings. If missingness differs across latent patient subgroups, then conditioning on pseudo labels can help the imputation model learn within-subgroup relationships. This does not solve all MNAR problems, but it gives the model a structured way to use heterogeneity that a standard GAIN model may not explicitly preserve (Wang et al. 2021).

Figure 2 shows how PC-GAIN adds clustering and a classifier to the GAIN workflow. Pseudo labels create a constraint signal so the generator can condition imputation on latent subgroup structure.

DATA SOURCES

The evaluation used three complete public clinical datasets for controlled missingness simulation and one public clinical dataset with real missingness. The complete datasets allowed direct RMSE comparison because the true values were known before missingness was introduced.

Although these public datasets are not CDISC datasets, they were selected to approximate common clinical trial data structures. The Parkinsons longitudinal voice measurements can be viewed as analogous to repeated clinical outcome or questionnaire records, such as ADQS-like data. The WDBC morphology features provide a biomarker-style complete dataset similar in spirit to laboratory or derived biomarker parameters, while the Heart Disease dataset provides a real-missingness example based on diagnostic procedure variables.

Table 1. Datasets Used in the Evaluation

Dataset	Role in evaluation	Clinical context	Notes
WDBC	Complete-data simulation	Breast cancer diagnosis	Cell nucleus morphology features from fine needle aspirate images.
Parkinsons	Complete-data simulation	Remote monitoring of Parkinson disease	Longitudinal voice measurements from 42 early-stage patients.
Diabetes	Complete-data simulation	Diabetes health indicators	Public health indicators from CDC BRFSS-derived data.
Heart Disease	Real missingness check	Cardiac diagnostic indicators	297 patient records with structured missingness in test-related variables.

MISSINGNESS SIMULATION

Unless otherwise noted, 25% missingness was introduced for controlled experiments. MCAR was generated by randomly masking entries. MAR was generated by allowing missingness in target variables to depend on observed control variables. MNAR was approximated through a within-cluster MAR design: latent classes were first defined from selected variables, and then class-specific MAR rules were applied. This design reflects the practical idea that patient subgroups can have different missingness behavior.

The Heart Disease dataset was used as a real missingness check rather than as an RMSE benchmark. In that dataset, baseline variables such as age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiogram result, and diagnosis were complete, while variables related to exercise testing and specialized diagnostic procedures had high missingness. This pattern is clinically plausible because not every patient is selected for every diagnostic test.

METHODS COMPARED

- Mean imputation, used as a simple low-information baseline.
- KNN imputation, using local similarity across observed features.
- MICE, implemented as iterative multivariate imputation by chained equations.
- Standard GAIN with zero initialization.
- Warm-start GAIN with mean, KNN, or MICE pre-filling.
- PC-GAIN with clustering-based pseudo labels and conditional generation.

EVALUATION METRIC

The primary metric was root mean square error (RMSE) between the original complete values and the imputed values at masked positions. Lower RMSE indicates closer recovery of the original data. RMSE was chosen because it gives a direct accuracy measure for controlled simulations, although interpretation should be paired with distributional checks and clinical plausibility review (Chai and Draxler 2014).

IMPLEMENTATION NOTES FROM THE PYTHON WORKFLOW

The implementation used Python with NumPy, pandas, scikit-learn, TensorFlow compatibility mode, and openpyxl for result output. The core workflow loaded a public dataset, generated the missingness mask, trained PC-GAIN, and compared RMSE values against traditional baselines. Displays 1 and 2 show the MNAR configuration, pseudo-label training logic, and RMSE comparison used in the code.

The displayed code uses TensorFlow compatibility mode because the implementation follows the original GAIN-style graph-execution framework. This choice supports reproducibility of the experimental code used in this evaluation; a production implementation could be refactored into TensorFlow 2.x or PyTorch after the statistical behavior is validated.

```
diabetes_mnar_params = {
    'latent_class_def': {'col_idx': 5, 'quantiles': [0.5]},
    'class_mar_configs': [
        {'target_cols': [1, 2, 3, 4], 'control_cols': [6, 7, 8], 'miss_rate': 0.25},
        {'target_cols': [1, 2, 3, 4], 'control_cols': [6, 7, 8], 'miss_rate': 0.30}
    ]
}
data_x, miss_data_x, data_M = data_loader(
    data_name='Diabetes', miss_rate=0.2, missingness_type='MNAR',
    mnar_params=diabetes_mnar_params)
imputed_data = PC_GAIN(miss_data_x, gain_parameters, data_M)
```

Display 1. Example MNAR Configuration and PC-GAIN Invocation

The PC-GAIN function normalizes the incomplete data, pre-fills missing entries using an iterative imputer, selects lower-missingness records for pre-training, clusters the pre-imputed records, trains a linear SVM to assign pseudo labels, and then updates the generator and discriminator with pseudo-label supervision. This workflow is practical for experimentation but should be hardened before regulated use, including fixed random seeds, version-controlled environments, independent validation, and documented quality checks.

Display 2 highlights the part of the implementation that converts clustering results into pseudo-label supervision and then compares PC-GAIN with traditional baselines.

```
# Cluster pre-imputed records and build pseudo labels
if cluster_species == 'KM':
    data_c, data_class = KM(imputed_data_L, k)
elif cluster_species == 'AC':
    data_c, data_class = AC(imputed_data_L, k)

coder = preprocessing.OneHotEncoder()
model = svm.SVC(kernel='linear', decision_function_shape='ovo')
coder.fit(data_class.reshape(-1, 1))
model.fit(imputed_data_L, data_class)

# Add pseudo-label supervision during generator updates
Hat_X_curr = sess.run(Hat_X, feed_dict={X: X_mb, M: M_mb, H: H_mb})
y_pred = model.predict(Hat_X_curr)
sample_prob = coder.transform(y_pred.reshape(-1, 1)).toarray()
sess.run(G_solver, feed_dict={X: X_mb, M: M_mb, H: H_mb, Y: sample_prob})

# Compare imputation accuracy at masked positions
rmse_pc_gain = rmse_loss(data_x, imputed_data, data_M)
rmse_mice = rmse_loss(data_x, mice_imputed_data, data_M)
rmse_knn = rmse_loss(data_x, knn_imputed_data, data_M)
```

Display 2. Pseudo-Label Training and RMSE Comparison Logic

RESULTS

BASELINE IMPUTATION UNDER MCAR

Table 2 compares basic imputation methods after introducing 25% MCAR missingness into the complete datasets. Mean imputation was generally weaker than the more structured approaches. MICE performed best on WDBC and Diabetes, while KNN performed best on Parkinsons. Standard GAIN outperformed mean imputation but did not consistently outperform KNN or MICE.

Table 2. Baseline RMSE Under 25% MCAR Missingness

Method	WDBC	Parkinsons	Diabetes
Mean imputation	0.2088	0.1741	0.3255
KNN	0.1872	0.1277	0.3389
MICE	0.1338	0.1513	0.3067
Zero-initialized GAIN	0.1519	0.1371	0.3210

WARM-START STRATEGIES FOR GAIN

Table 3 and Figure 3 show that replacing zero initialization with pre-filled values improved GAIN performance in most settings. KNN-initialized GAIN produced the lowest RMSE across all three complete datasets in the 25% MCAR comparison. MICE initialization was also strong, especially for WDBC.

Table 3. RMSE for GAIN With Different Initial Values

Method	WDBC	Parkinsons	Diabetes
GAIN	0.1519	0.1371	0.3210
Mean-initialized GAIN	0.1471	0.1452	0.3156
KNN-initialized GAIN	0.1197	0.1244	0.2980
MICE-initialized GAIN	0.1219	0.1235	0.3006

Warm-Start GAIN RMSE

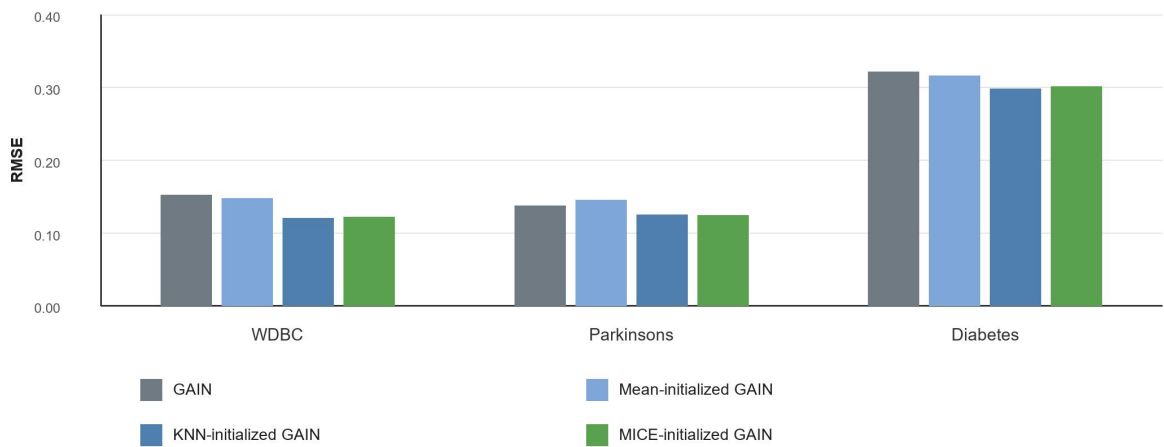


Figure 3. Warm-Start GAIN RMSE Across Complete Datasets

The result supports a practical recommendation: if an adversarial imputation model is used, it is worth testing a clinically interpretable pre-filling method rather than relying only on zero initialization. Warm-starting can give the generator a more informative starting point and may improve stability during training.

ROBUSTNESS ACROSS MISSINGNESS RATES

Table 4 evaluates WDBC at 10%, 25%, and 50% MCAR missingness. As expected, RMSE generally increased as the missingness rate increased. The best initialization strategy varied by rate: MICE-initialized GAIN was strongest at 10%, while KNN-initialized GAIN was strongest at 25% and 50%. This pattern suggests that local-neighbor information may be more useful when the dataset becomes more sparse.

Table 4. RMSE on WDBC Across MCAR Missingness Rates

Method	10%	25%	50%
GAIN	0.1408	0.1519	0.1684
Mean-initialized GAIN	0.1231	0.1471	0.1731
KNN-initialized GAIN	0.1326	0.1197	0.1433
MICE-initialized GAIN	0.1026	0.1219	0.1533

PC-GAIN UNDER MAR AND MNAR

Tables 5 and 6 compare PC-GAIN with MICE and KNN-initialized GAIN under simulated MAR and MNAR mechanisms. PC-GAIN had the lowest RMSE across all three datasets in both settings. The advantage was especially important because MAR and MNAR are closer to many clinical trial missingness patterns than simple MCAR masking.

Table 5. RMSE Under Simulated MAR Missingness

Method	WDBC	Parkinsons	Diabetes
MICE	0.1922	0.2659	0.3609
KNN-initialized GAIN	0.1876	0.3141	0.4739
PC-GAIN	0.1626	0.2165	0.2942

Table 6. RMSE Under Simulated MNAR Missingness

Method	WDBC	Parkinsons	Diabetes
MICE	0.5468	0.4364	0.3684
KNN-initialized GAIN	0.5642	0.4684	0.3940
PC-GAIN	0.4430	0.3636	0.3489

PC-GAIN Under MAR and MNAR

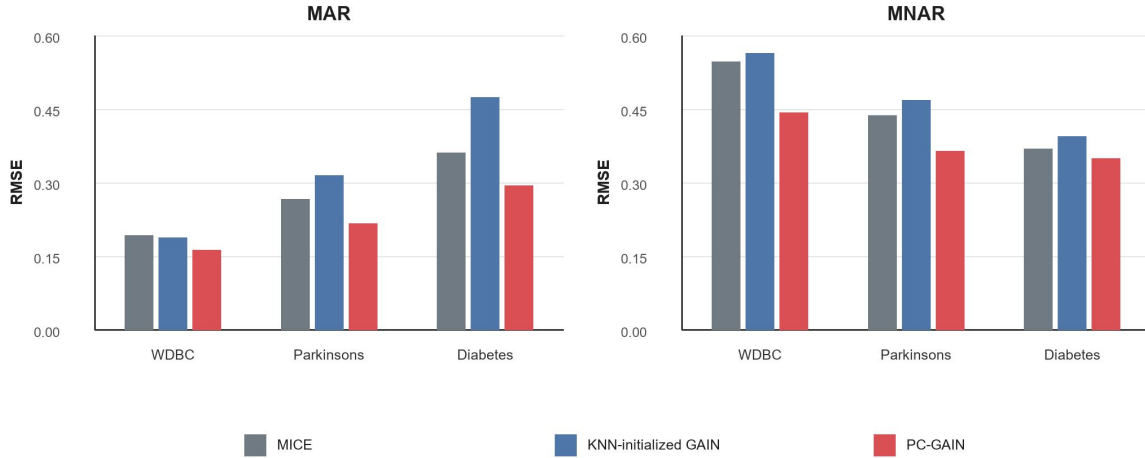


Figure 4. PC-GAIN Compared With MICE and Warm-Start GAIN Under MAR and MNAR

The MAR result suggests that pseudo-label conditioning can help when missingness is related to observed variables. The MNAR result should be interpreted more carefully because MNAR assumptions cannot be verified from observed data alone. Still, within the controlled simulation used here, PC-GAIN was better able to recover masked values than the tested traditional and warm-start GAIN alternatives.

DISTRIBUTIONAL CHECK ON REAL MISSINGNESS

The real-missingness dataset was used as a distributional diagnostic rather than an RMSE benchmark. For selected variables with substantial observed missingness, the observed values before imputation were compared with the completed values after GAIN imputation. This check is not a substitute for inferential validation, but it helps identify whether the imputation process introduces obvious marginal distribution shifts.

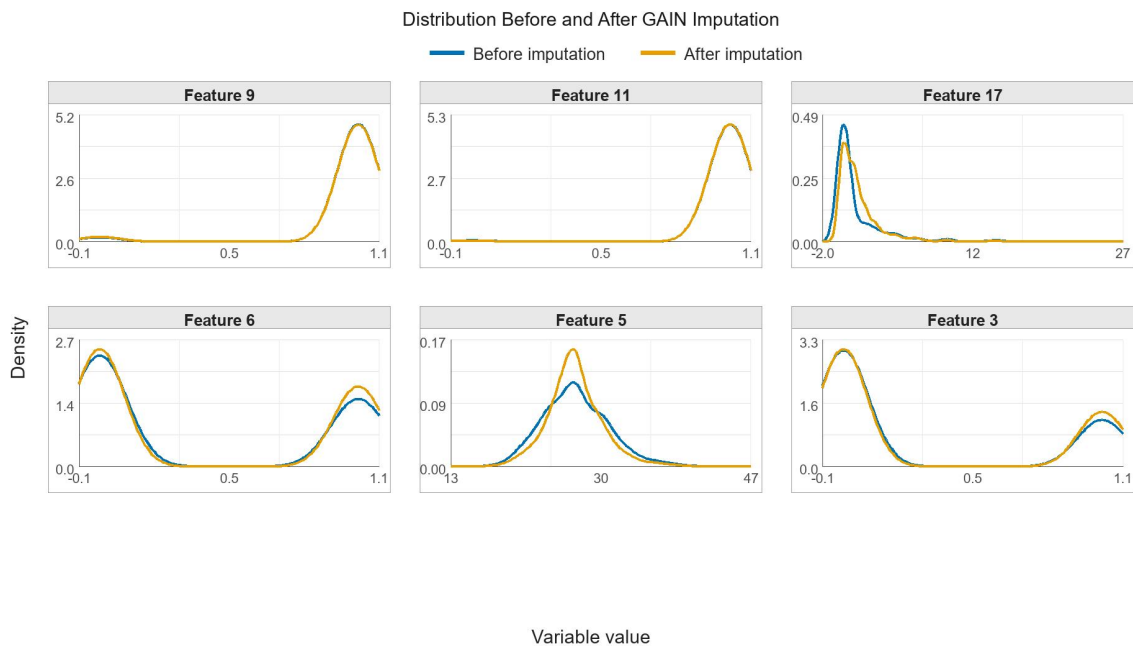


Figure 5. Distribution Before and After GAIN Imputation for Selected Variables

As shown in Figure 5, the before- and after-imputation density curves are generally aligned for several selected variables, while some variables show moderate smoothing or peak changes after completion. This pattern is expected when high-missingness variables are completed and reinforces the practical value of using distributional plots alongside RMSE and sensitivity analyses when evaluating imputation methods.

PRACTICAL CONSIDERATIONS FOR CLINICAL PROGRAMMING AND STATISTICAL REVIEW

Machine-learning imputation should be evaluated as part of a controlled analysis workflow. The method may be useful for data exploration, simulation, sensitivity analysis planning, or as a candidate imputation approach in research settings. In regulated clinical trial analysis, however, any use would need prespecified assumptions, reproducible implementation, traceable outputs, and statistician review.

- Prespecify the missingness assumptions and justify them clinically.
- Compare machine-learning imputation with established statistical methods rather than using it in isolation.
- Use multiple diagnostics, including RMSE in simulation, distributional plots, correlation preservation, and downstream analysis sensitivity.
- Fix random seeds and software versions, and retain logs of imputation parameters, model settings, and generated outputs.
- Document how imputed values enter analysis datasets, especially when ADaM-style traceability is required.
- Treat MNAR results as sensitivity evidence, not proof that the true missingness mechanism has been recovered.

Recent PharmaSUG work on MI ADaM datasets reinforces this traceability point. Even when the imputation method is statistically sound, programmers must still make the imputed records transparent, validation-ready, and consistent with the intended analysis structure (Zhao, Li, and Shi 2025; Xie 2025).

LIMITATIONS

This study has several limitations. First, the core RMSE comparisons were based on simulated missingness applied to public datasets. These simulations make objective evaluation possible, but they simplify the clinical and operational causes of missingness in real trials. Second, PC-GAIN was evaluated as an imputation accuracy method rather than as a full regulatory analysis strategy. Additional work is needed to connect imputed values to downstream treatment effect estimates, confidence intervals, and sensitivity analyses.

Third, the MNAR experiment used a within-cluster MAR approximation. This design is useful for evaluating whether latent subgroup structure helps imputation, but it does not remove the fundamental untestability of MNAR assumptions. Fourth, GAN-based methods are more complex than MICE or KNN and may require more training diagnostics, computational resources, and explanation before they can be accepted by clinical reviewers.

Finally, this evaluation focused on pointwise imputation accuracy using RMSE. It did not evaluate variance estimation, Rubin's rules, Type I error, or treatment effect inference after multiple imputation. In the current implementation, GAIN and PC-GAIN are used primarily as single completed-data generators. A future extension could generate multiple completed datasets through repeated noise sampling or model refitting and then evaluate whether pooled inference behaves similarly to established multiple-imputation workflows.

CONCLUSION

This paper presented a practical evaluation of traditional and GAN-based imputation methods for missing clinical trial data. In simple MCAR settings, KNN and MICE remained strong baselines, and standard GAIN did not automatically outperform them. Warm-start strategies improved GAIN performance, especially when KNN or MICE was used to provide initial values.

The most important finding was that PC-GAIN achieved lower RMSE than MICE and KNN-initialized GAIN across all tested datasets under simulated MAR and MNAR settings. This suggests that incorporating latent subgroup information through pseudo-label conditioning can improve imputation quality when missingness is associated with heterogeneous patient characteristics. For clinical programming and statistical practice, the recommended position is balanced: GAN-based imputation can complement established missing data strategies, but it should be implemented with transparent assumptions, reproducible code, and careful validation.

DATA AND CODE AVAILABILITY

The evaluation used public clinical datasets described in the paper and Python code prepared by the author for missingness simulation, GAIN, PC-GAIN, and RMSE comparison. The code and thesis source materials are stored with the draft package in the working project folder.

REFERENCES

- Chai, T. and R. R. Draxler. 2014. "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Arguments Against Avoiding RMSE in the Literature." *Geoscientific Model Development*, 7:1247-1250. <https://doi.org/10.5194/gmd-7-1247-2014>.
- Feng, J. and J. Liu. 2023. "Sensitivity Analysis for Missing Data Using Control-based Pattern Imputation." *Proceedings of PharmaSUG 2023*, Paper SA-188. Available at <https://pharmasug.org/proceedings/2023/SA/PharmaSUG-2023-SA-188.pdf>.
- Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative Adversarial Nets." *Advances in Neural Information Processing Systems* 27.
- International Council for Harmonisation. 2019. "E9(R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials." Available at https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf.
- Li, L. 2019. "SAS V9.4 MNAR Statement for Multiple Imputations for Missing Not at Random in Longitudinal Clinical Trials." *Proceedings of PharmaSUG 2019*, Paper ST-103. Available at <https://pharmasug.org/proceedings/2019/ST/PharmaSUG-2019-ST-103.pdf>.
- Little, R. J. A. and D. B. Rubin. 2019. *Statistical Analysis with Missing Data*. 3rd ed. Hoboken, NJ: John Wiley & Sons.
- Qian, X. and C. Lu. 2023. "Statistical Considerations and Methods for Handling Missing Outcome Data During the Era of COVID-19." *Proceedings of PharmaSUG 2023*, Paper SA-269. Available at <https://pharmasug.org/proceedings/2023/SA/PharmaSUG-2023-SA-269.pdf>.
- Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika*, 63(3):581-592. <https://doi.org/10.1093/biomet/63.3.581>.
- Smith, C. and S. Kosten. 2017. "Multiple Imputation: A Statistical Programming Story." *Proceedings of PharmaSUG 2017*, Paper SP01. Available at <https://pharmasug.org/proceedings/2017/SP/PharmaSUG-2017-SP01.pdf>.
- Sterne, J. A. C., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. 2009. "Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls." *BMJ*, 338:b2393. <https://doi.org/10.1136/bmj.b2393>.
- van Buuren, S. and K. Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, 45(3):1-67. <https://doi.org/10.18637/jss.v045.i03>.
- Wang, Y., D. Li, X. Li, and M. Yang. 2021. "PC-GAIN: Pseudo-label Conditional Generative Adversarial Imputation Networks for Incomplete Data." *Neural Networks*, 141:395-403. <https://doi.org/10.1016/j.neunet.2021.05.033>.
- Wittle, A. and M. Stackhouse. 2019. "Cluster Analysis: What It Is and How to Use It." *Proceedings of PharmaSUG 2019*, Paper ST-183. Available at <https://pharmasug.org/proceedings/2019/ST/PharmaSUG-2019-ST-183.pdf>.

Wood, A. M., I. R. White, and S. G. Thompson. 2004. "Are Missing Outcome Data Adequately Handled? A Review of Published Randomized Controlled Trials in Major Medical Journals." *Clinical Trials*, 1(4):368-376. <https://doi.org/10.1191/1740774504cn032oa>.

Xie, K. 2025. "Approaches to Developing Multiple Imputation ADaM Dataset." Proceedings of PharmaSUG 2025, Paper SI-084. Available at <https://pharmasug.org/proceedings/2025/SI/PharmaSUG-2025-SI-084.pdf>.

Yoon, J., J. Jordon, and M. van der Schaar. 2018. "GAIN: Missing Data Imputation Using Generative Adversarial Nets." Proceedings of the 35th International Conference on Machine Learning, 5689-5698. Available at <https://proceedings.mlr.press/v80/yon18a.html>.

Zhao, S., L. Li, and Y. Shi. 2025. "Multiple Imputation Techniques in the Context of ADaM Datasets." Proceedings of PharmaSUG 2025, Paper DS-067. Available at <https://pharmasug.org/proceedings/2025/DS/PharmaSUG-2025-DS-067.pdf>.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Yalin Yin from the Clinical Research Center for her professional guidance on clinical research considerations and to Zhengfang Du from the Medical Development Center, Data Science Department, for her valuable advice on data science workflow and validation considerations. We also thank colleagues from Xiamen Amoytop Biotech Co., LTD for their support on clinical trial data processing, missing data methodology, and programming implementation, and acknowledge the public dataset providers used in this evaluation.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jiajun Wen

Xiamen Amoytop Biotech Co., LTD

E-mail: wenjiajun@amoytop.com

Any brand and product names are trademarks of their respective companies.