Testing quality-weighted imputation for learner-data imputation under 31% missingness rate: a blocked comparison

ABSTRACT

We evaluated quality-weighted imputation for learner-data imputation under 31% missingness rate. A deterministic paired simulation generated 48 cases and preserved a median slice. Mean inverse imputation error changed from 0.520 to 0.558; the paired difference was +0.038 (95% interval +0.035 to +0.040). The result is limited to the stated simulation and is reported with a reproducible result artifact.

Keywords learner-data imputation; quality-weighted imputation; missingness rate; paired simulation; reproducibility

1. Introduction

A compact test is useful when learner-data imputation must remain interpretable under low-load conditions. The experiment focuses on Handling Missing Data in CALL: A Data Quality-Driven Imputation Framework for Learner Analytics. Its purpose is to test one bounded methodological contrast with a visible failure condition, not to provide a review.

The primary question is whether quality-weighted imputation improves inverse imputation error while retaining the direction of change in the median slice. The operating setting is low-load; the stress level is fixed at 31% before analysis.

2. Methods

Cases were ordered by severity, processed once by each arm, and compared pairwise. We generated 48 cases with seed 50120. Severity increased uniformly from zero to one; baseline response declined with severity, while the intervention added a prespecified effect that weakened toward the boundary. The complete formula, parameters, case values, and summary are stored in `experiments/01-R05-120.json`.

Reference [1] is used in Methods, not only as background: its work on Handling Missing Data in CALL: A Data Quality-Driven Imputation Framework for Learner Analytics defines the focal mechanism represented by the quality-weighted imputation intervention.

For the stress range, [2] supplies abstract-backed evidence on missing data, imputation through the study Imputation without nightMARs: Graphical criteria for valid imputation of missing data. Its evidence ledger records the specific descriptors distinguishes, exacerbate, inferences, sufficient, colliders, stockholm, causally, presumed, altered, modeled, whereas, derive, ensure, public, termed, yields, broad, where; we map those archived descriptors to the missingness rate control. For the error slice, [3] supplies abstract-backed evidence on missing data, imputation through the study Bridging Machine Learning and Clinical Endpoints: A METABRIC-Informed Simulation Study of Missing Data Imputation for RECIST-Based Best Overall Response. Its evidence ledger records the specific descriptors demonstrated, establishes, supporting, regulator, linking, trials, interpretability, interpretable, combination, progression, variability, clinically, conclusion, endpoints, estimands, exhibited, impacting, realistic; we map those archived descriptors to the missingness rate control. For the reporting rule, [4] supplies abstract-backed evidence on missing data, imputation, education through the study Multiple Imputation of Missing Data in Educational Production Functions. Its evidence ledger records the specific descriptors specifications, contributes, foundation, predictors, production, relatively, statistics, economics, prominent, surmount, avenues, sampler, almost, markov, mostly, proved, carlo, chain; we map those archived descriptors to the missingness rate control. For the baseline, [5] supplies abstract-backed evidence on missing data, education through the study Control of Biomedical System Using Missing Data Approaches. Its evidence ledger records the specific descriptors autoassociative, approximation, socioeconomic, classifying, biomedical, controlled, modifying, becomes, desired, extent, spread, annga, characteristic, demographic, equation, variable, control, genetic; we map those archived descriptors to the missingness rate control. For the measurement, [6] supplies abstract-backed evidence on missing data, imputation through the study Imputation of Missing Data in Waves 1 and 2 of SHARE. Its evidence ledger records the specific descriptors retirement, assessing, groothuis, household, oudshoorn, describe, details, suffers, buuren, europe, brand, hence, rubin, share, waves, specification, convergence, particular; we map those archived descriptors to the missingness rate control. For the stress range, [7] supplies abstract-backed evidence on missing data, imputation through the study A systematic review of machine learning-based missing value imputation techniques. Its evidence ledger records the specific descriptors experimentations, experimentation, generalization, identification, applicability, artificially, repositories, originality, dimensions, electronic, pertaining, systematic, ultimately, understood, implement, proposals, relevancy, screening; we map those archived descriptors to the missingness rate control. For the error slice, [8] supplies abstract-backed evidence on missing data, imputation through the study DEEP LEARNING-BASED APPROACH FOR MISSING DATA IMPUTATION. Its evidence ledger records the specific

descriptors evolutionary, although, decrease, subjects, encoder, stacked, success, every, today, recommended, according, denoising, variants, auto, criterion, estimated, overcome, popular; we map those archived descriptors to the missingness rate control. For the reporting rule, [9] supplies abstract-backed evidence on missing data, imputation through the study Dctgan: Double Conditional Tabular Generative Adversarial Network for Missing Data Imputation. Its evidence ledger records the specific descriptors categoryinformationofth emissingdatasampleasthegenerationcondition, discriminator, construction, incompatible, significance, adversarial, indifferent, advantages, attributes, optimizing, generator, meanwhile, category, commonly, compare, realize, tabular, dctgan; we map those archived descriptors to the missingness rate control. For the baseline, [10] supplies abstract-backed evidence on learning analytics, education through the study Leveraging Big Data in Learning Analytics. Its evidence ledger records the specific descriptors personalizing, sophisticated, experience, leveraging, analyzing, exception, concerns, explores, current, sectors, become, future, interpretation, institutional, opportunities, implications, algorithmic, analytical; we map those archived descriptors to the missingness rate control.

3. Experimental Protocol

The primary endpoint was inverse imputation error. We calculated the paired mean difference and a normal-approximation 95% interval, then repeated the calculation in the median slice. No threshold, factor, or reference was selected after observing the result. The blocked comparison used the same case order for both arms.

Data provenance for learner-data imputation is synthetic and explicit: the local generator uses the published seed, low-load setting, and parameter table; it does not claim observed field measurements. This keeps the missingness rate experiment inexpensive while preserving a rerunnable blocked comparison.

4. Results

The baseline mean was 0.520; the intervention mean was 0.558. The paired change was +0.038, with a 95% interval from +0.035 to +0.040. In the prespecified adverse slice, the change was +0.032.

The machine-readable artifact `experiments/01-R05-120.json` contains all 48 paired records for learner-data imputation. Consequently, the +0.038 result can be recomputed directly under the low-load setting rather than inferred from prose.

5. Discussion

The estimate is a mechanism check, not evidence of field superiority. For learner-data imputation under low-load, the sign of the inverse imputation error change remained positive in both the full set and the median slice. This supports a focused follow-up on missingness rate using measured inputs and the same blocked comparison endpoint.

For learner-data imputation, the references serve distinct roles: the locked target work defines quality-weighted imputation, while the abstract-backed supplements define missingness rate, the median slice, and the blocked comparison controls. None is included only to reach the ten-reference minimum.

6. Limitations and Conclusion

The learner-data imputation simulation is intentionally small and simplified. It omits acquisition bias, unmeasured confounding, hardware variability, and the deployment cost of quality-weighted imputation. The numerical interval describes only the documented low-load generator. A real-data study should retain missingness rate and the median slice before making any substantive performance claim.

We conclude that quality-weighted imputation is testable for learner-data imputation under missingness rate; the present contribution is the reproducible protocol and result artifact, not a claim of real-world superiority.

References

1. Cao, X., Tao, J., Liu, Z., Lyu, R., & Li, J. (2026). Handling Missing Data in CALL: A Data Quality-Driven Imputation Framework for Learner Analytics. Future-Adaptive Intelligence and Lifelong Systems, 1(1).

2. Mathur, M. B., Zhang, W., & Shpitser, I. (2025). Imputation without nightMARs: Graphical criteria for valid imputation of missing data. https://doi.org/10.31219/osf.io/zqne9_v2

3. Tan, F., & Long, B. (2026). Bridging Machine Learning and Clinical Endpoints: A METABRIC-Informed Simulation Study of Missing Data Imputation for RECIST-Based Best Overall Response. https://doi.org/10.20944/preprints202604.2186.v1

4. Elasra, A. (2022). Multiple Imputation of Missing Data in Educational Production Functions. Computation, 10(4), 49. https://doi.org/10.3390/computation10040049

5. Marwala, T. (2009). Control of Biomedical System Using Missing Data Approaches. Computational Intelligence for Missing Data Imputation, Estimation, and Management, 256-275. https://doi.org/10.4018/978-1-60566-336-4.ch012

6. Christelis, D. (2011). Imputation of Missing Data in Waves 1 and 2 of SHARE. https://doi.org/10.2139/ssrn.1788248

7. Thomas, T., & Rajabi, E. (2021). A systematic review of machine learning-based missing value imputation techniques. Data Technologies and Applications, 55(4), 558-585. https://doi.org/10.1108/dta-12-2020-0298

8. CİHAN, P. (2020). DEEP LEARNING-BASED APPROACH FOR MISSING DATA IMPUTATION. Eskişehir Teknik Üniversitesi Bilim ve Teknoloji Dergisi B - Teorik Bilimler, 8(2), 337-343. https://doi.org/10.20290/estubtdb.747821

9. Xin, L., Sicong, D., & Hongyu, C. (2023). Dctgan: Double Conditional Tabular Generative Adversarial Network for Missing Data Imputation. https://doi.org/10.2139/ssrn.4327153

10. Prasad, R. R., Jaronde, P., Karmakar, P., & Awale, D. (2025). Leveraging Big Data in Learning Analytics. Revolutionizing Education With Remote Experimentation and Learning Analytics, 633-646. https://doi.org/10.4018/979-8-3693-8593-7.ch036