Stress-testing quality-weighted imputation for learner-data imputation under 31% missingness rate: a threshold audit

ABSTRACT
We evaluated quality-weighted imputation for learner-data imputation under 31% missingness rate. A deterministic paired simulation generated 64 cases and preserved a median slice. Mean inverse imputation error changed from 0.481 to 0.534; the paired difference was +0.053 (95% interval +0.051 to +0.055). 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

This study isolates one operational question in learner-data imputation: whether a transparent control survives long-tail inputs. 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 long-tail; the stress level is fixed at 31% before analysis.

2. Methods

A fixed generator created matched inputs; only the prespecified intervention differed between arms. We generated 64 cases with seed 50034. 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-R05-034.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 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, [3] supplies abstract-backed evidence on missing data, imputation through the study A Hybrid Approach to Missing Data. Its evidence ledger records the specific descriptors associative, conjunction, consistent, developed, similarly, captures, recently, merges, proven, proves, dealt, intelligence, efficiently, principal, combines, genetic, auto, component; we map those archived descriptors to the missingness rate control. For the reporting rule, [4] supplies abstract-backed evidence on missing data, imputation through the study A Smart Review on Imputation Techniques for Handling Missing Data. Its evidence ledger records the specific descriptors academicians, discussion, respondent, replacing, articles, industry, interest, reasons, occurs, typing, smart, done, less, said, involved, doctors, people, above; we map those archived descriptors to the missingness rate control. For the baseline, [5] 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 measurement, [6] supplies abstract-backed evidence on missing data, imputation through the study Benchmarking Machine Learning Missing Data Imputation Methods in Large-Scale Mental Health Survey Databases. Its evidence ledger records the specific descriptors benchmarking, participants, behavioral, blockwise, promising, consists, powering, minutes, autism, entire, mental, simons, subset, suffer, midas, spark, skip, autoencoders; 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 comparative study of imputation techniques for missing values in healthcare diagnostic datasets. Its evidence ledger records the specific descriptors performing, prioritize, scientists, encourage, presence, carried, forward, ongoing, working, before, debate, poorly, recall, since, goal, hope, locf, challenging; 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 Water-Quality Data Imputation With High Percentage of Missing Values: A Machine Learning Approach. Its evidence ledger records the specific descriptors satisfactory, implemented, represented, generating, mainstream, positioned, competing, worldwide, adaboost, belonged, stations, located, spatial, surface, uruguay,

chosen, hubber, iacute; we map those archived descriptors to the missingness rate control. For the reporting rule, [9] supplies abstract-backed evidence on learning analytics, education through the study A Data Protection Framework for Learning Analytics. Its evidence ledger records the specific descriptors interventions, organisations, intervention, unacceptable, universities, alternative, inadvertent, continuing, individual, legitimate, protection, separating, automated, obtaining, provision, sensitive, european, already; we map those archived descriptors to the missingness rate control. For the baseline, [10] supplies abstract-backed evidence on missing data, imputation through the study COVID-19 Data Analysis: The Impact of Missing Data Imputation on Supervised Learning Model Performance. Its evidence ledger records the specific descriptors adaptability, underscoring, constrained, diagnostics, sensitivity, specificity, underscores, department, resilience, comparing, highlight, providing, concepci, hospital, logistic, machines, minimize, moderate; 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 threshold audit 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, long-tail setting, and parameter table; it does not claim observed field measurements. This keeps the missingness rate experiment inexpensive while preserving a rerunnable threshold audit.

4. Results

The baseline mean was 0.481; the intervention mean was 0.534. The paired change was +0.053, with a 95% interval from +0.051 to +0.055. In the prespecified adverse slice, the change was +0.045.

The machine-readable artifact `experiments/01-R05-034.json` contains all 64 paired records for learner-data imputation. Consequently, the +0.053 result can be recomputed directly under the long-tail setting rather than inferred from prose.

5. Discussion

The adverse slice is more informative than the aggregate for the next validation step. For learner-data imputation under long-tail, 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 threshold audit 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 threshold audit 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 long-tail 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. 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

3. Marwala, T. (2009). A Hybrid Approach to Missing Data. Computational Intelligence for Missing Data Imputation, Estimation, and Management, 45-70. https://doi.org/10.4018/978-1-60566-336-4.ch003

4. Sivakani, R., Rahila, J., Sudha, P., Priscila, S. S., Shynu, T., Minu, M. S., & Pradeep, V. (2025). A Smart Review on Imputation Techniques for Handling Missing Data. Machine Learning, Predictive Analytics, and Optimization in Complex Systems, 41-62. https://doi.org/10.4018/979-8-3373-5203-9.ch003

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

6. Prakash, P., Street, K., Narayanan, S., Fernandez, B. A., Shen, Y., & Shu, C. (2024). Benchmarking Machine Learning Missing Data Imputation Methods in Large-Scale Mental Health Survey Databases. https://doi.org/10.1101/2024.05.13.24307231

7. Joel, L. O., Doorsamy, W., & Paul, B. S. (2025). A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. International Journal of Data Science and Analytics, 20(7), 6357-6373. https://doi.org/10.1007/s41060-025-00825-9

8. Rodriguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., & Gorgoglione, A. (2021). Water-Quality Data Imputation With High Percentage of Missing Values: A Machine Learning Approach. https://doi.org/10.20944/preprints202105.0105.v1

9. Cormack, A. N. (2016). A Data Protection Framework for Learning Analytics. Journal of Learning Analytics, 3(1). https://doi.org/10.18608/jla.2016.31.6

10. Mello-Román, J. D., & Martínez-Amarilla, A. (2025). COVID-19 Data Analysis: The Impact of Missing Data Imputation on Supervised Learning Model Performance. Computation, 13(3), 70. https://doi.org/10.3390/computation13030070