STChealth

Leveraging Machine Learning to Reconcile Missing Demographic Data in IISs

AIRA 2026
National Meeting

Sara Brown, MPH, CHES
Epidemiologist II, Analytics

April 29, 2026

STChealth

The IIS of the Past, Present, and Future

Connected Intelligence

Predictive Insights

Improved Outcomes

Acknowledgements

Dr. Sam McGee – Data Scientist II, STChealth

Zoey Bulharowski – Data Scientist I, STChealth

Ousswa Kudia, MPH – Analytics Product Management Supervisor, STChealth

No conflicts of interest to disclose.

The Challenge

IIS data plays a critical role in guiding public health interventions and program development

Missing data is a widespread issue, compromising the validity of findings

If not properly handled, missing data can introduce bias, distort analysis outcomes, and misinform decisions

Traditional Methods

Omit or “Unknown”

Traditional analyses will omit records with missing data or label them as “unknown”

Computationally Expensive

In large, public health datasets, MICE utilizes a lot of time and computing resources

Multiple Imputation using Chained Equations (MICE)

Creates multiple complete datasets to reflect missing value ambiguity

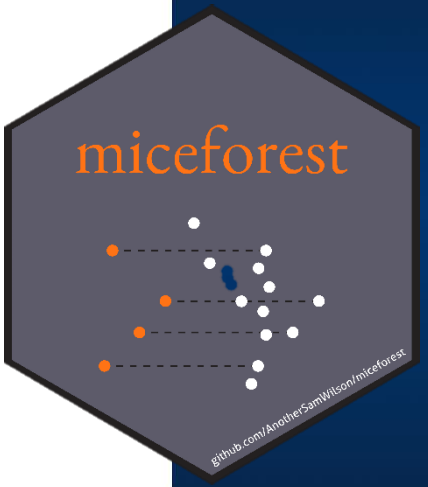
Managing Missing Data with Machine Learning

Machine learning (ML) offers robust methods for missing data reconciliation

Like MICE, ML depends on the type and pattern of missingness

Studies show that MICE has less bias and comparable standard errors compared to ML techniques

Combining MICE with ML can further reduce bias in imputed datasets

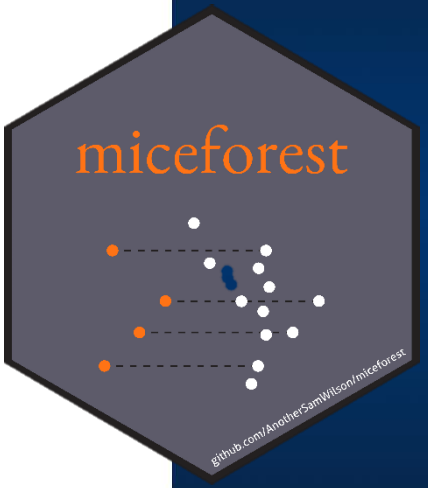


STChealth

Miceforest

Blends classical epidemiological methods with ML

Uses MICE and LightGBM, a gradient-boosted decision tree



Purpose

1. To evaluate the use of the ML technique Miceforest, in combination with cloud-based computing, for reconciling missing demographic data – compared to traditional MICE

2. Determine whether these methods impact flu vaccination coverage estimates

Materials & Methods

Data was obtained from a synthetic dataset based on census data, representing approximate IIS data distributions (N=4,472,365)

Variable	Missing (%)	Missing (n)
Gender	0.09%	4,154
Ethnicity	0.03%	1,299
Age	0.17%	7,650

Materials & Methods

Multiple Imputation Using Chained Equations (MICE)

STATA 17

Informing variables: race, city, county, & previous flu vaccine activity

15 imputations, 20 iterations

Miceforest

Python package, tree-based algorithm

Databricks platform

Informing variables: race, city, county, & previous flu vaccine activity

15 imputations, 20 iterations

Results

Through MICE and Miceforest, we obtained an additional 12,557 observations compared to the complete case analysis.

Flu coverage went from 49% to 24%

Computational time:

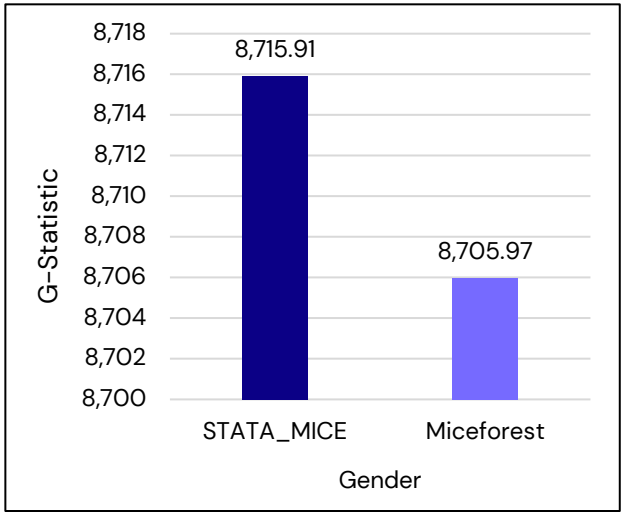
MICE = 8 hours

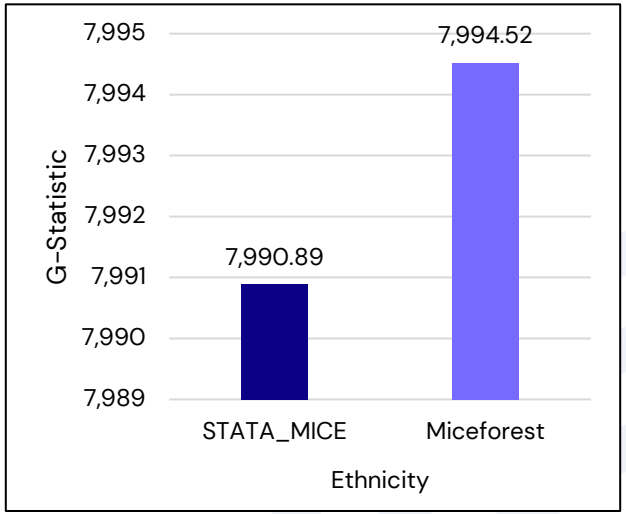
Miceforest = 10 minutes

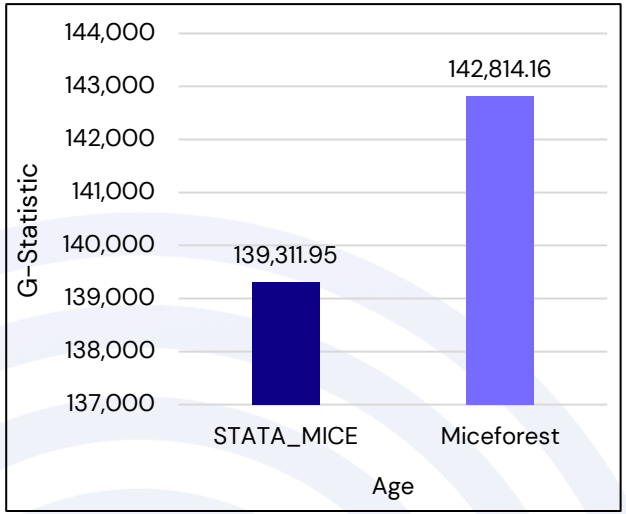
Likelihood Ratio Test Results

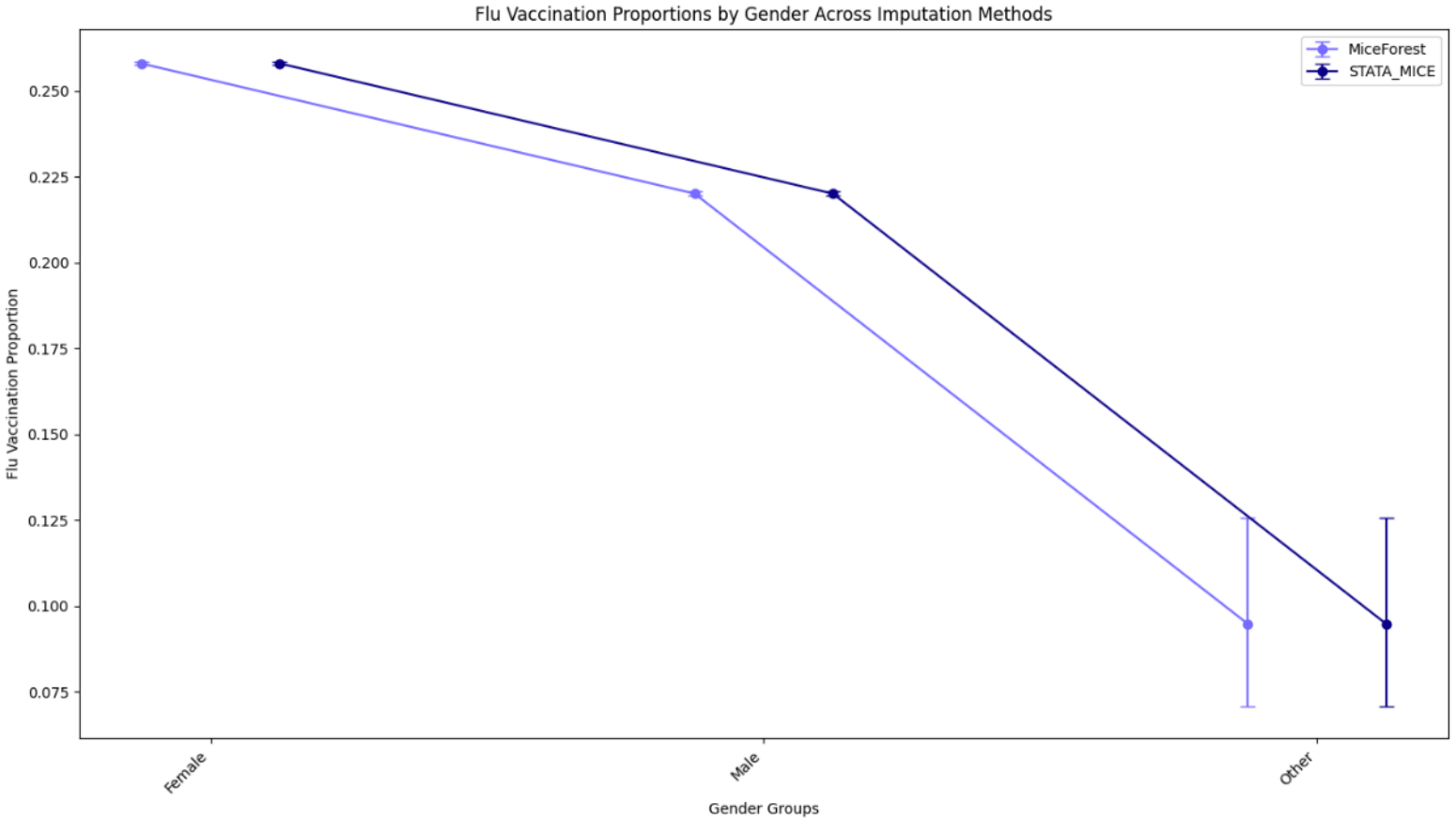
Variable	Method	G-Statistic	P-value	Degrees of Freedom	Between-Imputation Variance
Gender	STATA MICE	8715.905	<0.001	-	16.6847
	Miceforest	8705.9659	<0.001	2	-
Ethnicity	STATA MICE	7990.888	<0.001	-	4.1832
	Miceforest	7994.5242	<0.001	1	-
Age	STATA MICE	139311.949	<0.001	-	9588.596
	Miceforest	142814.1625	<0.001	5	-

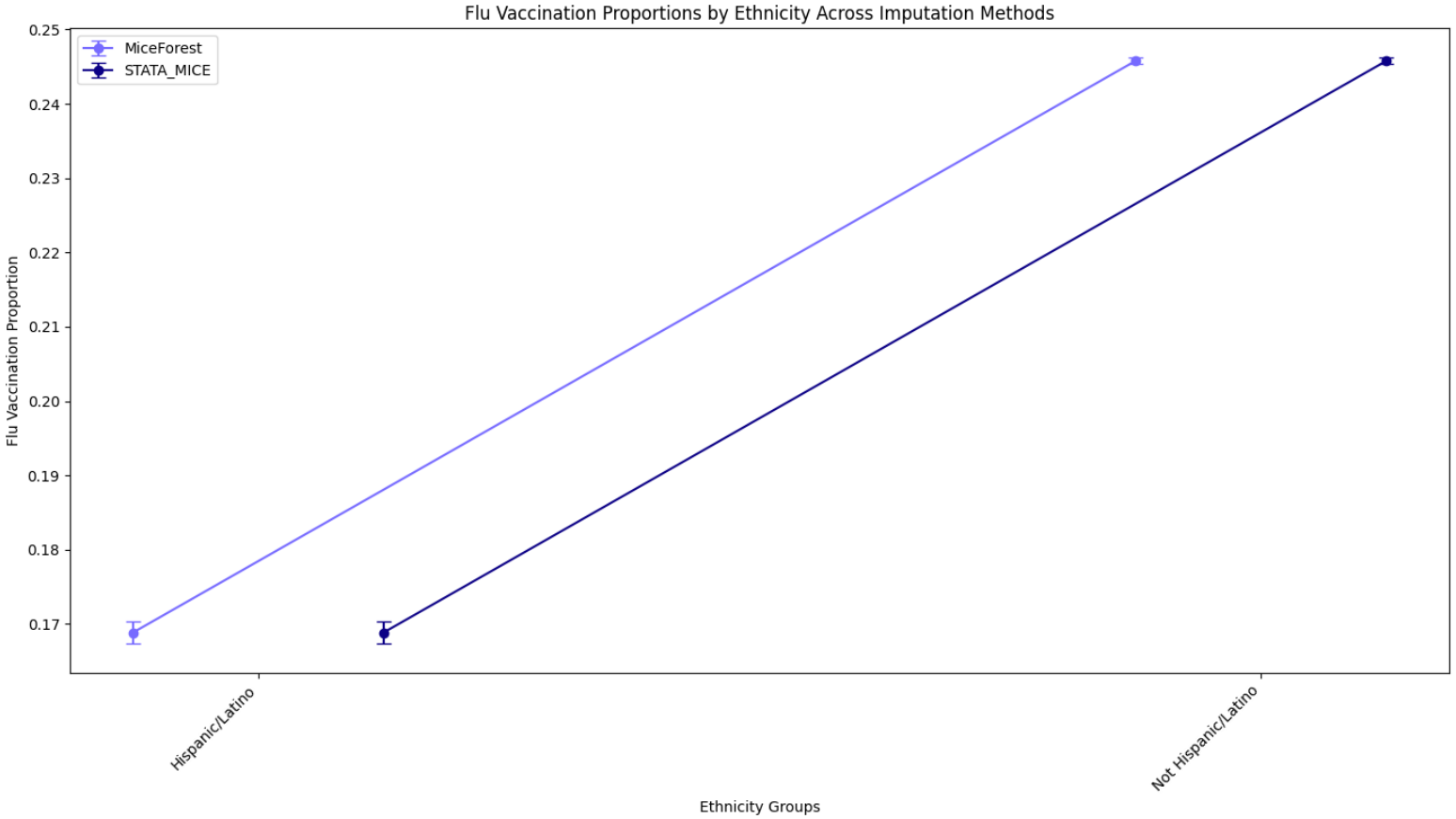
Likelihood Ratio Test Results

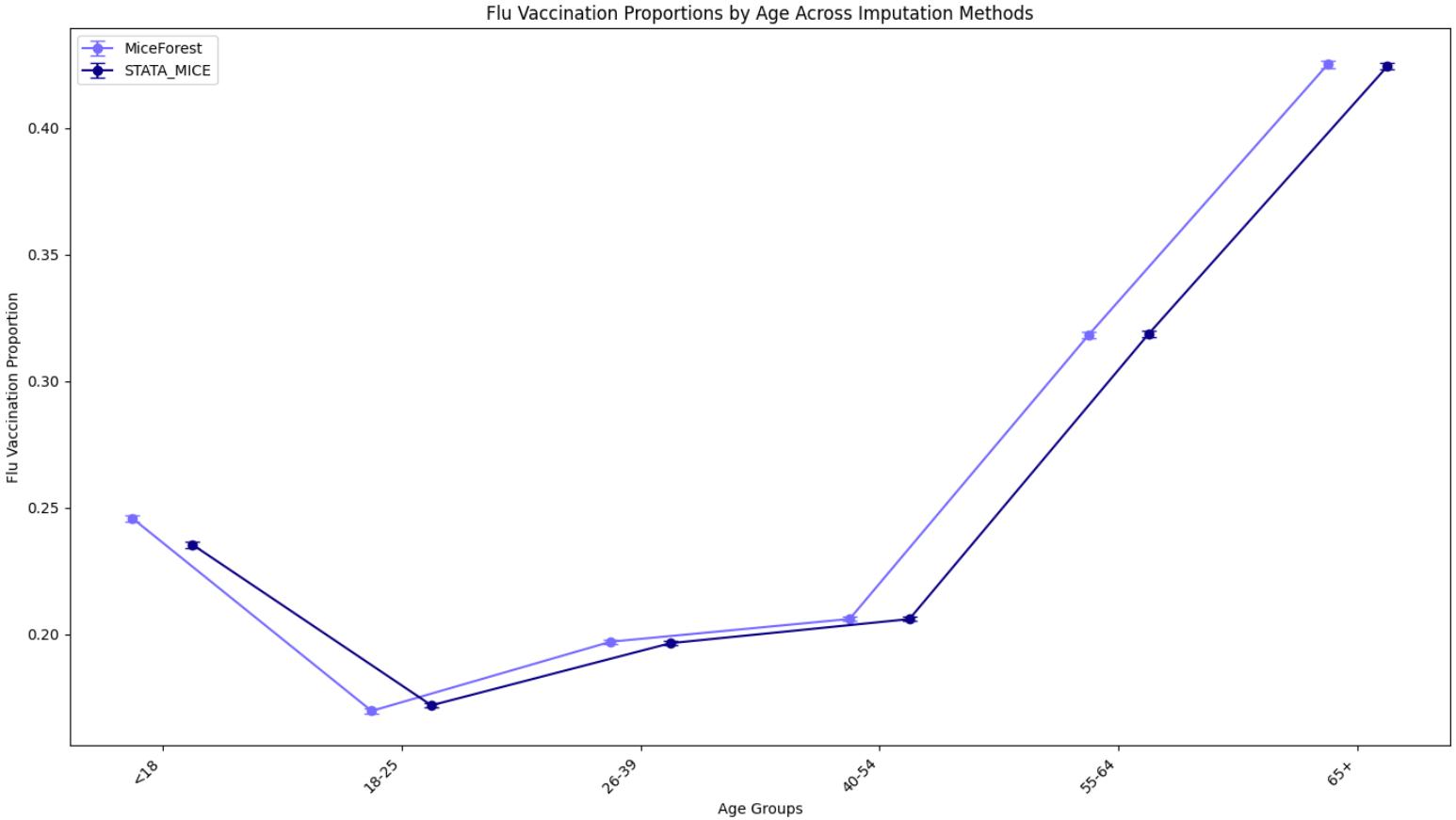




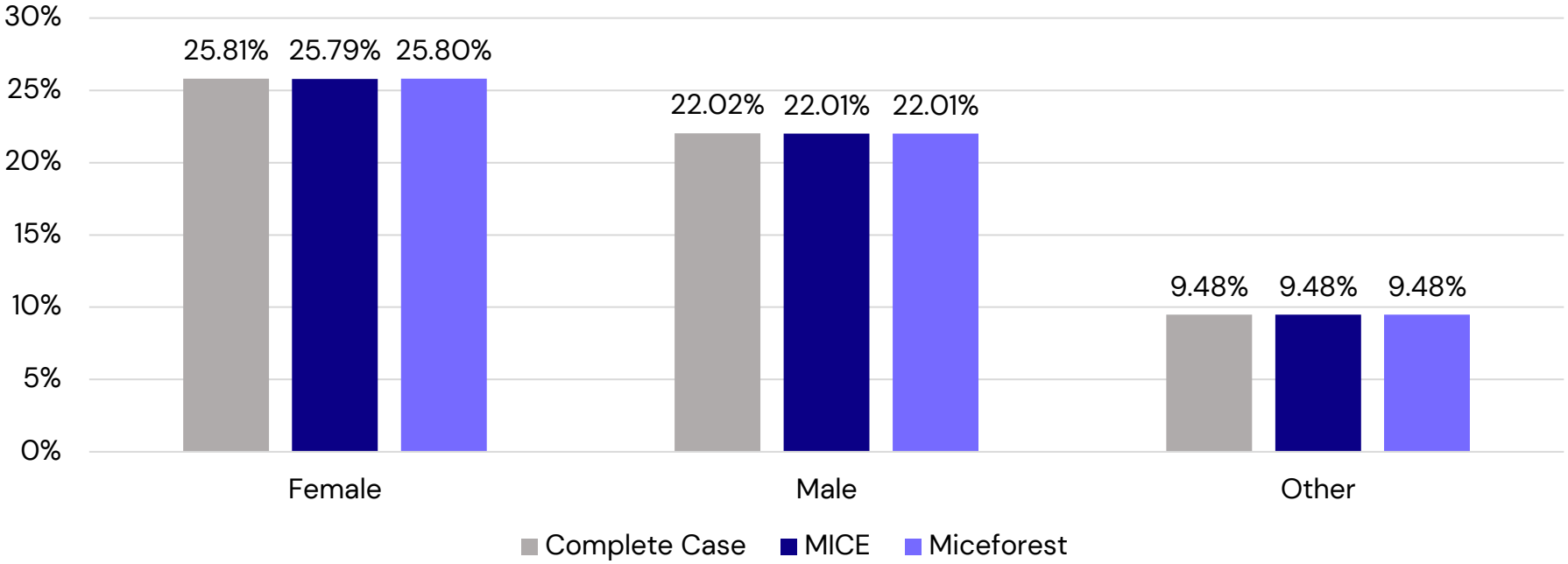



Flu Vaccination Proportions by Gender Across Imputation Methods
MiceForest
STATA_MICE
Flu Vaccination Proportion
0.250
0.225
0.200
0.175
0.150
0.125
0.100
0.075
Female
Male
Other
Gender Groups

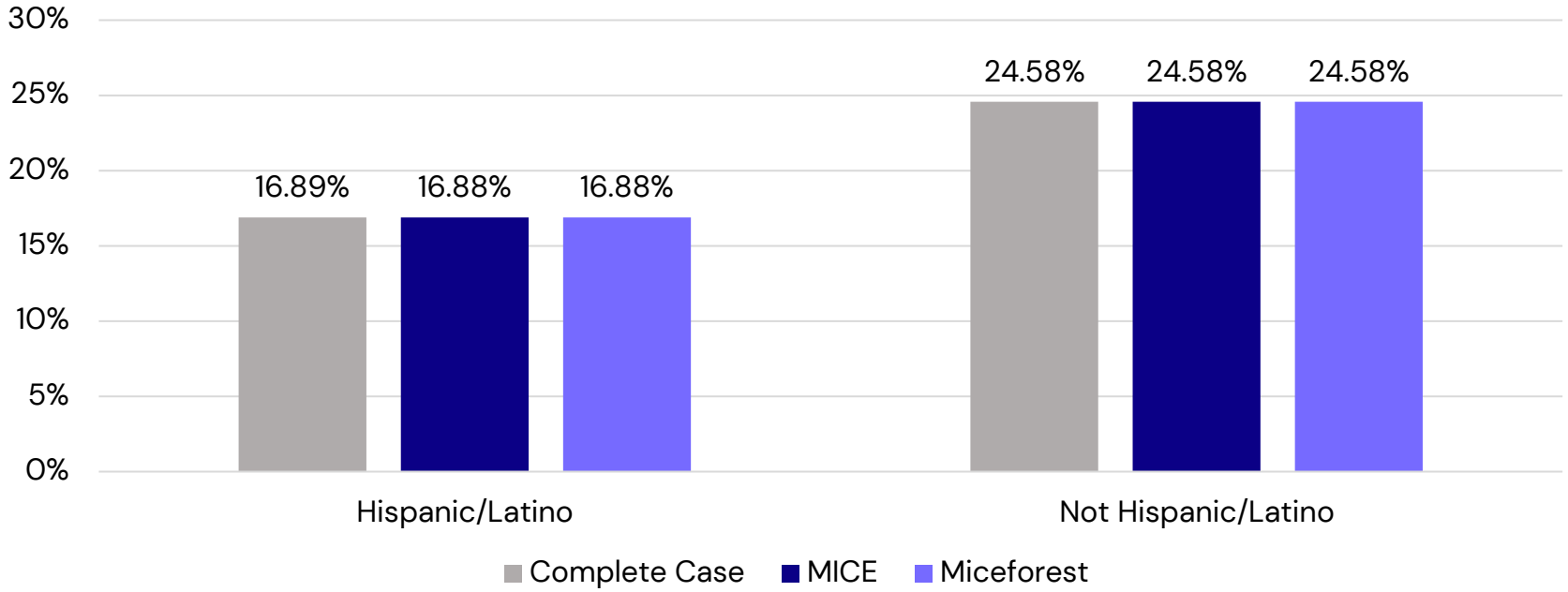

Flu Vaccination Proportions by Ethnicity Across Imputation Methods
MiceForest
STATA_MICE
Flu Vaccination Proportion
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17
Hispanic/Latino
Not Hispanic/Latino
Ethnicity Groups


Flu Vaccination Proportions by Age Across Imputation Methods
MiceForest
STATA_MICE
Flu Vaccination Proportion
0.40
0.35
0.30
0.25
0.20
<18
18-25
26-39
40-54
55-64
65+
Age Groups

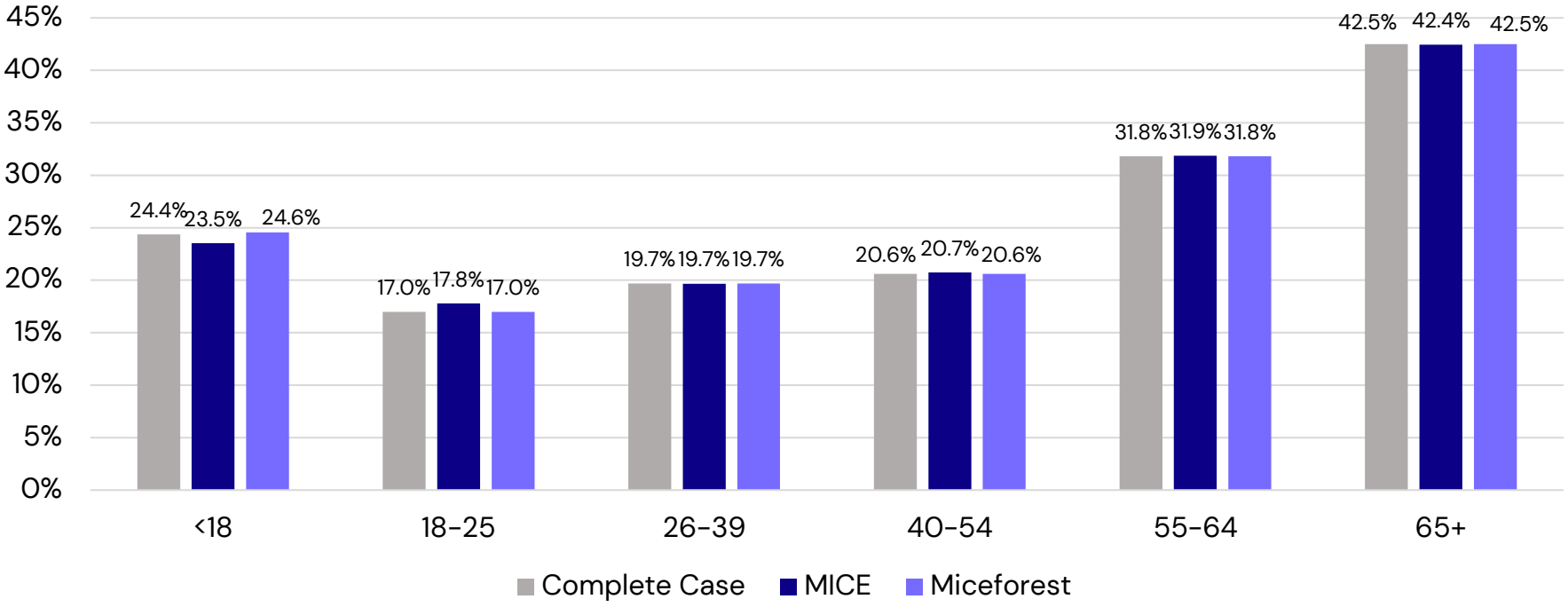
Flu Vaccination Coverage Rates by Gender Across MICE & Miceforest



Flu Vaccination Coverage Rates by Ethnicity Across MICE & Miceforest



Flu Vaccination Coverage Rates by Age Across MICE & Miceforest



Discussion

Imputation method alters initial findings.

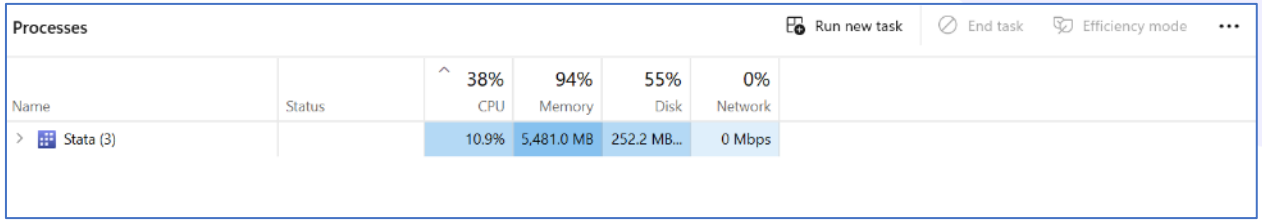
Ethnicity, gender, and age disparities in flu vaccination were consistently significant across all methods ($p\text{-}value < 0.001$).

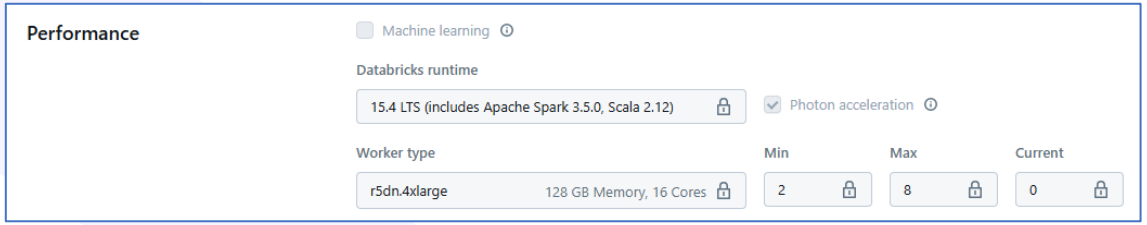
MICE and Miceforest performed similarly across all imputed variables (G-statistic).

Discussion

MICE (STATA 17): 8 hours, high CPU usage

Miceforest (Cloud): < 10 minutes, scalable





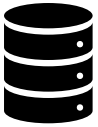
Implications

Imputation inflated dataset by 0.28%, reducing estimated flu coverage from 49% to 24%.

Stratified rates declined across all groups post-imputation, revealing potential hidden disparities.

Cloud-based tools offer speed and scalability for large datasets.

Conclusions



Accurate Data = Better Public Health Decisions

Even small amounts of missing data can obscure critical health disparities

Imputation is Essential

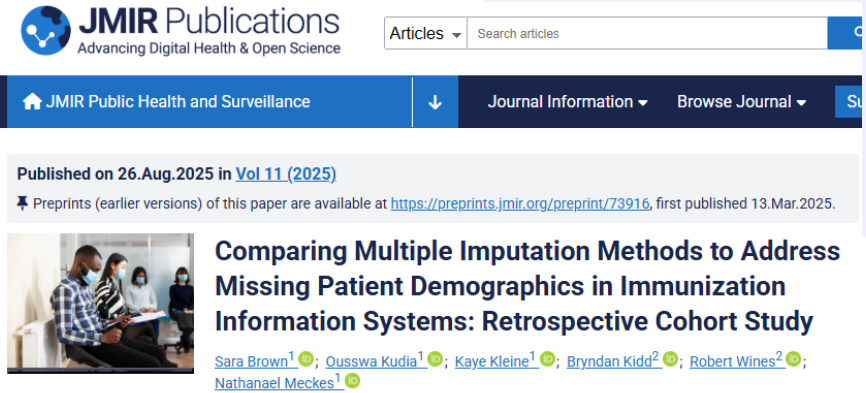
Proper imputation reveals hidden inequities and supports culturally responsible interventions

Cloud-Based Machine Learning = Scalable Solutions

Offloading to the cloud enables:

- Faster processing
- Reduced downtime
- Parallel task execution
- Greater productivity

Check us out!



Thank you

Email
analytics@stchealth.com

Phone
+1 (480) 745-8500

Website
stchealth.com

References

1. Noppert GA, Zalla LC. Who Counts and Who Gets Counted? Health Equity in Infectious Disease Surveillance. *Am J Public Health*. 2021;111(6):1004-1006. doi:10.2105/AJPH.2021.306249
2. Labgold K, Hamid S, Shah S, et al. Estimating the Unknown: Greater Racial and Ethnic Disparities in COVID-19 Burden After Accounting for Missing Race and Ethnicity Data. *Epidemiology*. 2021;32(2):157-161. doi:10.1097/EDE.0000000000001314
3. Kazeminia M, Afshar ZM, Rajati M, Saeedi A, Rajati F. Evaluation of the acceptance rate of covid-19 vaccine and its associated factors: A systematic review and meta-analysis. *Journal of Prevention*. 2022;43(4):421-467. doi:10.1007/s10935-022-00684-1
4. Yoon P, Hall J, Fuld J, et al. Alternative Methods for Grouping Race and Ethnicity to Monitor COVID-19 Outcomes and Vaccination Coverage. *MMWR Morb Mortal Wkly Rep*. 2021;70(32):1075-1080. Published 2021 Aug 13. doi:10.15585/mmwr.mm7032a2
5. Stokes EK, Zambrano LD, Anderson KN, et al. Coronavirus Disease 2019 Case Surveillance - United States, January 22-May 30, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(24):759-765. Published 2020 Jun 19. doi:10.15585/mmwr.mm6924e2
6. Wu SL, Mertens AN, Crider YS, et al. Substantial underestimation of SARS-CoV-2 infection in the United States. *Nat Commun*. 2020;11(1):4507. Published 2020 Sep 9. doi:10.1038/s41467-020-18272-4
7. Lu PJ, Hung MC, Srivastav A, et al. Surveillance of Vaccination Coverage Among Adult Populations -United States, 2018. *MMWR Surveill Summ*. 2021;70(3):1-26. Published 2021 May 14. doi:10.15585/mmwr.ss7003a1
8. Valente TW. Data Collection and Management. In: *Evaluating Health Promotion Programs*. Oxford University Press; 2002:123-146.
9. Weston BW. Blind spots: Biases in prehospital race and ethnicity recording. *Prehospital Emergency Care*. Published online 2023:1-4. doi:10.1080/10903127.2023.2175089
10. Sholle ET, Pinheiro LC, Adekkanattu P, et al. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *Journal of the American Medical Informatics Association*. 2019;26(8-9):722-729. doi:10.1093/jamia/ocz040
11. Rubin DB. Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. Proceedings of the Survey Research Methods Section of the American Statistical Association. 1978:20-28.
12. Rubin DB. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*. 1978;6(1). doi:10.1214/aos/1176344064
13. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581. doi:10.2307/2335739
14. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association*. 1996;91(434):473. doi:10.2307/2291635
15. Li KH, Raghunathan TE, Rubin DB. Large-sample significance levels from multiply imputed data using moment-based statistics and an *F* reference distribution. *Journal of the American Statistical Association*. 1991;86(416):1065-1073. doi:10.1080/01621459.1991.10475152
16. Meng X-L, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*. 1992;79(1):103. doi:10.2307/2337151

References

17. Chan KW, Meng X-L. Multiple improvements of multiple imputation likelihood ratio tests. *Statistica Sinica*. Published online 2022. doi:10.5705/ss.202019.0314
18. Chan KW. General and feasible tests with multiply-imputed datasets. *The Annals of Statistics*. 2022;50(2). doi:10.1214/21-aos2132
19. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. *Int J Methods Psychiatr Res*. 2011;20(1):40-49. doi:10.1002/mpr.329
20. Bouhlila DS, Sellaouti F. Multiple imputation using chained equations for missing data in TIMSS: A case study. *Large-scale Assessments in Education*. 2013;1(1). doi:10.1186/2196-0739-1-4
21. Wang H, Tang J, Wu M, Wang X, Zhang T. Application of machine learning missing data imputation techniques in clinical decision making: Taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Medical Informatics and Decision Making*. 2022;22(1). doi:10.1186/s12911-022-01752-6
22. Getz K, Hubbard RA, Linn KA. Performance of multiple imputation using modern machine learning methods in electronic health records data. *Epidemiology*. 2022;34(2):206-215. doi:10.1097/ede.0000000000001578
23. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. *Journal of Big Data*. 2021;8(1). doi:10.1186/s40537-021-00516-9
24. Zhang X, Yan C, Gao C, Malin BA, Chen Y. Predicting missing values in medical data via XGBoost regression. *Journal of Healthcare Informatics Research*. 2020;4(4):383-394. doi:10.1007/s41666-020-00077-1
25. Wang H, Tang J, Wu M, Wang X, Zhang T. Application of machine learning missing data imputation techniques in clinical decision making: Taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Medical Informatics and Decision Making*. 2022;22(1). doi:10.1186/s12911-022-01752-6
26. Zhao X, Shen W, Wang G, Schwenker F, Friedhelm Schwenker. Early Prediction of Sepsis Based on Machine Learning Algorithm. *Computational intelligence and neuroscience*. 2021;2021(1):6522633-6522633. doi:10.1155/2021/6522633
27. Shao L, Chen W. Coal and gas outburst prediction model based on miceforest filling and PHHO–kelm. *Processes*. 2023;11(9):2722. doi:10.3390/pr11092722
28. AnotherSamWilson. miceforest: Fast, Memory Efficient Imputation with LightGBM. GitHub. August 30, 2020. Accessed November 13, 2024. https://github.com/AnotherSamWilson/miceforest.
29. Pham HT, Do T, Baek J, et al. Handling Missing Data in COVID-19 Incidence Estimation: Secondary Data Analysis. *JMIR Public Health Surveill*. 2024;10:e53719. Published 2024 Aug 20. doi:10.2196/53719
30. Feng S, Hategeka C, Grépin KA. Addressing missing values in routine health information system data: an evaluation of imputation methods using data from the Democratic Republic of the Congo during the COVID-19 pandemic. *Popul Health Metr*. 2021;19(1):44. Published 2021 Nov 4. doi:10.1186/s12963-021-00274-z
31. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*. 2007;8(3):206-213. doi:10.1007/s11121-007-0070-9

References

32. Al-Jumaili AHA, Muniyandi RC, Hasan MK, Paw JKS, Singh MJ. Big Data Analytics Using Cloud Computing Based Frameworks for Power Management Systems: Status, Constraints, and Future Recommendations. *Sensors (Basel)*. 2023;23(6):2952. Published 2023 Mar 8. doi:10.3390/s23062952
33. Lebeda FJ, Zalatoris JJ, Scheerer JB. Government Cloud Computing Policies: Potential Opportunities for Advancing Military Biomedical Research. *Mil Med*. 2018;183(11-12):e438-e447. doi:10.1093/milmed/usx114
34. Ahmadi M, Aslani N. Capabilities and Advantages of Cloud Computing in the Implementation of Electronic Health Record. *Acta Inform Med*. 2018;26(1):24-28. doi:10.5455/aim.2018.26.24-28