Comparing Multiple Imputation Methods to Address Missing Patient Demographics in Immunization Information Systems: Retrospective Cohort Study

Sara Brown, Ousswa Kudia, Kaye Kleine, Bryndan Kidd, Robert Wines, Nathanael Meckes

Submitted to: JMIR Public Health and Surveillance
on: March 13, 2025

Table of Contents

Comparing Multiple Imputation Methods to Address Missing Patient Demographics in Immunization Information Systems: Retrospective Cohort Study

Sara Brown[1] MPH, CHES; Ousswa Kudia[1] MPH; Kaye Kleine[1] MPH, MS; Bryndan Kidd[2] MS; Robert Wines[2]; Nathanael Meckes[1] PhD

[1] Scientific Technologies Corporation (United States) Phoenix US
[2] West Virginia Department of Health and Human Services, Immunization Services Charleston US

Corresponding Author:
Sara Brown MPH, CHES

Scientific Technologies Corporation (United States)
411 S 1st St
Phoenix
US

Abstract

Surveillance data are essential for public health initiatives; however, missing data is often a challenge, impacting the ability to assess accurate vaccine coverage by introducing bias, particularly when addressing disparities. While methods like multiple imputation using chain equations (MICE) are robust, they can be computationally expensive when applied to large datasets. We explored the use of the machine learning techniques Iterative-Imputer and miceforest and cloud-based computing to reconcile missing demographic data. 2021-2022 flu vaccination and demographic data came from the WV Immunization Information System (N=2,302,036) where race (15%) and ethnicity (34%) were missing. We utilized MICE, Iterative-Imputer, and miceforest, where we jointly imputed missing variables and created 15 datasets each. After imputations, we obtained an additional 780,339 observations compared to the complete case. MICE and miceforest best preserved the proportional distribution of demographics relative to the complete case. MICE required 14 hours to complete 15 imputations, while Iterative-Imputer took 2 minutes and miceforest took in 10 minutes to complete the same number of imputations. After applying post-imputation estimates to flu data, vaccination coverage rates dropped between 0.87-18%. Utilizing miceforest to reconcile missing demographic data poses as a potential solution, offering a flexible, fast, and iterative approach that can improve data completeness while preserving underlying distributions and mitigating potential bias. By offloading resource-intensive tasks to a cloud-based server, public health officials can mitigate processing constraints, enabling parallel execution of multiple tasks while minimizing downtime. This enhanced efficiency accelerates analyses, facilitates culturally responsive decision-making, and optimizes organizational performance, ultimately improving communication and productivity by eliminating prolonged processing delays.

(JMIR Preprints 13/03/2025:73916)
DOI: https://doi.org/10.2196/preprints.73916

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Comparing Multiple Imputation Methods to Address Missing Patient Demographics in Immunization Information Systems: Retrospective Cohort Study

Abstract

Surveillance data are essential for public health initiatives; however, missing data is often a challenge, impacting the ability to assess accurate vaccine coverage by introducing bias, particularly when addressing disparities. While methods like multiple imputation using chain equations (MICE) are robust, they can be computationally expensive when applied to large datasets. We explored the use of the machine learning techniques Iterative-Imputer and miceforest and cloud-based computing to reconcile missing demographic data. 2021-2022 flu vaccination and demographic data came from the WV Immunization Information System (N=2,302,036) where race (15%) and ethnicity (34%) were missing. We utilized MICE, Iterative-Imputer, and miceforest, where we jointly imputed missing variables and created 15 datasets each. After imputations, we obtained an additional 780,339 observations compared to the complete case. MICE and miceforest best preserved the proportional distribution of demographics relative to the complete case. MICE required 14 hours to complete 15 imputations, while Iterative-Imputer took 2 minutes and miceforest took in 10 minutes to complete the same number of imputations. After applying post-imputation estimates to flu data, vaccination coverage rates dropped between 0.87-18%. Utilizing miceforest to reconcile missing demographic data poses as a potential solution, offering a flexible, fast, and iterative approach that can improve data completeness while preserving underlying distributions and mitigating potential bias. By offloading resource-intensive tasks to a cloud-based server, public health officials can mitigate processing constraints, enabling parallel execution of multiple tasks while minimizing downtime. This enhanced efficiency accelerates analyses, facilitates culturally responsive decision-making, and optimizes organizational performance, ultimately improving communication and productivity by eliminating prolonged processing delays.

Keywords

Multiple imputation, missing data, imputation methods, data science, machine learning, statistical modeling, immunization information system

Introduction

The usage of large datasets obtained from surveillance data and immunization information systems (IIS) have held a vital role in recognizing and comprehending the extent of health disparities and inequities within a population.[1-3] Previous research has shown that there is an association between race and ethnicity and vaccine acceptance and uptake.[4-7] However, as exemplified by the COVID-19 pandemic, race and ethnicity fields are historically underpopulated, thereby limiting a full understanding of the extent of vaccine inequities.[2, 3, 8, 9] Missing data can affect the capability to effectively describe vaccine coverage and may introduce bias into epidemiologic analyses, particularly when attempting to estimate and address racial and ethnic health inequities and disparities.

Studies exploring racial and ethnic inequalities traditionally omit individuals with missing demographic information or have classify them as “unknown.”[10-13] However, this can miscalculate vaccine coverage rates when stratified by race and ethnicity, leading to biased analyses between different racial and ethnic groups. This selection bias dampens capacities to accurately measure the true vaccine uptake among underserved populations that are more likely to have racial and ethnic

data missing in surveillance datasets.[3, 14-16]

Imputing values for missing data, alternatively, has been shown to be a better method for eliminating missing data and preserving the greatest amount of data. [14] Multiple imputation using chained equations (MICE) is a statistically sound method for managing missing data and has been used in the context of medical research and large, national datasets.[17,18] MICE employs the distribution of the original data to estimate values that signify the ambiguity of the true missing value. This can yield unbiased approximations after an adequate number of imputations, which is contingent based on the quality of the dataset.

Machine learning (ML) is increasingly being utilized to reconcile missing data by offering robust and sophisticated solutions for improving data quality. Some of these techniques include random forest (RF)[19-24], support vector machine (SVM)[19,25], neural networks (NN)[19,22,26-28], ensemble learning (EL)[19,29,30], K-nearest neighbor (KNN)[19,25,31], and eXtreme gradient boosting (XGBoost)[32-35], all with varying levels of accuracy. The accuracy and performance of machine learning algorithms to impute missing data relies heavily on the type of missing data being evaluated.

When compared to various machine learning techniques, MICE had less bias and similar standard errors for parameter estimates, with additional studies demonstrating combining MICE with machine learning yields less biased results.[36,37] Miceforest is a combination of classical epidemiological techniques and machine learning algorithms. This algorithm utilizes MICE with light gradient-boosting, a tree-based algorithm, to provide a flexible and powerful, and still efficient and reliable, solution for managing missing data.[38-40]

In this study, we investigate racial and ethnic disparities in flu vaccine uptake in West Virginia for the 2021-2022 flu season. We sought to address potential bias due to missing race and ethnicity data through multiple imputation and test the efficiency and accuracy of 3 established methods to address missing data in IIS and the implications of imputations on influenza vaccination coverage data. To date, there is no published peer-reviewed literature on methods to address missing data in IIS.

Methods

Patient data for 2021-2022 flu vaccination (June 1 – June 30) was obtained from the West Virginia Immunization Information System (WVSIIS). Geo-demographic data, such as urbanicity and SVI status, were calculated using address data obtained from the IIS. After a preliminary data cleaning process on a clone of the WVSIIS database, there were a resulting 2,302,036 records. 15% of records were missing patient race (n=347,633) and 34% (n=780,339) of records were missing patient ethnicity.

We first addressed missing race and ethnicity data by employing multiple imputation by chained equations (MICE) in Stata 17 using the mi impute chained command, where we jointly imputed patient race and ethnicity using age, urbanicity, SVI status, county, and flu vaccination status. We created 15 imputed datasets to stabilize our variance estimates.[41,42]

Python 3.11 was used as the primary engine for the second and third method of addressing missing data. All Python computations were performed on a cloud-based computing cluster with 16-core processors and 128 GB of RAM. Packages used include scikit-learn’s Iterative-Imputer and miceforest. Iterative-Imputer operates similarly to MICE by utilizing available data in other features to estimate missing values. The imputation is performed in an iterative, round-robin manner, with a regressor to predict the missing values.[43] Miceforest is designed for performing MICE using random forests. Rather than assume a linear relationship between variables, miceforest can capture complex, non-linear patterns and imputes values iteratively, unlike traditional MICE.[40]

Once missing data had been reconciled, post-imputation estimates were applied to WVSIIS flu vaccination data for the 2021-2022 flu season.

Results

MICE

347,633 race categories that were previously missing were imputed after completing MICE.

There were 16.5% additional White records, 16.8% additional Black records, 16.3% additional Asian records, 18.5% additional Indigenous records, 16.8% additional Native Hawaiian/Pacific Islander records, 17.4% additional Multi-racial records, and 15.3% additional Other records after imputations (Table 1).

780,339 ethnicity categories that were previously missing were imputed after completing MICE. There were 52.4% additional Hispanic/Latino records and 40.5% additional Not Hispanic/Latino records after imputations (Table 1).

After MICE, individual demographics remained proportional to the original dataset distributions. A chi-square test illustrated that there was a statistically significant difference between the complete case and MICE estimates ($P < .001$) (Table 2). Overall computational time was approximately 14 hours.

Iterative-Imputer

347,633 race categories were successfully imputed after completing Iterative-Imputer. There were 0.26% additional White, 142.5% additional Black, 50.4% additional Asian, 0% additional Indigenous, 95.2% additional Native Hawaiian/Pacific Islander records, 10.6% additional Multi-racial, and 0% additional Other records after imputations (Table 1).

Like MICE, 780,339 ethnicity categories that were previously missing were imputed after completing Iterative-Imputer. There were 0% additional Hispanic/Latino and 4.5% Not Hispanic/Latino records after imputations (Table 1).

After Iterative-Imputer, individual demographics did not remain proportional to the original dataset distributions. A chi-square test illustrated that there was a statistically significant difference between the complete case and Iterative-Imputer estimates ($P < .001$) (Table 2). Overall computational time was 2 minutes.

Miceforest

347,633 race categories that were previously missing were imputed after completing Miceforest. There were 16.5% additional White records, 17.9% additional Black records, 15.9% additional Asian records, 27.7% additional Indigenous records, 26.0% additional Native Hawaiian/Pacific Islander records, 13.6% additional Multi-racial records, and 15.1% additional Other records after imputations (Table 1).

780,339 ethnicity categories that were previously missing were imputed after completing Miceforest. There were 76.1% additional Hispanic/Latino records and 39.8% additional Not Hispanic/Latino records after imputations (Table 1).

After Miceforest, individual demographics remained proportional to the original dataset distributions. A chi-square test illustrated that there was a statistically significant difference between the complete case and Miceforest estimates ($P < .001$) (Table 2). Overall computational time was 10 minutes.

Flu Coverage with Miceforest Imputations

For the 2022-2023 flu season, the complete case analysis flu vaccine coverage rate was 26%. Using miceforest, after imputations, WV had an overall flu vaccine coverage rate of 19%, a 16% decrease. Flu vaccine coverage rates decreased when stratified by race and ethnicity when compared to a complete case analysis, with the most significant decreases observed in NH White (5%), NH Black (4%), Hispanic/Latino (7%), Asian (7%), Native Hawaiian/Pacific Islander (7%), and Other (18%) (Figure 1). After reconciling missing race/ethnicity, an additional 63,984 individuals were included in the analysis that were previously excluded from stratified analyses.

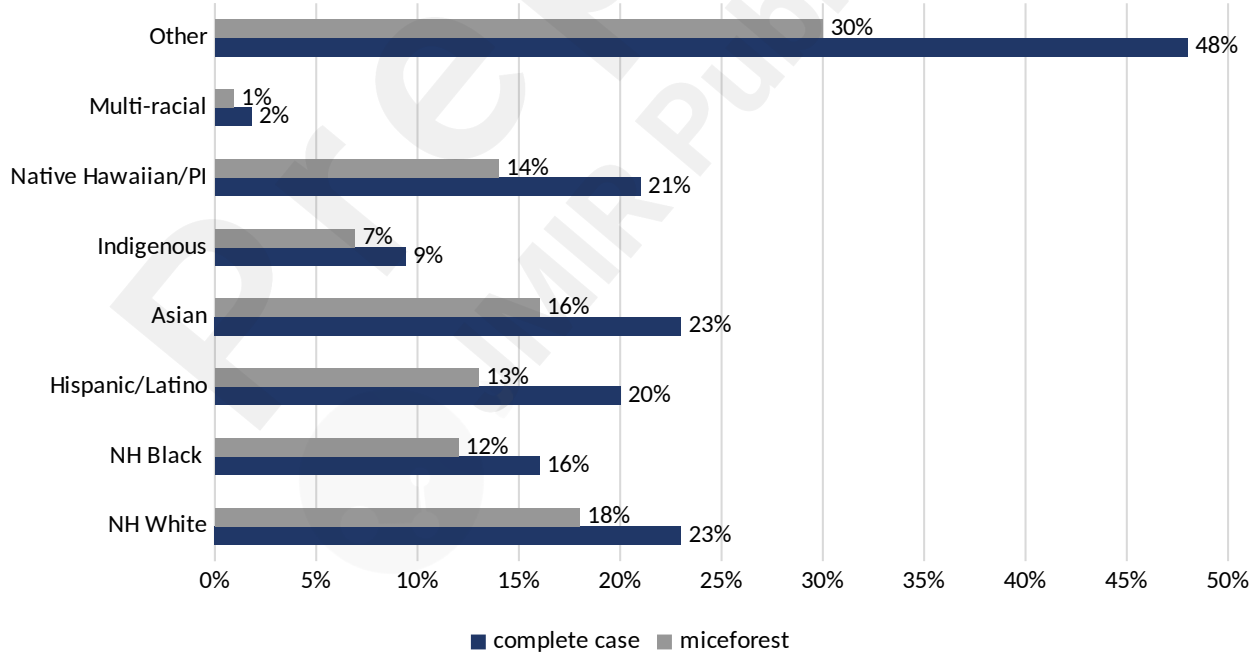
Table 1. Complete case, MICE, Iterative-Imputer, and Miceforest outputs

Variable	Missing %	Complete Case n	Complete Case %	MICE n	MICE %	Iterative-Imputer n	Iterative-Imputer %	Miceforest n	Miceforest %
Race	**15%**								
White		1,519,326	77.7%	1,793,167	77.8%	1,527,855	66.4%	1,792,102	77.8%
Black		65,716	3.36%	77,767	3.38%	391,756	17.0%	78,624	3.42%%
Asian		14,334	0.73%	16,886	0.734%	24,003	1.04%	16,813	0.730%
Indigenous		7,110	0.36%	8,561	0.372%	7,110	0.309%	9,399	0.408%
Native Hawaiian/ Pacific Islander		1,720	0.09%	2,036	0.088%	4,846	0.211%	2,235	0.097%
Multi-racial		2,405	0.12%	2,863	0.124%	2,674	0.116%	2,755	0.112%
Other		343,792	17.6%	400,757	17.4%	343,792	14.9%	400,108	17.4%
TOTAL		**1,954,403**		**2,302,036**		**2,302,036**		**2,302,036**	
Ethnicity	**34%**								
Hispanic/Latino		33,652	2.21%	57,552	2.50%	33,652	98.5%	75,031	3.26%
Not Hispanic/Latino		1,488,045	97.8%	2,244,484	97.5%	2,268,384	1.46%	2,227,005	96.7%
TOTAL		**1,521,697**		**2,302,036**		**2,302,036**		**2,302,036**	

Table 2. Chi-square test for complete case versus MICE, Iterative-Imputer, and Miceforest

Imputation Method	Pearson chi^2	*P*
MICE	1.1e+04	<0.001
Iterative-Imputer	1.8e+05	<0.001
Miceforest	1.3e+04	<0.001

Figure 1. Complete case versus miceforest flu vaccine coverage rates



Discussion

Of the three methods utilized to address missing data, we found that Miceforest had the best-balanced model fit and scalability. Miceforest was able to reconcile missing data quickly with little bias and allows for more sophisticated post-hoc analyses. Iterative-Imputer exhibited the most substantial deviations from the complete case analysis. While the majority of race and ethnicity categories retained proportions similar to those observed prior to imputation, the Black race category had a notable 142% increase post-Iterative-Imputer. The significant change within a singular racial group carries several potential implications: it may suggest under-reporting of race by Black individuals in the original dataset, or it may reflect a greater degree of bias introduced by the Iterative-Imputer.

The most striking difference between these methods was overall computational time. Large, population surveillance-based datasets can be computationally expensive to impute. While traditional MICE was proportionally similar to the complete case analysis and Miceforest, the computational load was a significant limitation for future epidemiological analyses. Since Stata 17 operates locally, executing MICE required approximately 14 hours, utilizing between 92% and 98% of the total CPU capacity of the local system. In contrast, as Iterative-Imputer and Miceforest were processed on a cloud-based drive, the total computational time was reduced to under 10 minutes. Cloud-based computing extends several advantages, including enhanced scalability, reduced dependance on local servers, and the capability to allocate computational jobs efficiently.[44,45] The larger and more complex datasets become, the requirements for larger memory and CPUs become necessary, and consequently, will have computational limitations. By divesting resource-intensive tasks to the cloud, researchers and public health officials can alleviate processing constrictions, enabling analogous performance of multiple tasks, and minimizing downtime. This improved efficiency not only accelerates analyses but improves overall productivity by removing lengthy processing delays.

Compared to the complete case analysis, the imputed dataset was 45% larger, and as a result, we observed greater differences in flu vaccination rates compared with a complete case analysis. Stratified flu vaccination coverage rates declined across all racial and ethnicity categories following imputation. This is both expected, due to denominator inflation, and consistent with existing literature, which suggests that enhancements in data completeness often uncover underlying racial inequities.[2,3,8-10,12,16,46] These results suggest that the Miceforest estimates were statistically significant compared to the complete case counts. This indicates that the complete case analysis estimates may be biased, potentially overstating flu vaccination coverage, particularly among vulnerable groups.

Additionally, the presence of the ‘Other’ category may be artificially influencing race missingness rate. Reports indicate that the increasing size of the ‘Other’ category may be influenced by individuals choosing not to disclose their race, varying cultural perceptions of race, or vaccine providers selecting ‘Other’ rather than leaving the race field blank.[47,48] Those who select ‘Other’ category get combined into a single, dissimilar category that does not reflect accurate identities and makes results for this group difficult to accurately decipher.

The absence of demographic data carries significant public health implications. Missing data limits our capability to correctly assess vaccination coverage, potentially leading to an incomplete or inaccurate understanding of disparities in vaccination rates among different vulnerable groups. Consequently, public health interventions and policies may be based on flawed assumptions, undermining their effectiveness. Moreover, the inability to identify and address disparities hinders efforts to promote equity within communities. Poor data quality can

also result in inefficient allocation of resources, both human and financial, further exacerbating inequities. Finally, inadequate data can erode public trust in health systems, as decisions based on incomplete information may be perceived as unreliable or inequitable. However, Miceforest successfully imputed missing values a large public health dataset in the most time efficient manner, and therefore, we were able to mitigate potential bias and increase our statistical power in our analyses.

There were several limitations to our approach. Due to high levels of missingness of variables in the dataset, there were limited informing variables for the imputations. The WVSIIS population denominator is higher than the census population, which can skew flu vaccination rates lower than they are. Race and ethnicity are self-reported, and it may not reflect the reality of demographic distribution in the state, especially with growing nuances of racial and ethnic identity.

Data is the foundation of all effective public health interventions, and missing data can reduce our comprehension of vaccination coverage, recognizing disparities, and creating successful public health programs. Without properly addressing missing data, vulnerable groups continue to be underserved. As demonstrated, different methods for reconciling missing data results in different assumptions regarding the data and the process. However, utilizing Miceforest to reconcile missing demographic data poses as a potential solution, offering a flexible, fast, and iterative approach that can improve data completeness while preserving underlying distributions and mitigating potential bias.

Acknowledgments

The results presented in this manuscript represent the collaborative efforts across the West Virginia Department of Health and Human Services, the West Virginia Immunization Department, and STChealth, West Virginia's IIS technology partner. The authors would like to extend their gratitude to Sawyer Koops, MS, Ilyssa Simmons, MPH, Kyle Freese, PhD, and countless other state, contract, federal, and STChealth staff whose expertise, leadership, and dedication were instrumental in this effort.

Conflicts of Interest

The authors have no conflict of interest to disclose.

Abbreviations

Immunization Information Systems (IIS)
Multiple Imputation using Chained Equations (MICE)
Central Processing Unit (CPU)
Machine Learning (ML)
Random forest (RF)
Support Vector Machine (SVM)
Neural networks (NN)
Ensemble learning (EL)
K-nearest neighbor (KNN)
eXtreme gradient boosting (XGBoost)

References

1. Wynia MK, Ivey SL, Hasnain-Wynia R. Collection of data on patients' race and ethnic group by physician practices. *New England Journal of Medicine*. 2010;362(9):846-850. doi:10.1056/nejmsb0910799
2. Grundmeier RW, Song L, Ramos MJ, et al. Imputing missing race/ethnicity in pediatric electronic health records: Reducing bias with use of U.S. Census location and surname Data. *Health Services Research*. 2015;50(4):946-960. doi:10.1111/1475-6773.12295

3. 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/ethnicity data. *Epidemiology*. 2020;32(2):157-161. doi:10.1097/ede.0000000000001314
4. 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
5. Shui IM, Weintraub ES, Gust DA. Parents concerned about vaccine safety. *American Journal of Preventive Medicine*. 2006;31(3):244-251. doi:10.1016/j.amepre.2006.04.006
6. Andersen JA, Gloster E, Hall S, et al. Associations between COVID-19 vaccine uptake, race/ethnicity, and political party affiliation. *Journal of Behavioral Medicine*. 2022. doi:10.1007/s10865-022-00379-2
7. Mahmud SM, Xu L, Hall LL, et al. Effect of race and ethnicity on influenza vaccine uptake among older US medicare beneficiaries: A record-linkage cohort study. *The Lancet Healthy Longevity*. 2021;2(3). doi:10.1016/s2666-7568(20)30074-x
8. Spangler KR, Levy JI, Fabian MP, et al. Missing race and ethnicity data among COVID-19 cases in Massachusetts. *Journal of Racial and Ethnic Health Disparities*. Published online 2022. doi:10.1007/s40615-022-01387-3
9. Krieger N, Testa C, Hanage WP, Chen JT. US racial and ethnic data for covid-19 cases: Still missing in action. *The Lancet*. 2020;396(10261). doi:10.1016/s0140-6736(20)32220-0
10. 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
11. 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
12. 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
13. 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
14. Valente TW. Data Collection and Management. In: *Evaluating Health Promotion Programs*. Oxford University Press; 2002:123-146.
15. 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
16. 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
17. 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

18. 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

19. 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

20. Breiman L, Cutler A. Random forests. Random forests - classification description. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#missing1. Published 2001. Accessed April 3, 2023.

21. Waljee AK, Mukherjee A, Singal AG, et al. Comparison of imputation methods for missing laboratory data in Medicine. *BMJ Open*. 2013;3(8). doi:10.1136/bmjopen-2013-002847

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. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. *American Journal of Epidemiology*. 2014;179(6):764-774. doi:10.1093/aje/kwt312

24. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*. 2020;20(1). doi:10.1186/s12874-020-01080-1

25. 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

26. Choudhury SJ, Pal NR. Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*. 2019;182:104838. doi:10.1016/j.knosys.2019.07.009

27. Petrozziello A, Jordanov I, Sommeregger C. Distributed Neural Networks for missing big data imputation. *2018 International Joint Conference on Neural Networks (IJCNN)*. 2018. doi:10.1109/ijcnn.2018.8489488

28. Cheng C-Y, Tseng W-L, Chang C-F, Chang C-H, Gau SS-F. A deep learning approach for missing data imputation of rating scales assessing attention-deficit hyperactivity disorder. *Frontiers in Psychiatry*. 2020;11. doi:10.3389/fpsyt.2020.00673

29. Tran CT, Zhang M, Andreae P, Xue B, Bui LT. Multiple imputation and ensemble learning for classification with Incomplete Data. *Proceedings in Adaptation, Learning*

and Optimization. 2016:401-415. doi:10.1007/978-3-319-49049-6_29
30. Carvalho AL, Ameyed D, Cheriet M. Ensemble learning for heterogeneous missing data imputation. *Big Data – BigData 2020*. 2020:127-143. doi:10.1007/978-3-030-59612-5_10

31. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*. 2016;16(S3). doi:10.1186/s12911-016-0318-z
32. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. doi:10.1145/2939672.2939785

33. Madhu G, Bharadwaj BL, Nagachandrika G, Vardhan KS. A novel algorithm for missing data imputation on machine learning. *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*. November 2019. doi:10.1109/icssit46314.2019.8987895

34. 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

35. Rusdah DA, Murfi H. XGBoost in handling missing values for life insurance risk prediction. *SN Applied Sciences*. 2020;2(8). doi:10.1007/s42452-020-3128-y

36. 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

37. 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
38. 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
39. 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
40. AnotherSamWilson. miceforest: Fast, Memory Efficient Imputation with LightGBM. GitHub. August 30, 2020. Accessed November 13, 2024. https://github.com/AnotherSamWilson/miceforest.
41. Raghunathan TE, Solenberger PW, Van Hoewyk J. IVEware imputation and Variance Estimation Software User Guide. Survey Research Center, Institute for Social Research University of Michigan. https://www.src.isr.umich.edu/wp-content/uploads/iveware/v0.1/Documentation/ive_user.pdf. Published March 2002. Accessed January 4, 2023.
42. 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
43. scikit-learn. Iterativeimputer. scikit learn. Accessed November 14, 2024. https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html#sklearn.impute.IterativeImputer.
44. 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
45. 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
46. Coelho R, Rocha R, Hone T. Improvements in data completeness in health information systems reveal racial inequalities: Longitudinal national data from hospital admissions in Brazil 2010–2022. *International Journal for Equity in Health*. 2024;23(1). doi:10.1186/s12939-024-02214-3
47. Holloway KR, Radack J, Barreto A, et al. The "Other" race category on birth certificates and its impact on analyses of preterm birth inequity. *J Perinatol*. Published online September 20, 2024. doi:10.1038/s41372-024-02123-x
48. Woolverton GA, Marks AK. "I just check 'other'": Evidence to support expanding the measurement inclusivity and equity of ethnicity/race and cultural identifications of U.S. adolescents. *Cultur Divers Ethnic Minor Psychol*. 2023;29(1):64-73. doi:10.1037/cdp0000360

Supplementary Files

Figures

Complete case versus miceforest flu coverage rates.