Methods for Imputing Race and Ethnicity in Real-World Data: A Scoping Review

Cassandra Lynch, BS, PharmD Candidate[1]; Jenna Wildeman, BS, PharmD Candidate[1]; Pam Pawloski, PharmD, BCOP, FCCP[2,3]; Cate Lockhart, PharmD, PhD[2,3]
1. University of Washington, 2. AMCP Research Institute 3. Biologics and Biosimilars Collective Intelligence Consortium

Background

- Real-world data (RWD) offers critical insights into patient outcomes and treatment patterns.[1]
- **RWD is limited by missing race/ethnicity data**, which hinders its use in identifying disparities and advancing equity-focused research.[2]
- Various imputation methods have been developed to estimate missing race and ethnicity data but each carries **risk of bias** if not validated and applied thoughtfully.
- Methods include complete case analysis, Bayesian surname and geocoding (BSG/BISG), multinomial logistic regression, and machine learning models.[3]
- It is essential to understand **which methods are being used, how accurate they are, and whether they are validated**, especially given the absence of a gold standard, to improve the quality and equity of healthcare research using RWD.

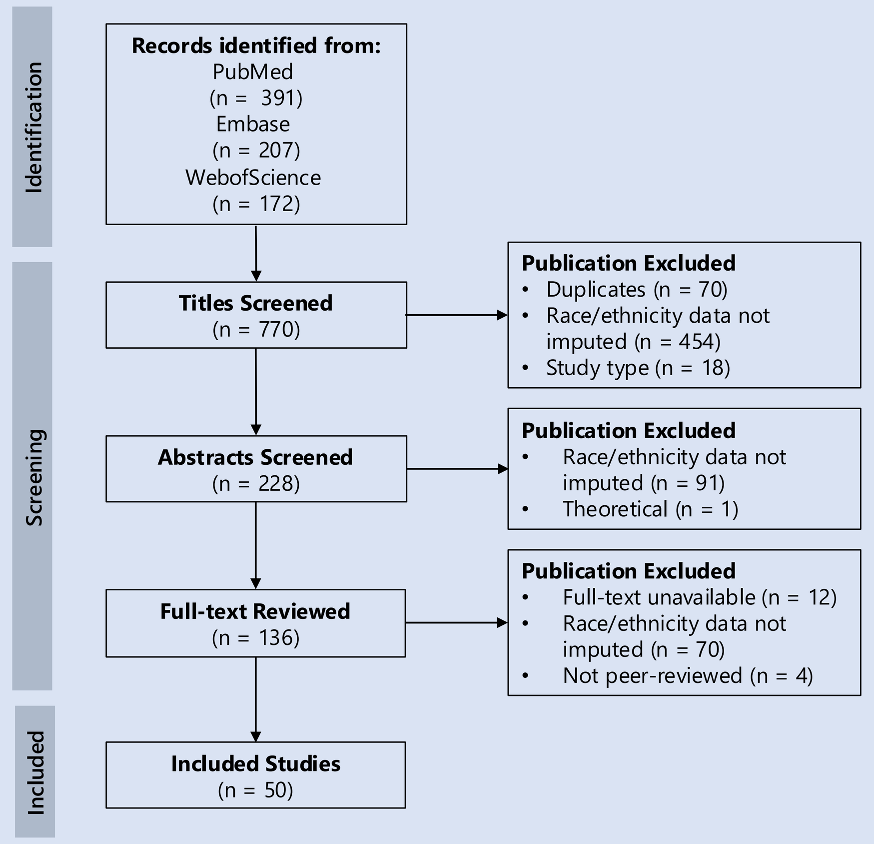
Objectives

- **Primary Objective:** Identify and categorize the methods used to impute missing race and ethnicity information in real-world healthcare data sources.
- **Secondary Objectives**: Assess the populations where these methods are applied, evaluating the frequency and types of imputation methods used, and examining how often these methods are validated and what metrics are used to assess their performance.

Methods

- We conducted a scoping review, following PRISMA-P guidelines, to identify studies on race and ethnicity imputation methods in real-world healthcare data.
- Peer-reviewed, English-language articles published between 2005 and 2025 were included.
- Searches were conducted in PubMed, Embase, and Web of Science, with additional reference list checks.
- Two reviewers independently screened and extracted data.
- Discrepancies were resolved through discussion.

Figure 1: PRISMA Flow Diagram


Identification
Screening
Included
Records identified from:
PubMed
(n = 391)
Embase
(n = 207)
WebofScience
(n = 172)
Titles Screened
(n = 770)
Publication Excluded
Duplicates (n = 70)
Race/ethnicity data not imputed (n = 454)
Study type (n = 18)
Abstracts Screened
(n = 228)
Publication Excluded
Race/ethnicity data not imputed (n = 91)
Theoretical (n = 1)
Full-text Reviewed
(n = 136)
Publication Excluded
Full-text unavailable (n = 12)
Race/ethnicity data not imputed (n = 70)
Not peer-reviewed (n = 4)
Included Studies
(n = 50)

Charts and Tables

Table 1: Demographics of Included Studies

Characteristic	N = 50 (%)
Study Focus	
Race/ethnicity imputation applied to RWD studies	38 (76%)
Race/ethnicity imputation methods evaluated	12 (24%)
Publication Year	
2021-2025	26 (52%)
2016-2020	16 (32%)
2011-2015	7 (14%)
2006-2010	1 (2%)
Geographical Location	
United States	34 (68%)
United Kingdom	15 (30%)
Europe	1 (2%)

Figure 2: Frequency of Race/Ethnicity Imputation Methods Used and Evaluated in Included Studies (N=50)

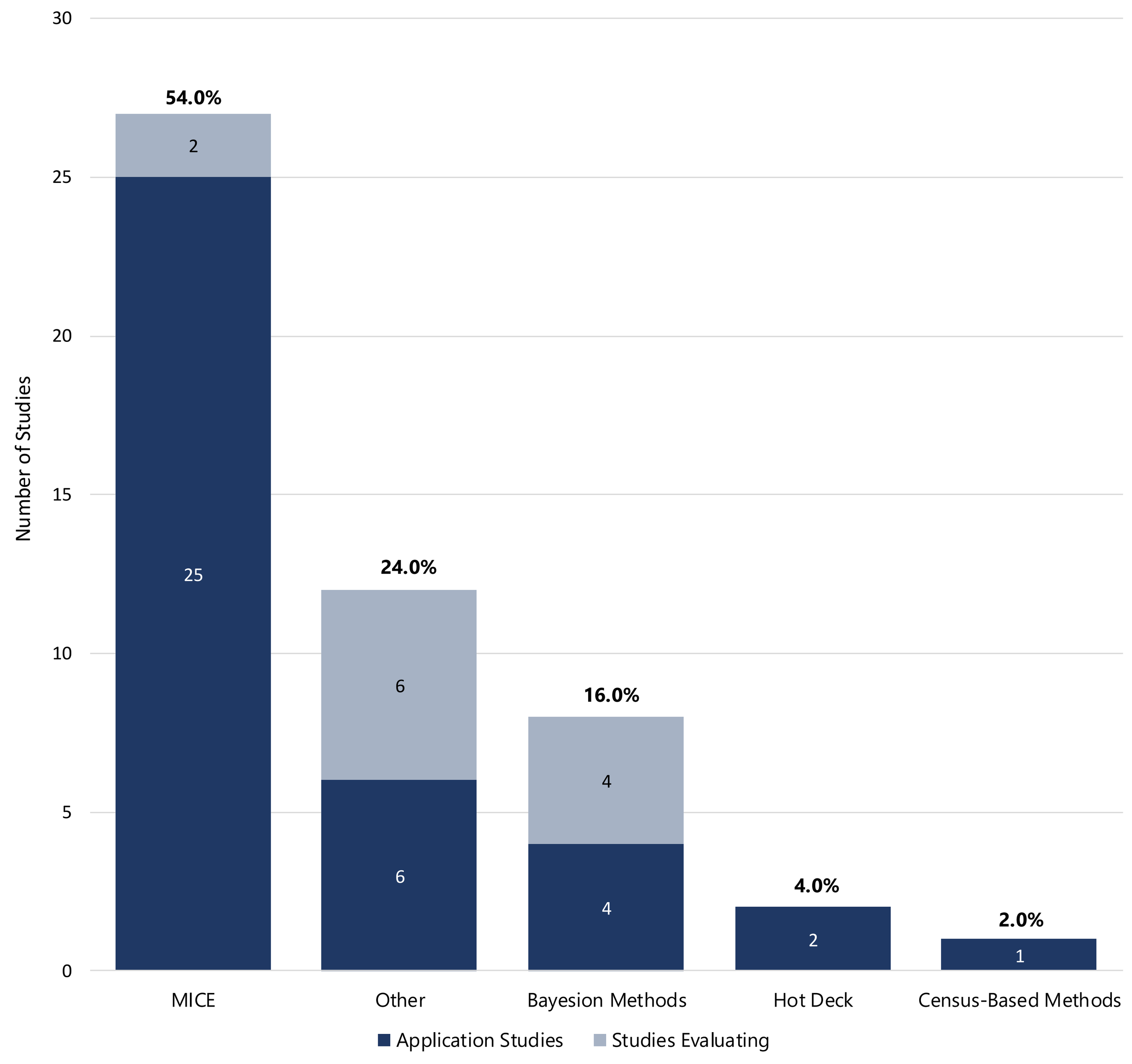

30
25
20
15
10
5
0
Number of Studies
54.0%
2
25
24.0%
6
6
16.0%
4
4
4.0%
2
2.0%
1
MICE
Other
Bayesion Methods
Hot Deck
Census-Based Methods
Application Studies
Studies Evaluating

Figure 3: Validation of Imputation Methods for Missing Race/Ethnicity (N=38)

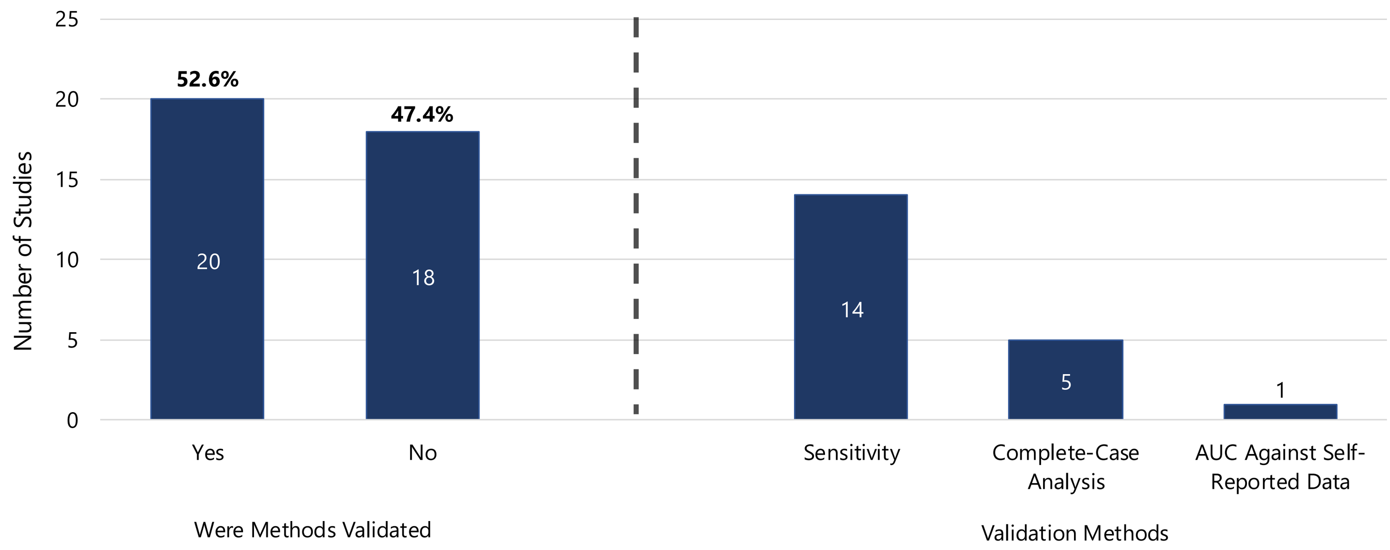

25
20
15
10
5
0
Number of Studies
52.6%
20
47.4%
18
14
5
1
Yes
No
Sensitivity
Complete-Case Analysis
AUC Against Self-Reported Data
Were Methods Validated
Validation Methods

Figure 4: Reporting of Missing Race/Ethnicity Data in Applied Studies (N=38)

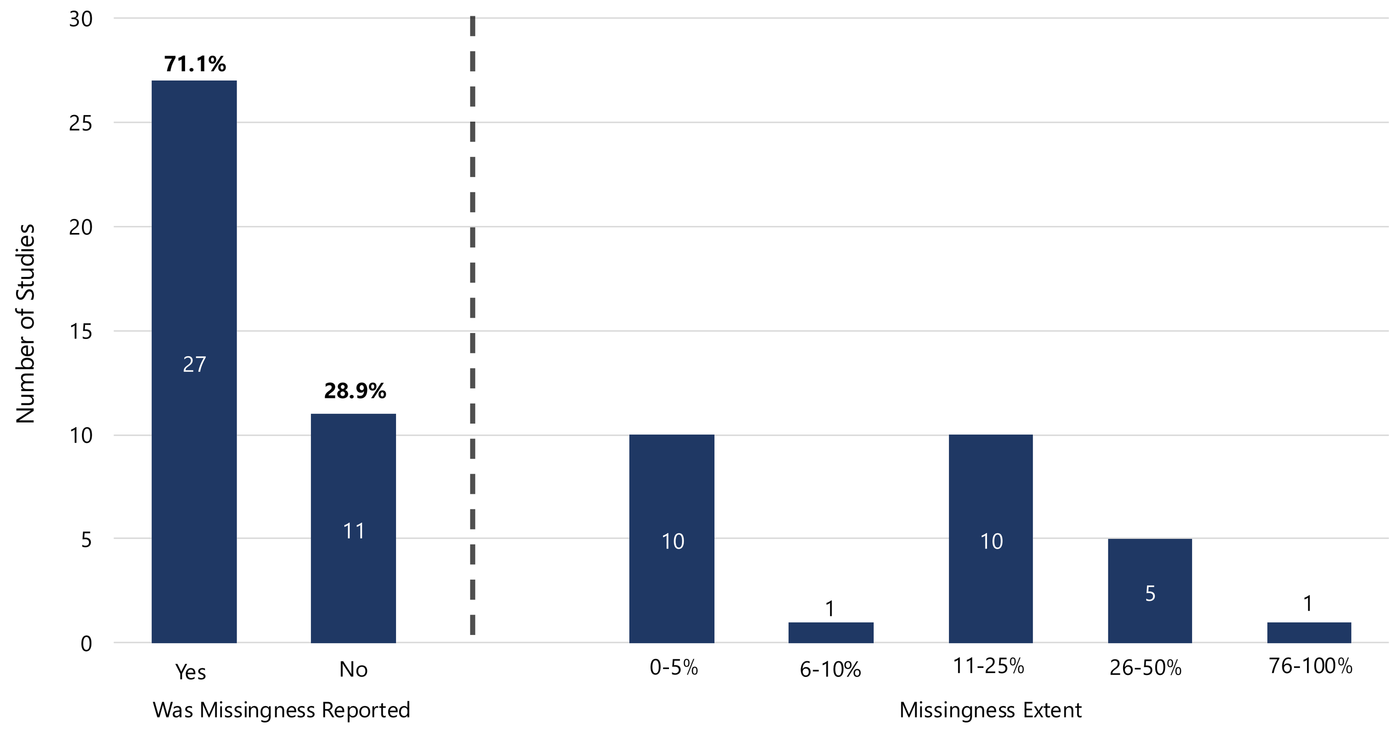

30
25
20
15
10
5
0
Number of Studies
71.1%
27
28.9%
11
10
1
10
5
1
Yes
No
0-5%
6-10%
11-25%
26-50%
76-100%
Was Missingness Reported
Missingness Extent

Table 2: Summary of Reporting Practices in Applied Studies

Characteristic	N = 38 (%)
Missing Race/Ethnicity Data Reporting	
Reported % of Missing Race/Ethnicity	27 (71.05%)
Range of Reported Missingness	<5% to >75%
Assumed Missingness Mechanism	
Missing At Random (MAR)	15 (39.47%)
Missing Not At Random (MNAR)	2 (5.26%)
Not Reported	21 (55.26%)
Validation or Sensitivity Analysis Reporting	
Reported any validation or sensitivity analysis	20 (52.63%)
Race/Ethnicity Classification Practices	
Used full OMB categories	4 (10.53%)
Used non-standard terms (e.g., South Asian, Latinx)	34 (89.47%)
Included "Other" category	31 (81.58%)
Dichotomized (e.g., White vs. non-White)	3 (7.90%)

Results

- **50 studies** met the pre-specified inclusion criteria; **38 applied** imputation methods in research while **12 evaluated** them
- Of the 12, 4 (33%) evaluated the Bayesian surname-geocoding methods (e.g., BISG, BIFSG), 2 (17%) assessed Multiple Imputation by Chained Equations (MICE), and others evaluated reallocation (1), deep learning (1), calibrated-δ adjustment (1), multinomial regression (1), or compared methods (2).
- Among the 38 application studies, MICE was used in 25 (66%), followed by other multiple imputation (6; 16%), Bayesian methods (4; 11%), hot deck (2; 5%), and census-based imputation (1; 3%).
- Among 27 studies reporting missingness percentages, levels ranged from <5% to >75%
- Only 4 studies (11%) used full Office of Management and Budget (OMB) categories for race and ethnicity; 34 (89%) used non-standard terms (e.g., South Asian, Latinx). "Other" was used in 31 (82%) of studies, and 3 (8%) dichotomized race as White vs. non-White.

Discussion and Conclusion

- Imputing race and ethnicity is increasingly common, reflecting the need to address missing data in **equity-focused research.**
- While widely used, imputation carries risks of bias, especially when methods and assumptions (e.g., MAR, MNAR) are not clearly reported.
- **MICE** is the dominant method, likely due to its practicality, but may not suit all data contexts or missingness mechanisms.
- Inconsistent race/ethnicity categories and limited adherence to OMB standards hinder comparability and may mask disparities.
- Overuse of broad or binary categories (e.g., "Other," White vs. non-White) reduces the nuance needed to capture inequities.
- **Validation** and **sensitivity analyses** were rare, despite their importance in assessing imputation reliability and impact.
- Without a **gold standard**, transparency, consistent classification, and rigorous validation are critical.
- **Standardized, equity-focused guidelines** are needed to ensure responsible imputation practices in RWD studies.
- Improved methods will lead to more trustworthy, inclusive, and actionable research to advance health equity.

References

1. Dang, A. Real-World Evidence: A Primer. Pharm. Med. 37, 25–36 (2023).
2. Studna, A. Executive Roundtable: The Rise of RWD in Clinical Research. Applied Clinical Trials https://www.appliedclinicaltrialsonline.com/view/executive-roundtable-the-rise-of-rwd-in-clinical-research (2023).
3. Xue, Y., Harel, O. & Aseltine, R. Comparison of Imputation Methods for Race and Ethnic Information in Administrative Health Data. 2019 13th Int Conf Sampl Theory Appl Sampta 00, 1– 4 (2019).

AMCP Research Institute™

BBCIC Biologics & Biosimilars Collective Intelligence Consortium