AMERICAN ACADEMY OF OPHTHALMOLOGY®

Factors Associated with Missing Sociodemographic Data in the IRIS® (Intelligent Research in Sight) Registry

*Connor Ross, BS,[1] Alexander Ivanov, MS,[1] Tobias Elze, PhD,[1] Joan W. Miller, MD,[1] Flora Lum, MD,[2] Alice C. Lorch, MD, MPH,[1] Isdin Oke, MD, MPH,[1,3] on behalf of the IRIS® Registry Analytic Center Consortium**

Purpose: To describe the prevalence of missing sociodemographic data in the IRIS® (Intelligent Research in Sight) Registry and to identify practice-level characteristics associated with missing sociodemographic data.

Design: Cross-sectional study.

Participants: All patients with clinical encounters at practices participating in the IRIS Registry prior to December 31, 2020.

Methods: We describe geographic and temporal trends in the prevalence of missing data for each sociodemographic variable (age, sex, race, ethnicity, geographic location, insurance type, and smoking status). Each practice contributing data to the registry was categorized based on the number of patients, number of physicians, geographic location, patient visit frequency, and patient population demographics.

Main Outcome Measures: Multivariable linear regression was used to describe the association of practice-level characteristics with missing patient-level sociodemographic data.

Results: This study included the electronic health records of 66 477 365 patients receiving care at 3306 practices participating in the IRIS Registry. The median number of patients per practice was 11 415 (interquartile range: 5849−24 148) and the median number of physicians per practice was 3 (interquartile range: 1−7). The prevalence of missing patient sociodemographic data were 0.1% for birth year, 0.4% for sex, 24.8% for race, 30.2% for ethnicity, 2.3% for 3-digit zip code, 14.8% for state, 5.5% for smoking status, and 17.0% for insurance type. The prevalence of missing data increased over time and varied at the state-level. Missing race data were associated with practices that had fewer visits per patient ($P < 0.001$), cared for a larger nonprivately insured patient population ($P = 0.001$), and were located in urban areas ($P < 0.001$). Frequent patient visits were associated with a lower prevalence of missing race ($P < 0.001$), ethnicity ($P < 0.001$), and insurance ($P < 0.001$), but a higher prevalence of missing smoking status ($P < 0.001$).

Conclusions: There are geographic and temporal trends in missing race, ethnicity, and insurance type data in the IRIS Registry. Several practice-level characteristics, including practice size, geographic location, and patient population, are associated with missing sociodemographic data. While the prevalence and patterns of missing data may change in future versions of the IRIS registry, there will remain a need to develop standardized approaches for minimizing potential sources of bias and ensure reproducibility across research studies.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100542 © 2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

The American Academy of Ophthalmology IRIS® (Intelligent Research in Sight) Registry includes the electronic health record (EHR) data from the majority of ophthalmology practices in the United States.[1] Data from the IRIS Registry have been used to answer clinical questions related to the practice patterns of ophthalmologists, the visual outcomes of rare ocular diseases, and the complications of surgical interventions.[2−8] A major challenge in working with a large EHR-based registry is the presence of missing data.[9] Analyses involving missing data are susceptible to selection bias and often require additional analytic considerations.[10] The prevalence of missing data in the IRIS Registry has not been previously described. In considering the patterns of missing data, it is important to recognize potential variations in data collection methods, including distinctions between patient registration and clinical visits. Complexities in the data curation and integration processes may contribute to missing data, as we cannot differentiate between data that was originally missing in the practices' raw data versus data that were

https://doi.org/10.1016/j.xops.2024.100542
ISSN 2666-9145/24

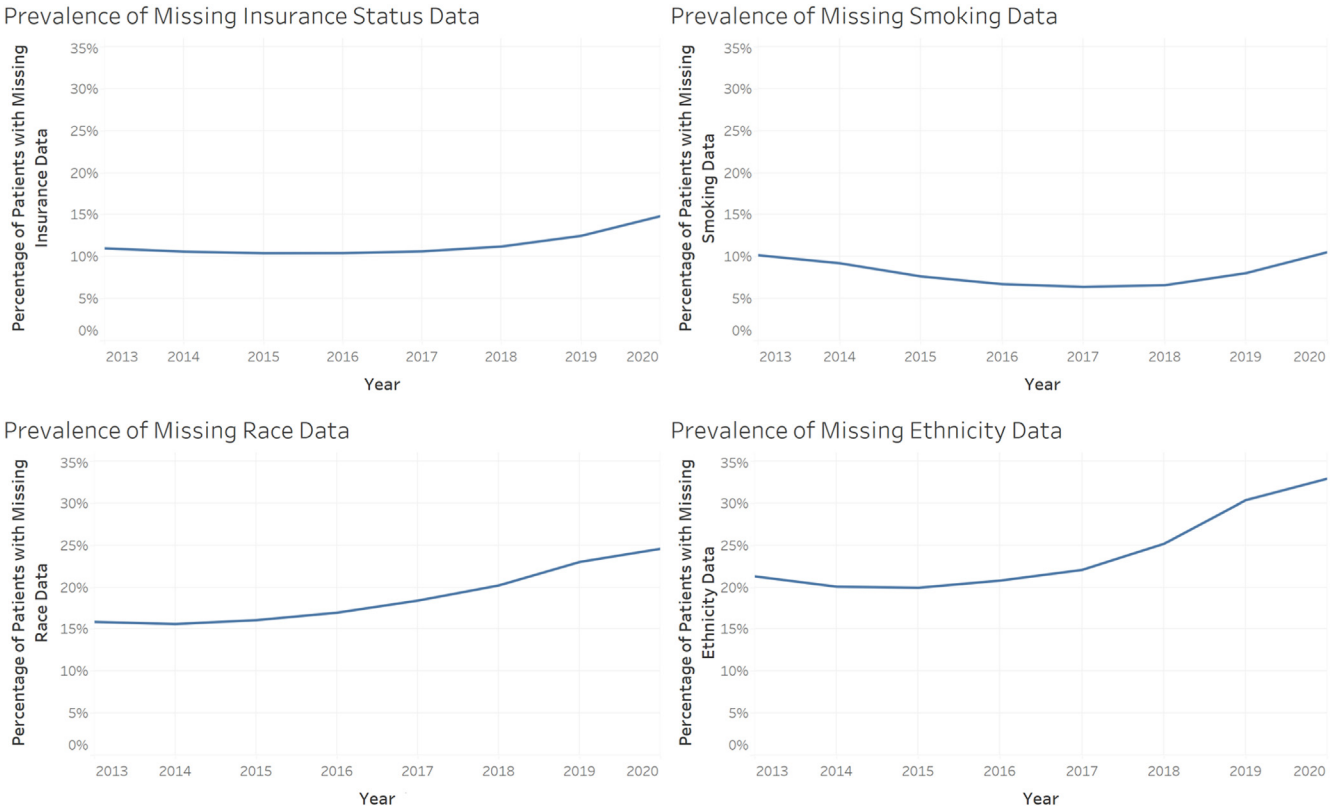

Figure 1. The percentage of patients with missing sociodemographic data and ≥1 clinical visit in each calendar year.

relabeled as missing at some stage during the data curation. A greater understanding of the trends in missing data may help inform the interpretation of clinical findings and limitations from registry-based studies.[11] This report aims to describe the prevalence of missing sociodemographic data in the IRIS Registry, as these variables are frequently used in clinical research. We hypothesize that practice-level characteristics may be associated with the prevalence of missing data and these factors may need to be considered in the design of studies using registry data.

Methods

This cross-sectional study was deemed exempt by the Massachusetts General Brigham Institutional Review Board and informed consent was not required given analysis of deidentified EHR data. The study adhered to the tenets of the Declaration of Helsinki and the Strengthening the Reporting of Observational Studies in Epidemiology guidelines.[12] We included all patients followed at practices participating in the IRIS registry. The database was frozen on December 24, 2021 and accessed on July 16, 2022. The data collection methodology of the IRIS Registry has been described previously.[1] We collected data at the patient-level and at the practice-level. For each patient we collected birth year, sex, race, ethnicity, insurance type, 3-digit zip code, state of residence, and smoking status. The IRIS Registry provides all sociodemographic data as static values located in a patient demographic table. Smoking status and insurance type have more granular records than the other sociodemographic information, for example, start date and stop date. The most recent nonnull data were extracted for these variables. If there was not a nonnull record for a patient, then this sociodemographic column was labeled unknown. For each practice we collected data on the practice size, geographic location, and patient population characteristics.

We described the prevalence of missing data for each sociodemographic variable at the patient-level between 2013 and 2020. Next, we described geographic trends in the prevalence of missing data across the United States at the state level. Finally, we identified all practices contributing data to the IRIS Registry and categorized them into quartiles (Qs) based on each characteristic over the 2013 to 2020 study interval (total number of physicians conducting patient visits, diagnoses, or procedures, total number of patients that recorded a visit, encounters per patient, proportion of pediatric patients, and proportion of privately insured patients). We also categorized practices based on the date of first available data in the registry and into rural versus urban location using the Rural-Urban Commuting Area codes from the United States census.[13] We used multiple linear regression to evaluate the association of practice-level characteristics with missing patient-level sociodemographic data. Statistical analyses were performed using R, version 4.2.0 (R Foundation for Statistical Computing), and data visualizations were created using Tableau, version 2020.3 (Tableau Software, LLC). All statistical tests were 2-tailed with significance defined as $P < 0.05$.

Results

This study included the EHR data of 66 477 365 patients followed at 3306 practices participating in the IRIS Registry between 2013 and 2020. The median number of patients per

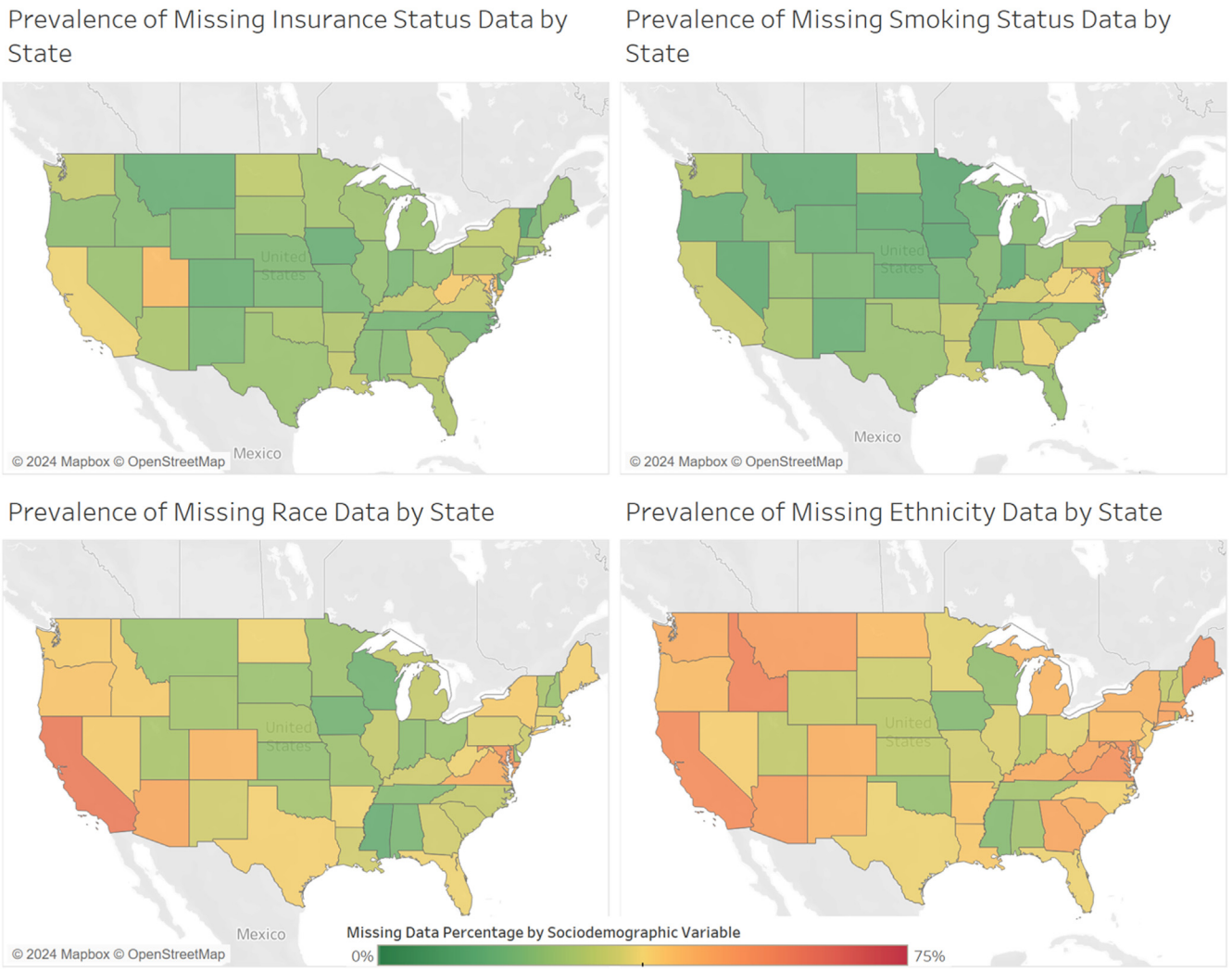

Figure 2. Geographic distribution of missing data for each sociodemographic variable.

practice was 11 415 (interquartile range: 5849−24 148) and the median number of physicians per practice was 3 (interquartile range: 1−7). The prevalence of missing patient-level sociodemographic data was 0.1% for birth year, 0.4% for sex, 24.8% for race, 30.2% for ethnicity, 2.3% for 3-digit zip code, 14.8% for state, 5.5% for smoking status, and 17.0% for insurance type. There was an increasing trend in the prevalence of missing data for race, ethnicity, and insurance type as a function of time (Fig 1). There was also evidence of geographic variation in the prevalence of missing data at the state-level (Fig 2). The distribution of missing data percentages across practices exhibits a right-skewed pattern, characterized by a lengthy tail. Most practices had a low prevalence of missing sociodemographic data, while few practices had a higher percentage of missing data (Fig 3).

At the practice-level, we compared the largest Q (Q4) to the smallest Q (Q1) for each continuous measure to capture the contrast between the upper and lower ends of each scale. Larger practices had a slightly lower prevalence of patients with missing race data (20% vs. 21%; $P = 0.011$), though it should be noted that the prevalence was even lower for Q2 and Q3, 17% and 18% respectively. Practices with many visits per patient had a lower prevalence of missing race (15% vs. 25%; $P < 0.001$), ethnicity (20% vs. 27%; $P < 0.001$), and insurance type (10% vs. 15%; $P < 0.001$), but a slightly higher prevalence of missing data for smoking status (3.3% vs. 5.5%; $P < 0.001$). Practices caring for a predominantly publicly insured patient population had a greater prevalence of missing race (23% vs. 18%; $P < 0.001$) and missing insurance (29% vs. 7%, $P < 0.001$). Practices with a greater proportion of pediatric patients had a higher prevalence of missing insurance variables (14% vs. 11%; $P < 0.001$) and a lower prevalence of missing data for smoking status (5.4% vs. 3.6% $P < 0.001$). Practices whose first record in the IRIS Registry was after 2013 had a higher prevalence of missing race (26% vs. 18%; $P < 0.001$) and ethnicity (33% vs. 21%; $P < 0.001$) data, but lower prevalence of missing smoking status (3.6% vs. 4.6%; $P < 0.001$) data (Table 1).

In the multiple linear regression analysis, missing race was associated with practices located in the West region

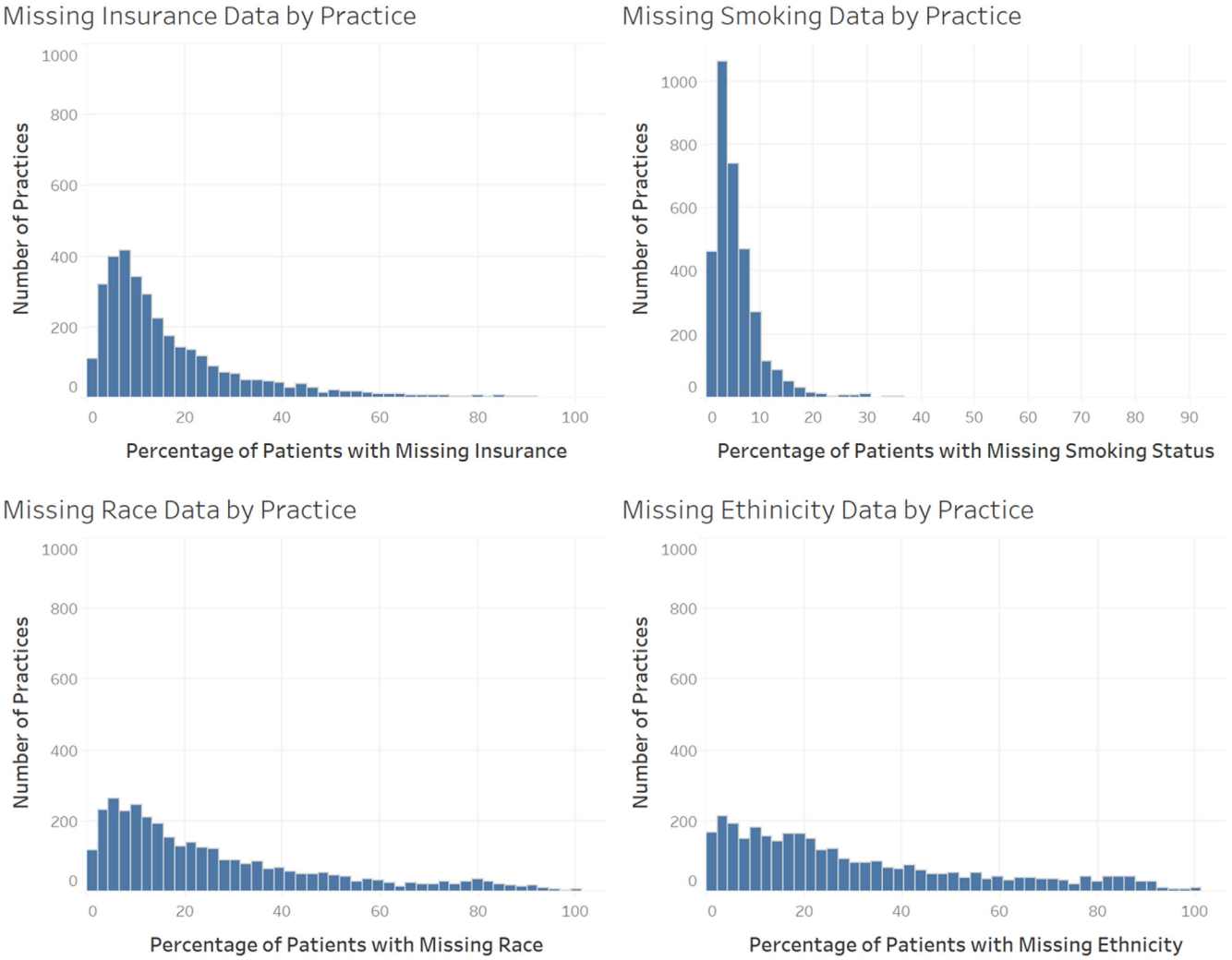

Figure 3. The distribution of missing data for each sociodemographic variable at the practice-level.

($P < 0.001$) and practices with unknown regional data ($P < 0.001$), in urban areas and areas with unknown community type ($P < 0.001$), with a lower proportion of privately insured patients ($P = 0.002$), with fewer visits per patient ($P < 0.001$), and with a first IRIS Registry record beyond the year 2013 ($P < 0.001$). Missing ethnicity data were associated with practices in the West region ($P < 0.001$) and those with unknown region ($P = 0.02$), in urban areas ($P = 0.004$), with fewer visits per patient ($P < 0.001$), and practices whose first record was after 2013 ($P < 0.001$). Missing smoking status data were associated with practices with a lower proportion of pediatric patients ($P < 0.001$), with fewer visits per patient ($P < 0.001$), and practices whose first record precedes the year 2013 ($P < 0.001$). Missing insurance status data were associated with practices located in the Northeast ($P = 0.001$), South ($P = 0.004$), and Unknown ($P = 0.005$) regions, with a higher proportion of pediatric patients ($P < 0.001$), in urban areas ($P < 0.001$), with a lower proportion of privately insured patients out of the patients with known insurance status ($P < 0.001$), with fewer visits per patient ($P < 0.001$), and practices whose first record precedes the year 2013 ($P = 0.02$) (Table 2).

Discussion

Missing sociodemographic data in the IRIS Registry are prevalent and associated with several practice-level characteristics. Although several sociodemographic variables such as birth year, sex, and zip code rarely contain missing data, others including race, ethnicity, and insurance status are missing for up to a third of patients in the registry. The temporal and geographic trends in missing sociodemographic data as well as practice-level associations deserve careful analytic consideration to minimize the potential for biased effect estimates.

The overall rates of missing sociodemographic data in the IRIS Registry are comparable to other large EHR-derived registries.[14,15] Several potential sources may contribute to missing data in the registry. Data were collected in the EHR of each participating practice and

Table 1. The Prevalence of Missing Patient Sociodemographic Data by Practice Characteristics

Sociodemographic Variable*	Number of Practices†	Percentage of Patients with Missing Race		Percentage of Patients with Missing Ethnicity		Percentage of Patients with Missing Insurance Status		Percentage of Patients with Missing Smoking Status	
	N = 3306	%	P *Value*§	%	P *Value*§	%	P *Value*§	%	P *Value*§
Practice size (patients)			0.011		0.10		0.6		0.5
Q1 (<5848)	841 (25%)	21 (9, 47)		24 (9, 57)		12 (6, 24)		4.5 (2.7, 7.2)	
Q2 (5849–11 410)	842 (25%)	17 (9, 38)		21 (9, 46)		12 (7, 21)		4.4 (2.7, 7.3)	
Q3 (11 420–24 125)	841 (25%)	18 (8, 33)		22 (11, 43)		12 (7, 23)		4.3 (2.7, 7.0)	
Q4 (≥24 171)	842 (25%)	20 (10, 33)		24 (13, 40)		12 (8, 20)		4.6 (3.1, 6.9)	
Practice size (physicians)			0.2		0.15		0.011		**<0.001**
Q1 (1)	872 (26%)	20 (9, 42)		21 (8, 50)		12 (7, 25)		4.2 (2.5, 6.9)	
Q2 (2–4)	826 (25%)	17 (8, 38)		21 (9, 50)		11 (6, 23)		4.5 (2.7, 7.5)	
Q3 (4–7)	847 (25%)	18 (9, 35)		23 (11, 44)		11 (6, 20)		4.8 (3.0, 7.5)	
Q4 (≥8)	821 (24%)	20 (10, 34)		24 (14, 41)		12 (7, 21)		4.3 (3.0, 6.9)	
Visits per patient			**<0.001**		**<0.001**		**<0.001**		**<0.001**
Q1 (<4.2)	841 (25%)	25 (12, 54)		27 (12, 64)		15 (8, 30)		3.3 (1.9, 5.6)	
Q2 (4.2–5.5)	842 (25%)	20 (10, 40)		25 (11, 47)		14 (7, 24)		4.0 (2.6, 6.4)	
Q3 (5.5–7.2)	841 (25%)	16 (8, 32)		20 (9, 40)		11 (7, 20)		4.8 (3.2, 7.7)	
Q4 (≥7.2)	842 (25%)	15 (8, 29)		20 (10, 36)		10 (6, 16)		5.5 (3.9, 8.3)	
US census region			**<0.001**		**<0.001**		**<0.001**		**<0.001**
South	783 (23%)	14 (7, 29)		19 (8, 38)		10 (6, 18)		5.0 (3.2, 7.6)	
Northeast	304 (9.0%)	17 (8, 36)		20 (10, 46)		14 (8, 27)		4.3 (2.5, 6.9)	
North Central	417 (12%)	15 (8, 30)		21 (10, 41)		12 (7, 18)		4.4 (3.1, 7.1)	
West	438 (13%)	32 (17, 51)		32 (16, 50)		14 (7, 25)		4.4 (3.0, 6.8)	
Unknown	1424 (42%)	20 (10, 38)		23 (11, 47)		12 (7, 23)		4.1 (2.5, 7.0)	
Community type			**<0.001**		**<0.001**		**0.009**		**<0.001**
Urban	1723 (51%)	19 (9, 38)		23 (11, 43)		12 (7, 22)		4.6 (3.1, 7.3)	
Rural	219 (6.5%)	10 (5, 25)		16 (5, 33)		11 (6, 18)		4.5 (2.7, 7.1)	
Unknown	1424 (42%)	20 (10, 38)		23 (11, 47)		12 (7, 23)		4.1 (2.5, 7.0)	
Private-to-public insurance ratio			**<0.001**		**0.061**		**<0.001**		**<0.001**
Q1 (<55.3%)	841 (25%)	23 (10, 47)		24 (11, 54)		29 (15, 44)		3.8 (2.3, 7.0)	
Q2 (55.3%–66.2%)	842 (25%)	18 (9, 36)		23 (12, 43)		14 (8, 22)		4.9 (3.1, 7.5)	
Q3 (66.2%–75.7%)	841 (25%)	17 (8, 35)		22 (9, 45)		10 (6, 15)		4.7 (3.0, 7.2)	
Q4 (≥75.7%)	842 (25%)	18 (9, 33)		22 (10, 41)		7 (4, 12)		4.3 (3.0, 6.8)	
Pediatric population			0.5		0.3		**<0.001**		**<0.001**
Q1 (<0.59%)	842 (25%)	17 (8, 37)		21 (10, 45)		11 (6, 21)		5.4 (3.5, 8.2)	
Q2 (0.59%–1.65%)	841 (25%)	19 (9, 37)		23 (11, 45)		11 (7, 22)		4.8 (3.0, 7.3)	
Q3 (1.65%–4.18%)	842 (25%)	19 (9, 38)		23 (11, 47)		12 (7, 21)		4.1 (2.7, 6.8)	
Q4 (≥4.18%)	841 (25%)	20 (10, 36)		22 (9, 42)		14 (8, 24)		3.6 (2.2, 5.8)	
Earliest record in registry			**<0.001**		**<0.001**		0.2		**<0.001**
≤2013	2831 (84%)	18 (9, 34)		21 (10, 41)		12 (7, 22)		4.6 (2.9, 7.4)	
2014 to ≤2020	535 (16%)	26 (12, 60)		33 (13, 69)		12 (7, 23)		3.6 (2.2, 5.7)	

Q = quartile; US = United States.
*Continuous variables are split into quartiles.
†n (% of total practices).
§*P* values and statistical significance for the Kruskal–Wallis test (continuous values) and Wilcoxon signed-rank test (discrete values) are based on an alpha of 0.05 in bold.

subsequently uploaded to the registry database.[16] Sociodemographic data collected during patient registration often have a high prevalence of missing data.[17–19] Race and ethnicity, for example, are missing for an average of 25% of patients in other United States-based EHR databases.[18] There may be differences in how these data are collected (at time of patient registration vs. during clinical visits) that may explain some of the differences in missing data across practices. The prevalence of missing insurance type data may reflect differences in the coding of insurance information across practices and challenges in aggregation across multiple EHR systems. The IRIS Registry is unique in that, unlike payer claims databases, it includes publicly insured, privately insured, and uninsured patients. It is possible that some of the missing data represent patients without insurance since the same classification system for uninsured patients may not be used by all practices. Smoking status information was collected frequently across practices participating in the registry, likely reflecting incentive structures to collect these data. Interestingly, smoking status was less likely to be reported among practices with many visits per patient. This finding may suggest that these data are not collected or updated from patients who are frequent visitors to a practice, and this association should be considered in analyses involving the smoking status variable.

Table 2. Modeling Missing Sociodemographic Data Using Multivariable Linear Regression with Practice-Level Characteristics

Predictors	Proportion of Missing Race Data			Proportion of Missing Ethnicity Data			Proportion of Missing Smoking Status Data			Proportion of Missing Insurance Status		
	Estimates	*CI*	P *Value*	*Estimates*	*CI*	P *Value*	*Estimates*	*CI*	P *Value*	*Estimates*	*CI*	P *Value*
Practice size (number of distinct patients, reference = Q1)												
Q2	−0.02	−0.04 to 0.00	0.129	−0.02	−0.04 to 0.01	0.174	−0.00	−0.01 to 0.00	0.554	−0.01	−0.02 to 0.01	0.330
Q3	−0.02*	−0.05 to −0.00	0.036	−0.01	−0.04 to 0.01	0.312	−0.00	−0.01 to 0.00	0.053	0.00	−0.01 to 0.01	0.958
Q4	−0.01	−0.04 to 0.01	0.217	−0.01	−0.03 to 0.02	0.527	−0.01*	−0.01 to −0.00	0.042	−0.01	−0.02 to 0.01	0.313
Region (reference = North Central)												
Northeast	0.01	−0.03 to 0.04	0.734	0.01	−0.03 to 0.04	0.740	−0.00	−0.01 to 0.01	0.663	0.03‡	0.01–0.05	0.001
South	−0.01	−0.04 to 0.01	0.283	−0.02	−0.05 to 0.01	0.145	0.00	−0.00 to 0.01	0.326	−0.02†	−0.04 to −0.01	0.004
Unknown	0.07‡	0.04–0.11	<0.001	0.05*	0.01–0.09	0.020	−0.00	−0.01 to 0.01	0.986	0.03†	0.01–0.05	0.005
West	0.12‡	0.10–0.15	<0.001	0.06‡	0.03–0.10	<0.001	−0.00	−0.01 to 0.00	0.368	0.00	−0.01 to 0.02	0.600
Community type (reference = rural)												
Urban	0.06‡	0.03–0.10	<0.001	0.05†	0.02–0.09	0.004	0.00	−0.01 to 0.01	0.726	0.03‡	0.02–0.05	<0.001
Pediatric patient proportion (reference = Q1)												
Q2	0.01	−0.02 to 0.03	0.607	0.01	−0.01 to 0.03	0.423	−0.00	−0.01 to 0.00	0.092	0.02‡	0.01–0.03	0.001
Q3	0.01	−0.01 to 0.03	0.514	0.02	−0.01 to 0.04	0.128	−0.01†	−0.01 to −0.00	0.002	0.03‡	0.02–0.04	<0.001
Q4	−0.00	−0.03 to 0.02	0.874	−0.01	−0.03 to 0.02	0.645	−0.01‡	−0.02 to −0.01	<0.001	0.04‡	0.03–0.05	<0.001
Privately insured patient proportion (reference = Q1)												
Q2	−0.03†	−0.05 to −0.01	0.003	−0.01	−0.04 to 0.01	0.281	0.00	−0.00 to 0.01	0.446	−0.15‡	−0.16 to −0.14	<0.001
Q3	−0.05‡	−0.07 to −0.03	<0.001	−0.02	−0.04 to 0.01	0.132	0.00	−0.00 to 0.01	0.636	−0.20‡	−0.21 to −0.19	<0.001
Q4	−0.03†	−0.06 to −0.01	0.002	−0.02	−0.04 to 0.01	0.120	−0.00	−0.01 to 0.00	0.232	−0.23‡	−0.25 to −0.22	<0.001
Visits per patient (reference = Q1)												
Q2	−0.05‡	−0.07 to −0.03	<0.001	−0.04‡	−0.07 to −0.02	<0.001	0.00	−0.00 to 0.01	0.665	−0.02‡	−0.03 to −0.01	0.001
Q3	−0.09‡	−0.11 to −0.07	<0.001	−0.08‡	−0.10 to −0.05	<0.001	0.01†	0.00–0.01	0.005	−0.03‡	−0.04 to −0.02	<0.001
Q4	−0.11‡	−0.14 to −0.09	<0.001	−0.09‡	−0.12 to −0.06	<0.001	0.01‡	0.01–0.02	<0.001	−0.04‡	−0.05 to −0.02	<0.001
Earliest record in registry (reference ≤2013)												
First record after 2013	0.07‡	0.05–0.09	<0.001	0.09‡	0.06–0.11	<0.001	−0.02‡	−0.02 to −0.01	<0.001	−0.01*	−0.03 to −0.00	0.020

CI = confidence interval; Q = quartile.
*$P < 0.05$.
†$P < 0.01$.
‡$P < 0.001$.

Based on the results of this study, we propose several recommendations for working with missing sociodemographic data in the IRIS Registry. First, the prevalence of missing data for geographic location is greater at the state-level than the zip code-level. We recommend using zip code data primarily and mapping to larger geographies (cities, states, regions) as needed. Second, missing race and ethnicity data are more prevalent in practices that care for publicly-insured patient populations. We do not recommend excluding unknown values in regression models using these variables as this will result in selection bias. Third, given changes in the prevalence of missing data over time, sensitivity analyses should be considered to confirm robustness of results across different study time periods. Finally, frequent patient visits were associated with lower rates of missing sociodemographic data—likely because of increased opportunities to collect these data overtime. This relationship should be considered in the design of studies involving conditions or procedures with variable length of follow-up (or high frequency of loss to follow-up).

This study has several limitations. We focused our analyses on the sociodemographic variables in the registry. This was done intentionally since most studies rely on these data; however, further work is needed to understand the prevalence of missing data in diagnostic and procedural coding, clinical variables, and medication data available in the registry. At the moment, we cannot differentiate between data that was originally missing in the practices' raw data versus data that were relabeled as missing at some stage during the data curation. Relatedly, findings based on practice-level variables with a high prevalence of null values in the database, including practice location, should be interpreted with caution in the case that the data are not missing completely at random. It should be noted that there are optimizations being made to the ingestion pipelines that are intended to increase transparency in data processing. However, while the data processing pipelines change in future versions of the database, it remains important to mitigate potential sources of bias by understanding the extent and distribution of missing information in the dataset. In acknowledging the potential impact of missing data on prior studies utilizing the IRIS Registry, it is crucial to emphasize how addressing this limitation can enhance our ability to interpret findings and shed light on potential biases that may have influenced previous research outcomes.

In conclusion, there is evidence of geographic and temporal variation in the prevalence of missing sociodemographic data in the IRIS Registry. Several practice-level characteristics—including size, visit frequency, and patient population—were associated with missing data at the patient-level. Large EHR registries offer a unique opportunity to investigate clinical outcomes and physician practice patterns. However, there is a need to develop standardized approaches for handling missing data in the IRIS Registry to minimize potential sources of bias and ensure reproducibility across research studies.

Footnotes and Disclosures

Originally received: January 8, 2024.
Final revision: April 17, 2024.
Accepted: April 23, 2024.
Available online: April 30, 2024. Manuscript no. XOPS-D-24-00005.

[1] Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts.

[2] American Academy of Ophthalmology, San Francisco, California.

[3] Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts.

*See Appendix for members of the IRIS® Registry Analytic Center Consortium.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

T.E.: Support − National Eye Institute (P30 EY003790); Grants − Genentech, National Science Foundation, and Brightfocus Foundation.

J.W.M.: Consultant − Genentech/Roche, Sunovion, KalVista Pharmaceuticals, Ltd, and ONL Therapeutics; Payment or honoraria − Heidelberg Engineering; Patents − US 7,811,832 (with royalties paid by ONL Therapeutics to Massachusetts Eye and Ear), and US 5,798,349; US 6,225,303; US 6,610,679; CA 2,185,644; CA 2,536,069 (with royalties paid by Valeant Pharmaceuticals to Massachusetts Eye and Ear); Shares − Lowy Medical Research Institute, Ltd Mactel Study, Ciendias Bio Equity; Others - Personal fees from Aptinyx, Inc board of directors.

A.C.L.: Consultant − Regeneron.

I.O.: Grants - Research to Prevent Blindness, Knights Templar Eye Foundation, Boston Children's Hospital.

Financial support was provided by the Massachusetts Eye and Ear Clinical Data Science Fund, National Institutes of Health grant number P30 EY003790, Agency for Healthcare Research and Quality grant number T32HS000063 (I.O.), National Center for Advancing Translational Sciences (K12TR004381 [I.O.]), and the Children's Hospital Ophthalmology Foundation, Inc., Boston, MA (I.O.). The sponsor or funding organizations had no role in the design or conduct of this research.

HUMAN SUBJECTS: No human subjects were included in this study. This cross-sectional study was deemed exempt by the Massachusetts General Brigham Institutional Review Board and informed consent was not required given analysis of deidentified electronic health record data. The study adhered to the tenets of the Declaration of Helsinki and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Ross, Ivanov, Lorch, Oke

Data collection: Ross, Elze, Oke

Analysis and interpretation: Ross, Ivanov, Elze, Lorch, Miller, Lum, Oke

Obtained funding: Oke

Overall responsibility: Ross, Lorch, Miller, Lum, Oke

Abbreviations and Acronyms:
EHR = electronic health record; **IRIS** = Intelligent Rearch in Sight; **Q** = quartile.

Keywords:
Electronic health records, IRIS registry, Missing data, Sociodemographic data.

Correspondence:
Isdin Oke, MD, MPH, Boston Children's Hospital, 300 Longwood Avenue, Boston, MA 02115. E-mail: Isdin.Oke@childrens.harvard.edu.

References

1. Chiang MF, Sommer A, Rich WL, et al. The 2016 American Academy of ophthalmology IRIS® registry (intelligent research in Sight) database. *Ophthalmology*. 2018;125:1143−1148.
2. Ciociola EC, Yang SA, Hall N, et al. Effectiveness of trabeculectomy and tube shunt with versus without concurrent phacoemulsification: intelligent research in Sight registry longitudinal analysis. *Ophthalmol Glaucoma*. 2023;6:42−53.
3. Yang S-A, Mitchell W, Hall N, et al. Trends and usage patterns of minimally invasive glaucoma surgery in the United States: IRIS® registry analysis 2013-2018. *Ophthalmol Glaucoma*. 2021;4:558−568.
4. Oke I, Hall N, Elze T, et al. Risk factors associated with pterygium reoperation in the IRIS registry. *JAMA Ophthalmol*. 2022;140:1138−1141.
5. Oke I, Hall N, Elze T, et al. Adjustable suture technique is associated with fewer strabismus reoperations in the intelligent research in Sight registry. *Ophthalmology*. 2022;129:1028−1033.
6. Repka MX, Li C, Lum F. Multivariable analyses of amblyopia treatment outcomes from A clinical data registry. *Ophthalmology*. 2022;S0161-6420:00692−00693. https://doi.org/10.1016/j.ophtha.2022.09.005.
7. Oke I, Reshef ER, Elze T, et al. Smoking is associated with a higher risk of surgical intervention for thyroid Eye disease in the IRIS registry. *Am J Ophthalmol*. 2023;249:174−182.
8. Oke I, Elze T, Miller JW, et al. Factors associated with nasolacrimal duct probing failure among children in the intelligent research in Sight registry. *JAMA Ophthalmol*. 2023;141:342−348.
9. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. *JAMA Netw Open*. 2021;4:e210184.
10. Henry AJ, Hevelone ND, Lipsitz S, Nguyen LL. Comparative methods for handling missing data in large databases. *J Vasc Surg*. 2013;58:1353−1359.e6.
11. Haneuse S, Bogart A, Jazic I, et al. Learning about missing data mechanisms in electronic health records-based research: a survey-based approach. *Epidemiology*. 2016;27:82−90.
12. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. 2007;147:573−577.
13. U.S. Census Bureau. *2020 Census Redistricting Data (Public Law 94-171): Data, tools, and reports for the 2020 Census*. 2020.
14. Yang DX, Khera R, Miccio JA, et al. Prevalence of missing data in the national cancer database and association with overall survival. *JAMA Netw Open*. 2021;4:e211793.
15. Koffman L, Levis AW, Arterburn D, et al. Investigating bias from missing data in an electronic health records-based study of weight loss after bariatric surgery. *Obes Surg*. 2021;31:2125−2135.
16. Pershing S, Lum F. The American Academy of Ophthalmology IRIS Registry (Intelligent Research in Sight): current and future state of big data analytics. *Curr Opin Ophthalmol*. 2022;33:394−398.
17. Sholle ET, Pinheiro LC, Adekkanattu P, et al. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *J Am Med Inform Assoc*. 2019;26:722−729.
18. Polubriaginof FCG, Ryan P, Salmasian H, et al. Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc*. 2019;26:730−736.
19. Zingmond DS, Parikh P, Louie R, et al. Improving hospital reporting of patient race and ethnicity–approaches to data auditing. *Health Serv Res*. 2015;50(Suppl 1):1372−1389.