Assessing Disparities Through Missing Race and Ethnicity Data: Results from a Juvenile Arthritis Registry

Katelyn Banschbach

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2024

Committee:

Peter Tarczy-Hornoch

Esi Morgan

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

Assessing Disparities Through Missing Race and Ethnicity Data: Results from a Juvenile Arthritis Registry

Katelyn Banschbach

Chair of the Supervisory Committee:

Peter Tarczy-Hornoch

Biomedical and Health Informatics

Ensuring high quality race and ethnicity data within the electronic health record (EHR) and across linked systems, such as patient registries, is necessary to achieve a goal of inclusion of racial and ethnic minorities in scientific research and detect disparities associated with race and ethnicity. The project goal was to improve race and ethnicity data completion within the Pediatric Rheumatology Care Outcomes Improvement Network (PR-COIN) and assess impact of improved data completion on conclusions drawn from the registry. The project consisted of 5 parts: (1) Identifying baseline missing race and ethnicity data, (2) REDCap survey of current collection and entry, (3) Data completion through audit and feedback cycles, (4) Assessment of impact on outcome measures, and (5) Participant interviews and thematic analysis. REDCap survey (Supplementary Materials A) and participant interviews (Supplementary Materials B) are available in the supplementary materials. Across 6 participating centers, 29% of patients were missing race and 31% were missing ethnicity, with most patients missing both. Rates of missingness varied by data entry method (electronic vs manual). Recovered data had a higher

percentage of patients with Other race or Hispanic/Latino ethnicity compared to patients with non-missing race and ethnicity at baseline. Black patients had a significantly higher odds ratio of having a clinical juvenile arthritis disease activity score (cJADAS10) of ≥5 at first follow up compared to White patients. There was no significant change in odds of cJADAS10 ≥5 for race and ethnicity after data completion. Patients missing race and ethnicity were more likely to be missing cJADAS values which may affect the ability to detect changes in odds of cJADAS ≥5 after completion. About 1/3 of patients in a pediatric rheumatology registry were missing race and ethnicity data. After three audit and feedback cycles, centers decreased missing data by 94%, primarily via data recovery from the EHR. In this sample, completion of missing data did not change the findings related to differential outcomes by race. Recovered data was not uniformly distributed compared to those with non-missing race and ethnicity at baseline, suggesting that differences in outcomes after completing race and ethnicity data may be seen with larger sample sizes.

Table of Contents

Publication Notes:

This thesis and a corresponding manuscript were prepared in parallel. The manuscript was submitted on May 5, 2024 to Frontiers in Pediatrics. The journal was informed that the same content was submitted as a thesis and approved the submission to the journal. The thesis was submitted to the University of Washington ProQuest system on May 22, 2024.

1 Introduction

I hypothesized that improved race and ethnicity data quality will enhance the accuracy of observational studies and the inclusion of minority patients.

Learning health systems are cited as a tool which can help identify and address disparities in real time.[1,2] Registries, such as PR-COIN are components of the learning health system which can help achieve these equity goals. However, to achieve equity and inclusion, the equity metrics must have quality data. Consider using race data as a measure of racism and systemic inequities within the health system. The registry can be used to track differences in disease outcomes by race and ongoing quality improvement interventions can be created from registry data. However, if you are stratifying by race and 30% of patients are missing that data, these patients are now excluded from analysis. Can you trust the outcomes you are tracking?

To answer that question, some additional research questions needed to be addressed which formed the basis of the project aims.

Q1. What is the extent of missing data in the registry?

Q2. How is information getting from the patient, into the EHR, then into PR-COIN?

Q3. Can we improve completion of race and ethnicity?

Q4. How does completion affect outcome measures results from the registry?

The project consisted of four aims:

(1) To understand the extent of missing race and ethnicity data in PR-COIN database for patients with JIA. Baseline demographics including missing race and ethnicity data was pulled from the registry for participating sites.

(2) Survey of data flow and collection methods for race and ethnicity for each participating center to identify best practices and barriers to optimal data quality in these fields. Data was gathered via REDCap survey of collection for each center.

(3) Improve data completion via audit and feedback reports. Centers were sent reports of missing race and ethnicity data with request for completion.

(4) Assessment of impact of data completion on outcome measurements from the registry. A chosen outcome measurement (cJADAS ≥5) was measured before and after data completion and compared to assess for change.

2 Background

Secondary use of electronic health record (EHR) data holds great potential for understanding patient populations, choosing interventions, and facilitating real-time research, overall pushing institutions towards becoming true learning health systems.[3,4] As we develop these learning health systems and large clinical and research databases, ensuring data quality becomes even more important.[4] This is of particular importance in foundational areas on which further analyses will be performed, such as race and ethnicity data, especially given their known association with healthcare disparities.

While there is not a single standardized way of evaluating data quality, Feder has described a set of common domains which can be used to evaluate and improve data quality including data accuracy, completeness, consistency, credibility, and timeliness.[4] The literature

suggests three main threats to high quality race and ethnicity data including accuracy, completeness, and consistency.[5–7] Accuracy is defined as "the degree to which the value in the EHR is a true representation of the real-world value," completeness describes missing data, and consistency reflects truth of the value across multiple sources.[4]

Reliable, culturally conscious ascertainment of race and ethnicity data and completeness of entry are crucial for inclusion of minority populations in health systems research and to mitigate inherent systemic bias.[8,9,10] While race and ethnicity are social constructs, they serve as important markers for disparities and social determinants of health.[11,12] These concepts reflect a person's identity rather than a genetic or phenotypic basis, making self-reporting the gold standard for accurate race and ethnicity data.

Racial and ethnic minorities remain underrepresented in research despite similar willingness to participate.[8] Incomplete race and ethnicity data can lead to exclusion from disparities analysis. Moreover, those missing this data are more likely to be Black or Hispanic, further worsening disparities and exclusion of minority patients from research.[13,14] Research and secondary analytics done with incomplete race and ethnicity can unintentionally worsen disparities.[14–17] Alternatively, missing data may obscure disparities which are already present.[14] Ensuring high quality race and ethnicity data within the EHR and across linked systems, such as patient registries, allows identification of disparities and is necessary to achieve a goal of inclusion of racial and ethnic minorities in scientific research.[5,15]

We describe the iterative process of identifying and completing missing race and ethnicity data at six centers within the Pediatric Rheumatology Care Outcomes Improvement Network (PR-COIN). The PR-COIN database contains over 7,200 active patients with juvenile idiopathic arthritis (JIA) spanning 50,000 encounters with plans to add more pediatric rheumatologic

diseases over time. Completing missing race and ethnicity data will help avoid unintentionally building inequitable algorithms and system structures. Furthermore, research done with incomplete data may make invalid inferences on disparities and stratification by race due to exclusion of patients with missing data. This study provides a framework for addressing missing data and explores the impact of filling in missing data on conclusions drawn from the registry.

3 Methods

This study was approved by the Seattle Children's Institutional Review Board and was conducted using data obtained through PR-COIN, collected by the physicians, providers and families participating in this multicenter quality improvement collaborative.[18]

The project consisted of 5 parts: (1) Identifying baseline missing race and ethnicity data, (2) Survey of current collection and entry, (3) Data completion (filling in missing race/ethnicity values) through audit and feedback cycles, (4) Assessment of impact of additional race and ethnicity values on outcome measures, and (5) Participant interviews and thematic analysis. PR-COIN centers that were actively submitting data to the registry were eligible to participate. Eligible centers were issued an email invitation for voluntary participation in the research.

Baseline aggregate patient demographic and diagnosis data were obtained for the participating PR-COIN centers, and descriptive analyses were performed. Amount of missing race and ethnicity data was calculated by center. Only patients present in baseline data were included in the subsequent rounds of data completion and final data analysis. I did not incorporate new patients enrolled into the registry during the study period. Due to very small numbers of patients, 3 race categories independently defined in the registry were aggregated as 'Other' for purpose of analysis, these were Asian, Native Hawaiian or Other Pacific Islander, and American Indian or Alaska Native. To maximize opportunities for data completion and

accuracy, patients with designated registry categories of 'Unknown', 'Not Reported', and 'Other' selected for race in the registry were aggregated with patients with the race field left blank to form the 'Missing' category for requested completion. For ethnicity, any patients with registry categories of 'Unknown' or 'Not Reported' selected were aggregated with patients with the ethnicity field left blank to form the 'Missing' category for this study. 'Unknown' represents data not available in the EHR and 'Not reported' represents patients who have chosen not to disclose their race and/or ethnicity. A REDCap survey on race and ethnicity collection and upload methods was administered at each center prior to starting data completion and could be answered by the centers primary investigator, the research coordinator, or both. Survey questions are available in the Supplementary Materials.

Audit and feedback cycles were performed by creating and sending reports of patients with 'Missing' race and/or ethnicity data to each center. Centers were requested to complete the missing data fields within the registry using data already available in the EHR. After allowing a period for completion, new reports were generated and sent again with request for completion for a total of 3 cycles over 6 months. No new patients were added with the audit and feedback cycles, and any duplicate patient records were deleted from the registry. Data was obtained before completion (time 0), after round 1 of data completion (time 1), after round 2 of data completion (time 2), and after round 3 of data completion (time 3 or after completion). For round 1, centers were asked to focus on identifying and addressing any systematic reasons for missing data such as incomplete mapping or electronic transfer of data. If no such problems could be corrected, the center would manually complete data where possible. For round 2, centers were requested to manually fill in remaining missing data in the registry that was available in the EHR. For round 3, centers were requested to convert remaining 'Missing' to either 'Unknown'

or ‘Not Reported’, as appropriate. No patients were contacted for updating of race and ethnicity data.

We obtained clinical juvenile arthritis disease activity scores (cJADAS10) at first registry follow up visit within 2-6 months of enrollment. CJADAS10 was chosen as an outcome measure due to prevalent use in the registry. It also contains components which are considered critical data elements with respect to data quality. CJADAS10 is a continuous disease activity measure which is more sensitive to detect change than the dichotomous ACR criteria for inactive disease.[19] I used a threshold of cJADAS10 ≥5 for all JIA subtypes using the cJADAS10 as this reflects greater than low disease activity for both oligoarticular and polyarticular arthritis. Odds ratio (OR) of cJADAS10 ≥5 at first visit after enrollment was compared before data completion and after data completion to assess how data completion changes in odds of cJADAS ≥5.

“We conducted two separate analyses: first using the initial data set with missing race/ethnicity values, and second with the updated data set that included observations with recovered missing values of race and ethnicity. For each analysis, we estimated the crude (univariable) odds ratio (OR) of disease activity score, cJADAS10 ≥5, for age, gender, race, ethnicity, and JIA subtype. Then we used a multivariable logistic regression model to estimate the adjusted ORs for race and ethnicity, while accounting for differences between race and ethnicity groups in distribution of age and gender. Our interest was in the difference in ORs for race and ethnicity before and after recovering missing values of race and ethnicity. All analyses were performed in R studio.” (Analysis methods written and performed by Jade Singleton)

Semi-structured, exploratory group interviews were conducted over two sessions with five out of six centers. The interviews were conducted to provide feedback on user experience with report format, to understand reasons for missing data, and identify best practice

recommendations for completeness based on participant experiences. Interview questions are available in the Supplementary Materials. The first author (KMB) was the moderator and concurrently took notes during the interviews. Interviews were not recorded. Interviews were followed by inductive thematic analysis conducted according to methodology and steps outlined by Braun and Clarke and are described below.[20] Coding was reviewed for agreement by a single second reviewer by another physician and last author on the paper, and any disagreement resolved via discussion (EMM).

1. **Familiarizing yourself with the data:** Notes from interviews were reviewed multiple times followed by a written summary and key points. (KMB)
2. **Generating initial codes**: Notes were reviewed line by line with codes assigned. Some lines were assigned multiple codes. This was performed twice with adjustment of codes during the second coding session. (KMB)
3. **Searching for themes**: Note segments were organized based on coding and used to identify themes or key concepts. (KMB)
4. **Reviewing themes**: Themes were compared to interview questions and goals for alignment, both reviewers established themes. (KMB, EMM)
5. **Define themes:** Meaning and patterns associated with themes and relationships between themes were identified. Discussion between reviewers used to arrive at a consensus. (KMB, EMM)
6. **Writing up:** Description of themes is written in the results section. (KMB)

4 Results

4.1 Identifying Baseline Missing Data

A total of 2359 patients with JIA were included across six PR-COIN centers. Table 1 depicts demographics of the baseline population prior to data completion.

Table 1: Patient Demographics

Age	Frequency
Mean (sd)	11.4 (5)
Gender	
Female	1653 (70%)
Male	706 (30%)
Race	
Black	105 (4%)
White	1430 (61%)
Other	141 (6%)
Missing	683 (29%)
Ethnicity	
Hispanic/Latino	159 (7%)
Not Hispanic/Latino	732 (31%)
Missing	1468 (62%)
ILAR Code	
Oligoarticular (Persistent and Extended)	716 (30%)
Polyarticular (RF+ and RF-)	579 (25%)
Enthesitis Related Arthritis	218 (9%)
Psoriatic Arthritis	113 (5%)
Systemic JIA	109 (5%)
Undifferentiated Arthritis	63 (3%)
Unknown	561 (24%)
Insurance	
Commercial/Private	1009 (43%)
Medicare/Medicaid	238 (10%)
Other	232 (10%)
Self-pay/None	163 (7%)
Missing	717 (30%)

Sd: standard deviation, ILAR: International League of Associations for Rheumatology, RF: rheumatoid factor

At baseline, race was missing in 29% of patients and ethnicity was missing in 31%. Of the 683 patients missing race, 669 (98%) of patients were also missing ethnicity. Percent of patients missing race or ethnicity by center ranged from 0.5% to 99%. Patients with missing race were more likely to be missing other metrics including International League of Associations for Rheumatology (ILAR) subtype as well as cJADAS10 and its components. Around 50% of patients with missing race or ethnicity were also missing ILAR subtype or cJADAS, compared to around 12% missing cJADAS or ILAR subtype in patients with non-missing race or ethnicity at baseline.

4.2 Survey of Current Collection and Entry

A REDCap survey was administered to each center to gather data on collection methods and practices for registry data. The survey also included questions about race and ethnicity collection at the institution and methods of input into the EHR. Lastly, data was collected on race and ethnicity options within each EHR for comparison to registry options. Full survey is available in the Supplementary Materials.

Table 2 depicts survey results. Five of 6 centers cited registration as the primary staff for collecting race and ethnicity data for the EHR. Data collection for the EHR occurs through a variety of methods across institutions including verbal reporting, direct entry online, and paper form. Most centers (4/6) have a research coordinator that inputs data, including race and ethnicity data into the registry. If race and ethnicity data is missing from the registry, no additional attempt is made to fill in data for 5 of 6 centers.

Table 2: Center REDCap Survey Data

Centers	A	B	C	D	E	F
Registry Data Entry Method	Manual	Manual	Electronic Data Transfer	Manual	Manual	Manual
Registry Data Entry Personnel	Not answered	Research coordinator, student	Research coordinator	Research coordinator, Other	Other	Research coordinator
Master List?	Yes	No	Yes	Yes	Yes	Yes
Master List with race and ethnicity?	No	Not applicable	No	Yes	No	No
Master List Updates	New enrollments	Not applicable	Monthly	Quarterly	Every other year	Weekly
Race/ethnicity data collection	Verbal collection	Direct entry, electronic form	Verbal collection, direct entry	Verbal collection, direct entry	Verbal collection, direct entry, paper	Direct entry, paper form
Who inputs in race and ethnicity in EHR	Registration	Registration, other - parent	Registration	Unknown	Registration, scheduling	Registration
Who inputs race and ethnicity into PR-COIN?	Provider	Research coordinator, other	Research coordinator	Research coordinator	Other	Research coordinator
Is there a process for identifying missing race or ethnicity in PR-COIN?	No	No	No	No	No	Yes - demographic form at visit

One center cited difference in race and ethnicity categories between the institution and registry as a barrier to accurate data collection and entry. One center uploads data via electronic data transfer (EDT) from the EHR, all other centers enter data manually. The center uploading data to the registry via EDT has the highest percent of missing race and ethnicity compared to other sites because the demographic data was not mapped from the EHR to the registry fields. The center

with the lowest amount of missing data also notes use of race and ethnicity in a 'Master List'. The Master List is a network recommended procedure in which centers create a list of all patients eligible for participation in the registry to monitor that registry enrollment is complete and reflective of the entire clinical patient population. Historically, the minimum data elements recommended for the Master List were patient name, MRN, date of birth, gender, ILAR code, diagnostic code, date of diagnosis, first, last and next visit date, and provider, as described in a network Change Package (or instruction on keeping a Master List). Prior to this project, race/ethnicity were considered optional in construction of the Master List.

All sites have the NIH five minimum categories for race including American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White.[11] The PR-COIN registration form includes these categories as well as Other, Unknown, and Not Reported with the ability to check multiple options to represent multiracial individuals. Two centers can select multiple races, 4 centers have Not Reported as an option, 4 have Other as an option, and Unknown is an option for 1 center. One center documents Hispanic/Latino as part of race, all others have a separate ethnicity category with Hispanic/Latino and Not Hispanic/Latino options.

4.3 Data Completion via Audit and Feedback Cycles

Throughout this section 'baseline non-missing' will refer to patients whose race and ethnicity were present before completion. Percent baseline non-missing represents the proportion of a given race or ethnicity as a percent of the total patients without missing race or ethnicity at baseline. Lastly, 'recovered' represents patients with missing race or ethnicity at baseline that was completed through audit and feedback.

Both missing race and ethnicity decreased by 94% over the course of the project, (from race missing in 29% of patients down to 2% missing and ethnicity missing in 31% down to 2%). Rounds 1 and 2 of audit and feedback cycles showed the largest reductions in missing race and ethnicity data, as shown in Figure 1. There was a 45% decrease in missing race after round 1. An additional 39% of missing race was completed with round 2 and 10% in round 3. There was a 46% decrease in missing ethnicity after round 1, a 33% decrease after round 2 and a 14% decrease after round 3. One center did not perform data completion during round 1 attributed to insufficient time to complete the task.

Figure 1: Percent Change in Race and Ethnicity by Round of Audit and Feedback

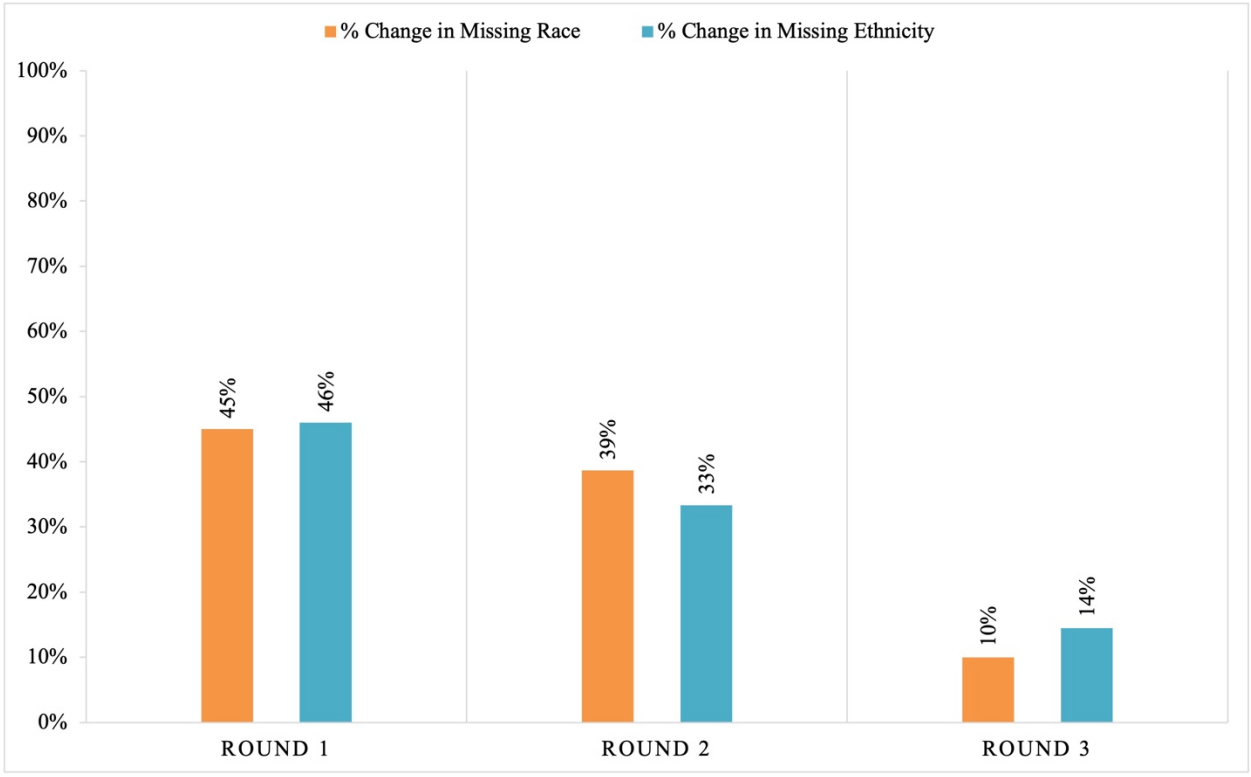


Figure 1 depicts the percent decrease in missing race and ethnicity across each round of data completion.

Figure 2 shows the distribution of race and ethnicity as a percent of total patients, comparing before and after completion. The population distribution of race and ethnicity was consistent across all time points.

Figure 2: Population Distribution of Race and Ethnicity Before and After Data Completion

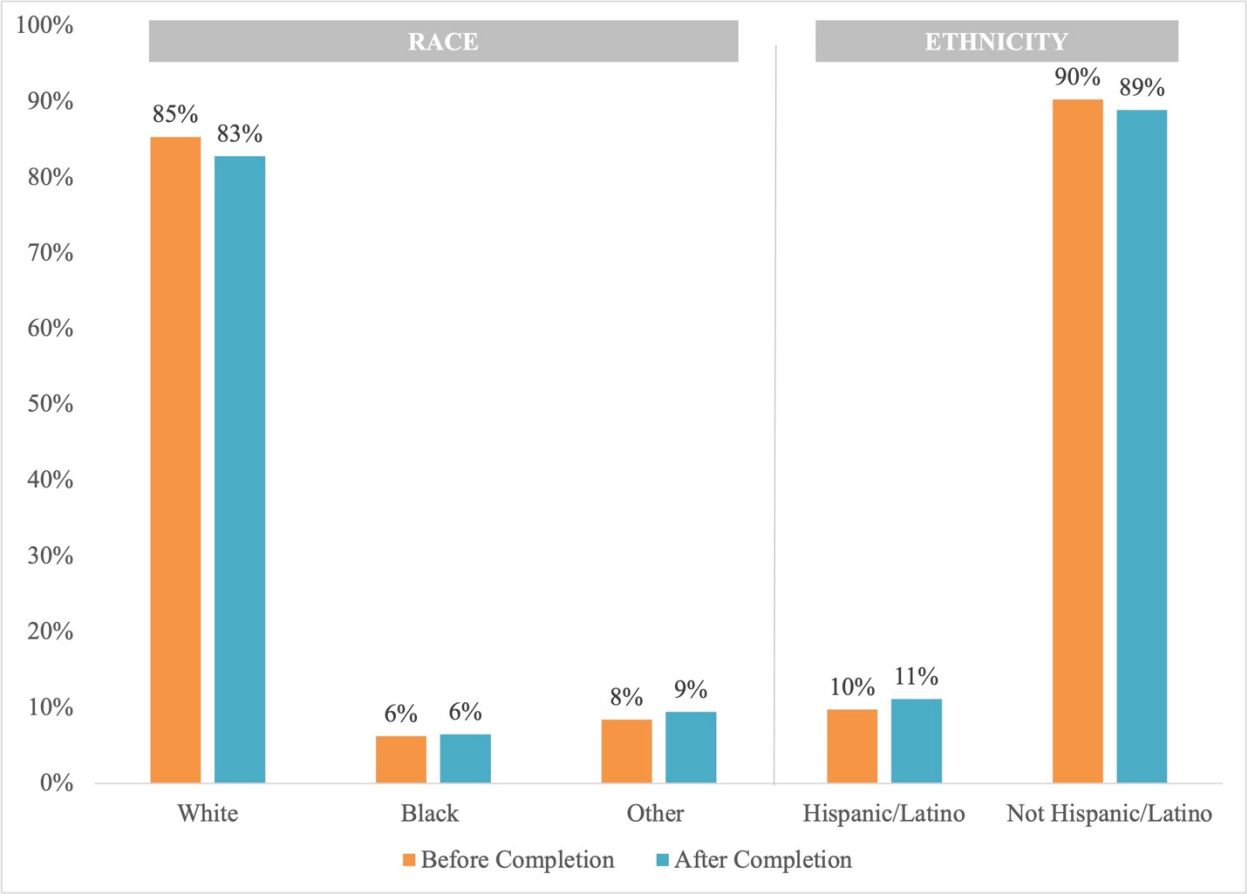
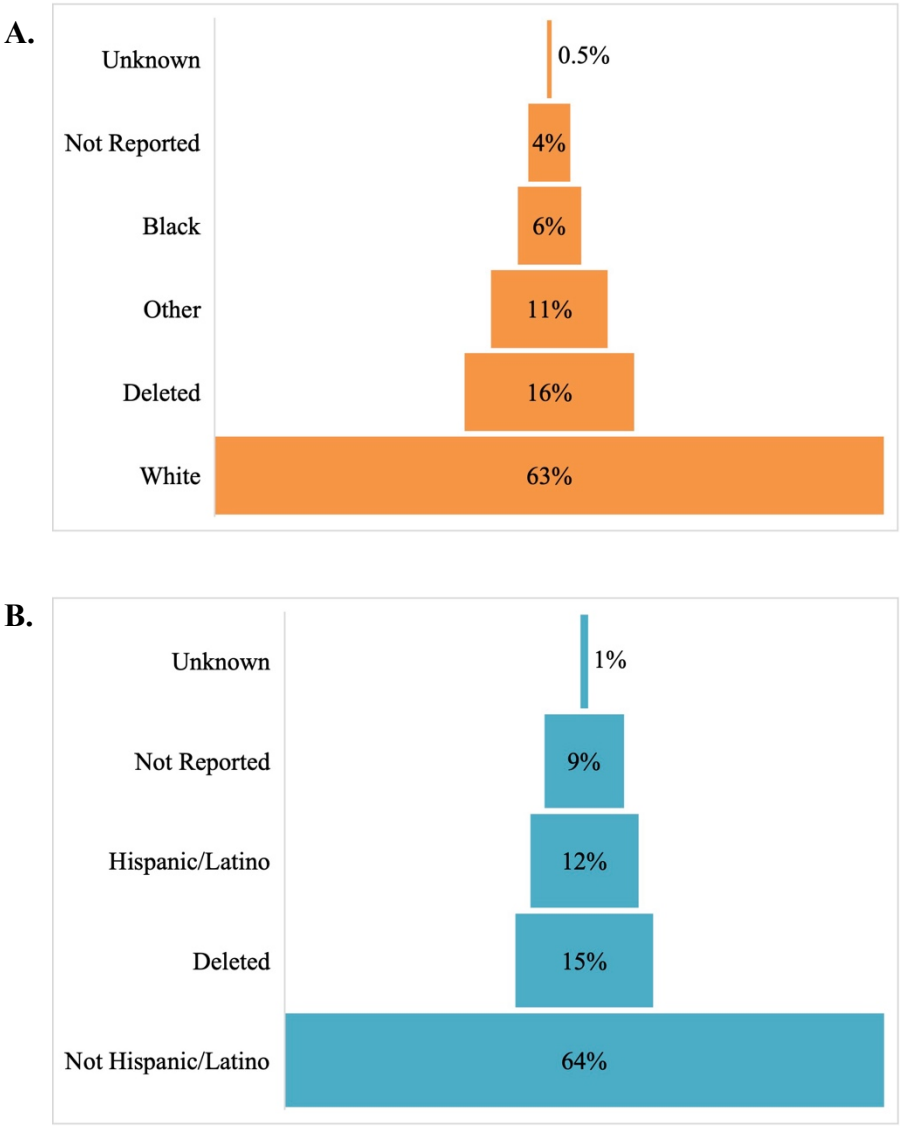


Figure 2 represents the distribution of race and ethnicity before completion and after completion as a percent of the total population of patients.

Distribution of recovered race and ethnicity data is depicted by Figure 3. Recovered data was primarily White and Not Hispanic/Latino. 'Deleted' represents patient entries that were identified as duplicate and deleted during the first round of data completion.

Figure 3: Distribution of Missing Data by Race (Figure A) and Ethnicity (Figure B)

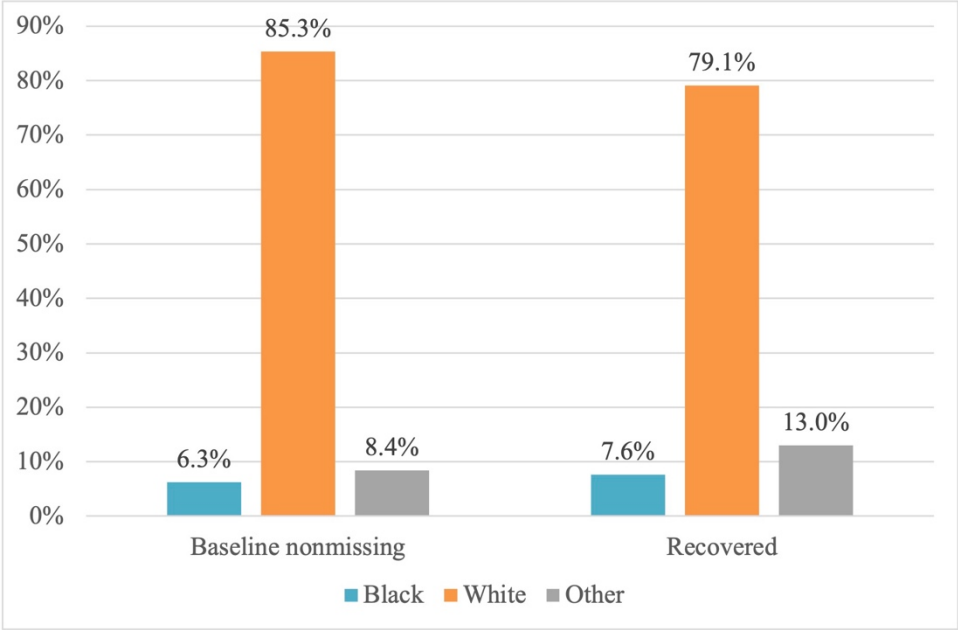


Graphical representation of the distribution of race (A) and ethnicity (B) in patients with completed data represented as a percent of total patients with missing data at baseline.

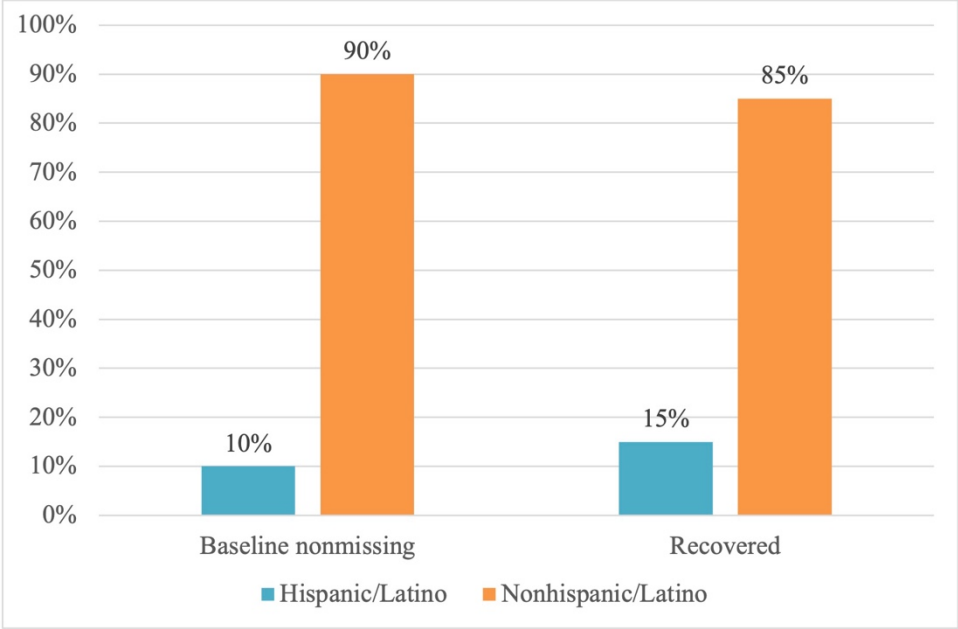
Of those with race data that was recovered during the three rounds of audit and feedback, 63% were identified as White, 6% were identified as Black, and 11% were identified as Other (Figure 3A). Sixteen percent of patients were found to be duplicate entries and deleted. For patients with ethnicity data missing at baseline that was completed during the study, 64% were identified as Not Hispanic/Latino and 12% were identified as Hispanic/Latino (Figure 3B).

Figure 4: Distribution of Patients with Non-Missing Data at Baseline vs. Recovered Patients

A.



B.



Race (A): Baseline nonmissing is the distribution of race for patients whose race was present in the dataset before completion. Recovered represents the distribution of race for patients whose race was recovered and input into the registry during data completion, expressed as a percent of total patients with recovered ethnicity. Ethnicity (B): Baseline nonmissing is the distribution of ethnicity in patients whose ethnicity was present in the dataset before completion. Recovered is the distribution of ethnicity for patients whose ethnicity was recovered and input into the registry during data completion, expressed as a percent of total patients with recovered ethnicity.

Figure 4 shows the distribution of race and ethnicity in patients as a percent of total patients with non-missing values at baseline and is compared to the race and ethnicity distribution in patients as a percent of total patients with recovered race and/or ethnicity. Race designated as Other was 55% higher in patients with missing race at baseline that was subsequently recovered (13%), compared to patients with non-missing race at baseline (8.4%) (Figure 4A). Hispanic ethnicity was 50% higher in patients with missing ethnicity at baseline that was subsequently recovered (15%), compared to patients with non-missing ethnicity at baseline (10%) (Figure 4B).

Table 3 shows the change in missing data by Center. Centers A-C and E had a completion rate of 98% or higher for race. Center F was able to complete two-thirds of their missing race. Center D decreased missing race by 33%, decreasing patients missing race from 3 to 2 patients. Centers B-E completed 100% of those missing ethnicity. Center A decreased missing ethnicity by 89% and center F decreased by 66%. Of note, site C was missing 99% of race and ethnicity before completion and was also the only center uploading data to the registry via EDT.

Table 3: Missing Data by Center

Centers	**A**	**B**	**C**	**D**	**E**	**F**
Missing Race						
Before completion	47 (24%)	171 (37%)	248 (99%)	3 (0.5%)	160 (38%)	54 (13%)
After completion	1 (1%)	2 (1%)	18 (7%)	2 (0.3%)	0 (0%)	18 (4%)
Percent recovered	98%	99%	93%	33%	100%	67%
Missing Ethnicity						
Before completion	70 (36%)	173 (38%)	248 (99%)	4 (0.6%)	166 (39%)	71 (18%)
After completion	8 (4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	24 (6%)
Percent recovered	89%	100%	100%	100%	100%	66%

4.4 Assessment of Impact on Outcome Measures

4.4.1 Unknown cJADAS10

CJADAS10 from first registry follow-up 2 to 6 months after enrollment was obtained. The distribution of cJADAS10 ≥5, cJADAS10 ≤5, and unknown cJADAS10 before and after data completion is shown in Table 4 for race and in Table 5 for ethnicity. Before completion, 50% (341/683) of patients with Missing race and 47% (341/732) with Missing ethnicity had unknown cJADAS10. Meanwhile, cJADAS10 was unknown for 16% (17/105) of Black patients, 20% (28/141) of patients with Other race, and 12% (167/1430) White patients. For ethnicity before completion, cJADAS10 was unknown in 16% (25/159) of Hispanic/Latino patients and 13% (187/1468) of Not Hispanic/Latino patients.

Table 4: cJADAS10 Distribution Among Race Before and After Completion

	White	Black	Other	Missing Race
Before completion				
cJADAS10 ≥5	438 (30%)	43 (41%)	41 (29%)	97 (14%)
cJADAS10 <5	825 (58%)	45 (43%)	72 (51%)	245 (36%)
Unknown cJADAS10	167 (12%)	17 (16%)	28 (20%)	341 (50%)
After completion				
cJADAS10 ≥5	494 (27%)	49 (34%)	70 (30%)	22 (37%)
cJADAS10 <5	999 (54%)	60 (42%)	112 (48%)	16 (26%)
Unknown cJADAS10	341 (19%)	35 (24%)	53 (22%)	22 (37%)

cJADAS10: clinical juvenile disease activity score

Unknown cJADAS10 was seen more frequently in those with Missing race with 50% unknown cJADAS10 before completion and 49% unknown cJADAS10 after completion. Unknown cJADAS10 in those with Missing ethnicity increased from 47% to 65% from before

completion to after completion. When race and ethnicity were known, unknown cJADAS10 ranged from 12-20% before completion and from 19-25% after completion.

Table 5: cJADAS10 Distribution Among Ethnicity Before and After Completion

	Not Hispanic/Latino	Hispanic/Latino	Missing Ethnicity
Before completion			
cJADAS10 ≥5	459 (31%)	48 (30%)	112 (15%)
cJADAS10 <5	822 (56%)	86 (54%)	279 (38%)
Unknown cJADAS10	187 (13%)	25 (16%)	341 (47%)
After completion			
cJADAS10 ≥5	528 (28%)	67 (28%)	4 (14%)
cJADAS10 <5	1034 (54%)	115 (48%)	5 (21%)
Unknown cJADAS10	348 (18%)	5 (24%)	23 (65%)

cJADAS10: clinical juvenile disease activity score

4.4.2 Comparing cJADAS10 Before and After Completion

Tables 4 and 5 also show cJADAS10 ≥5 for race and ethnicity before and after data completion. Before completion, cJADAS10 was ≥5 for 31% (438/1430) of White patients and 41% (43/105) of Black patients and 29% (41/141) of patients with Other race. CJADAS10 was ≥5 for 14% (97/683) of patients with Missing race and 15% (112/732) of patients with Missing ethnicity. For ethnicity before completion, 30% (48/159) of Hispanic/Latino and 31% (459/1468) of Not Hispanic/Latino patients had cJADAS10 ≥5.

After completion (time 3), cJADAS10 was ≥5 in 27% (494/1834) of White patients, 28% (59/206) of Other patients, and 34% (49/144) of Black patients. CJADAS10 was ≥5 in 28% (67/239) Hispanic/Latino patients and 28% (528/1910) Not Hispanic/Latino patients. The proportion of cJADAS10 ≥5 was decreased in all races and ethnicities after completion.

Patients with Missing race had the lowest frequency of cJADAS10 ≥5, present in 14% of patients before completion and 15% after completion. Findings were similar for those with

Missing ethnicity, cJADAS10 ≥5 was seen in 15% before completion and 14% of patients after completion. In patients with known race and ethnicity 29-41% had cJADAS10 ≥5 before completion and 27-34% had cJADAS10 ≥5 after completion.

4.4.3 Odds of cJADAS10 ≥5

Table 6 shows the adjusted OR of cJADAS10 ≥5 at first registry follow up for race and ethnicity comparing results before and after completion. Adjusted odds ratios control for patient age, gender, race, and ethnicity. Before data completion, the odds of cJADAS10 ≥5 was noted to be significantly higher for Black patients compared to White patients with odds increased by 76% (p=0.011). The odds of cJADAS10 ≥5 for patients of Other races (OR=1.12, p=0.596) or those with Missing race (OR=0.97, p=0.916) were not significantly different compared to White patients. The odds of cJADAS10 ≥5 at first registry follow up for Hispanic/Latino patients or Missing ethnicity were not statistically different than the odds for Not Hispanic/Latino patients.

Table 6: Odds Ratio of cJADAS10 ≥5 for Race and Ethnicity Before and After Data Completion

Odds of cJADAS10† ≥5 Before Completion (N=1806)			**Odds of cJADAS10† ≥5 After Completion (N=1806)**		
Predictors	*Odds Ratios*	p	*Predictors*	*Odds Ratios*	*p*
Ethnicity			**Ethnicity**		
Not Hispanic/Latino	*Reference*		Not Hispanic/Latino	*Reference*	
Hispanic/Latino	0.99	0.972	Hispanic/Latino	1.11	0.554
Missing	0.82	0.431	Missing	1.02	0.939
Race			**Race**		
White	*Reference*		White	*Reference*	
Black	1.76	**0.011**	Black	1.61	**0.019**
Other	1.12	0.596	Other	1.19	0.347
Missing	0.97	0.916	Missing	1.39	0.352

† cJADAS10 is defined as cJADAS10 score at the first registry follow-up visit (2 to 6 months after enrollment)

After data completion, controlling for patient age, gender, race and ethnicity, the odds of cJADAS10 ≥5 was significantly higher with a 61% (p=0.019) increase for Black patients compared to White patients. The odds of cJADAS10 ≥5 for patients of Other races (OR=1.19, p=0.347) or Missing race (OR=1.39, 0.352) were not significantly different than the odds of cJADAS10 ≥5 for White patients. For ethnicity after completion, the odds of cJADAS10 ≥5 at first registry follow up for Hispanic/Latino patients or patients Missing ethnicity were not statistically different than the odds for Not Hispanic/Latino patients.

The estimated odds ratio for cJADAS10 ≥5 at first registry follow up (2 to 6 months after enrollment) was higher for Black patients before completion compared to after completion. After completion the OR of cJADAS ≥5 decreased from 1.76 to 1.61, a relative decrease of 8.5%. Odds of cJADAS10 ≥5 was not statistically significant when comparing White to patients with Other or Missing race after data completion. Estimated OR of cJADAS10 ≥5 for Hispanic/Latino patients changed from 0.99 to 1.11, after data completion, a 12% relative increase. However, there was no statistically significant difference in the odds of cJADAS10 ≥5 for Hispanic/Latino patients when compared to Not Hispanic/Latino patients.

4.5 Interviews Analysis

Initial coding performed by KMB based on interview notes. After initial coding, both reviewers (KMB and EMM) established themes and resolved discrepancies via discussion to establish the final emergent themes. Three themes emerged from inductive thematic analysis of the post completion interview sessions including: project experience, variation in reporting and data collection, and defining data processes. I also gathered participant recommendations with regards to improving data collection moving forward.

4.5.1 Project Experience

For project experience, participants noted that data completion process was manageable and sustainable. Use of an audit report was noted to be helpful in identifying and completing missing race and ethnicity data. Most sites completed registry data via demographics data present within the EHR entered during the clinic registration process. Three sites reported that portions of missing data were not able to be identified within the EHR. Duplicate data was identified in one site resulting in working with registry platform for resolution. Another site worked with the registry platform manager, to troubleshot EDT and data migration issues. One center initiated a site-specific quality improvement project to educate staff on appropriate collection and self-reporting of race and ethnicity data.

4.5.2 Variation in Reporting and Data Collection

Multiple sites noted confusion and inconsistent documentation practices around 'Unknown' versus 'Not Reported' as options and appreciated education around this distinction, recommending adjustment of these terms within the registry. One site noted that many marked as 'Not Reported' had data present within the EHR. Meanwhile, another site hypothesized that their large number of 'Unknowns' may reflect a lack of options with which a patient identified. The separation of Hispanic/Latino ethnicity from racial group is also noted as an area of confusion for some patients. One site also documents Hispanic/Latino as race which can result in difficulty with data reconciliation as the patient may not identify a race category separate from their ethnicity. Multiracial is also a source of difficulty for data mapping, multiple sites have multiracial as a single select option. PR-COIN allows for multiselect to document two or more races but does not have a multiracial, single select option. Sites also noted ongoing changes in

their data collection practices including processes and options which result in ongoing challenges for data mapping and upload.

4.5.3 Defining Data Processes

Many sites commented on lack of understanding or transparency of institutional race and ethnicity data collection practices. Multiple sites used this project as a starting point for improving overall registry data entry, staff education, as well as understanding and improving data collection practices at the institution level. The site uploading via electronic data transfer identified that race and ethnicity were not part of transfer, resulting in 99% missing race and ethnicity. Strategies for manual verification were suggested including using a site Master List with race and ethnicity to identify those missing data and frequent audit of race and ethnicity for new enrollments.

4.5.4 Participant Recommendations

1. Race and ethnicity should be considered critical data elements.
2. Adjustment of wording for Unknown and Not Reported options to improve consistency with documentation.
3. Develop a tip sheet on best practices for race and ethnicity data collection and entry.
4. Identify which elements are/are not included in electronic data transfer.

5 Discussion

5.1 Summary

Amongst the 6 participating centers, a mean of one-third of race and ethnicity data was missing within the PR-COIN registry, with substantial variability across centers. This mean number is consistent with previous reports of missing race and ethnicity data in other databases.[14,15,21] When considering use of patient registry data for disparities research or equity related quality improvement, complete and accurate data is important to prevent exclusion of these patients in analysis due to missing data. This project has demonstrated that race and ethnicity data quality can be improved through manual completion from the EHR where most of the missing data can be found. In this scenario, data can be improved via audit and feedback cycles through EHR data which may ultimately lead to improved completion of race and ethnicity data. Future, registry-wide data completion efforts could reasonably be completed in 1-2 rounds given signs of diminishing returns for this cohort after the second round of completion. Alternatively, data improvements could be accomplished via the Master List by adding race and ethnicity data to create a self-reporting mechanism to maintain data completion.

Previous reports have suggested that missing data is often disproportionately Black and Hispanic/Latino.[13,14] As a result of this data completion effort, there are now over 600 patients with completed race and/or ethnicity data that will be included in future disparities assessments. I found higher proportions of Hispanic/Latino ethnicity and Other races in recovered data compared to the baseline population of patients with non-missing race or ethnicity. However, the population distribution remained stable. While other studies have identified new or worsened disparities with completion of race and ethnicity data, I found no difference in the odds ratio of

having a cJADAS10 ≥5 at first registry follow up after data completion. However, 50% of patients with missing data were also missing cJADAS10.

This project has informed improvements and best practice recommendations for the registry moving forward. Multiple centers have embarked on formal or informal education and quality improvement initiatives to understand and optimize data collection into the EHR and entry into the registry. These are the first steps to determine data accuracy which must be validated and improved at each institution. I identified that the center entering registry data via EDT was missing 98% of race and ethnicity due to data mapping and transfer issues. Mapping issues also exist for sites with manual entry due to discordance between registry options and options for race and ethnicity. Specifically, Hispanic/Latino and multiple races, via multiselect or single select options, are noted to increase difficulties with data reconciliation which can compromise data accuracy.

5.2 Limitations and Future directions

The ability to detect changes in cJADAS10 after data completion may have been limited bt the small sample size as well as the high degree of missing cJADAS10 values in patients with missing race and ethnicity at baseline. The slightly skewed distribution of recovered data towards Hispanic/Lation ethnicity and Other races, suggests that additional data completion at a larger scale may reveal changes in the population distribution. However, given the concordance between missing race and ethnicity and other missing data elements such as cJADAS10 and its components, missing race and ethnicity data may identify patients with larger data quality problems. It is also possible that, due to this missing data, I could still be missing small changes in disparities assessments for cJADAS10. Although there was not an identified impact on our

outcome assessment before and after data completion, the completion of this data remains an important priority.

Future work within the registry may need to include mapping for EDT of race and ethnicity to improve completion, given the presence of data within the EHR. There is ongoing work for standardization and implementation of race and ethnicity along with other social determinants of health which may provide helpful guidance for data mapping in the future.[22] Moving forward, I recommend that race and ethnicity be included as critical data elements to prioritize input during registration and provide ongoing data quality feedback.

As of March 2024, the Office of Management and Budget (OMB) standards has published new recommendations for race and ethnicity data with two major changes: (1) Hispanic/Latino will now be part of race with no ethnicity category. (2) There will be an additional minimum racial category of Middle Eastern or North African which may similarly provide mapping and data challenges across different sites as these new recommendations are implemented across different institutions.[23] This has implications that registries may need to consider on future data capture, especially if health systems update their collection of this data into the EHR to reflect these changes.

When using a registry or learning health system to monitor and address disparities, having complete race and ethnicity data is extremely important for accurate assessments. Prior to data completion, disparities assessments would have excluded almost 1/3 of patients due to missing data. Thus, learning health systems with missing race and ethnicity data are at risk of widening disparities through exclusion from research and inaccurate assessment of disparities. Addressing race and ethnicity data quality should be a component of equity work within learning health systems. This project provides a baseline assessment of missing data and outlines a data

completion process which can be applied to all sites and new disease additions to the registry moving forward.

5.3 Key Findings

- Missing race and ethnicity data can be improved through audit and feedback reports and is largely present within the electronic health record.
- Sites using EDT should be aware that race and ethnicity is not included and create a workflow for inputting this data in the registry.
- In this small sample, completion of race and ethnicity did not affect the chosen outcome measure.
- Recovered data was not uniform and was more likely to contain Hispanic/Latino ethnicity or Other races.

6 Reference

1. Enticott J, Johnson A, Teede H. Learning health systems using data to drive healthcare improvement and impact: a systematic review. *BMC Health Services Research*. 2021;21(1):200. doi:10.1186/s12913-021-06215-8

2. Foley T, Horwitz L, Zahran R. LEARNING HEALTH SYSTEMS. :101.

3. Sarwar T, Seifollahi S, Chan J, et al. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. *ACM Comput Surv*. 2022;55(2):33:1-33:40. doi:10.1145/3490234

4. Feder SL. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. *West J Nurs Res*. 2018;40(5):753-766. doi:10.1177/0193945916689084

5. Vega Perez RD, Hayden L, Mesa J, et al. Improving Patient Race and Ethnicity Data Capture to Address Health Disparities: A Case Study From a Large Urban Health System. *Cureus*. 2022;14(1):e20973. doi:10.7759/cureus.20973

6. Sohn MW, Zhang H, Arnold N, et al. Transition to the new race/ethnicity data collection standards in the Department of Veterans Affairs. *Popul Health Metrics*. 2006;4(1):7. doi:10.1186/1478-7954-4-7

7. Jarrín OF, Nyandege AN, Grafova IB, Dong X, Lin H. Validity of Race and Ethnicity Codes in Medicare Administrative Data Compared With Gold-standard Self-reported Race Collected During Routine Home Health Care Visits. *Medical Care*. 2020;58(1):e1. doi:10.1097/MLR.0000000000001216

8. George S, Duran N, Norris K. A Systematic Review of Barriers and Facilitators to Minority Research Participation Among African Americans, Latinos, Asian Americans, and Pacific Islanders. *Am J Public Health*. 2014;104(2):e16-e31. doi:10.2105/AJPH.2013.301706

9. Hasson Charles RM, Sosa E, Patel M, Erhunmwunsee L. Health Disparities in Recruitment and Enrollment in Research. *Thoracic Surgery Clinics*. 2022;32(1):75-82. doi:10.1016/j.thorsurg.2021.09.012

10. Bailey ZD, Feldman JM, Bassett MT. How Structural Racism Works — Racist Policies as a Root Cause of U.S. Racial Health Inequities. *New England Journal of Medicine*. 2021;384(8):768-773. doi:10.1056/NEJMms2025396

11. Explanation of the Standards - The Office of Minority Health. Accessed October 30, 2022. https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=3&lvlid=54

12. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. Conceptualizing, Contextualizing, and Operationalizing Race in Quantitative Health Sciences Research. *Ann Fam Med*. 2022;20(2):157-163. doi:10.1370/afm.2792

13. Branham DK, Finegold K, Chen L, et al. Trends in Missing Race and Ethnicity Information After Imputation in HealthCare.gov Marketplace Enrollment Data, 2015-2021. *JAMA Network Open*. 2022;5(6):e2216715. doi:10.1001/jamanetworkopen.2022.16715

14. 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*. 2021;32(2):157-161. doi:10.1097/EDE.0000000000001314

15. Yee K, Hoopes M, Giebultowicz S, Elliott MN, McConnell KJ. Implications of missingness in self-reported data for estimating racial and ethnic disparities in Medicaid quality measures. *Health Services Research*. 2022;57(6):1370-1378. doi:10.1111/1475-6773.14025

16. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342

17. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. *J Am Med Inform Assoc*. 2018;25(8):1080-1088. doi:10.1093/jamia/ocy052

18. Bingham CA, Harris JG, Qiu T, et al. Pediatric Rheumatology Care and Outcomes Improvement Network's Quality Measure Set to Improve Care of Children With Juvenile Idiopathic Arthritis. *Arthritis Care & Research*. 2023;75(12):2442-2452. doi:10.1002/acr.25168

19. Consolaro A, Negro G, Chiara Gallo M, et al. Defining Criteria for Disease Activity States in Nonsystemic Juvenile Idiopathic Arthritis Based on a Three-Variable Juvenile Arthritis Disease Activity Score. *Arthritis Care & Research*. 2014;66(11):1703-1709. doi:10.1002/acr.22393

20. Braun V, Clarke V. Using thematic analysis in psychology: Qualitative Research in Psychology. *Qualitative Research in Psychology*. 2006;3(2):77-101. doi:10.1191/1478088706qp063oa

21. 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(8-9):730-736. doi:10.1093/jamia/ocz113

22. Gravity Project. Gravity Project. Accessed April 10, 2024. https://thegravityproject.net/

23. Revisions to OMB's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. Federal Register. Published March 29, 2024. Accessed April 10, 2024. https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and