THERAPEUTIC ADVANCES in
Rare Disease

Original Research

Toward representative genomic research: the children's rare disease cohorts experience

Ther Adv Rare Dis

2023, Vol. 4: 1–12

DOI: 10.1177/26330040231181406

Zoë J. Frazier, Eurnestine Brown, Shira Rockowitz, Ted Lee, Bo Zhang, Abigail Sveden, Nancy L. Chamberlin, Kira A. Dies, Annapurna Poduri, Piotr Sliz and Maya Chopra on behalf of the CRDC Consortium

Abstract

Background: Due to racial, cultural, and linguistic marginalization, some populations experience disproportionate barriers to genetic testing in both clinical and research settings. It is difficult to track such disparities due to non-inclusive self-reported race and ethnicity categories within the electronic health record (EHR). Inclusion and access for all populations is critical to achieve health equity and to capture the full spectrum of rare genetic disease.

Objective: We aimed to create revised race and ethnicity categories. Additionally, we identified racial and ethnic under-representation amongst three cohorts: (1) the general Boston Children's Hospital patient population (general BCH), (2) the BCH patient population that underwent clinical genomic testing (clinical sequencing), and (3) Children's Rare Disease Cohort (CRDC) research initiative participants.

Design and Methods: Race and ethnicity data were collected from the EHRs of the general BCH, clinical sequencing, and CRDC cohorts. We constructed a single comprehensive set of race and ethnicity categories. EHR-based race and ethnicity variables were mapped within each cohort to the revised categories. Then, the numbers of patients within each revised race and ethnicity category were compared across cohorts.

Results: There was a significantly lower percentage of Black or African American/African, non-Hispanic/non-Latine individuals in the CRDC cohort compared with the general BCH cohort, but there was no statistically significant difference between the CRDC and the clinical sequencing cohorts. There was a significantly lower percentage of multi-racial, Hispanic/Latine individuals in the CRDC cohort than the clinical sequencing cohort. White, non-Hispanic/non-Latine individuals were over-represented in the CRDC compared to the two other groups.

Conclusion: We highlight underrepresentation of certain racial and ethnic populations in sequencing cohorts compared to the general hospital population. We propose a range of measures to address these disparities, to strive for equitable future precision medicine-based clinical care and for the benefit of the whole rare disease community.

Plain language summary

Racial and ethnic representation amongst general clinics, clinics that provide genetic testing, and genomic-based research at Boston Children's Hospital

Background: Individuals who identify as belonging to a race or ethnicity that has been historically excluded from mainstream cultural, political, and economic activities ('historically marginalized') experience barriers to clinical care. These barriers are

Correspondence to:
Maya Chopra
Rosamund Stone Zander Translational Neuroscience Center, Boston Children's Hospital, Harvard Medical School, Boston, 2 Brookline Place, 7th Floor, Brookline, Boston, MA 02445, USA

Department of Neurology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA.
Maya.Chopra@childrens.harvard.edu

Zoë J. Frazier
Rosamund Stone Zander Translational Neuroscience Center, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Eurnestine Brown
Rosamund Stone Zander Translational Neuroscience Center, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Brazelton Touchpoints Center, Division of Developmental Medicine, Boston Children's Hospital, Boston, MA, USA

Shira Rockowitz
Division of Genetics and Genomics, Research Computing, Information Technology, The Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA, USA

Ted Lee
Department of Urology, Boston Children's Hospital, Boston, MA, USA

Bo Zhang
Rosamund Stone Zander Translational Neuroscience Center, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Biostatistics and Research Design Center, Institutional Centers for Clinical & Translational Research, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Abigail Sveden
Nancy L. Chamberlin
Kira A. Dies
Rosamund Stone Zander Translational Neuroscience Center, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Annapurna Poduri
Department of Neurology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Piotr Sliz
Division of Molecular Medicine, Research Computing, Information Technology, The Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA, USA

CRDC Consortium
Boston Children's Hospital, Boston, MA, USA

further complicated for families touched by rare genetic conditions. Obstacles can present as accessibility issues (transportation, financial, linguistic), low-quality medical care, or inadequate inclusion in research. It is important to have representation within rare disease research so that the full scope of these conditions is understood, leading to better patient care for all, and for health equity.

Objective: We aimed to (1) to create new and inclusive race and ethnicity categories for the electronic health record (EHR) and (2) identify differences in racial and ethnic representation amongst patients generally seen at Boston Children's Hospital (general BCH), those who received genetic testing in a clinic at Boston Children's Hospital (clinical sequencing), and participants who enrolled in the CRDC research project at Boston Children's Hospital (CRDC).

Design and Methods: We combined race and ethnicity categories to make more inclusive options than existing EHR categories. Differences in race and ethnicity representation were observed when looking at the three different patient groups (general BCH, clinical sequencing, and CRDC).

Results: We observed a lower percentage of individuals who self-identify as Black or African American/African, non-Hispanic/non-Latine in the genetic testing groups (both research and clinical) than in the general BCH group. Individuals who self-identify as multi-racial, Hispanic/Latine are also under-represented in the CRDC research compared to the two other groups. The highest population percentage seen in all groups was that of patients who identify as White, non-Hispanic/non-Latine. This group was over-represented in the research CRDC group compared to the two others.

Conclusion: Our study found that patients who are historically marginalized are underrepresented in clinical genetic testing and genomic research studies compared to their White counterparts. In order to benefit all patients with rare genetic conditions, these differences must be addressed by improving access to specialty physicians/researchers and incorporating inclusive language in the EHR, clinics, and research protocols.

Keywords: electronic medical record, ethnicity, genetic testing, genomic research, health equity, race, racial disparity, representation

Received: 6 February 2023; revised manuscript accepted: 23 May 2023.

Introduction

There are approximately 7000 known rare disorders, 71% of which are believed to have a genetic etiology.[1,2] Affected individuals and their caregivers often undergo long diagnostic odysseys characterized by a series of specialist referrals, years of testing, emotional turmoil, and non-specific or suboptimal care.[3,4] While whole exome sequencing (WES) and whole genome sequencing (WGS) have a high diagnostic yield for children with rare disorders, these technologies are unattainable for many in the United States due to inequities based on location, high cost, and insurance barriers.[5–7]

For families and communities who are historically under-resourced and experience racial, cultural, and linguistic marginalization, navigating the rare disease diagnostic odyssey can be especially challenging, with barriers at each step of the path to a genetic diagnosis.[8] The cost of clinical care and other barriers to genetics services disproportionately impacts patients who self-report as belonging to historically marginalized racial and ethnic backgrounds, which include, but are not limited to African American, Asian, Hispanic, Latino/Latina, and Native American.[9–11] Such communities may additionally face a lack of provider recognition of the likelihood of an underlying genetic diagnosis, and lack of culturally and linguistically responsive and respectful care.[12] Structural racism and racial biases held by providers often form the basis of inadequate medical care and delayed or missed diagnosis of pediatric genetic conditions.[13–15] These, and other issues that limit clinical and research opportunities for a subset of the rare disease population, translate

into limited racial and ethnic representation in rare disease genomic data. This in turn leads to lost opportunities for children of the global majority (Black, Brown, multi-racial, indigenous to the southern hemisphere peoples labeled as 'ethnic minorities' who comprise the majority percentage of the world population)[16] to experience health equity.

In addition to inherent inequities, genomic data that lacks diverse representation is to the detriment of all individuals with rare genetic diseases. Such data fails to capture the full spectrum of psychosocial and comorbid influences on the natural history and phenotypic spectrum of rare disorders.[15,17] To the extent that certain ancestries are underrepresented, a subset of genetic variants may not be captured. An important component of genetic variant assessment involves analysis of non-disease/unaffected populations (e.g. gnomAD).[18] The lack of diverse ancestries in these resources means that underrepresented ancestries are less likely to achieve diagnostic certainty even when testing is done. Further, lack of representative data limits the relevance and the pace of translational efforts to develop comprehensive clinical care guidelines, prognoses, and possibly novel therapies. In this light, efforts to actively promote the inclusion of all racial and ethnic populations in rare disease genomic research benefits not only prospective participants and their families, but also the rare disease community in general. Unfortunately, there is a paucity of data on the race and ethnicity representation in pediatric rare disease genomic research populations. To shed light on these issues, and in an effort to encourage other rare disease research groups to do the same, we undertook an exploratory analysis of demographics of clinical and research sequencing cohorts at Boston Children's Hospital (BCH).

The mission of the Children's Rare Disease Cohort (CRDC) initiative at BCH is to better understand the association between genotype and phenotype through genomic-based research studies for a variety of rare diseases.[19] At the time of writing, 45 research cohorts focused on rare pediatric-onset Mendelian conditions are participating in the CRDC, including Unexplained Epilepsies, Congenital Sensorineural Hearing Loss, and Very Early Onset Inflammatory Bowel Disease. Together, at the time of writing, more than 5242 probands are enrolled in the CRDC initiative cohorts. Through this initiative, WES, WGS, and transcriptomic data are analyzed through the CRDC Investigator Tools, providing researchers and clinicians the opportunity for novel gene discovery, phenotype expansion, and rare variant classification. In addition to advancing rare disease research, results from the CRDC may impact clinical care by facilitating genetic diagnoses for participants.[19] Although the CRDC is a mechanism by which individuals with rare disorder presentations may access genomic sequencing free of cost, many families who meet inclusion criteria may not participate. Reasons may include unsuccessful recruitment strategies, lack of patient-centric benefits at the study site, or failure to access prerequisite specialty clinics.[20]

There are also no standardized methods to categorize race and ethnicity, let alone define what constitutes fair or adequate representation. We did not find any other publications integrating self-reported race and ethnicity in pediatric rare disease genomic research. In many genomics research studies, analyses include ancestry groupings deduced from genomic datasets, but estimated genetic ancestry does not correlate with self-perceived race and ethnicity, although there may be some overlap in these concepts. We analyzed the electronic health record (EHR) at BCH for self-reported racial and ethnic demographics to calculate representation among three populations: (1) participants enrolled in the CRDC research initiative, (2) patients who underwent whole exome or whole genome genomic testing for clinical diagnostic purposes at BCH, and (3) all patients seen at BCH. We hypothesized differences in self-reported racial and ethnic representation between the three groups. We present our approach categorizing race and ethnicity in this study, highlighting the importance of language. We propose possible reasons that may explain the disparities we identified between the groups and suggest solutions to increase research inclusion, belonging, and accessibility for all.

Methods

This was a retrospective analysis of de-identified, aggregate data where we (1) revised race and ethnicity categories and (2) used the revised categories to examine representation between the general BCH patient population, BCH patients

who received clinical genetic testing, and participants who enrolled in the CRDC. The analysis in this report was completed as part of a quality improvement initiative for the capture of ethnicity and race categories from patient cohorts at BCH. No identifiable or personal health information was collected from the EHR.

Creation of race and ethnicity categories: language matters

The BCH EHR records race and ethnicity data from two multi-select picklist fields. Details are in the Supplemental section titled 'Race and Ethnicity Categories in the BCH EHR'. Briefly, there are nine multi-select options under the 'Race' category, including 'other', 'unable to answer' and 'unknown'. In the ethnicity category, there are 149 multi-select options encompassing many countries and ethnic populations, but excluding others. For example, although 23 Asian ethnicities were listed, Hmong, Uyghur and Mongolian were among the many notable ethnicity options not included. There was better representation across ethnic groups in Europe (Spanish Basque, Castilian, Catalonian, and Andalusian were all included) than Africa (where only 8/54 countries were listed) and the Middle East (4/22 countries listed). We also noted that many individuals may have not seen an option that aligns with their identity, so they may have selected 'other' or 'not reported'. We observed this phenomenon particularly in the Latine population and in the Middle Eastern/North African populations, where neither was listed as an option for race in the current BCH EHR configuration, but an aligned country was selected under ethnicity. In addition to limited inclusivity of the categories, we also identified the language as being potentially problematic (e.g. use of 'other' may lead to individuals feeling like they don't belong). In view of these issues, it was decided that for our analysis, a revision of these categories was required.

A review the literature failed to reveal any widely accepted categorizations of race and ethnicity.[17] While the NIH style guide outlines proposed categories,[21] these groups may not reflect the identities of many populations. For this reason, we performed an intentional recategorization of BCH EHR data with a team of CRDC investigators, data scientists, genetic counselors, and a research scholar with expertise in equity, diversity and inclusion. Our objective was to be intentional in devising inclusive categories, appropriately named as possible, while retaining, as much as possible, the integrity of the raw data.[22]

We constructed a single comprehensive set of race and ethnicity categories based on the combined format (combining race and ethnicity) and a mutually decided set of principles in contemplation of the cited literature.[23,24] We created categories for those who identify as bi-racial or multi-racial, rejecting use of the term 'mixed', which may be considered disrespectful.[25] 'Latine' was utilized in lieu of 'Latino', 'Latina' (which are gender-specific terms), or 'LatinX' (while gender-inclusive, this term is unpronounceable in Spanish).[26,27] We intentionally avoided the terms 'minority', 'non-White', and Black and Indigenous People Of Color (BIPOC) as the meanings are neither specific nor uniformly understood.[28,29] For records where race/ethnicity data was missing, we intentionally did not use genomic ancestry data in an attempt to infer race/ethnicity. While genomic ancestry is grounded in geographical origin, race is a social construct and not necessarily correlated.[30] Our categories are listed in alphabetical order.[25]

Cohorts and statistical analysis

Race and ethnicity data were extracted from the EHRs of three groups: (1) participants enrolled in the CRDC research initiative (CRDC), (2) the BCH patient population that underwent genomic testing for clinical diagnostic purposes (clinical sequencing), and (3) the general BCH patient population (general BCH).

The CRDC cohort was composed of 3627 probands with a BCH medical record number (MRN) enrolled in CRDC initiative protocols between October 2019 and 8 August 2022. The clinical sequencing cohort was composed of the 1791 probands who had undergone clinical exome sequencing at GeneDx Laboratory (Gaithersburg, Maryland) since April 2019 and had data returned to BCH by 20 June 2022 (regardless of indication). There were two probands shared by CRDC cohort and clinical sequencing cohort. The general BCH cohort was composed of 3,067,921 records reflecting all patients encountered at BCH. This includes all individuals with a MRN in the BCH EHR, including those who were seen in satellite and community-based clinics under the BCH health system. Almost all of the data for the general

BCH group corresponded to entries between 1988 and 22 August 2022, with <1% of the records being before that time period corresponding to paper records.

A coding schema was designed (see Supplemental Methodology and Supplemental Table 1) to map race and ethnicity variables from the general BCH cohort to the revised categories. The Supplemental methodology contains the race and ethnicity variables available to patients in the general BCH cohort as well as pseudocode mapping the race and ethnicity variables to the revised categories. Records that did not fit were assigned 'Not aligned with above categories'. For records where no race/ethnicity data were recorded, the group 'Unknown race/ethnicity/declined to report' was assigned. For validation, hundreds of CRDC cohort patient records were manually reviewed to ensure consistency and accuracy, with updates to the coding schema made in an iterative fashion.

The numbers of patients that fit within each revised race and ethnicity category were calculated for each cohort and compared using Z_1 test (a practical test for comparing two proportions with overlapping observations) introduced by Choi and Stablein,[31] to assess the difference within revised racial and ethnic categories for every pair of cohorts. To correct for the multiple hypotheses tested, p-values were adjusted for false discovery rate by the Benjamini-Hochberg procedure. A significant level of 0.05 was pre-specified for the adjusted p-values.[32,33]

Results

Race and ethnicity categories

We developed 12 revised combined race and ethnicity categories that expand on the races and ethnicities published by the National Institution of Health[34] (American Indian or Alaskan Native; Asian; Black or African American/African, Hispanic/Latine; Black or African American/African, non-Hispanic/non-Latine; Middle Eastern/North African; Multi-racial, Hispanic/Latine; Multi-racial, non-Hispanic/non-Latine; Race/ethnicity not aligned with revised categories; Pacific Islander/Native Hawaiian; Race/ethnicity not aligned with revised categories; Unknown race/ethnicity or declined to report; White, Hispanic/Latine; White, non-Hispanic/non-Latine).

Comparison of the three datasets

Descriptive statistics of the CRDC ($n=3627$), clinical sequencing ($n=1791$), and general BCH ($n=3{,}067{,}921$) cohorts by racial and ethnic categories are provided in Table 1. Over half of the general BCH EHR records had unreported race and ethnicity (53.14%) compared to the CRDC and clinical sequencing cohorts (11.94% and 12.73%, respectively).

CRDC versus *general BCH*

The following race/ethnicity categories had lower representation in the CRDC cohort compared to general BCH cohort: Black or African American/African, non-Hispanic/non-Latine (4.60% *versus* 9.72%, $p<0.0001$); Multi-racial, Hispanic/Latine (7.36% *versus* 9.26%, $p=0.0003$); Race/ethnicity not aligned with above categories (4.26% *versus* 6.48%, $p<0.0001$). The following race/ethnicity categories had higher representation in the CRDC cohort compared to general BCH cohort: White, Hispanic/Latine (3.66% *versus* 1.14%, $p<0.0001$); Middle Eastern/North African (1.60% *versus* 0.55%, $p<0.0001$); Multi-racial, non-Hispanic/non-Latine (1.69% *versus* 0.87%, $p<0.0001$); White, non-Hispanic/non-Latine (71.88% *versus* 66.80%, $p<0.0001$).

Clinical sequencing versus *general BCH cohort*

The following race/ethnicity category had lower representation in the clinical sequencing cohort compared to general BCH cohort: Black or African American/African, non-Hispanic/non-Latine (5.82% *versus* 9.72%, $p<0.0001$). The following race/ethnicity categories had higher representation in the clinical sequencing cohort compared to general BCH cohort: White, Hispanic/Latine (2.62% *versus* 1.14%, $p<0.0001$); Multi-racial, Hispanic/Latine (11.52% *versus* 9.26%, $p=0.0054$); Middle Eastern/North African (2.05% *versus* 0.55%, $p<0.0001$).

CRDC versus *clinical sequencing cohort*

The following race/ethnicity category had lower representation in the CRDC cohort compared to clinical sequencing cohort: Multi-racial, Hispanic/Latine (7.36% *versus* 11.52%, $p<0.0001$). The following race/ethnicity category had higher representation in the CRDC cohort compared to clinical sequencing cohort: White, non-Hispanic/non-Latine (71.88% *versus* 65.96%, $p=0.0002$).

Table 1. Comparison of distribution (percentage and count) of race and ethnicity categories among CRDC, clinical sequencing, and general BCH cohorts.

Revised combined race and ethnicity category	CRDC cohort distribution in percentage (*n*)	Clinical sequencing cohort distribution in percentage (*n*)	General BCH cohort distribution in percentage (*n*)	CRDC *versus* general BCH cohorts *p*-value	Clinical sequencing *versus* general BCH cohorts *p*-value	CRDC *versus* Clinical sequencing cohorts *p*-value
American Indian or Alaska Native	0.22% (7)	0.19% (3)	0.18% (2578)	0.7462	1.0000	1.0000
Asian	4.29% (137)	4.67% (73)	4.74% (68,201)	0.2674	1.0000	0.7319
Black or African American/African, Hispanic/Latine	0.44% (14)	0.32% (5)	0.26% (3713)	0.8170	0.9988	0.7878
Black or African American/African, non-Hispanic/non-Latine	4.60% (147)	5.82% (91)	9.72% (139,756)	<0.0001	<0.0001	0.1794
Middle Eastern/North African	1.60% (51)	2.05% (32)	0.55% (7910)	<0.0001	<0.0001	0.5844
Multi-racial, Hispanic/Latine	7.36% (235)	11.52% (180)	9.26% (133,197)	0.0003	0.0054	<0.0001
Multi-racial, non-Hispanic/non-Latine	1.69% (54)	1.41% (22)	0.87% (12,564)	<0.0001	0.0605	0.7319
Pacific Islander/Native Hawaiian	0.00% (0)	0.00% (0)	0.00% (0)	NA	NA	NA
Race/ethnicity not aligned with revised categories	4.26% (136)	5.44% (85)	6.48% (93,125)	<0.0001	0.1660	0.1794
White, Hispanic/Latine	3.66% (117)	2.62% (41)	1.14% (16,350)	<0.0001	<0.0001	0.1794
White, non-Hispanic/non-Latine	71.88% (2296[a])	65.96% (1031[a])	66.80% (960,386)	<0.0001	0.6886	0.0002
Total, reported	100.00% (3194)	100.00% (1563)	100.00% (1,437,780)	NA	NA	NA
Race/ethnicity not reported (unknown or decline to report)	11.94% (433[a])	12.73% (228[a])	53.14% (1,630,141)	<0.0001	<0.0001	0.6714

[a]There were two overlapping patients in the CRDC and clinical sequencing cohorts (one patient from the White, non-Hispanic/non-Latine category; one patient from the race/ethnicity not reported category).
BCH, Boston Children's Hospital; CRDC, Children's Rare Disease Cohort.

Discussion

Following an intentional revision of EHR-based race and ethnicity categories, we compared the demographics of three cohorts, the CRDC cohort (representing those undergoing research-based sequencing), the clinical sequencing cohort (representing patients that had undergone exome sequencing as a part of their standard clinical care

within the preceding 3 years) and the general BCH cohort (representing the general patient population at BCH). This latter group served as a comparator group, and allowed us to identify relative lack of diversity in populations undergoing research and/or clinical sequencing.

Most notably, there was a significantly lower percentage of Black or African American/African, non-Hispanic/non-Latine individuals in the CRDC cohort compared with the general BCH cohort (4.6% *versus* 9.72%, $p < 0.0001$). However, for this group, there was no statistically significant difference between the CRDC and the clinical sequencing cohorts ($p = 0.1794$), suggesting that the barriers to sequencing in this population are systemic rather than specific to research participation. Under-representation may reflect limited access to highly specialized clinics, which often provide avenues for sequencing in both clinical and research contexts.

Systemically restricted access to high quality medical care for Black patients has been previously well documented. Racial discordance between the patient and provider, lack of race-conscious physician training, and false beliefs regarding biological differences between White and Black patients negatively impacts quality of care.[12,35,36] Lower quality of care can also be exemplified by lack of access to necessary, specialized diagnostic services. Barriers to clinical genetic testing for historically marginalized racial and ethnic populations have also been previously documented.[37]

For Black or African American individuals, views on genetic testing have likely been shaped by a multi-generational history of exploitation by physicians and researchers. The Tuskegee Syphilis Study is a prominent, but not isolated, example of Black or African American patient ill-treatment, which continued to stoke fear and mistrust of medical establishments.[38] The history of unethical experimentation has had a strong impact on perceptions of the medical system within the Black or African American community, reflected in under-representation in clinical trials and precision medicine initiatives.[39,40] Factors contributing to under-enrollment into sequencing initiatives include distance and/or socioeconomic barriers, lack of a diagnosis, low referral rates to by primary care physicians, and recruitment tactics that lack cultural competency.[39,41,42]

A significantly higher percentage of Middle Eastern/North African, White, Hispanic/Latine, and White, non-Hispanic/non-Latine patients were enrolled in the CRDC and the clinical sequencing cohorts compared to those in the general BCH population. The relatively higher representation of Middle Eastern/North African participants for those being sequenced in both contexts probably reflects higher rates of autosomal recessive Mendelian genetic conditions in these populations, in which consanguineous marriages are more frequent.[43,44]

The highest population demographic across all cohorts was the White, non-Hispanic/non-Latine group, who were over-represented in the CRDC cohort compared to both the clinical sequencing ($p = 0.0002$) and general BCH ($p < 0.0001$) cohorts. Relative ease of access to high-quality healthcare and specialty clinics may account for the over-representation of White (both Hispanic/Latine and non-Hispanic/non-Latine) participants in the CRDC cohort. The over-representation of this group in research highlights the need for focused efforts to ensure that populations recruited for research are representative of the diversity of the target population.

There was a significantly lower representation of multi-racial, Hispanic/Latine individuals in the CRDC (7.36%) cohort compared to the general BCH cohort (9.26%, $p = 0.0003$) and compared to the clinical sequencing cohort (11.52%, $p < 0.0001$). For this group, there is a suggestion of a specific disparity at the level of research sequencing. Language and cultural barriers, and researcher bias may be contributing factors to the under-representation.[8,11] However, when factoring in representation from Black, Hispanic/Latine individuals, multi-racial, Hispanic/Latine individuals, and White, Hispanic/Latine individuals, the overall percentage of individuals identifying as Latine was similar in the CRDC cohort (11.46%) compared to the general BCH cohort (10.66%) and lower than the representation in the clinical sequencing cohort (14.46%).

Over half of the patients (53.14%) within the general BCH cohort did not report their race and ethnicity, compared to 11.94% in the CRDC cohort and 12.73% in the clinical sequencing cohort. This difference may reflect a participant population with complex needs that interface regularly with BCH. Patients diagnosed with rare

conditions may benefit from routine, specialized care and more frequent points of contact to medical care.[45] We suspect that regular clinical follow-up may lead to more opportunities to collect demographic data, such as the racial and ethnic identities of said patients. Furthermore, for individuals undergoing sequencing, race and ethnicity information may be pro-actively collected in order to guide interpretation and indeed such data is routinely collected on the test requisition form for genetic testing companies, such as GeneDx.

Next steps

We propose a range of measures to address the disparities identified in this study. Developing and evaluating the effectiveness of any efforts to mitigate inequities requires a method to quantify such imbalances. Therefore, we advocate for standardized, comprehensive EHR population categories. The aim is to convey respect and create a sense of belonging for all individuals, while allowing a more accurate system for tracking demographic data.[25] If EHR systems are to continue collecting detailed demographic data, concerted efforts to include comprehensive racial and ethnic subgroups need to be made. Our distilled categories, and more importantly, the intentional, collaborative systematic approach to creating them, can serve as an example of how such data can be grouped for the purposes of tracking and accountability. In addition to internal efforts, we propose cross-institutional discussions to better understand, and potentially harmonize how different EHR systems (Epic and Cerner) collect demographic data.

We submit two approaches aimed to address the barriers to clinical and research sequencing identified in this study. We recognize that enrollment in the CRDC is by clinician referral. First, we propose broad initiatives to improve referrals and access to specialty providers, which include engagement with primary care providers, incorporation of genetic counselors in a variety of subspecialities throughout the hospital, and coordination with our own quaternary care research institution. Secondly, as a field, we must confront underlying racial biases that may impact clinical practices. Curriculum aimed to enhance medical student, clinician, and researcher cultural competences may critique race-based medicine practices, challenge the notion that race is a biomarker for health outcomes, encourage diversification of textbook images and patient resources, and teach historical cases that show the consequences of structural racism.[46] The integration of these practices and perspectives will improve patient–physician/researcher interactions and ideally help reduce inequities in care.[47]

We intend to broaden community partnerships with participants of all demographics and work to ensure that research staff reflect the diversity of the community. Racial and ethnic representation within the research or care team has been shown to strongly impact participant perceptions. Historically marginalized populations have more trust in health care professionals from their own cultural background.[48] Similarly, participants are more motivated to enroll in a study if there is race concordance within the research team.[49,50] Efforts underway at our institution include the fostering of training opportunities for students to pursue medicine and genomics, and employing diverse recruitment strategies. Additionally, collaborative partnerships with members of the study population ensures that the community is engaged and benefits from research results.[17] We therefore suggest that research protocols promote trust within the community through community-engaged, transparent research approaches.[49,51] While incorporating the aforementioned steps, it is vital that research initiatives and clinical genomics divisions routinely and transparently re-examine the impact of these changes. Reports should include percentages of clinical sequencing and research enrollment amongst the revised race and ethnic categories as clear metrics for the impact of these initiatives.

Impact of next steps

It is well-established that clinical and genomic data that is not representative of the breadth of the global population has negative consequences. Lack of representation in sequencing initiatives results in restricted access to the opportunities for natural history and clinical trial participation, misdiagnoses and substandard care, and limited knowledge of how these therapies may impact the health of specific populations.[39,50] Failure to capture the full spectrum of human diversity in genomic research will perpetuate the disparities in future precision medicine-based clinical care.[52]

In order to benefit the rare disease research community and future precision medicine initiatives

in the spirit of pediatric health equity, we aspire to practice inclusivity in our clinical and research practices. Our intention behind publishing the general BCH, clinical sequencing, and CRDC cohort data was twofold. Firstly, we hope to accelerate existing hospital-wide efforts to actively mitigate the barriers to clinical care and research enrollment for individuals with rare disease. Additionally, by publishing our institutional data, we hope to encourage other rare disease research groups to similarly examine and transparently share their race and ethnicity enrollment data.

Study limitations

A limitation of this study was that our data was restricted to the self-reported race and ethnicity categories that exist in the current EHR. These existing racial and ethnic categories are not inclusive, making it challenging to decipher and recategorize data. Conversely, some categories, such as Asian are heterogenous despite geographic proximity. We created inclusive racial and ethnic categories and attempted to rearrange existing EHR data according to these new categories. In doing so, the sample size of our data may be diminished, or certain populations may be under-reported.

Additionally, as this study was a quality improvement initiative, we were only authorized to utilize de-identified, aggregate data. Our analysis was therefore limited to the use of a small number of de-identified categories available through the EHR. This impeded our ability to comment upon other factors or barriers that may contribute to the under-representation seen across the three cohorts.

We also recognize the CRDC is composed of a diverse group of projects. Not all CRDC cohorts recruit patients with disorders that equally impact patients across all populations. Some investigators recruit for conditions that have a higher prevalence within certain ancestries, which may overlap with specific race and ethnic populations. Additionally, some cohorts, such as severe COVID-19 recruit for conditions have been shown to impact historically marginalized communities at a higher rate.[53,54]

Conclusion

Upon thorough review and revision of racial and ethnic data captured in a single tertiary/quaternary pediatric institution, population groups belonging to the global majority were under-represented in sequencing cohorts compared to the general hospital population. This is consistent with the lack of equitable racial and ethnic representation in rare disease genomics in both clinical care and research. We propose a range of measures to address disparities, including recategorization of demographic data using inclusive language, community partnerships in genomics research, initiatives to improve access to specialty providers and confrontation of underlying racial biases that may impact clinical practices and future discovery. Finally, we advocate for careful and transparent metrics to measure the impact of such initiatives.

Declarations

Ethics approval and consent to participate

This project utilized de-identified, aggregate data and was conducted as a Quality Improvement initiative; neither ethics review nor consent was required nor obtained.

Consent for publication

Not applicable.

Author contributions

Zoë J. Frazier: Investigation; Methodology; Project administration; Writing – original draft; Writing – review & editing.

Eurnestine Brown: Conceptualization; Methodology; Supervision; Writing – review & editing.

Shira Rockowitz: Data curation; Formal analysis; Methodology; Software; Writing – review & editing.

Ted Lee: Formal analysis; Methodology; Writing – review & editing.

Bo Zhang: Formal analysis; Methodology; Validation; Writing – review & editing.

Abigail Sveden: Investigation; Resources; Writing – review & editing.

Nancy L. Chamberlin: Methodology; Resources; Writing – review & editing.

Kira A. Dies: Conceptualization; Funding acquisition; Resources; Supervision; Writing – review & editing.

Annapurna Poduri: Investigation; Writing – review & editing.

Piotr Sliz: Conceptualization; Data curation; Formal analysis; Project administration; Supervision; Writing – review & editing.

Maya Chopra: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Supervision; Writing – original draft; Writing – review & editing.

CRDC Consortium: Conceptualization; Writing – review & editing.

Acknowledgements

The authors thank the Research Computing Department, specifically Luis Villa, for consulting with them regarding the coding scheme to map race and ethnicity variables and statistical analysis. Additionally, we thank the CRDC Consortium Members, who include Annapurna Poduri, Scott Snapper, Margaret Kenna, Eliot Shearer, Joel Hirschhorn, Janet Chou, Benjamin Raby, Jay Thiagarajah, Pankaj Agrawal, Jia Zhu, Jessica Kremen, Siddharth Srivastava, Maya Chopra, Ingrid Ganske, Alan Beggs, Joseph Gonzalez-Heydrich, Catherine Brownstein, Ryan Doan, Adrienne Randolph, Philip Boone, Matthew Sampson, Christina Jacobsen, David Glahn, Charles Berde, Mark Fleming, Amar Majmundar, Akiko Shimamura, Christopher Walsh, Diane Shao, Mary Whitman, Monica Wojcik, Stephanie Roberts, Darius Ebrahimi-Fakhari, Ted Lee, Nina Mann, Piotr Sliz.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was conducted with support from the Rosamund Stone Zander Translational Neuroscience Center, Boston Children's Hospital and from the Boston Children's Hospital IDDRC (NIH P50 HD105351).

Competing interests

The authors declare that there is no conflict of interest.

Availability of data and materials

Not applicable.

ORCID iD

Maya Chopra https://orcid.org/0000-0002-2574-3375

Supplemental material

Supplemental material for this article is available online.

References

1. Global Genes. RARE disease facts, https://globalgenes.org/rare-disease-facts/ (2018, accessed 30 November, 2022).
2. Haendel M, Vasilevsky N, Unni D, *et al*. How many rare diseases are there? *Nat Rev Drug Discov* 2020; 19: 77–78.
3. Carmichael N, Tsipis J, Windmueller G, *et al*. "Is it going to hurt?": The impact of the diagnostic odyssey on children and their families. *J Genet Counsel* 2015; 24: 325–335.
4. Miller D. The diagnostic odyssey: our family's story. *Am J Hum Genet* 2021; 108: 217–218.
5. Retterer K, Juusola J, Cho MT, *et al*. Clinical application of whole-exome sequencing across clinical indications. *Genet Med* 2016; 18: 696–704.
6. Clark MM, Stark Z, Farnaes L, *et al*. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom Med* 2018; 3: 16.
7. Reuter CM, Kohler JN, Bonner D, *et al*. Yield of whole exome sequencing in undiagnosed patients facing insurance coverage barriers to genetic testing. *J Genet Couns* 2019; 28: 1107–1118.
8. Hernandez SM and Sparks PJ. Barriers to health care among adults with minoritized identities in the United States, 2013–2017. *Am J Public Health* 2020; 110: 857–862.
9. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Population Health and Public Health Practice, et al. *Communities in action: pathways to health equity*. Washington, DC: National Academies Press (US), http://www.ncbi.nlm.nih.gov/books/NBK425848/ (2017, accessed 30 January 2023).
10. Wojcik MH, Bresnahan M, Del Rosario MC, *et al*. Rare diseases, common barriers: disparities

in pediatric clinical genetics outcomes. *Pediatr Res* 2023; 93: 110–117.

11. D'Angelo CS, Hermes A, McMaster CR, *et al.* Barriers and considerations for diagnosing rare diseases in indigenous populations. *Front Pediatr* 2020; 8: 579924.
12. Hoffman KM, Trawalter S, Axt JR, *et al.* Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proc Natl Acad Sci U S A* 2016; 113: 4296–4301.
13. Koretzky M, Bonham VL, Berkman BE, *et al.* Towards a more representative morphology: clinical and ethical considerations for including diverse populations in diagnostic genetic atlases. *Genet Med* 2016; 18: 1069–1074.
14. Bailey ZD, Krieger N, Agénor M, *et al.* Structural racism and health inequities in the USA: evidence and interventions. *Lancet* 2017; 389: 1453–1463.
15. Landry LG, Ali N, Williams DR, *et al.* Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Aff (Millwood)* 2018; 37: 780–785.
16. McBaiden R. Not 'ethnicity minority', try Global Majority. *Open Palm*, https://www.openpalm.org.uk/post/not-ethnic-minority-try-global-majority (2022, accessed 23 January 2023).
17. Lemke AA, Esplin ED, Goldenberg AJ, *et al.* Addressing underrepresentation in genomics research through community engagement. *Am J Hum Genet* 2022; 109: 1563–1571.
18. Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020; 581: 434–443.
19. Rockowitz S, LeCompte N, Carmack M, *et al.* Children's rare disease cohorts: an integrative research and clinical genomics initiative. *NPJ Genom Med* 2020; 5: 29.
20. Augustine EF, Adams HR and Mink JW. Clinical trials in rare disease: challenges and opportunities. *J Child Neurol* 2013; 28: 1142–1150.
21. Office of Research on Women's Health. Office of Management and Budget (OMB) Standards, https://orwh.od.nih.gov/toolkit/other-relevant-federal-policies/OMB-standards (1997, accessed 30 January 2023).
22. Mauro M, Allen DS, Dauda B, *et al.* A scoping review of guidelines for the use of race, ethnicity, and ancestry reveals widespread consensus but also points of ongoing disagreement. *Am J Hum Genet* 2022; 109: 2110–2125.
23. The White House Office of Management and Budget. Standards for the classification of federal data on race and ethnicity, https://obamawhitehouse.archives.gov/omb/fedreg_race-ethnicity#:~:text=If%20a%20combined%20format%20is%2CBlack%2C%20not%20of%20Hispanic%20origin (1995, 30 November 2022).
24. American Medical Association and Association of American Medical Colleges. Advancing health equity: guide on language, narrative and concepts, https://www.ama-assn.org/system/files/ama-aamc-equity-guide.pdf (2021, accessed 20 January 2023).
25. Flanagin A, Frey T, Christiansen SL, *et al.* Updated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA* 2021; 326: 621–627.
26. Ochoa M. Stop using 'Latinx' if you really want to be inclusive. The Conversation, https://theconversation.com/stop-using-latinx-if-you-really-want-to-be-inclusive-189358 (2022, accessed 6 December 2022).
27. Iruka IU, Bathia D, Suggs M, *et al. Are black and Latine families with babies feeling relief from the child tax credit?* Chapel Hill, NC: Equity Research Action Coalition, Frank Porter Graham Child Development Institute, The University of North Carolina at Chapel Hill; Equity Coalition Rescue Plan Microbrief Series 21-002, 2021.
28. Sotto-Santiago S. Time to reconsider the word minority in academic medicine. *J Best Pract Health Prof Divers* 2019; 12: 72–78.
29. Grady C. Why the term 'BIPOC' is so complicated, explained by linguists, https://www.vox.com/2020/6/30/21300294/bipoc-what-does-it-mean-critical-race-linguistics-jonathan-rosa-deandra-miles-hercules (2020, accessed 30 November 2022).
30. Khan AT, Gogarten SM, McHugh CP, *et al.* Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: experiences from the NHLBI TOPMed program. *Cell Genomics* 2022; 2: 100155.
31. Choi SC and Stablein DM. Practical tests for comparing two proportions with incomplete data. *J R Stat Soc C* 1982; 31: 256–262.
32. Althouse AD. Adjust for multiple comparisons? It's not that simple. *Ann Thorac Surg* 2016; 101: 1644–1645.

33. Li G, Taljaard M, Van den Heuvel ER, *et al.* An introduction to multiplicity issues in clinical trials: the what, why, when and how. *Int J Epidemiol* 2017; 46: 746–755.

34. National Institutes of Health (NIH). Racial and ethnic categories and definitions for NIH diversity programs and for other reporting purposes, https://grants.nih.gov/grants/guide/notice-files/not-od-15-089.html#:~:text=NOT%2DOD%2D15%2D089,and%20for%20Other%20Reporting%20Purposes&text=On%20January%2012%2C%202015%2C%20an,%2DOD%2D15%2D053 (2015, accessed 20 January 2023).

35. Cerdeña JP, Plaisime MV and Tsai J. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *Lancet* 2020; 396: 1125–1128.

36. Shen MJ, Peterson EB, Costas-Muñiz R, *et al.* The effects of race and racial concordance on patient-physician communication: a systematic review of the literature. *J Racial Ethn Health Disparities* 2018; 5: 117–140.

37. Chapman-Davis E, Zhou ZN, Fields JC, *et al.* Racial and ethnic disparities in genetic testing at a hereditary breast and ovarian cancer center. *J Gen Intern Med* 2021; 36: 35–42.

38. Gamble VN. Under the shadow of Tuskegee: African Americans and health care. *Am J Public Health* 1997; 87: 1773–1778.

39. Brown RF, Cadet DL, Houlihan RH, *et al.* Perceptions of participation in a phase I, II, or III clinical trial among African American patients with cancer: what do refusers say? *J Oncol Pract* 2013; 9: 287–293.

40. Walley NM, Pena LDM, Hooper SR, *et al.* Characteristics of undiagnosed diseases network applicants: implications for referring providers. *BMC Health Serv Res* 2018; 18: 652.

41. Shaya FT, Gbarayor CM, Yang HK, *et al.* A perspective on African American participation in clinical trials. *Contemp Clin Trials* 2007; 28: 213–217.

42. Fraiman YS and Wojcik MH. The influence of social determinants of health on the genetic diagnostic odyssey: who remains undiagnosed, why, and to what effect? *Pediatr Res* 2021; 89: 295–300.

43. Hoodfar E and Teebi AS. Genetic referrals of Middle Eastern origin in a western city: inbreeding and disease profile. *J Med Genet* 1996; 33: 212–215.

44. Jaouad IC, Elalaoui SC, Sbiti A, *et al.* Consanguineous marriages in Morocco and the consequence for the incidence of autosomal recessive disorders. *J Biosoc Sci* 2009; 41: 575–581.

45. Heuyer T, Pavan S and Vicard C. The health and life path of rare disease patients: results of the 2015 French barometer. *Patient Relat Outcome Meas* 2017; 8: 97–110.

46. Green K-A, Wolinsky R, Parnell SJ, *et al.* Deconstructing racism, hierarchy, and power in medical education: guiding principles on inclusive curriculum design. *Acad Med* 2022; 97: 804–811.

47. Nielsen DS, Korsholm KM, Mottelson I, *et al.* Cultural competences gained through an education program as ethnic patient coordinator: a qualitative study. *J Transcult Nurs* 2019; 30: 394–402.

48. Kennedy BR, Mathis CC and Woods AK. African Americans and their distrust of the health care system: healthcare for diverse populations. *J Cult Divers* 2007; 14: 56–60.

49. Statler MC, Wall BM, Richardson JW, *et al.* African American perceptions of participating in health research despite historical mistrust. *ANS Adv Nurs Sci* 2023; 46: 41–58.

50. Shah-Williams E, Levy KD, Zang Y, *et al.* Enrollment of diverse populations in the INGENIOUS pharmacogenetics clinical trial. *Front Genet* 2020; 11: 571.

51. Patel YR, Carr KA, Magjuka D, *et al.* Successful recruitment of healthy African American men to genomic studies from high-volume community health fairs: implications for future genomic research in minority populations. *Cancer* 2012; 118: 1075–1082.

52. Fisher ER, Pratt R, Esch R, *et al.* The role of race and ethnicity in views toward and participation in genetic studies and precision medicine research in the United States: a systematic review of qualitative and quantitative studies. *Mol Genet Genomic Med* 2020; 8: e1099.

53. Price-Haywood EG, Burton J, Fort D, *et al.* Hospitalization and mortality among black patients and white patients with COVID-19. *N Engl J Med* 2020; 382: 2534–2543.

54. Millett GA, Jones AT, Benkeser D, *et al.* Assessing differential impacts of COVID-19 on black communities. *Ann Epidemiol* 2020; 47: 37–44.