PLOS GLOBAL PUBLIC HEALTH

REVIEW

Use of race, ethnicity, and ancestry data in health research

Clara Lu[1], **Rabeeyah Ahmed**[2], **Amel Lamri**[1], **Sonia S. Anand**[1,3]*

1 Department of Medicine, McMaster University, Hamilton, Ontario, Canada, **2** Arts and Science Program, McMaster University, Hamilton, Ontario, Canada, **3** Department of Health Research Methods, Evidence & Impact, McMaster University, Hamilton, Ontario, Canada

* anands@mcmaster.ca

Abstract

Race, ethnicity, and ancestry are common classification variables used in health research. However, there has been no formal agreement on the definitions of these terms, resulting in misuse, confusion, and a lack of clarity surrounding these concepts for researchers and their readers. This article examines past and current understandings of race, ethnicity, and ancestry in research, identifies the distinctions between these terms, examines the reliability of these terms, and provides researchers with guidance on how to use these terms. Although race, ethnicity, and ancestry are often treated synonymously, they should be considered as distinct terms in the context of health research. Researchers should carefully consider which term is most appropriate for their study, define and use the terms consistently, and consider how their classification may be used in future research by others. The classification should be self-reported rather than assigned by an observer wherever possible.

Citation: Lu C, Ahmed R, Lamri A, Anand SS (2022) Use of race, ethnicity, and ancestry data in health research. PLOS Glob Public Health 2(9): e0001060. https://doi.org/10.1371/journal.pgph.0001060

Editor: Amy Yuk-Ying Tan, The University of British Columbia, CANADA

Published: September 15, 2022

Funding: SA is supported by a Canada Research Chair in Ethnic Diversity and Cardiovascular Disease, and the Michael DeGroote Heart and Stroke Foundation Chair in Population Health. These funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: Dr. Anand is the Associate Chair of Equity and Diversity for the Department of Medicine in McMaster University. All other authors declare no competing interests.

Introduction

Over the past two decades, ethnicity and race-based health research has accelerated [1, 2]. In parallel, an increasing number of population genomics studies are being conducted in which the term ancestry is often used. Currently there is confusion regarding which is the optimal term to use in health research–ethnicity, race, or ancestry, partly because they lack standard definitions. Conducting such research necessitates being aware of the differences between the concepts of ethnicity, race, and ancestry, and requires that clear definitions of these terms be made explicit by the researcher [3]. This confusion is fuelled by three contemporary perspectives: i. the contention that human populations are more alike than different, ii. genome-wide and whole genome sequencing technology which generates ancestral genetic information, and iii. the societal reckoning with structural racism in healthcare provision and practices. These perspectives justify a re-examination of the concepts and terms of ethnicity, race, and ancestry. This article will review evidence that informs our understanding of these terms, identify gaps in our understanding, and provide direction to researchers regarding the use of these terms in health research.

Methods

We conducted literature searches in PubMed to identify published manuscripts in the English language for i. ethnicity and race and health research, ii. ethnicity and race with ancestry and genetics, and iii. ethnicity and religion. All search strategies are described in S1 Table. Key articles were then selected and reviewed in depth, and others identified through a review of the citations to generate this narrative review [4].

1. Use of "race" and "ethnicity" terminology in health research

Race and ethnicity are often used as classification variables in health research. However, the terms are frequently misused and invoked inconsistently [5, 6]. Race is a socio-political construct that divides people into groups based on perceived physical differences [7, 8]. For more than two centuries, this classification has widely influenced medical education and led to racial discrimination in healthcare [9]. The persistent use of race categories in medicine has had two significant impacts, leading to controversy regarding the ongoing utility of this classification in health research. On one hand, the use of the term race has perpetuated the incorrect belief that race categories reflect biological "including genetic" differences, which may reinforce harmful stereotypes and perpetuate "race science" [3, 10–14]. This calls into question whether measuring race is an effective way for health researchers to analyze differences between various groups and whether overall findings from studies that assess race can be generalized to the broader population. On the other hand, race indicators can help identify and address inequalities arising from discrimination [8, 15, 16]. As categorization of human populations by race fuels racism, and people affected by racism are referred to as "racialized," the reporting of racial differences may illuminate the powerful effects of social inequalities on health and healthcare. The use of the term 'race' therefore may be used as a proxy to account for social inequalities related to racism and its socio-political disparities. For the practical benefit of monitoring health inequities that arise from racism, some argue that we should continue to measure "race" [17]. This however is heavily dominated by the American perspective, and is not common in Europe, where the term race is rarely used [18]. However the European Union Commission has recently provided a framework for disaggregating data by ethnicity and race to quantify discrimination and inequity, with the goal of benefiting the groups they describe [19].

As race is widely acknowledged to be a social rather than a biological construct, the resulting challenges associated with racial classification and data collection have led some researchers to prefer the term "ethnicity." Ethnicity is a construct that encompasses common cultural characteristics including language, religion, dietary practices, and nationality; it *may* also reflect common ancestry or geographic origin [8, 15, 19]. These characteristics create a sense of collective identity that is often carried forward between generations [20]. Ethnicity categories can be used to design "culturally appropriate health services" and investigate clinical-biological differences in risk factors for diseases and responses to therapies [8, 21]. Some religious groups for example Jewish populations can also be included as ethnicity categories, as this can be the dominant identity for some individuals more so than country of origin or skin colour in the case of race-based classifications. To account for various cultural and socioeconomic factors that may influence population health outcomes, some researchers may choose to define ethnicity in terms of religious affiliation and categorize the Jewish population as a distinct ethnic group [22].

Although some researchers advocate for using the term ethnicity instead of race, using the term ethnicity alone without a precise definition can also cause confusion [23]. Ethnicity is a complex, multidimensional construct that may change over time, and this contributes to its within-group heterogeneity. While the construct of race relies on perceived physical

characteristics, ethnicity also captures elements of an individual's identity beyond the physical, as previously described. As a result, the use of each term is subject to various interpretations by researchers, study participants, and readers [23]. Researchers have noted that the racial groups traditionally classified by the United States (US) Census Bureau–such as Black or African American, Asian, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and White–can themselves include multiple ethnic groups. For example, Black, White, and Asian races can all include individuals belonging to Hispanic ethnicity [23]. Recently, calls for a re-evaluation of this antiquated definition and use of race categories by the US Government have been made [20].

Finally, there has been no standard agreement on the distinction between the terms ethnicity and race in the context of research [3, 7, 24]. No formal consensus has been reached regarding which term is preferable to the other, and in which circumstances. Despite substantial literature emphasizing the consequences of using these terms interchangeably, in practice, the two terms are often treated as synonyms [5, 24–26]. As such, it is essential for researchers to clearly define their chosen terms.

2. Implications of misuse and confusion of terms

A 2011 systematic review highlighted studies that have misused and confused the terms race, ethnicity, and ancestry, and noted that researchers often did not justify why recording these variables was necessary [27]. Many studies featured in the review did not define these terms in the context of their research [28–33]. In addition to inconsistencies within individual manuscripts, authors often used varying terminology to describe the same concepts. For instance, in manuscripts focused on genetics, authors variably described their findings in terms of ancestry [29], race [30], and ethnicity [31]. In some publications, the descriptors used in the main article were different from those in the supplementary text [31, 33]. Finally, some authors introduced multiple terms without defining or explaining whether each of the terms are distinct or synonymous in the context of their study [34, 35]. It is unlikely that comparing populations in this manner is meaningful and this ultimately causes substantial confusion amongst readers [35].

When researchers fail to define "race" and "ethnicity" in their research, they reinforce the idea that racial and ethnic classifications are unchanging and well-defined [36]. This also fuels the assumption that biological differences may exist when racial differences are reported regarding differences in medication efficacy, for example [37], thereby perpetuating the belief that racial differences reflect biological differences. While some biological-race examples appear to be firmly ensconced in the practice of medicine, a recent report demonstrated that risk models that do not include race can be more precise and can replace such race-based formulae [38].

The internalization of racial stereotypes by healthcare professionals can lead to disparities in healthcare delivery for non-white people. Physician's referral bias for cardiac catheterization based on gender and race after assessing actors describing anginal pain. was well illustrated by Schulman et al. [39]. The observation that using race as a descriptor in medical histories can invoke such bias has led to calls for the removal of this characteristic from case presentations [40]. In contrast, when ethnicity or race is not collected, or when under-reporting occurs in situations where individuals want to avoid stigmatization, this can lead to under-recognition of ethnic and racial health disparities and discrimination when they do exist [41].

Precise and accurate assessments of race and ethnicity are important if they are to be used in health services research. The race-and ethnicity-specific data that emerge from clinical studies are often used by governments, private and public institutions, and businesses to create

health interventions and to distribute healthcare resources based on priority. Since race and ethnicity are widely used in health services research, consistent terminology and measurements are essential for identifying health services gaps that may be race-and/or ethnicity-specific [23]. However, ethnicity and/or race data can be variably collected in electronic health records (EHR) and are not routinely collected in paper-based records in some countries like Canada. This lack of standardized ethnicity data has prompted the development of surrogate measures, such as unique surname analysis to assign ethnicity for some ethnic groups like South Asian and Chinese people [42, 43]. This practice, however, cannot be readily applied to most ethnic groups. Furthermore, some analyses of EHR have identified serious gaps in completion of the race and ethnicity fields when these characteristics require "assignment" by an observer, rather than by patient self-report. The race and ethnicity categories available, and whether these fields are assigned or self-reported, also need to be carefully considered as outlined below.

Discrepancies between self-reported and observer-classified race or ethnicity. The validity of race and ethnicity classification may depend on whether it is *self-reported* by a research participant or patient or *assigned* by a research assistant or healthcare worker (i.e. observer-classified). The relationship between self-reported and observer-classified race and ethnicity has been described in several studies using population-based survey data from the US and Latin America [16, 44–48]. Some of these surveys transitioned from observer-recorded to self-reported race in response to federal guidelines, for example from the US Office of Management and Budget in 1997 [47]. Observers within these studies included public health researchers, government officials, census enumerators, medical examiners, hospital or clinic personnel, and next of kin [49–51] with wide variations in observer training on race and ethnicity classification [16, 52–54]. An analysis of the US health survey data from the Behavioral Risk Factor Surveillance System found that agreement between self- and observer-identified race varied across racial and ethnic groups. Higher agreement rates existed among self-identified Black (96% agreement) and White (98% agreement) participants, with lower agreement rates among non-Black minority groups (35% agreement among Native Hawaiian and other Pacific Island participants) [45]. An analysis of US Veterans Affairs healthcare users also showed high agreement rates among White (98%) and Black (94%) racial groups and low agreement rates among non-Black minority groups [47], with similar patterns demonstrated across studies [52, 55–57]. These trends are not unique to the US. An analysis from the Project on Ethnicity and Race in Latin America reported that only 61% of classifications by interviewers matched those of respondents [46] while an overall agreement rate of 75% between participants and observers was reported in the Brazilian Social Survey [44]. Again, in Latin American populations, high agreement rates were found for White respondents but lower reliability existed between self and observer classification for minority respondents [16, 44, 48]. "Passing" is a commonly described phenomenon in which an individual assumes the identity of a different racial or ethnic group to increase socioeconomic status [53, 58]. Socially-assigned race refers to the concept of an observer making assumptions of an individual based on physical appearance in particular skin colour [45]. Being socially-assigned as "White" is associated with higher health status, even among those who self-identify as non-White [45]. Other factors beyond skin colour can affect observer classification of race and ethnicity. In addition to an observer's interpretation of both visual cues (e.g. skin tone or hair texture), auditory cues (e.g. accent or vernacular) may affect their assignment of race or ethnicity [53, 59].

Discrepancies between self-reported and observer classification of race and ethnicity can have substantial implications on health research findings. For example, observer-assigned compared to self-identification of race has led to underestimations of infant mortality and cancer incidence of Native Americans [51]. On the other hand, individuals who are classified by

others as White, irrespective of their self-identified race, are more likely to report their health status as being very good or excellent compared to individuals who are not socially-assigned as White [45]. It is plausible that being classified by others as White could positively bias observer assessments of other health-related outcomes, such as patient health literacy and adherence to interventions, thereby reinforcing racial stereotypes and hierarchies [6, 48].

For all of the above reasons, we conclude that self-reported identity is preferred over observer classification [17]. However we contend that in some circumstances, measuring both self-reported and observer assigned race and ethnicity may be informative. For example, observer classification of perceived race and ethnicity may illuminate the impacts of discrimination within and outside healthcare, particularly when outcomes vary substantially between self and observer classification [44, 56]. Further research is needed to discern the impact of different race and ethnicity classification methods on findings within health research [45].

Reliability of self-reported ethnicity or race over time. Evidence that self-reported ethnicity and race can change over time challenges the assumption that these variables remain fixed across the life span [51, 53, 56, 58, 60]. Census studies demonstrating limited reliability of ethnicity and race data date back to 1974, when a third of the US population reported a different racial or ethnic status one year after their initial interview [61]. A more recent analysis of US census data found that 6.1% of respondents–approximately 9.8 million people–reported a different race in 2010 compared to 2000 [58]. This phenomenon of racial fluidity appears most pronounced among minority and multiracial populations. Saperstein et al. describe such phenomena as "emblematic of how racial fluidity interacts with racial inequality," with the reporting of race closely tied to historical and ongoing racial discrimination within society [53, 62].

Only 40% of people who reported multiple races in the 2000 US Census subsequently reported multiracial status in a follow-up survey, and among minority respondents, consistent reporting of race ranged between 55% and 78% [63]. Similar racial fluidity was reported among people identifying as American Indian, Pacific Islander, and multiracial between the 2000 and 2010 US Census. The highest rate of response changes occurred among double-minority groups; for example, among non-Hispanic individuals reporting multiracial American Indian and Pacific Islander background in 2000, only 10% selected the same two categories in 2010 [58].

Several factors that influence the reliability of race and ethnicity reporting over time have been identified at the individual, societal, and study level. First, an individual's ethnic and racial identity may evolve over time and space–that is, across life stages and social contexts. A "simplification" phenomenon has been described among people of mixed heritage, whereby a multiracial individual may simplify their background into a single race or ethnicity identity when they marry or leave their childhood home [60, 64].

The context in which the question is asked may also affect an individual's response. An individual may disclose different backgrounds while at home, at work, and during travel, depending on comfort, perceived safety, or benefit of disclosure [53, 58, 60, 62, 65]. In the 1994–1995 US National Longitudinal Study of Adolescent Health, for example, 12% of adolescents reported a different race at home versus at school [66]. In addition to the phenomenon of "passing," a "chameleon change" experience has been described among multiracial individuals who preferentially identify with different racial groups at different times [67]. Finally, there is increasing concern among Indigenous communities regarding "ethnic fraud" whereby false self-identification of Indigenous heritage is committed by non-Indigenous people in order to gain resources or opportunities [68].

Lastly, study design may influence the reliability of race and ethnicity data. The mode of data collection, such as written surveys or face-to-face interviews, may also yield differing responses [60]. Other factors influencing responses include question wording, response

choices, and the instructions and examples available to respondents [50, 53, 60]. Finally, changes in the classification and language of racial and ethnic categories–such as the option of reporting mixed races and ethnicities introduced on the 2000 US and 2001 United Kingdom censuses–may change an individual's classification over time [54, 56, 62].

Confusion and conflation of terms. We have described how race and ethnicity are fluid concepts, and while various definitions for both have been proposed, there is no standard agreement on these definitions. Efforts to collect race- and ethnicity-based data should be iterative, taking caution not to assume these variables are fixed [53]. Interpretation of research findings based on race and ethnicity data must also ensure that results can be appropriately generalized over time [26]. The confusion about the use of these terms has led some researchers to altogether avoid defining race and ethnicity as distinct categories, and instead adopt "race/ethnicity" as a category [69]. Trends in the bibliographic database, MEDLINE, show that researchers are using the term "ethnicity" or "race/ethnicity" instead of "race" more often [70]. While some argue that "race/ethnicity" should be the preferred phrase among researchers, others claim that this categorization is still imprecise, since there are no universally accepted definitions for either concept. As well, while "race/ethnicity" is broader, its use is limited for groups and individuals who do not fall under traditional categories, such as Hispanic populations [69]. The conflation of the terms race and ethnicity can also lead to the misinterpretation that race and ethnicity findings have a biological explanation, without considering the social determinants that usually explain between group differences. Another source of confusion comes from the lack of guidance from gatekeepers of research scholarship, including granting agencies and journal editors regarding which terminology should be used (e.g. race, ethnicity, or ancestry) for the health and biomedical research studies they fund or publish. While journal instructions for authors often advocate for including diverse populations in clinical studies, guidelines are less clear about specific race and ethnicity descriptors [1]. Recently, however, the Journal of the American Medical Association (JAMA) created guidelines on reporting race and ethnicity, noting that "terminology, usage, and word choice are critically important, especially when describing people and when discussing race and ethnicity" [17]. JAMA recommends that authors use specific racial and ethnic classifications rather than collective terms [17] and suggests avoiding the "race/ethnicity" classification, and instead recommending using "race and ethnicity." This terminology recognizes the diverse populations that exist within racial and ethnic groups [17]. A requirement for precise definitions by granting agencies and journals may also engage researchers to reflect more deeply on these constructs and what roles they play in health research. As societal understandings of race and ethnicity continue to evolve, the use and application of these terms must continue to be reassessed regularly [17].

3. Genetically-inferred ethnicity and ancestry

Ancestry is another way to characterize individuals beyond race or ethnicity. Ancestry can be defined geographically, genealogically, or genetically, and can suffer limitations similar to race or ethnicity. Geographic ancestry refers to ancestors originating from similar geographic regions. Genealogical ancestry refers to one's ancestral pedigree, and genetic ancestry refers to ancestors from whom one is biologically descended. It is commonly understood that humans from different ethnic groups are more similar genetically than different; so why is research of genetic differences between diverse populations even undertaken [10–14, 71]? Although the human genome of people from different racial, ethnic, or ancestral groups is ~99.9% identical, some variation is present in a typical human genome compared to a reference genome at 3.7 to 5 million sites representing 0.1% of the entire genome (Fig 1). The vast majority of these

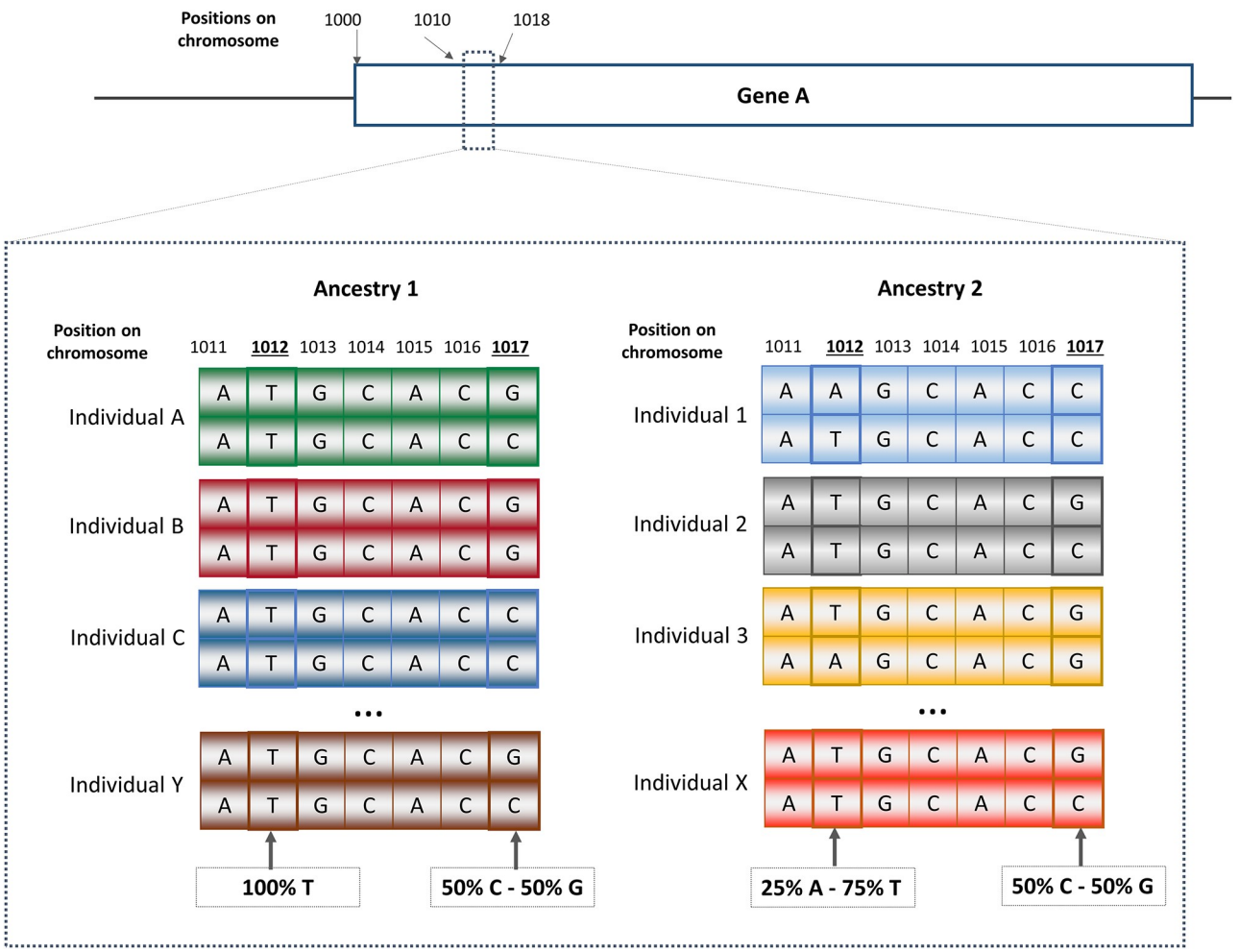

Fig 1. Variation involving 0.1% of the human genome. (A) Proportion of human genome sequence that is similar (blue) and varies (orange) across individuals. (B) A close examination of the DNA sequence (between positions 11 and 17 included) of gene A reveals 2 genetic variants: Variant #1 (position 1012 on chromosome): All participants from Ancestry 1 have a **T** allele, while participants of ancestry 2 have both **T** and **A** allele. This variant could be ethnic specific and could potentially be used as an ancestry marker; Variant #2, (position 1017): the two alleles **G** and **C** are observable in both ancestry groups with similar frequencies, hence, this genetic variant is not an ancestry informative.

https://doi.org/10.1371/journal.pgph.0001060.g001

genetic variants belong to a class referred to as single nucleotide variants [72, 73], which can be used to determine genetically-inferred ethnicity and/or ancestry. This is because the number of varying sites differs greatly between genetically distant populations (i.e. ~86% of variants are unique to a single ethnic continental group) [72], with populations of African descent typically displaying the largest number of polymorphic sites [72, 73]. In addition, genetic variants that are shared across different ancestral groups also contribute to population differentiation by displaying large differences in allele frequencies. These differences reflect the history and movement of human populations. Although the ancestry of human populations is a continuum, current genetic studies typically stratify populations at the continental level. The term genetic ancestry refers to the outcome of tracing back the DNA of contemporary populations to extinct common ancestors. This is typically done either by comparing contemporary to ancient DNA samples, or by using complex mathematical models.

Multiple programs can be used to determine the genetically-inferred ancestry of study participants. These programs can accommodate different study designs and population types (i.e.

unrelated [74] vs. related participants [75]; homogenous vs. admixed populations [76, 77]. Among the most commonly used methods are: *i)* Principal component analysis (PCA) [74], a multi-dimensional scaling method implemented in multiple software [75, 78, 79]; and *ii)* model-based clustering approaches [76, 80] used to characterize and visualize genetic ancestry. Genetic ancestry is typically combined with self-reported race or ethnicity data for validation and *vice versa*. In the absence of self-reported data, the information can be inferred using genetic ancestry. This requires the use of samples with validated/harmonized genetic and self-reported data as a reference (Fig 2). A similar approach can also be used to correct or remove data for problematic samples where self-reported race or ethnicity is deemed to be either inaccurate or inconsistent (across time or multiple sources). In an attempt to automate this process of imputation and correction in a large multi-ethnic longitudinal study, a program that combines genetic and self-reported data from multiple sources to generate a "harmonized ancestry and race/ethnicity" variable was recently developed using a machine learning approach [81].

The availability of large reference populations characterizing the genetic structure of non-European populations plays a key role in the accurate inference/validation of genetic ancestry in multi-ethnic and non-European populations, especially when working at a finer scale than

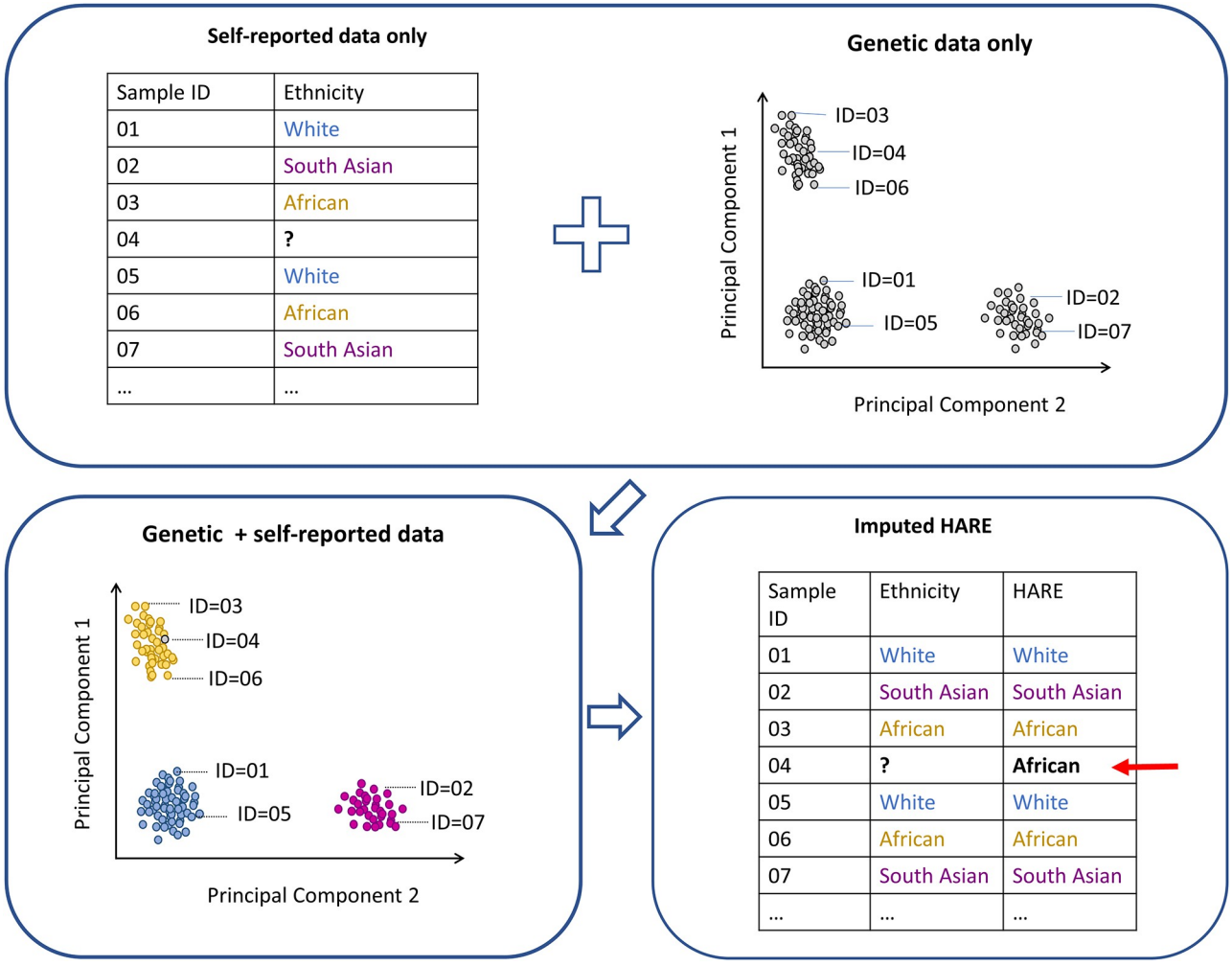

Fig 2. Validation and imputation of self-reported ethnicity using genetic data.

https://doi.org/10.1371/journal.pgph.0001060.g002

the continental level (regions or subregions). However, there has been a lag in whole-genome sequencing and genotyping studies funded, conducted, and published in non-White populations. This has hindered both the ability to refine currently known GWAS signals and the chances of discovering new associations, hence limiting our understanding of the biological mechanisms involved. For example, a recent study revealed massive disparities in predicting accurate genetic risk scores when applying European-derived scores to non-European ancestry populations in UK Biobank, the largest biobank available in the field [82]. A growing number of non-European and multi-ethnic studies and biobanks with available whole-genome/exome sequencing projects have emerged in the last decade to overcome the above limitations, including the 1000Genomes project [72], TOPmed program [73], African Pan-Genome [83], GenomeAsia 100K Project [84], and Singapore WGS project [85, 86]. Large databases aggregating data from multiple sources and ethnic groups are also being developed [87]. However, more efforts are needed to increase diversity and representation in genetic studies.

Validity of genetic ethnicity and ancestry data. Contrary to self-reported ethnicity, genetically-inferred ethnicity/ancestry can provide a biologically-based and measurable estimate of an individual's ancestry. Genetically-inferred ethnicity/ancestry can also provide crucial information that may not be available from self-reported ethnicity, especially among individuals with unknown family ancestry. This information can be informative in genetic investigations, but far less so in non-genetic research. An analysis of genetic studies recording self-reported ethnicity and a harmonized definition of self-reported and genetically-inferred ethnicity concluded that self-reported and genetically-inferred ethnicity have complementary strengths. This variable, termed HARE (harmonized ancestry and race/ethnicity), appears to be a reliable classification method in genetic associations studies [81] and uses genetically-inferred ancestry to refine self-reported race/ethnicity for genetic association studies in three ways: i. to identify individuals whose self-reported race/ethnicity is discordant with genetic ancestry, ii. to reconcile conflicts among multiple sources, and iii. to impute missing racial and ethnic information when the predictive confidence is high. Thus, when an individual's self-reported ethnicity is ambiguous, either because there are no data or inconsistent responses from multiple sources, genetic information can identify the stratum that most resembles the individual with respect to genetic ancestry as a surrogate for self-reported race or ethnicity [81]. This may be adequate for primarily genetic studies but is not recommended for social sciences and other health studies. For example, an individual living in the US with mixed ancestry for whom self-reported race or ethnicity is missing may be classified as primarily African origin by imputing their genetically-inferred ancestry. If a database combining genetic, social, and health data used genetically-inferred ancestry as a surrogate for self-reported race and ethnicity, then this individual may be grouped with others of self-reported African origin, although this may be completely erroneous, and add substantial noise to social and health-based research questions.

In 2017, a survey of clinical genetics professionals and researchers revealed that a lack of definitions of race, ethnicity, and ancestry in medical research may contribute to inconsistencies in data collection, missing or inaccurate classifications, and misleading or inconclusive results. The authors called for standardization and harmonization of race/ethnicity/ancestry data collection in clinical genetics and precision medicine research–referring to how race, ethnicity, and ancestry are perceived, defined, and measured [88].

4. Going forward: Suggestions to improve choice and reporting of terms

In 2022, it is clear that the terms race, ethnicity, and ancestry are distinct but overlapping concepts that imperfectly relate to social factors, physical features, cultural characteristics, and

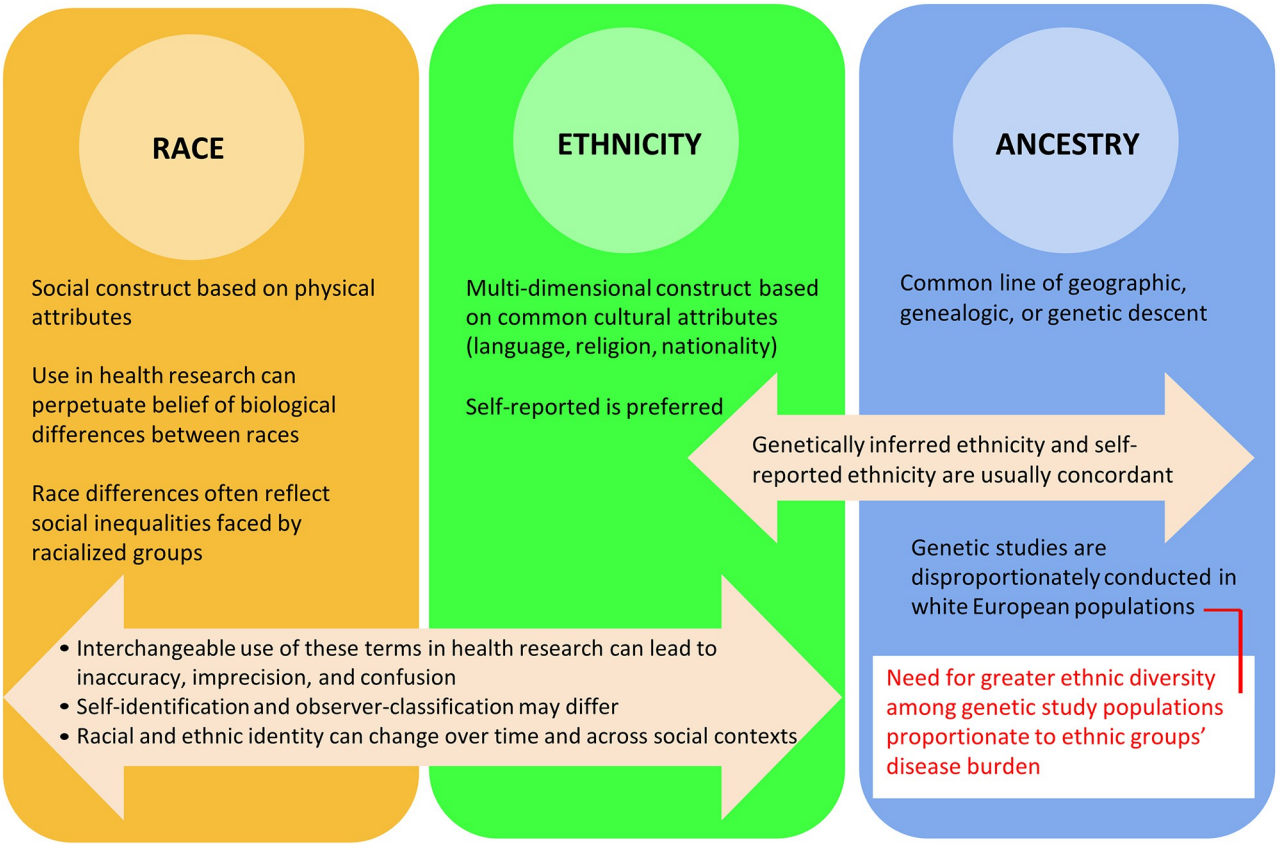

Fig 3. Race, ethnicity, and ancestry considerations in health research.

https://doi.org/10.1371/journal.pgph.0001060.g003

ancestral backgrounds. Fig 3 depicts these interrelated concepts. Race is a social construct carrying historical divisions made purely on the basis of physical characteristics, while ethnicity describes a shared cultural background which may include nationality, language, religion, and sometimes common biological characteristics. Ancestry refers to lineage and can refer to common geographic origin, genealogic, or genetic characteristics. Table 1 provides examples of research questions focused on type 2 diabetes mellitus for which using the variables of race,

Table 1. Examples of type 2 diabetes mellitus (T2DM) questions and suggested classification/terms.

	Research question	Suggested Classification/ Term*	Rationale
1	Among patients with T2DM, do referrals to diabetic care clinics differ by race?	Race	Implicit bias and structural racism may influence referral patterns.
2	What foods are associated with T2DM in a multicultural population?	Ethnicity	Ethnic groups share common cultural characteristics such as common foods and meal preparation techniques.
3	Does GLP-1 receptor agonist effectiveness in T2DM differ by ethnicity?	Ethnicity	Since race is a social and not a biological construct, ethnicity is preferred.
4	Do genetic variants for T2DM differ by ethnicity?	Ethnicity and genetically-inferred ancestry	Ethnicity and genetically inferred ancestry can be used together to compare ethnic groups.

*For all variables, self-report is preferred over observer classification.

T2DM, type 2 diabetes mellitus; GLP-1, glucagon-like peptide-1.

https://doi.org/10.1371/journal.pgph.0001060.t001

ethnicity, or ancestry may be most appropriate. Although race and ethnicity data are more reliable when self-reported by an individual, even self-reported identities can change over time depending on the individual's affiliation with their heritage community, perceptions of acceptance, advantage or disadvantage, and concerns for safety. Genetic researchers may use self-reported and/or genetically-inferred ethnicity in scenarios which best suit their research questions, but should be cautious about using genetically-inferred ethnicity alone (especially when investigating gene-environment interactions) or as a proxy for self-reported ethnicity (e.g. when addressing a social science or health services research question) [89]. Observer classification has potential harms such as misclassification, stigmatization, and perpetuating structural racism, but also potential applications–through ethnicity inferences using unique surname classification and through use of discordance between self-report and observer classification–to explore the impacts of biases in healthcare delivery. Considering all of these factors, in order to improve choice of terms and consistency of reporting, we recommend that health researchers: 1) carefully consider the objectives of the research and choose the most appropriate term (i.e. ethnicity, race, or ancestry) and use it consistently; 2) collect self-reported ethnicity, race, or ancestry as preferred over observer-assigned; 3) ask for self reported ethnicity, race or ancestry at each assessment in prospective studies; 4) use self-reported ethnicity together with genetically-inferred ancestry in genetic association studies, and 5) frame data in a manner that maximizes the potential benefit to their study participants, and carefully consider if harm could occur through stigmatization or by perpetuating racism.

Conclusion

There is no consensus definition of race or ethnicity. Every decision to collect race, ethnicity, or ancestry data in the context of health research must involve specific definitions of these terms, justification for collecting and analyzing these data, and wherever possible, prioritization of self-report over observer classification. As our understanding of these concepts evolves, their ongoing refinement is of integral importance: when used appropriately in health research, their collection and analysis can provide invaluable information for health researchers and health equity advocates.

Supporting information

S1 Table. Search strategies.
(DOCX)

References

1. Kanakamedala P, Haga SB. Characterization of clinical study populations by race and ethnicity in biomedical literature. Ethn Dis. 2012; 22(1):96–101. PMID: 22774316
2. Williams DR, Lawrence J, Davis B. Racism and Health: Evidence and Needed Research. Annu Rev Public Health. 2019 Apr 1; 40:105–25. https://doi.org/10.1146/annurev-publhealth-040218-043750 PMID: 30601726
3. Anand SS. Using ethnicity as a classification variable in health research: perpetuating the myth of biological determinism, serving socio-political agendas, or making valuable contributions to medical sciences? Ethn Health. 1999 Nov; 4(4):241–4. https://doi.org/10.1080/13557859998029 PMID: 10705561
4. Greenhalgh T, Thorne S, Malterud K. Time to challenge the spurious hierarchy of systematic over narrative reviews? Eur J Clin Invest. 2018 Jun; 48(6):e12931. https://doi.org/10.1111/eci.12931 PMID: 29578574
5. Ross PT, Hart-Johnson T, Santen SA, Zaidi NLB. Considerations for using race and ethnicity as quantitative variables in medical education research. Perspect Med Educ. 2020 Oct; 9(5):318–23. https://doi.org/10.1007/s40037-020-00602-3 PMID: 32789666

6. Ma IWY, Khan NA, Kang A, Zalunardo N, Palepu A. Systematic review identified suboptimal reporting and use of race/ethnicity in general medical journals. J Clin Epidemiol. 2007 Jun; 60(6):572–8. https://doi.org/10.1016/j.jclinepi.2006.11.009 PMID: 17493512
7. Witzig R. The medicalization of race: scientific legitimization of a flawed social construct. Ann Intern Med. 1996 Oct 15; 125(8):675–9. https://doi.org/10.7326/0003-4819-125-8-199610150-00008 PMID: 8849153
8. Canadian Institute for Health Information. Proposed Standards for Race-Based and Indigenous Identity Data Collection and Health Reporting in Canada. 2020;32.
9. Gans HJ. Racialization and racialization research. Ethnic and Racial Studies. 2017 Feb 19; 40(3):341–52.
10. Brown RA, Armelagos GJ. Apportionment of racial diversity: A review. Evolutionary Anthropology: Issues, News, and Reviews. 2001; 10(1):34–40.
11. Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, et al. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res. 2002 Apr; 12(4):602–12. https://doi.org/10.1101/gr.214902 PMID: 11932244
12. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science. 2002 Dec 20; 298(5602):2381–5. https://doi.org/10.1126/science.1078311 PMID: 12493913
13. Witherspoon DJ, Wooding S, Rogers AR, Marchani EE, Watkins WS, Batzer MA, et al. Genetic Similarities Within and Between Human Populations. Genetics. 2007 May; 176(1):351–9. https://doi.org/10.1534/genetics.106.067355 PMID: 17339205
14. Yu N, Chen FC, Ota S, Jorde LB, Pamilo P, Patthy L, et al. Larger genetic differences within africans than between Africans and Eurasians. Genetics. 2002 May; 161(1):269–74. https://doi.org/10.1093/genetics/161.1.269 PMID: 12019240
15. Bonham VL, Green ED, Pérez-Stable EJ. Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research. JAMA. 2018 Oct 16; 320(15):1533–4. https://doi.org/10.1001/jama.2018.13609 PMID: 30264136
16. Bastos JL, Peres MA, Peres KG, Dumith SC, Gigante DP. [Socioeconomic differences between self- and interviewer-classification of color/race]. Rev Saude Publica. 2008 Apr; 42(2):324–34.
17. Flanagin A, Frey T, Christiansen SL, Bauchner H. The Reporting of Race and Ethnicity in Medical and Science Journals: Comments Invited. JAMA. 2021 Mar 16; 325(11):1049–52. https://doi.org/10.1001/jama.2021.2104 PMID: 33616604
18. Jugert P, Kaiser MJ, Ialuna F, Civitillo S. Researching race-ethnicity in race-mute Europe. Infant and Child Development. 2022; 31(1):e2260.
19. Subgroup on Equality Data. ECHLG onNon-discrimination Equality and Diversity. Guidance note on the collection and use of equality data based on racial or ethnic origin. European Commission; 2021 Sep.
20. Yudell M, Roberts D, DeSalle R, Tishkoff S, 70 signatories. NIH must confront the use of race in science. Science. 2020 Sep 11; 369(6509):1313–4. https://doi.org/10.1126/science.abd4842 PMID: 32913094
21. Anand SS, Yi Q, Gerstein H, Lonn E, Jacobs R, Vuksan V, et al. Relationship of metabolic syndrome and fibrinolytic dysfunction to cardiovascular disease. Circulation. 2003 Jul 29; 108(4):420–5. https://doi.org/10.1161/01.CIR.0000080884.27358.49 PMID: 12860914
22. Pearson JA, Geronimus AT. Race/ethnicity, socioeconomic characteristics, coethnic social ties, and health: evidence from the national Jewish population survey. Am J Public Health. 2011 Jul; 101(7):1314–21. https://doi.org/10.2105/AJPH.2009.190462 PMID: 21164093
23. Ford ME, Kelly PA. Conceptualizing and Categorizing Race and Ethnicity in Health Services Research. Health Serv Res. 2005 Oct; 40(5 Pt 2):1658–75. https://doi.org/10.1111/j.1475-6773.2005.00449.x PMID: 16179001
24. Sankar P, Cho MK, Mountain J. Race and Ethnicity in Genetic Research. Am J Med Genet A. 2007 May 1; 143(9):961–70. https://doi.org/10.1002/ajmg.a.31575 PMID: 17345638
25. Sheldon TA, Parker H. Race and ethnicity in health research. J Public Health Med. 1992 Jun; 14(2):104–10. PMID: 1515192
26. Senior PA, Bhopal R. Ethnicity as a variable in epidemiological research. BMJ. 1994 Jul 30; 309(6950):327–30. https://doi.org/10.1136/bmj.309.6950.327 PMID: 8086873
27. Ali-Khan SE, Krakowski T, Tahir R, Daar AS. The use of race, ethnicity and ancestry in human genetic research. Hugo J. 2011 Dec; 5(1–4):47–63. https://doi.org/10.1007/s11568-011-9154-5 PMID: 22276086

28. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature. 2009 May 28; 459(7246):528–33. https://doi.org/10.1038/nature07999 PMID: 19404256

29. Treviño LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet. 2009 Sep; 41(9):1001–5. https://doi.org/10.1038/ng.432 PMID: 19684603

30. Yeager M, Chatterjee N, Ciampa J, Jacobs KB, Gonzalez-Bosquet J, Hayes RB, et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. Nat Genet. 2009 Oct; 41(10):1055–7. https://doi.org/10.1038/ng.444 PMID: 19767755

31. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature. 2009 May 28; 459(7246):569–73. https://doi.org/10.1038/nature07953 PMID: 19404257

32. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science. 2008 Apr 25; 320(5875):539–43. https://doi.org/10.1126/science.1155174 PMID: 18369103

33. Nath SK, Han S, Kim-Howard X, Kelly JA, Viswanathan P, Gilkeson GS, et al. A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. Nat Genet. 2008 Feb; 40(2):152–4. https://doi.org/10.1038/ng.71 PMID: 18204448

34. Perry AC, Rosenblatt EB, Wang X. Physical, behavioral, and body image characteristics in a tri-racial group of adolescent girls. Obes Res. 2004 Oct; 12(10):1670–9. https://doi.org/10.1038/oby.2004.207 PMID: 15536231

35. Neal KC. Use and Misuse of 'Race' in Biomedical Research. Journal of Health Ethics [Internet]. 2008 [cited 2021 Dec 5]; 5(1). Available from: https://aquila.usm.edu/ojhe/vol5/iss1/8/

36. Sankar P, Cho MK, Monahan K, Nowak K. Reporting Race and Ethnicity in Genetics Research: Do Journal Recommendations or Resources Matter? Sci Eng Ethics. 2015 Oct; 21(5):1353–66. https://doi.org/10.1007/s11948-014-9596-y PMID: 25407312

37. Exner DV, Dries DL, Domanski MJ, Cohn JN. Lesser response to angiotensin-converting-enzyme inhibitor therapy in black as compared with white patients with left ventricular dysfunction. N Engl J Med. 2001 May 3; 344(18):1351–7. https://doi.org/10.1056/NEJM200105033441802 PMID: 11333991

38. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, et al. New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race. N Engl J Med. 2021 Nov 4; 385(19):1737–49. https://doi.org/10.1056/NEJMoa2102953 PMID: 34554658

39. Schulman KA, Berlin JA, Harless W, Kerner JF, Sistrunk S, Gersh BJ, et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. N Engl J Med. 1999 Feb 25; 340(8):618–26. https://doi.org/10.1056/NEJM199902253400806 PMID: 10029647

40. Brett AS, Goodman CW. First Impressions—Should We Include Race or Ethnicity at the Beginning of Clinical Case Presentations? N Engl J Med. 2021 Dec 30; 385(27):2497–9. https://doi.org/10.1056/NEJMp2112312 PMID: 34951753

41. Anand SS, Arnold C, Bangdiwala S, Bolotin S, Bowdish D, Chanchlani R, et al. What factors converged to create a COVID-19 hot-spot? Lessons from the South Asian community in Ontario [Internet]. medRxiv; 2022 [cited 2022 Jun 12]. p. 2022.04.01.22273252. Available from: https://www.medrxiv.org/content/10.1101/2022.04.01.22273252v1

42. Shah BR, Chiu M, Amin S, Ramani M, Sadry S, Tu JV. Surname lists to identify South Asian and Chinese ethnicity from secondary data in Ontario, Canada: a validation study. BMC Med Res Methodol. 2010 May 15; 10:42. https://doi.org/10.1186/1471-2288-10-42 PMID: 20470433

43. Sheth T, Nargundkar M, Chagani K, Anand S, Nair C, Yusuf S. Classifying ethnicity utilizing the Canadian Mortality Data Base. Ethn Health. 1997 Nov; 2(4):287–95. https://doi.org/10.1080/13557858.1997.9961837 PMID: 9526691

44. Bailey SR, Loveman M, Muniz JO. Measures of "Race" and the analysis of racial inequality in Brazil. Soc Sci Res. 2013 Jan; 42(1):106–19. https://doi.org/10.1016/j.ssresearch.2012.06.006 PMID: 23146601

45. Jones CP, Truman BI, Elam-Evans LD, Jones CA, Jones CY, Jiles R, et al. Using "socially assigned race" to probe white advantages in health status. Ethn Dis. 2008; 18(4):496–504. PMID: 19157256

46. Perreira KM, Telles EE. The color of health: skin color, ethnoracial classification, and discrimination in the health of Latin Americans. Soc Sci Med. 2014 Sep; 116:241–50. https://doi.org/10.1016/j.socscimed.2014.05.054 PMID: 24957692

47. Sohn MW, Zhang H, Arnold N, Stroupe K, Taylor BC, Wilt TJ, et al. Transition to the new race/ethnicity data collection standards in the Department of Veterans Affairs. Popul Health Metr. 2006 Jul 6; 4:7. https://doi.org/10.1186/1478-7954-4-7 PMID: 16824220

48. Telles EE, Lim N. Does it matter who answers the race question? Racial classification and income inequality in Brazil. Demography. 1998 Nov; 35(4):465–74. PMID: 9850470

49. Janka EA, Vincze F, Ádány R, Sándor J. Is the Definition of Roma an Important Matter? The Parallel Application of Self and External Classification of Ethnicity in a Population-Based Health Interview Survey. Int J Environ Res Public Health. 2018 Feb 16; 15(2):E353. https://doi.org/10.3390/ijerph15020353 PMID: 29462940

50. Warren RC. Use of Race and Ethnicity in Public Health Surveillance. Summary of the DC/ATSDR workshop. Centers for Disease Control and Prevention; 1993 Mar p. 1–16. (MMWR. Morbidity and mortality weekly report.).

51. Williams DR. Race/ethnicity and socioeconomic status: measurement and methodological issues. Int J Health Serv. 1996; 26(3):483–505. https://doi.org/10.2190/U9QT-7B7Y-HQ15-JT14 PMID: 8840198

52. Hahn RA, Truman BI, Barker ND. Identifying ancestry: The reliability of ancestral identification in the United States by self, proxy, interviewer, and funeral director. Epidemiology. 1996 Jan; 7(1):75–80. https://doi.org/10.1097/00001648-199601000-00013 PMID: 8664405

53. Saperstein A, Penner AM. Racial Fluidity and Inequality in the United States. 1 | American Journal of Sociology. 2012; 118(3):676–727.

54. Saunders CL, Abel GA, El Turabi A, Ahmed F, Lyratzopoulos G. Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the English Cancer Patient Experience survey. BMJ Open. 2013 Jun 28; 3(6):e002882. https://doi.org/10.1136/bmjopen-2013-002882 PMID: 23811171

55. Boehmer U, Kressin NR, Berlowitz DR, Christiansen CL, Kazis LE, Jones JA. Self-reported vs administrative race/ethnicity data and study results. Am J Public Health. 2002 Sep; 92(9):1471–2. https://doi.org/10.2105/ajph.92.9.1471 PMID: 12197976

56. Kaplan JB, Bennett T. Use of race and ethnicity in biomedical publication. JAMA. 2003 May 28; 289 (20):2709–16. https://doi.org/10.1001/jama.289.20.2709 PMID: 12771118

57. Kressin NR, Chang BH, Hendricks A, Kazis LE. Agreement between administrative data and patients' self-reports of race/ethnicity. Am J Public Health. 2003 Oct; 93(10):1734–9. https://doi.org/10.2105/ajph.93.10.1734 PMID: 14534230

58. Liebler CA, Porter SR, Fernandez LE, Noon JM, Ennis SR. America's Churning Races: Race and Ethnicity Response Changes Between Census 2000 and the 2010 Census. Demography. 2017 Feb; 54(1):259–84. https://doi.org/10.1007/s13524-016-0544-0 PMID: 28105578

59. DeFrank JT, Bowling JM, Rimer BK, Gierisch JM, Skinner CS. Triangulating differential nonresponse by race in a telephone survey. Prev Chronic Dis. 2007 Jul; 4(3):A60. PMID: 17572964

60. Perez AD, Hirschman C. The Changing Racial and Ethnic Composition of the US Population: Emerging American Identities. Popul Dev Rev. 2009 Mar; 35(1):1–51. https://doi.org/10.1111/j.1728-4457.2009.00260.x PMID: 20539823

61. Johnson CE. Consistency of Reporting Ethnic Origin in the Current Population Survey. Washington, D.C.: Bureau of the Census, U.S. Department of Commerce; 1974. Report No.: 31.

62. Sondik EJ, Lucas JW, Madans JH, Smith SS. Race/ethnicity and the 2000 census: implications for public health. Am J Public Health. 2000 Nov; 90(11):1709–13. https://doi.org/10.2105/ajph.90.11.1709 PMID: 11076236

63. del Pinal JH. Race and ethnicity in Census 2000. Washington, D.C.: United States Census Bureau; 2004. (Census 2000). Report No.: 9.

64. Lieberson S, Waters MC. The Ethnic Responses of Whites: What Causes Their Instability, Simplification, and Inconsistency? Social Forces. 1993; 72(2):421–50.

65. Rezai MR, Maclagan LC, Donovan LR, Tu JV. Classification of Canadian immigrants into visible minority groups using country of birth and mother tongue. Open Med. 2013 Oct 1; 7(4):e85–93. PMID: 25237404

66. Harris DR, Sim JJ. Who Is Multiracial? Assessing the Complexity of Lived Race. American Sociological Review. 2002; 67(4):614–27.

67. Miville ML, Constantine MG, Baysden MF, So-Lloyd G. Chameleon Changes: An Exploration of Racial Identity Themes of Multiracial People. Journal of Counseling Psychology. 2005; 52(4):507–16.

68. McKay DL. Real Indians: Policing or Protecting Authentic Indigenous Identity? Sociology of Race and Ethnicity. 2021 Jan 1; 7(1):12–25.

69. Ford CL, Harawa NT. A new conceptualization of ethnicity for social epidemiologic and health equity research. Soc Sci Med. 2010 Jul; 71(2):251–8. https://doi.org/10.1016/j.socscimed.2010.04.008 PMID: 20488602

70. Afshari R, Bhopal RS. Ethnicity has overtaken race in medical science: MEDLINE-based comparison of trends in the USA and the rest of the world, 1965–2005. Int J Epidemiol. 2010 Dec; 39(6):1682–3. https://doi.org/10.1093/ije/dyp382 PMID: 20089694

71. Borrell LN, Elhawary JR, Fuentes-Afflick E, Witonsky J, Bhakta N, Wu AHB, et al. Race and Genetic Ancestry in Medicine—A Time for Reckoning with Racism. N Engl J Med. 2021 Feb 4; 384(5):474–80. https://doi.org/10.1056/NEJMms2029562 PMID: 33406325

72. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245

73. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021; 590(7845):290–9. https://doi.org/10.1038/s41586-021-03205-y PMID: 33568819

74. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006 Dec; 2(12): e190. https://doi.org/10.1371/journal.pgen.0020190 PMID: 17194218

75. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet Epidemiol. 2015 May; 39(4):276–93. https://doi.org/10.1002/gepi.21896 PMID: 25810074

76. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep; 19(9):1655–64. https://doi.org/10.1101/gr.094052.109 PMID: 19648217

77. Wang H, Sofer T, Zhang X, Elston RC, Redline S, Zhu X. Local Ancestry Inference in Large Pedigrees. Sci Rep. 2020 Jan 13; 10:189. https://doi.org/10.1038/s41598-019-57039-w PMID: 31932708

78. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug; 38(8):904–9. https://doi.org/10.1038/ng1847 PMID: 16862161

79. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4:7. https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852

80. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000 Jun; 155(2):945–59. https://doi.org/10.1093/genetics/155.2.945 PMID: 10835412

81. Fang H, Hui Q, Lynch J, Honerlaw J, Assimes TL, Huang J, et al. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. Am J Hum Genet. 2019 Oct 3; 105(4):763–72. https://doi.org/10.1016/j.ajhg.2019.08.012 PMID: 31564439

82. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019 Apr; 51(4):584–91. https://doi.org/10.1038/s41588-019-0379-x PMID: 30926966

83. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019 Jan; 51(1):30–5. https://doi.org/10.1038/s41588-018-0273-y PMID: 30455414

84. GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature. 2019 Dec; 576(7785):106–11. https://doi.org/10.1038/s41586-019-1793-z PMID: 31802016

85. Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, et al. Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. Cell. 2019 Oct 17; 179(3):736–749.e15. https://doi.org/10.1016/j.cell.2019.09.019 PMID: 31626772

86. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020 Mar 20; 367(6484): eaay5012. https://doi.org/10.1126/science.aay5012 PMID: 32193295

87. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May; 581(7809):434–43. https://doi.org/10.1038/s41586-020-2308-7 PMID: 32461654

88. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016 Oct; 538(7624):161–4. https://doi.org/10.1038/538161a PMID: 27734877

89. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Hum Genomics. 2015 Jan 7; 9:1. https://doi.org/10.1186/s40246-014-0023-x PMID: 25563503