Tracking the History of Race and Ethnicity Data Collection in Genomics Research

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Peneeta Wojcik
Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor
Kent Wayland, Department of Engineering and Society

Introduction

Prior to the Human Genome Project, it was widely believed that quantifiable genetic differences existed between different races. Its results instead showed remarkable human genetic similarity, with 99% of the human genome being identical and only 0.1% of the variation attributed to phenotypic differences (International Human Genome Sequencing Consortium, 2001). This miniscule amount of variation supported that race had no biological basis. Despite these findings, race, ethnicity, and ancestry (REA) data are widely used in conjunction with cell samples in genomic research.

Using REA data in biological research perpetuates harmful claims about certain racial groups. Race-as-biology ideologies are unconsciously embedded in many students and clinicians alike, many holding false beliefs about biological differences between races. In a recent survey involving 222 white American medical students and residents, 40% of first and second year medical students thought that the skin of a Black person was thicker than a white person, which is biologically untrue (Lujan & DiCarlo, 2024). These incorrect beliefs are concerning, especially when expressed by future medical staff, which adversely affects how treatment is administered to different racial groups. Statistics like these beg the question as to why REA data is still widely used in biological research today.

Government policies on REA data collection and use are major contributing factors to the use of racial data in conjunction with cell samples and clinical subjects in genomics. *The goal of this study is to paint the picture of mutual shaping of REA data collection policies, scientific advancements, and societal viewpoints.* To accomplish this, three policies on REA data collection and reporting, one issued by the United States Office of Management and Budget, one

from the Food and Drug Administration (FDA) and the other from the National Institutes of Health (NIH), will be examined chronologically. This analysis will be used to highlight that race is a social construct, not biological, and recommend changes to current policies in favor of specificity and standardization.

Background and Context

Deoxyribonucleic Acid (DNA), a set of genetic instructions comprised of a sequence of nucleotides, fundamentally changed the way researchers understood the human body. Once the structure of DNA was discovered in the 1950s (Pattan et al., 2021), a multitude of questions arose: what does it code for, and does its sequence differ between individuals? This resulted in the birth of genomics, or the study of the structure, function, editing, and mapping of the human genome. The field began as highly specialized, but with the introduction of high-throughput sequencing and messenger RNA vaccines, many began to realize the clinical potential of genomics (Pattan et al., 2021).

It is a common misconception that all science, including genomics, is independent of societal opinions and biases. The United States government has sought to understand its population by categorizing individuals based on race and ethnicity through the census, the first of which was collected in the year 1790 (Race, Ethnicity, and Genetics Working Group, 2005). This long-standing use of racial categorization bled into healthcare and biomedical research. Before modern methods of genomic sequencing, scientists and clinicians relied on physical appearance to group individuals together as a means of convenience without presenting quantitative evidence that this method was valid (Cooper, 2013). Marginalization and incorrect beliefs of certain races, especially during the eugenics movement, fueled race-as-biology ideals. High-profile scientists

during the time authored articles and books claiming human races had fundamental genetic differences (Roberts, 2011). An example of this is shown in *The Origin of Races* published in 1962 by University of Pennsylvania professor Carleton Coon. He wrote that racial subgroups of the human population evolved at different rates and were distinct biologically from one another due to skull shape and capacity, asserting that certain racial groups were more civilized than others (Jackson, 2001).

The Human Genome Project held much anticipation from researchers and the public alike. Many wondered if racial differences would be seen within genetic code and whether the claims that certain races were genetically inferior, or superior, were true. The results of the project demonstrated that all humans were remarkably similar. With 99% genetic similarity, it was finally quantifiably shown that race was not a valid biological construct at all, despite the claims that have been made for centuries (International Human Genome Sequencing Consortium, 2001). Many researchers envisioned a post-racial society free of prejudice in discrimination given its lack of biological validity. This has not been achieved. Race is still used in clinical and research settings today, leading people to believe that it does have biological validity (Schaare et al., 2023).

What distinguishes "race", "ethnicity", and "ancestry"?

Currently, genomics researchers and clinicians use REA data as diversity measures (Popejoy et al., 2020). The main issues are the lack of standardization of REA data collection and the interchangeable uses of the terms "race", "ethnicity", and "ancestry" in scientific literature (Race, Ethnicity, and Genetics Working Group, 2005). There is no clear consensus on the exact definitions of these terms (Lewis et al., 2023). Some researchers reject race as a concept

altogether, arguing many races only differ by one or two genes (Roberts, 2011). Others view race as a purely political category, and instead use ancestry in its place, because it implies inheritance of traits from a person's lineage (Roberts, 2011). To convolute these definitions further, some scientists have redefined race to represent genetic differences due to evolutionary pressure, coinciding with the previous definitions of ancestry (Baye & Wilke, 2010). This lack of agreement contributes to clinical misdiagnoses, inaccurate classification of cell samples or patients, and misleading results.

What does diversity mean?

Diversity has a different definition to different researchers. Genomic researchers are interested in genetic diversity, or finding regions in the genome that differ between disease states and normal states to develop new therapies (Roth, 2019). These differences are on the order of single base pairs throughout the genome. Clinical researchers are interested in diversity to ensure that medications or devices do not adversely affect certain members of the population (National Institute on Minority Health and Health Disparities, 2024). The metrics they are interested in are age, race, ethnicity, sex, or presence of comorbidities (Office of the Commissioner, 2023). A study conducted by the NIH shows that 85% of genetic diversity occurs within racial groups, while only 15% occurs between racial groups (Health (US) & Study, 2007). This evidence completely rejects race as a valid biological construct and renders it meaningless for measuring diversity. If this is the case, why is this data still being collected in clinical research?

Completely disregarding REA may have negative consequences. REA categories are a valid social construct, and using these categories demonstrate the effects of sociocultural status or marginalization. Most chronic diseases are the result of environmental impacts, which are

closely tied to socioeconomic status, geographic region, and a multitude of other factors. For example, cardiovascular disease (CVD) is significantly higher in low-income populations due to the absence of healthy food options and increased stress, and these populations mainly contain racial and ethnic minorities (Minhas et al., 2023). This does not mean that racial and ethnic minorities are biologically predisposed to CVD; the conclusion is that marginalization and socioeconomic status are the sole drivers of CVD. REA data retains significance as a way to demonstrate social effects contributing to disease (Cooper, 2013).

Other researchers have brought up the issue of representation of genetic samples in online databases. These databases are commonly used in current computational research and contain data that is heavily biased towards certain population groups. As of 2016, 81% of samples available for genome-wide association database were of European descent (Popejoy & Fullerton, 2016). Some researchers argue that removing REA considerations may hinder progress in increasing representation in these studies. Given that REA categories are social constructs and not biological, there is an apparent contradiction in this argument. The purpose of genome-wide association studies is to test genetic mutations across thousands of genomes to find mutations associated with certain diseases (Uffelmann et al., 2021). REA categories are purely social and are not accurate measures of diversity. They cannot be used to increase representation in these databases without simultaneously perpetuating race-as-biology ideologies.

Clinical decision-making versus scientific accuracy

Clinicians argue that there is a clear distinction between using REA data in clinical settings versus research settings. To gauge professional opinions on REA data in genetics, researchers from various medical institutions in the United States conducted a qualitative content

analysis of a survey sent to geneticists (Catherine Nelson et al., 2018). A genetic epidemiologist respondent wrote: "*I have somehow managed to hold seemingly mutually exclusive views that 1) races don't exist and are biologically meaningless and 2) races have a genetic basis and biological influences on health*". A clinical geneticist contrasted this, writing: "*As a physician and public health professional knowing about the geographic/ ancestry/ "ethnic" group /"whatever you want to call it" is useful in providing appropriate care and services to specific populations"* (Catherine Nelson et al., 2018). These two statements highlight the discrepancy between adequate clinical decision-making and scientific accuracy. Though not a biological category, REA data is used to estimate societal factors that may contribute to biological outcomes.

Despite the claim that REA data should be used as a proxy for the effects of societal factors in medicine, using race in clinical settings brings up more disturbing issues. Medical students are taught to take their patient's race into account when prescribing treatments or performing procedures (Roberts, 2011). In a review of 15 studies, researchers documented that racial biases were present during treatment decisions and patient-provider interactions, indicating that a large number of practicing physicians and medical students held incorrect beliefs of biological racial differences (Lujan & DiCarlo, 2024). This is especially noticeable when physicians prescribe medication for pain. Dr. Knox Todd, a doctor at Emory University School of Medicine, examined this discrepancy in the 1990s. 217 emergency room patients were analyzed, 127 black and 90 white, who had long bone fractures. He found that black patients were 66% less likely to be prescribed pain medication than white patients despite both black and white patients complaining of pain at the same rate (Todd et al., 2000). The belief that black patients have a higher pain tolerance than white patients is one of the many myths that contribute to

unfair health practices. If race is not biological, then patients should be treated based on their individual response, not their race.

Methods

REA guidelines, heavily influenced by societal opinions and biological advancements, are instantiated by the federal government. It is apparent that current definitions of race, ethnicity, and ancestry are blurred and a better way to represent human variability has not yet been devised. The key players in this system are societal views of race and ethnicity, genomics researchers using this data in studies, the funding institutions supporting the research such as the FDA and NIH, and the federal government creating these standards.

The primary federal REA data collection policy was the Office of Management and Budget (OMB) federal directive 15, which was enacted in 1977 (U.S. Office of Management and Budget, 2023). The FDA and NIH primarily base their REA data collection guidelines on this directive (Office of the Commissioner, 2023). The changes to this policy will be examined chronologically, as well as resulting amendments to FDA and NIH REA data collection policies. To track changes to this policy through time, the Code of Federal Regulations and entries in the Federal Register pertaining to OMB directive 15 will be used. Analyzing these guidelines will highlight the mutual shaping of scientific advancements and changing ideologies surrounding REA on policy guidelines.

Results

First Iterations of OMB Directive 15

REA data collection standards were first established in 1977, when the OMB federal directive 15 was passed. At the time, the sociopolitical climate heavily involved the Civil Rights Movement, and with that, the task of ensuring affirmative action to combat existing race and gender discrimination (Clinton White House Archives, n.d.). The goal of this directive was to ensure consistency in REA data collection and reporting (Schaare et al., 2023). In a clause dedicated to statistical reporting, the initial directive states that the racial and ethnic categories denoted are not necessary "*when the collection effort focuses on a specific racial or ethnic group*" and "*reporting…which uses more detail shall be organized…such that the additional categories can be aggregated into these basic racial/ethnic categories*" (U.S. Office of Management and Budget, 1977). The directive also adds *"in no case should the provisions of this Directive be construed to limit the collection of data"* (U.S. Office of Management and Budget, 1977), however the reporting categories and simplicity of them are extremely limited when applied to genomics research. One could argue that this document is supposed to be standardized across government entities and is not specific to research, however the NIH and FDA are federally funded agencies, and both follow the standards presented in this directive.

It was not until 1997, when Directive 15 was updated, that the Federal Government addressed the incorrect race-as-biology argument. An additional clause was added in the fine print enumerating the review process, stating *"racial and ethnic categories set forth in the standard should not be interpreted as being primarily biological or genetic in reference"* (U.S. Office of Management and Budget, 1997). More specific REA categories were added as well in response to the criticism of the initial categories, which many argued did *"not reflect the diversity of our Nation's population"* (U.S. Office of Management and Budget, 1997). Adding this distinction of race and biology was a large stride from the previous iteration; however, it did

not have immediate implications in research. This is notable in a study published in 2000 examining differences in birth weight between African American and Caucasian populations. In this study, researchers concluded there must be strong genetic differences between both groups and suggested additional research into understanding genetic differences (Frank, 2001). Despite these researchers incorrectly concluding that genetic differences existed between races, it was apparent that health disparities existed between them based on societal factors.

NIH Response

In 2001, the NIH updated its policy on reporting of REA data in clinical research in response to the 1997 OMB directive. They state that the OMB directive describes the minimum standards for data reporting, and further reinforce that *"categories in this classification are social-political constructs and should not be interpreted as being anthropological in nature"* (National Institutes of Health, 2001). This policy requires REA data collection for all clinical trials. The NIH further defines clinical research as any study conducted with human subjects or on material of human origin such as tissue samples. *In vitro* studies that use human tissues not directly linked to a living individual are excluded from this definition. The policy also encourages researchers to collect additional data *"that will provide additional insights into the relationships between race and ethnicity and health"* (National Institutes of Health, 2001). These guidelines, though more specific to clinical research, still assert that racial and ethnic differences play a role in clinical health outcomes. By encouraging researchers to extract insights into relationships between REA data and health outcomes, it encourages REA data to be used as a standalone entity, not tied to any societal factors. This policy does not make any strides to

address racial inequalities in clinical treatment and prescriptions, instead supporting the stratification of clinical trial participants by race.

FDA REA Collection Guidelines

How have directive 15 revisions and scientific advancements impacted REA data collection by the FDA? In a guide for FDA Staff in 2016, they provide two reasons why they recommend the use of OMB REA categories in clinical trials. Using REA data helps *"evaluate potential differences"* in drug effectiveness and safety across population groups and allows for demographic subset analyses due to the consistency of REA labels (Office of the Commissioner, 2016). The FDA further states that differential responses to drugs have been "*observed in racially and ethnically distinct subgroups*", and that they may be attributed to "*intrinsic factors (e.g. genetics, metabolism, elimination) extrinsic factors (e.g., diet, environmental exposure, sociocultural issues), or interactions between these factors*" (Office of the Commissioner, 2016). This passage suggests interplay between extrinsic and intrinsic factors, further distancing ideologies involving genetic differences between races. A subsequent clause in this document highlights federal REA categories as *"social-political constructs"* and how they *"should not be interpreted as being scientific or anthropological in nature"* (Office of the Commissioner, 2016).

Recent Updates to Directive 15

OMB Directive 15 has last been revised in March of 2024 to supersede the 1997 revision, mostly focused on adding specificity to race and ethnicity categories. In the revision, the clause reinforcing the sociopolitical construct of race is located at the beginning of the directive. The new directive also states that *"the standards do not require any agency or program to collect*

race and ethnicity data" (U.S. Office of Management and Budget, 2024), only to act as a guide for data standardization. Other updates include adding Middle Eastern or North African categories separate from the existing White category (U.S. Office of Management and Budget, 2024).

Discussion

REA data collection and reporting is still federally required for clinical trials based on the NIH and FDA guidelines above. These policies were enacted to combat racial health disparities in the United States by ensuring clinical trials are balanced across racial groups, making sure certain groups are not over or underrepresented. This idea of equal representation is still based on REA categories, which have no biological meaning and cannot be used to measure diversity. Using REA categories as proxy for social effects results in harmful generalizations. For example, the Office of Minority Health publishes population statistics comparing health outcomes of different racial and ethnic groups (U.S. Office of Minority Health, n.d.). These statistics are presented using race and ethnicity as standalone categories, which leads individuals to incorrectly believe that biology is to blame for the disease occurrence rather than social or environmental factors.

The NIH and FDA policies do not address racial discrepancies in clinical treatment; if anything, the requirement for REA data collection encourages finding racial differences in clinical treatment responses. The assertion that certain races react differently to treatment perpetuates biological differences between races; instead, treatment should occur at the individual level. Interestingly, the most recent update to OMB Directive 15 does not require any federal agency or program to collect REA data anymore. This could indicate that in the future,

federal agencies will not collect REA data for clinical trials and instead use other metrics centered around quantifiable socioeconomic factors or genetic factors. This is a small step in the direction of a society that disregards race in the clinical environment.

Conclusion

The ease and availability of genetic sequencing is rapidly expanding. As this trajectory continues, there will be no need to stratify humans based on phenotype. The use of REA data and data collection standards are a remnant of historical attitudes towards race and ethnicity, however, have been considerably modified since their inception. OMB Directive 15 remains the main policy governing federal research agencies. This has been modified through time to add more specific categories for REA self-identification, reflecting the increasing diversity in the United States. Health disparities exist in the United States between population groups, but it must be clearly shown that they are due to social and environmental factors and not racial or ethnic background. The FDA still recommends using REA in studies for lack of a better measure of social factors and often presents clinical data using REA as a feature.

Perhaps, in the future, we will achieve a kind of post-genomic society where more quantifiable ways have been developed to accurately measure human genetic diversity. This is promising with the latest OMB Directive 15 revision, but it is still too early for other research agencies to respond. One thing is clear: genomics research cannot be viewed as separate from the individuals that cell samples or DNA originate from and cannot be separated from federal data collection standards.

References

Baye, T. M., & Wilke, R. A. (2010). Mapping Genes that Predict Treatment Outcome in Admixed Populations. *The Pharmacogenomics Journal*, *10*(6), 465–477. https://doi.org/10.1038/tpj.2010.71

Catherine Nelson, S., Yu, J.-H., Wagner, J. K., Harrell, T. M., Royal, C. D., & Bamshad, M. J. (2018). A content analysis of the views of genetics professionals on race, ancestry, and genetics. *AJOB Empirical Bioethics*, *9*(4), 222–234. https://doi.org/10.1080/23294515.2018.1544177

Clinton White House Archives. (n.d.). *2. AFFIRMATIVE ACTION: HISTORY AND RATIONALE*. Clinton White House Archives. Retrieved April 4, 2024, from https://clintonwhitehouse3.archives.gov/WH/EOP/OP/html/aa/aa02.html

Cooper, R. S. (2013). Race in Biological and Biomedical Research. *Cold Spring Harbor Perspectives in Medicine*, *3*(11), a008573. https://doi.org/10.1101/cshperspect.a008573

Frank, R. (2001). The misuse of biology in demographic research on racial/ethnic differences: A reply to van den Oord and Rowe. *Demography*, *38*(4), 563–567. https://doi.org/10.1353/dem.2001.0034

Health (US), N. I. of, & Study, B. S. C. (2007). Understanding Human Genetic Variation. In *NIH Curriculum Supplement Series [Internet]*. National Institutes of Health (US). https://www.ncbi.nlm.nih.gov/books/NBK20363/

International Human Genome Sequencing Consortium. (2001). *Initial sequencing and analysis of the human genome*. https://www.nature.com/articles/35057062

Jackson, J. P. (2001). “In Ways Unacademical”: The Reception of Carleton S. Coon’s “The Origin of Races.” *Journal of the History of Biology*, *34*(2), 247–285.

Lewis, C., Cohen, P. R., Bahl, D., Levine, E. M., & Khaliq, W. (2023). Race and Ethnic Categories: A Brief Review of Global Terms and Nomenclature. *Cureus*, *15*(7), e41253. https://doi.org/10.7759/cureus.41253

Lujan, H. L., & DiCarlo, S. E. (2024). Misunderstanding of race as biology has deep negative biological and social consequences. *Experimental Physiology*, *n/a*(n/a). https://doi.org/10.1113/EP091491

Minhas, A. M. K., Jain, V., Li, M., Ariss, R. W., Fudim, M., Michos, E. D., Virani, S. S., Sperling, L., & Mehta, A. (2023). Family income and cardiovascular disease risk in American adults. *Scientific Reports*, *13*(1), 279. https://doi.org/10.1038/s41598-023-27474-x

National Institute on Minority Health and Health Disparities. (2024). *Diversity and Inclusion in Clinical Trials*. NIMHD. https://nimhd.nih.gov/resources/understanding-health-disparities/diversity-and-inclusion-in-clinical-trials.html

National Institutes of Health. (2001, August 8). *NOT-OD-01-053: NIH Policy on Reporting Race and Ethnicity Data: Subjects in Clinical Research*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-01-053.html

Office of the Commissioner. (2016, October). *Collection of Race and Ethnicity Data in Clinical Trials*. FDA. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/collection-race-and-ethnicity-data-clinical-trials

Office of the Commissioner. (2023, August 10). *Collection of Race and Ethnicity Data in Clinical Trials*. FDA. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/collection-race-and-ethnicity-data-clinical-trials

Pattan, V., Kashyap, R., Bansal, V., Candula, N., Koritala, T., & Surani, S. (2021). Genomics in medicine: A new era in medicine. *World Journal of Methodology*, *11*(5), 231–242. https://doi.org/10.5662/wjm.v11.i5.231

Popejoy, A. B., Crooks, K. R., Fullerton, S. M., Hindorff, L. A., Hooker, G. W., Koenig, B. A., Pino, N., Ramos, E. M., Ritter, D. I., Wand, H., Wright, M. W., Yudell, M., Zou, J. Y., Plon, S. E., Bustamante, C. D., & Ormond, K. E. (2020). Clinical Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures. *American Journal of Human Genetics*, *107*(1), 72–82. https://doi.org/10.1016/j.ajhg.2020.05.005

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. https://doi.org/10.1038/538161a

Race, Ethnicity, and Genetics Working Group. (2005). The Use of Racial, Ethnic, and Ancestral Categories in Human Genetics Research. *American Journal of Human Genetics*, *77*(4), 519–532.

Roberts, D. (2011). *Fatal Invention: How Science, Politics, and Big Business Re-Create Race in the Twenty-First Century*. New Press, The. http://ebookcentral.proquest.com/lib/uva/detail.action?docID=729441

Roth, S. C. (2019). What is genomic medicine? *Journal of the Medical Library Association : JMLA*, *107*(3), 442–448. https://doi.org/10.5195/jmla.2019.604

Schaare, D., Abenavoli, L., & Boccuto, L. (2023). Race: How the Post-Genomic Era Has Unmasked a Misconception Promoted by Healthcare. *Medicina*, *59*(5), 861. https://doi.org/10.3390/medicina59050861

Todd, K. H., Deaton, C., D'Adamo, A. P., & Goe, L. (2000). Ethnicity and analgesic practice. *Annals of Emergency Medicine*, *35*(1), 11–16. https://doi.org/10.1016/S0196-0644(00)70099-0

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1–21. https://doi.org/10.1038/s43586-021-00056-9

U.S. Office of Management and Budget. (1977). *OMB Directive 15: Race and Ethnic Standards for Federal Statistics and Administrative Reporting*. https://wonder.cdc.gov/wonder/help/populations/bridged-race/directive15.html

U.S. Office of Management and Budget. (1997). *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity*. The White House. https://obamawhitehouse.archives.gov/node/15626

U.S. Office of Management and Budget. (2023, January 25). *History of Statistical Policy Directive No. 15*. https://spd15revision.gov/content/spd15revision/en/history.html

U.S. Office of Management and Budget. (2024, March 29). *Revisions to OMB's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity*. Federal Register. https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and

U.S. Office of Minority Health. (n.d.). *Heart Disease and African Americans*. Retrieved April 4, 2024, from https://minorityhealth.hhs.gov/heart-disease-and-african-americans