REAGAN-UDALL FOUNDATION FOR THE FDA

RAISE
REAL-WORLD ACCELERATOR TO IMPROVE THE STANDARD OF COLLECTION AND CURATION OF RACE AND ETHNICITY DATA IN HEALTHCARE

Advanced Analytics - Interim Solutions When Race & Ethnicity are Missing

RAISE Community Workshop 9

Thursday, May 18, 2023, 2 – 3 PM ET

Summary

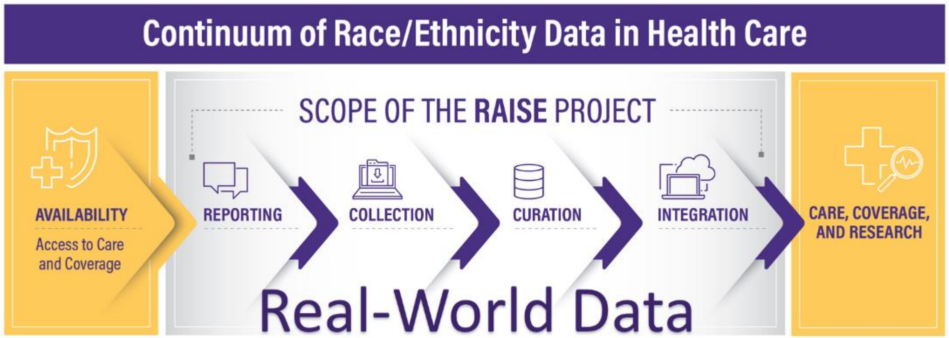
Overview of RAISE Community Workshop VIII

Susan Winckler, CEO of the Reagan-Udall Foundation opened the meeting, followed by remarks by RDML Richardae Araojo, FDA Associate Commissioner for Minority Health and Director of the Office of Minority Health and Health Equity. During the session we heard four presentations. First, Dr. Carla Rodriguez-Watson, RAISE PI, summarized our previous RAISE workshops and their connection to workshop 9. Next, Dr. Francisco De La Vega (Tempus) discussed the imputation of mutually exclusive race and ethnicity categories from genetic ancestry. Then, from Dr. Ali Mokdad (Institute for Health Metrics and Evaluation at the University of Washington), we learned about county-level stratifications in the U.S. and how to quantify disparities in health outcomes by race/ethnicity by building on their systematic effort to measure diseases, injuries, and risk factors by age and sex at the county level. Finally, from Dr. Krystal Tsosie (Arizona State University), we discovered that simply recruiting more Indigenous peoples into datasets is not going to solve the health equity problem and that we need to consider more structural power dynamics. After the presentations, our citizen voice Dr. Cleo A. Ryals (Flatiron Health) joined our speakers on the virtual stage to engage in a discussion led by Dr. Phil Febbo (Illumina, RAISE expert panel member and workshop champion).

Connecting the Dots

Carla Rodriguez-Watson, PhD, MPH
Director of Research, Reagan-Udall Foundation for the FDA

To level set, the RAISE project begins with the assumption that race and ethnicity (R&E) are critical for understanding population health and the real-world utilization and performance and medical products across racialized groups; and thus, the impact that has on the health of those racialized groups.


Continuum of Race/Ethnicity Data in Health Care
SCOPE OF THE RAISE PROJECT
AVAILABILITY
Access to Care and Coverage
REPORTING
COLLECTION
CURATION
INTEGRATION
CARE, COVERAGE, AND RESEARCH
Real-World Data

So, we are focused on the part of the data continuum that includes reporting, collection, curation, and integration of R&E data because this is where the corpus of Real-World Data (RWD) lives. We acknowledge that having R&E in the model doesn't answer all the questions – but it does address some critical questions of importance to the FDA. Which is why, though important, questions of when R&E is not the right variable are not in scope with RAISE. Similarly, the timeline for our discussions does not allow us to delve into issues of access to care.

As we're nearing our final workshops, we reflect on all the insightful presentations and wonderful conversations that have allowed us to think more deeply about the underlying ambiguity, trust, and benefit sharing issues along the continuum. What has been presented are a sample of how we can make progress in closing the gaps. RAISE is not endorsing any one approach but is simply putting them out for consideration. Our next workshop (June 1st) will lay out all those different approaches.

Over the last 2 workshops, we've focused on classification and measurement choices, and their impact on descriptive, predictive, and inferential models of disparity. Today, we're going to focus on what we can do in the meantime with the imperfect measure of R&E. We'll hear about imputation with alleles, how we can assess the representativeness and assumptions of more sophisticated techniques to predict R&E, and how technology can get us closer or further from benefit sharing with different communities.

Imputation of Race and Ethnicity Categories in RWD from Genetic Ancestry

Francisco De La Vega, DSc
Vice President of Hereditary Disease
Tempus

- Healthcare data, including EHR, insurance claims, molecular profiling, and research data, can be a great resource to investigate and address disparities in access to, utilization, and outcomes of care. Unfortunately, there are challenges associated with R&E metadata from RWD, as they are frequently plagued by high rates of missingness and inaccuracies.
- Tempus has created a multimodal research database generated from de-identified records of cancer tumor genomic profiles captured with the Tempus xT next-generation sequencing assay targeting 648 cancer related genes and associated clinical data. This de-identified RWD are being used for different research applications. As previously reported by others in the literature and the RAISE series, a significant missingness of R&E metadata exists in this RWD.
- Some of the potential consequences of missing R&E in RWD include lack of generalizability of analyses, biases in outcome predictors derived from this data, and inability to fully utilize the data in disparities studies. Remediation strategies include imputation, where multiple methods have been applied and are widely used in studies of equity. These typically involve using personal information on the patient, including address and surname. Benchmarking of the accuracy of such methods shows that sensitivity and specificity of imputed R&E data is good but limited and not equal across all groups.
- This is where imputation methods leveraging genetic information could be useful. Since the Tempus database includes genetic information from sequencing, one can use one of several methods to infer genetic ancestry.
- Tempus has implemented a method using ancestry informative markers (AIMs) on top of data from our Tempus xT assay. AIMs were selected to be informative for 5 super continental populations: Africa, the Americas, East Asia, Europe, and South Asia, based on allele frequency data available on the public domain.
- It is important to note that ancestry is not the same as R&E. Ancestry is information about the people that an individual is biologically descended from, which tends to correlate with R&E on a continental level. We can leverage this correlation to impute mutually exclusive R&E categories.

When evaluated against 20,000 records for which we have complete R&E metadata, there was very good sensitivity and specificity -- higher than other methods that rely on geocoding and surnames.

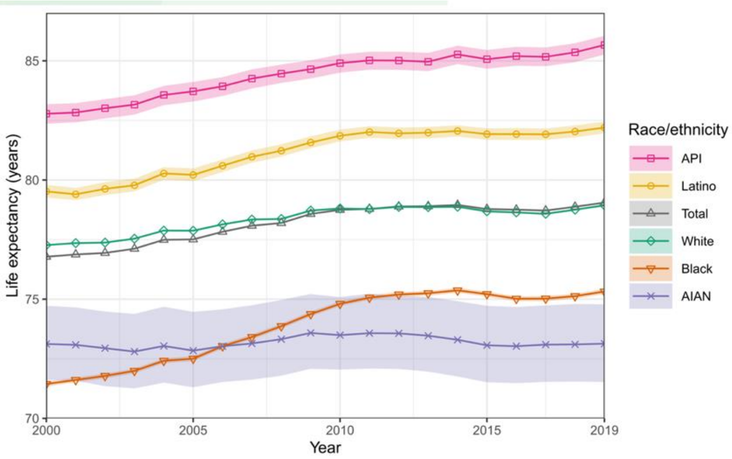
- Tempus' analysis also suggests that the missing data was not completely random. This is also why it is important to impute data in a way that doesn't introduce biases in downstream analysis.
- As a final note, there are limitations. There is not enough data to impute certain categories like American Indian or Alaska Native, and this method may not transfer well to other settings (outside of oncology setting or outside of the US). However, this can be an important piece of the puzzle to start including more populations in certain types of analyses.

Statistical Methods for Measuring Health Disparities in the Population

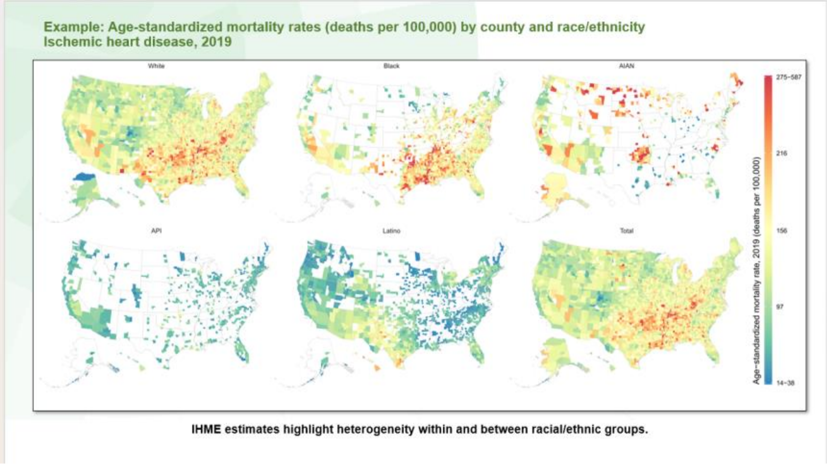
Ali Mokdad, PhD
Professor
Chief Strategy Officer of Population Health
Associate Chair for Collaboration, Associate Chair for Equity
Institute for Health Metrics and Evaluation (IHME) at the University of Washington

IHME is an academic center of excellence in health measurement housed within the University of Washington with a mission is to provide comprehensive health data and analytics to improve decision-making that will lead to health equity. IHME has 3 indicators: What are the health problems of a population or a geographic location? What is the society community government doing about these health problems. And the third one, the most important is, how could we maximize our input to produce the maximum output for our investment?

- IHME has partnered with the National Institute of Health (NIH) on a project to support work quantifying health disparities in the United States by providing mortality, and years of life lost (the burden of disease) at the county level by R&E, socioeconomic status (SES) and education.
- IHME has developed a robust approach that incorporates county-level R&E stratifications in the US to produce estimates for 3,110 stable county units. The death data comes from the National Center for Health Statistics (NCHS). Then, 'garbage coding' (Elizabeth Arias methodology) is used to account for misclassification of R&E on death certificates and incorporate population-level data on post-secondary education, income, poverty, nativity, and population density as covariates in the statistical model (e.g., decennial census and ACS). Publications and the methodology are available on the IHME website. The current challenges faced with estimating by R&E include the increased dimension of model sizes and outputs, small racial/ethnic group sample sizes, limited data inputs and the changes over time for both the way R&E are collected on death certificates and the interpretation of R&E.
- Results from the years 2000-2019 are shown on the slide below. Keep in mind that this data is pre-COVID, and there will likely be an increase in disparities when that data is incorporated. The Asian/ Pacific Islander (API) group has the highest life expectancy but have plateaued in terms of improvement over the years. There have been no real changes for Alaskan Indian/ Native American (AIAN).



- Using this data, IMHE can highlight the heterogeneity within and between racial and ethnic groups for specific causes of death. For example, the slide below shows mortality rate for ischemic heart disease. Here we can see that the rates for Black or African American population (top middle) in the Mississippi delta and parts of Texas are much higher than other areas of the country.



Example: Age-standardized mortality rates (deaths per 100,000) by county and race/ethnicity
Ischemic heart disease, 2019

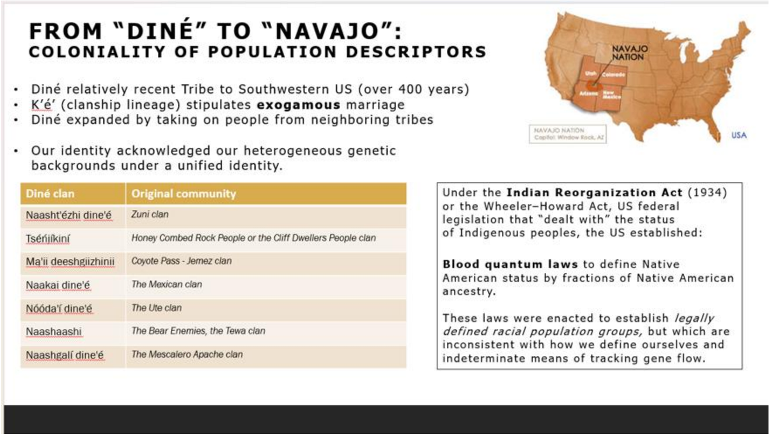
IHME estimates highlight heterogeneity within and between racial/ethnic groups.

- The work ahead is to complete this work with stratification by education and the analyses to include the data from 2020-2021; accounting for the impact of COVID-19.
- For any questions, please reach out via email at services@healthdata.org or http://ihmeclientservices.org.

Genomics for Everyone? Considerations for Equity and Benefit-Sharing for Indigenous Peoples

Krystal Tsosie, PhD, MPH, MA
Indigenous Geneticist-Bioethicist
Assistant Professor
School of Life Sciences
Arizona State University

- To level-set, colonial population descriptors of Indigenous peoples are often arbitrary and mediated by outside factors. Indigenous peoples have their own clanship systems that allow the acknowledgement of heterogenous genetic backgrounds under a unified identity.


NAVAJO NATION
Utah
Colorado
Arizona
New Mexico
NAVAJO NATION
Capitol: Window Rock, AZ
USA

FROM "DINÉ" TO "NAVAJO":
COLONIALITY OF POPULATION DESCRIPTORS

- Diné relatively recent Tribe to Southwestern US (over 400 years)
- K'é' (clanship lineage) stipulates **exogamous** marriage
- Diné expanded by taking on people from neighboring tribes

- Our identity acknowledged our heterogeneous genetic backgrounds under a unified identity.

Diné clan	Original community
Naasht'ézhi dine'é	*Zuni clan*
Tsénjíkiní	*Honey Combed Rock People or the Cliff Dwellers People clan*
Ma'ii deeshgiizhinii	*Coyote Pass - Jemez clan*
Naakai dine'é	*The Mexican clan*
Nóóda'í dine'é	*The Ute clan*
Naashaashi	*The Bear Enemies, the Tewa clan*
Naashgalí dine'é	*The Mescalero Apache clan*

Under the **Indian Reorganization Act** (1934) or the Wheeler–Howard Act, US federal legislation that "dealt with" the status of Indigenous peoples, the US established:

Blood quantum laws to define Native American status by fractions of Native American ancestry.

These laws were enacted to establish *legally defined racial population groups,* but which are inconsistent with how we define ourselves and indeterminate means of tracking gene flow.

- It wasn't until 1934 that rules were established to define a 'Native American' status, and these laws were enacted to legally define racial population groups. These definitions are often inconsistent with how Indigenous peoples define themselves and track gene flow.
- Many large-scale diversity projects including Human Genome Diversity Project, Genographic Project, 1000 Genomes and HapMap have unethically procured biomarkers and then allowed them to be utilized by for-profit partners. One example is 23andMe. Their datasets for imputation have less than 30 Indigenous individuals specific to Central and South America but use this data to infer genetic ancestry for all US indigenous peoples.
- Too often, Indigenous peoples (and others) are categorized as one Tribal affiliation. However, these populations are not stagnant. We do a disservice when we default to colonial definitions of indigeneity and fail to acknowledge multiple tribal identities. This also ignores the lived experience of many Indigenous peoples.
- In looking at statistics, in 2018 the Indian Health Service spent an average of 2.4 times less per patient compared with the national per capita average. Similarly, for DNA specific therapeutics, it is considered "not profit-generative" for pharmaceutical companies to use Indigenous peoples DNA. Simply recruiting more Indigenous peoples into datasets is not going to solve the health equity problem. We need to think more structurally about the power dynamics of the disciplines in which we inhabit.
- The Native BioData Consortium is a nonprofit started by Indigenous community members and scientists to keep the notion of data held locally to tribes, and that the best people to create relevant research health questions are community members themselves who are more cognizant about potential false narratives that can be drawn from DNA studies. For example, there have been over 140 studies in in PubMed that look at genetic factors as it relates to increased rates of alcoholism and Native Americans. However, epidemiological data shows that Native Americans have rates of alcoholism that are the same or less compared to the dominant population.
- In terms of moving forward, we need to:
  - ➢ Rethink informed consent, which has been problematic for Indigenous Peoples in the past and look more into dynamic consent and acknowledge group consent.
  - ➢ Acknowledge Indigenous data sovereignty to build equitable data partnerships and agreements.
  - ➢ Re-think data "ownership" and think instead of stewardship and responsibility to society and colonized peoples.

  - Think globally about health to contextualize whether the study or treatment is meaningful for underserved communities.
- To advance innovation, we need to advance benefit equity, decision-making equity, and engagement equity. Otherwise, we default to the status quo.

Moderated Discussion

Moderator: *Phil Febbo, MD*
Chief Medical Officer, Illumina
Expert Panel Member and Workshop Champion, RAISE

Discussants:

- *Citizen Voice: Cleo A. Ryals, PhD*
- *Francisco De La Vega. DSc*
- *Ali Mokdad, PhD*
- *Krystal Tsosie, PhD, MPH, MA*

The moderated discussion took questions from the chat as well as those posed by our moderator to further expand on the workshop's presentations. Our discussion emphasized imputation and other advanced analytic methods can assist when R&E are missing, but we need to be cognizant of our methods and assumptions; and include communities in the design to ensure that they can benefit and are not further disenfranchised. Highlights from the discussion:

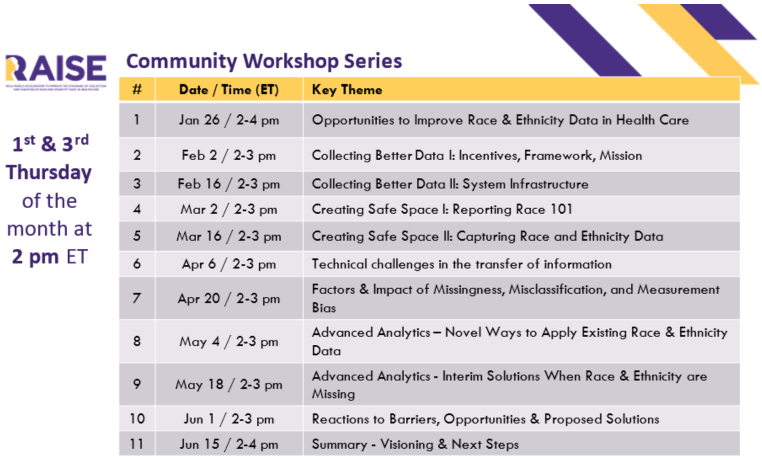
- The best way to get R&E data is to find ways to engage all communities, so that individuals feel empowered to provide data, they trust the individual or the entity asking for it, they trust that they will benefit from providing the data and they trust that they will not be discriminated against because of the data they provide.
- The life sciences industry does themselves a disservice when they find ways to circumvent consent. For example, during the pandemic there was a company that committed to recruiting Indigenous peoples before they actually had tribal permission, which is not respectful at all. They used instead the fact that 80% of Indigenous peoples reside in areas outside of tribal lands to circumvent tribal consent. We cannot view Indigenous, and other underrepresented groups, as sources for understanding undiscovered genetic variation. Simply recruiting more of these groups into datasets is not going to solve the health equity problem. We need to think more structurally about the power dynamics of the disciplines in which we inhabit.
- There really is no perfect approach to imputation but many traditional Bayesian imputation approaches assume that missing data occurs at random. We know that this is not the case and have to take that into consideration. All approaches need to advance health equity and not perpetuate inequities.
- The preferred approach is self-reported R&E data. R&E needs to be collected in a systematic way, when it's not that is a form of structural racism.
- It's critically important to evaluate the degree of interchangeability between R/E and ancestry. R/E are social constructs and self-ascribed. They're also socially assigned and can vary over time. They're directly linked to racial stratification and more specifically, racism. We need to think about this when we use the variable race in our analysis or models.

- In many cases, real data is used to train AI predictors of outcomes. If those AI predictors are not appropriately evaluated across different groups, this adds bias. To overcome some of these inequities, we need to understand how these inequities propagate across software systems.
- There should be ethical checks on imputation in general (see Urban Institute research paper).
- The individual versus group consent is often used by those outside of the Indigenous community as a narrative to create divisions that doesn't exist. We should instead focus on the ways to define data, governance rules, and consenting that is responsive to both individuals and the group.
- We really need to measure racism in this country. Race, racism, ancestry, and all related concepts all deserve their own separate attention careful consideration from a methodological and an ethical moral standpoint. It's not just how we look at genetics, it's how we are treated.
- When using methodological approaches, we need to keep the purpose in mind. It's so important to have multiple stakeholders at the table to provide different interdisciplinary perspectives that reflect the range of lived experiences so we can come to this work with a more holistic and solution-oriented approach rather than just a scientific approach.

Hot Takes and links from the Chat

- The agenda and other materials for Workshop 9 can be found on the FDA Foundation website: https://reaganudall.org/news-and-events/events/advanced-analytics-interim-solutions-when-race-ethnicity-are-missing
- Given how fraught Data Ownership is--a topic that has been wrangled and litigated to pieces--the concept of moving to "Data Responsibility" seems like a far more tractable approach and mindset.
- Data is not missing at random!
- Nothing About Us without Us: https://nativebio.org/
- https://www.urban.org/research/publication/ethics-and-empathy-using-imputation-disaggregate-data-racial-equity-recommendations-and-standards-guide
- Individual + Community Consent = Both, Not a This or That https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5075161/
- For Tribal community names, “don’t collect it if you don’t need it”
- Be Deliberate and Intentional with how we use Race & Ethnicity

Please join us for future RAISE Workshops! Only 2 Remain!



Community Workshop Series

1st & 3rd Thursday of the month at 2 pm ET

#	Date / Time (ET)	Key Theme
1	Jan 26 / 2-4 pm	Opportunities to Improve Race & Ethnicity Data in Health Care
2	Feb 2 / 2-3 pm	Collecting Better Data I: Incentives, Framework, Mission
3	Feb 16 / 2-3 pm	Collecting Better Data II: System Infrastructure
4	Mar 2 / 2-3 pm	Creating Safe Space I: Reporting Race 101
5	Mar 16 / 2-3 pm	Creating Safe Space II: Capturing Race and Ethnicity Data
6	Apr 6 / 2-3 pm	Technical challenges in the transfer of information
7	Apr 20 / 2-3 pm	Factors & Impact of Missingness, Misclassification, and Measurement Bias
8	May 4 / 2-3 pm	Advanced Analytics – Novel Ways to Apply Existing Race & Ethnicity Data
9	May 18 / 2-3 pm	Advanced Analytics - Interim Solutions When Race & Ethnicity are Missing
10	Jun 1 / 2-3 pm	Reactions to Barriers, Opportunities & Proposed Solutions
11	Jun 15 / 2-4 pm	Summary - Visioning & Next Steps