

Technical Challenges in the Transfer of Information

RAISE Community Workshop 7

Thursday, April 20, 2023, 2 – 3 PM ET

Summary

Overview of RAISE Community Workshop 7

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 8. Next, Andrea Thoumi and Kamaria Kaalund (Duke-Margolis Center for Health Policy) discussed the LATIN-19 project and their work to bridge the health care gap for the Latino/a/x community in North Carolina. Then, from Dr. Michele Jonsson-Funk (University of North Carolina (UNC)-Gillings School of Public Health), we learned that missing race and ethnicity data are often not missing at random and can yield non-generalizable results, poor prediction for individuals not well-represented, and biased estimates of treatment effects due to failure to achieve exchangeability. After the presentations, our citizen voice Dr. Aaron Kamaau (Ikaika Health) and expert panel member Dr. Anne-Marie Meyer (UNC-Gillings School of Public Health) joined our speakers on the virtual stage to engage in a discussion led by Dr. Carla Rodriguez-Watson.

Connecting the Dots: From Continuum to Impact

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.



RAISE is focused on the part of the data continuum that includes reporting, collection, curation and integration of race data because this is where the corpus of RWD lives. We acknowledge that having race in the model does not answer all the questions – but it does address some critical questions of importance to the FDA. Which is why, though important, questions of when race 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.

In workshops 1 through 6, we discussed issues along the continuum from data reporting, collection, and integration to data exchange. We also talked about the issues and solutions that affect multiple areas along this continuum. For the latter half of the RAISE workshop series, we'll be talking about the impact of missing data and how to deal with it. Workshop 7 opened with a discussion of how employment, access to care, immigration status, and multi-generational housing impacted risk for COVID-19 in the Latin community and how the LATIN-19 team got creative to bridge enrollment gaps and the associated data. Then Dr. Michele Jonsson-Funk from the UNC-Gillings School of Public Health discussed why it is important to address those gaps (as missing data are seldom at random), along with the impact of that missing data and associated bias.

Using Data to Inform Community-Engaged Interventions


*Andrea Thoumi, MPP, MSc
Health Equity Policy Fellow and Faculty Director
Health Equity Education
Duke-Margolis Center Health Policy*

*Kamaria Kaalund
Policy Analyst for Health Equity
Duke-Margolis Center Health Policy*


We'd like to acknowledge all the community members, organizations, and partners in [LATIN-19 project](#) that are doing the groundwork in North Carolina.

- In early 2020, Hispanic or Latino patients represented 77% of COVID-19 cases in Durham NC, five times greater than the share of the Hispanic or Latino population. For many of these patients, their COVID diagnosis led to their first interaction with the Duke Healthcare System.
- The reasons for those statistics are rooted in the systemic inequities listed on the slide below. For the purposes of this presentation, the focus will be on experiences of discrimination in the health care system which led to warranted mistrust.

Systemic Inequities



- **Overrepresentation** in frontline/essential jobs, without proper protection
- Multi-generational and multi-family homes that make **social distancing challenging**
- **Economic pressures and lack of policies** that have limited ability to take sick leave
- **Reduced access** to COVID-19 testing and vaccination (and other health care-related needs)
- Experiences of **discrimination** in the health system and other sectors leading to warranted mistrust
- **Lack of adequately-translated official information** leading to misinformation spreading through social networks



- One key factor discussed during focus groups with North Carolina Hispanic or Latino community members is that it was not often made clear why data, such as R&E and insurance status, were

being collected. This lack of communication on the part of the health system sparked concerns of immigration-related ramifications, and many declined providing this information as it was unclear where the data are going and how it's going to be used.

- The next two examples demonstrate why having R&E data is critical to identify the gaps in utilization and health outcomes:
 - R&E data were used to illustrate patterns of resource distribution in the early part of the pandemic (July 2020). There were disparities in the geographic distribution of COVID-19 testing sites. Eastern North Carolina, which is made up by primarily Black and Latino communities, had only 2 COVID-19 testing sites. In predominantly White areas, there were several COVID-19 testing sites. Here the R&E data were used to identify the gaps in utilization and health outcomes.
 - Similarly, Hispanic or Latino communities in North Carolina were under vaccinated in May 2021. The cumulative total of vaccinated individuals in the Hispanic or Latino community was 6.9 percent which was less than the population share of Hispanic or Latino individuals in the state. But by February 2022, the percentage increased to 9.
- Part of the reason the vaccination rate increased (in the above example) can be credited to LATIN-19. LATIN-19 is a coalition of partners that started in March of 2020 by Latina clinicians at Duke University. It has since grown to over 700 members, representing several diverse stakeholder groups, including public health departments, NC Health and Human Services, health system leaders, and community based-organizations (CBOs). LATIN-19 works closely with community health workers and CBOs through interdisciplinary meetings and partnerships to conduct research and policy analysis. As a result of the LATIN-19 partnership with Duke Health:
 - The percent of vaccination recipients of Hispanic or Latino ethnicity at Duke Health vaccination events from February 2021 –July 2021 increased.
 - The proportion of COVID-19 vaccination doses administered at COVID-19 vaccination events from February–July 2021 reached 90.5% of Hispanic or Latino individuals vaccinated (versus 5.6% at the Duke Health only events).
- There is also an issue of systemic exclusion discovered during early LATIN-19 events. Many Hispanic or Latino individuals did not have an existing electronic medical record (EMR), so the R&E data collected at registration could not be entered.
- For those that had an EMR, missing ethnicity data was also common. In many cases, the entry indicated “Spanish-speaking only.” Though the EMR system supports entry for country of origin, these data were often not collected. The country of origin could have been used to extrapolate ethnicity data if present.
- LATIN-19 is currently working on a project to bridge health insurance enrollment gaps for Hispanic or Latino communities in North Carolina who are eligible for Medicaid or an Affordable Care Act (ACA) health plan. One in three North Carolinians who identify as Hispanic or Latino are uninsured. For those not naturalized, 84% have no health insurance (note: 75% of Hispanic or Latino individuals are not naturalized). To estimate eligibility, LATIN-19 is using publicly available census data. Unfortunately, using this data poses 3 issues:
 - Missingness (e.g., many Hispanic or Latino individuals select “other” in race questions; fear of filling out demographic information)
 - Hard to estimate across 2+ demographic categories (e.g., race and ethnicity, enrollment rates, uninsured rates, documentation status)

- Data lag does not capture completeness (e.g. data collected every five years)
- Understanding the limitations of using census data, LATIN-19 is employing a mixed- methods approach to estimate enrollment gaps. The total number and percentage of Hispanic or Latino individuals, along with their poverty and current insurance status, is combined with community knowledge and focus group data to estimate eligibility by county. LATIN-19 can then prioritize which counties to focus enrollment interventions and partner with existing CBOs to add community engagement into the final prioritization process.
- To read more about LATIN-19 and other Duke-Margolis projects, please visit <https://healthpolicy.duke.edu/> or subscribe to the monthly newsletter at dukemargolis@duke.edu.

Missing, Misclassified, or Mismeasured? Beware Bias

Michele Jonsson-Funk, PhD

Associate Professor, Department of Epidemiology

Director, Center of Pharmacoepidemiology

UNC Gillings School of Global Public Health

We'd like to acknowledge the UNC-based 'Beyond the Boxes' team whose work helped frame this presentation including Natalie Smith, Rae Anne Martinez, Nafeesa Andrabi, Andrea (Andi) Goodwin, Rachel Wilbur and Paul Zivich.

- To do a better job using real-world data to try to understand drug safety and effectiveness, measuring R&E is critical. This begins with thinking broadly about where the impacts of missing data show up in various types of research questions to figure out solutions.
- Research questions fall into 3 buckets listed on the slide below.

The spectrum of research questions



Descriptive: detail the experience of health and health care for a particular group of people and compare how that experience differs between groups


Predictive: what are the likely future outcomes for a group of people given their observed characteristics, health condition(s) and treatment at a point in time

Causal: what are the expected future outcomes for a group of people given their observed characteristics and health condition(s) under two (or more) different treatments or treatment strategies

- Generally, research is not done to simply understand the experience of study participants, but to learn something that can be generalized in the real world. In descriptive analyses, the main concern about missing R&E data is that the missingness is not random and thus the study does not have a complete representation. That means missing out on the important information about patients who didn't feel safe, invited, or enabled to share that information about their identity along with the other kind of health care data and experiences the study seeks to understand.

- The impact of this missing R&E data is not just a momentary problem but makes it difficult to adequately characterize that longitudinal experience of individuals in descriptive work. To highlight an example of both the magnitude of the challenge and a potential solution, a [study](#) was published last year by Branham et al that reviewed missing R&E data in the Federal marketplace (www.heathcare.gov). Researchers used name, address, and Census block group information to estimate the probabilities of R&E categories for the thirty-some percent of individuals with missing R&E data.
- In predictive studies, the machines (computers or programs) used are only as good as the data that are used to train them. An [example study](#) was published last year looking at the use of pulse oximeters (use different waves of light and infrared to detect the amount of oxygenated hemoglobin in the blood via the finger). Because pigmentation has an impact on measurement accuracy, pulse oximeters systematically overestimate how much oxygen is in the blood of individuals of Asian, Black or Hispanic descent. The effects were detrimental in patients with COVID-19; leading to delay or failure to recognize eligibility for COVID-19 treatments.
- Another example [study](#) involved an algorithm to create a risk score used to predict who needed extra care. The algorithm used health care costs as a proxy for health needs. Since Black or African American patients have historically poor access to care, the algorithm falsely concluded that Black or African American patients didn't need care because they hadn't been receiving it. Artificial intelligence (AI) can exacerbate the existing health care disparities if fed data that already reflects those disparities.
- For causal studies, exchangeability is only as good as the models. When estimating a causal effect in the absence of randomization, we rely on the ability to use the measured data to find similar individuals – some treated with A and some treated with B - who can stand in for the experience of each other. The challenge here is that the ability to identify exchangeable individuals is only as good as the data. If a heterogeneous mix of racial and ethnic groups is combined into a category ‘other, unknown, missing, not reported’, the assumption is that all those are a good ‘stand in’ for the counterfactual experience of each other. The result is potentially greater confounding of treatment effects.

The spectrum of impacts due to measurement error, misclassification, and missing data



Descriptive: results that are not generalizable

Predictive: poor prediction for individuals who are not well-represented in the training data, leading to further disparities in care and outcomes

Causal: biased estimates of the treatment effect of interest due to failure to achieve exchangeability; inability to fully address other sources of measurement error that differ by race or ethnicity

THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL 43

- A summary of the spectrum of impacts due to measurement error, misclassification, and missing is listed on the slide above. Potential solutions are to:

- Supplement data with an understanding of who is missing; with special care when making decisions about combining groups.
- Use analytics such as multiple imputation and inverse probability of missingness weights with the caution that in nonrandomized settings, using complete case analysis or missing indicators is likely to be biased.
- A final note is that both R&E are fluid and contextually specific and unlike many things measured in health care, there may not be a gold standard. If identity is not fixed over time and between different settings, there may be several different ‘true’ responses that require greater care than a standard validation study.

Moderated Discussion

*Moderator: Carla Watson-Rodriguez, PhD
Director of Research, Reagan-Udall Foundation for the FDA
Principal Investigator, RAISE*

Discussants:

- *Workshop Champion: Anne-Marie Meyer, PhD*
- *Citizen Voice: Aaron Kamauu, MD, MS, MPH*
- *Michele Jonsson-Funk, PhD*
- *Kamaria Kaalund*
- *Andrea Thoumi, MPP, MSc*

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 (again) emphasized that trust is key for data collection and community engagement. We need to build trust to engage diverse patient populations so that we can have representation and treat the patients in need. Highlights from the discussion:

- Different populations of people engage in the health system very differently. Understanding how patients engage in healthcare could play a significant role in how data are understood and how research trials are run.
- Is there consensus in how and why R&E are collected in health care settings? There is a need to be explicit about defining race to create appropriate response values that capture relevant facets of race and ethnicity for self-identification.
- Researchers need to acknowledge the context and the purpose for which R&E data are captured in real-world data sets and refrain from using it in ways that are not appropriate.
- The real-world data community needs to standardize how to collect R&E data and transparency in how data are used once collected.
- Data *should* represent the patients that may ultimately be treated by these medical advancements and interventions.
- Healthcare systems could potentially supplement their EMR with census or community data to pinpoint which areas and communities need intervention. The key element is to build community engagement by working with CBOs or partners to build trust.
- There needs to be a broad and careful approach to engage diverse patient populations by helping them to understand exactly how data are being used, what the benefit will be to their community

and allowing the community to define what questions are important to them. There should be a platform for bidirectional communication with research projects or programs where there is an immediate feedback loop between community members, CBOs and health systems.

- AI tools can help with patient prioritization, but we need to be cautious that the inputs behind them are unbiased so that we're not exponentially increasing biases.
- From Beyond the Boxes: *"Racial and ethnic identities may be rooted in oppression. But here, too, are joy, love, and belonging. We're proud of our racial and ethnic identities because they honor our ancestors and the communities that cherished, supported, and raised us up."* We need to capture all of it, the pain and the beauty, to make sense of population health.

Hot Takes and links from the Chat

- LATIN-19 used publicly available data to prioritize which counties are exhibiting the greatest need by using 3 indicators: poverty status, uninsurance rates, and total population who identifies as Hispanic or Latino. Our project includes a substantial component to train community health workers to support community members navigate insurance enrollment. The goal is to close the enrollment gap of Hispanic or Latino community in the state. Hispanic or Latino individuals represent about 1 in 3 uninsured in NC and about 1 in 3 Hispanic or Latino individuals are uninsured.
- Measurement error has real consequences!
- Paper recommendation: Matthew K Chin et al. Methods for Retrospectively Improving Race/Ethnicity Data Quality: A Scoping Review, Epidemiologic Reviews, 2023, mxad002, <https://doi-org.libproxy.lib.unc.edu/10.1093/epirev/mxad002> or <https://pubmed.ncbi.nlm.nih.gov/37045807/>
- RE: under-represented groups, there are methods that have long been used in ecology (MARK; capture-recapture) to statistically characterize the "trap-shy" populations.
- Still need to work with communities, of course. But mental health, addictions, and other stigma-associated conditions are as confounded as capture-rates biased by race and ethnicity
<https://cran.r-project.org/web/packages/openCR/openCR.pdf>
<https://cran.r-project.org/web/packages/mra/mra.pdf>
- Diversity is very important, getting diverse data is important. Did we ask the minorities if they care about learning from their data/participation?
- We are trying to bring the AI and other method development community together to figure out ways to reduce biases in AI and other topics: <http://psb.stanford.edu/callfor/papers/ohdpm>
- Sometimes, we are so anxious to collect the data that we do not pay much attention to what the communities are anxious to get.

Please join us for future RAISE Workshops:



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