Collecting Better Data II: System Infrastructure

RAISE Community Workshop 3

Thursday, February 16, 2023, 2 – 3 PM ET

Summary

Overview of RAISE Community Workshop III

During RAISE workshop III, we had welcome remarks from Susan Winckler, CEO of the Reagan-Udall Foundation, and 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 three presentations. First, Dr. Carla Rodriguez-Watson, RAISE PI, summarized our previous RAISE workshops and their connection to workshop III. Then, Meredith Welsh described how SameSky Health works with diverse populations to help them engage with the health care system and the infrastructure needed to collect the necessary data to develop actionable insights. Finally, Andrew Kress of HealthVerity presented an overview of the infrastructure used to support the robust capture of multi-level and longitudinal race and ethnicity data (R&E). The session closed with a discussion moderated by Dr. Oscar Benavidez of Massachusetts General Hospital.

Connecting the Dots: From Opportunities, Incentives, to Infrastructure

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 the performance of medical products across racial and ethnic groups. We acknowledge that race and ethnicity alone do not answer all the questions, but are important demographic variables to the FDA. The question of whether or when race or ethnicity is, or is not, the appropriate variable is not in the scope of the RAISE project. The focus of RAISE is the continuum that includes reporting collection, curation, and integration of R&E data. So far, the workshops to date have laid the foundation for this continuum.

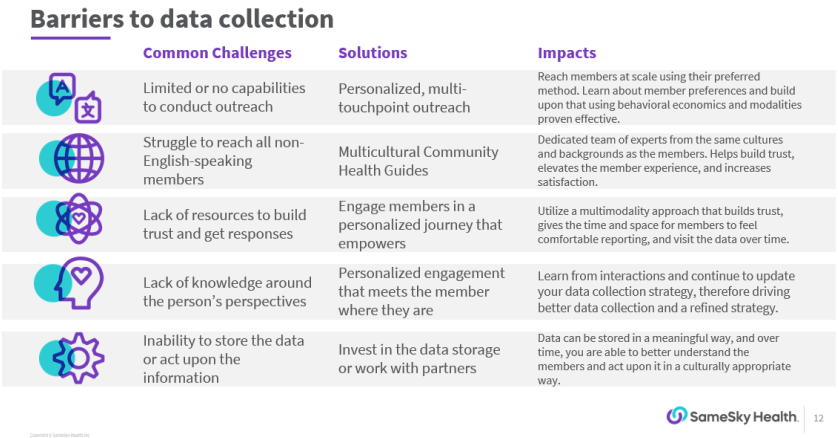
The presentations of Workshop I: "Improving Race and Ethnicity Data in Health Care" explored how to get R&E data in care delivery, the pairing of measuring and reducing inequities and how to build and keep patients' trust—with emphasis on ensuring that the access to R&E does not unintentionally wind-up marginalizing those we intend to help. We also examined how the current fragmented approach and business built around transacting health care does not follow the patient, and thus does not help us understand population health. We heard how one health care delivery system identified workflow, training, and user interface issues that limit the collection of race, ethnicity and language data in health care settings; and brass tacks on how to get R&E in care delivery and among insurers.
Our next workshop, "Collecting Better Data I: Incentives, Framework, Mission", built upon the first to discuss value-based payment models to incentivize health equity and the capture of R&E, keeping patients at the center of all investments, and how R&E in real-world data form the cornerstone of a framework to identify targets for more representative clinical trials. That brings us to today, workshop three, where we will discuss the system infrastructure needed to capture R&E across all those fragmented siloes to create a data view that keeps the patient as its focus.

The Human at The Center: The Data They Generate, and The Goal of Health Equity
Meredith Welsh, SVP Health Operations
SameSky Health

SameSky Health is a multicultural company with a goal of removing barriers to health care access and empowering members to engage in their health care. SameSky Health believes in meeting people where they are, with services that balance technology, reach, and keep the human component at the center of all interactions.

- We've heard a lot from health plans about obstacles in gathering data such as race, ethnicity, language, and social drivers of health. The more you can personalize a journey for a member, the more likely you are to build trust and connect in a way that facilitates the collection of stronger data.
- 90% of our interactions within our solution, called Culture Guide, are through SMS text message. As such, we must be very strategic about building trust to ensure that when we ask sensitive questions, members understand why we're asking and feel comfortable answering.

Commented [MM1]: Changed for consistency



- Some of the barriers with a large-scale approach to data collection, along with potential solutions, are listed on the slide above. To summarize:
    - Limited to no capabilities to conduct the outreach: use a personalized multi-touch approach to meet members where they are. For example, ask for preferred method for contact. Show caring and investment in members.
    - Struggle to reach non-English speaking members: work with community health guides from the same cultures and backgrounds as the members that we're outreaching to. They can really help build trust.
    - Lack of resources to build trust and get responses: give space for individuals to be able to answer in a way that makes sense for them. For example, if a person identifies as two or more races, we get as much information as possible, help map it to OMB standards and store it (if infrastructure isn't prepared for the extra information) until an IT solution is in place.

    - Lack of knowledge around perspectives: empower the person to be able to understand the importance of data collection within their culture. People can be skeptical when asked for sensitive personal data like sexual orientation and gender identity. It's important to make sure there is knowledge and understanding around the reasoning for the data collection.
    - The inability to store data or act upon the information: invest in infrastructure.
- A case study where SameSky Health (using Culture Guide) partnered with a California Medicaid plan really demonstrates how the technique around cultural engagement can influence an individual's willingness to provide sensitive data. The results of the partnership were that:
    - 78% of members moved from an unknown category to a known category for ethnicity.
    - 70% of respondents updated their race from an unknown category into a known category.
    - Members were willing to disclose social drivers of health through a discovery screener sent via text message. Every question was optional and there were community health guides that could help guide participants on phone calls if needed.
    - 94% of respondents answered whether they were a part of the LGBTQIA+ community.
- Lessons learned from millions of interactions though Culture Guide:
    - Be considerate. A member may be taking a nap, or on their way to work. Asking a simple question like "is now a good time?" can help build trust, and really increase response rates.
    - How questions are framed makes a difference. Is it to improve the services provided within a community? If the focus is on the community, the call to action needs to impact the community. This can vary significantly by culture. Taking the time to learn about what's important within a specific culture can help improve response rates and the quality of data collection.
    - Use prompts to tell people how long it is going to take to fill out this information. If they know exactly how long it's going to take, they're more likely to provide this information than if they don't even know what they're getting into.
    - Train community health guides by thinking about cultural context, barriers to providing information and lived experiences as a cultural group. That way, they know how to respond to common questions and issues. For example, a common response to "what is your preferred language" may be "we live in the US- what do you think my preferred language is?" Training community health guides to be empathetic in their responses and to say they're trained not to make any assumptions will yield better quality data.
    - Be proactive. For example, you might prompt someone and they're not comfortable giving information immediately. After you build trust over time, they may be more likely to respond. In follow-up interactions, ask if anything's changed. Though race or ethnicity don't necessarily change, there are a many personal reasons why someone might change their responses over time. Be proactive in getting information and giving individuals an opportunity to update this information.

Reconciling Race and Enhancing Ethnicity: Challenges in Real-World Data (RWD)
Andrew Kress, CEO
HealthVerity

HealthVerity provides technologies and software tools for the discovery and integration of optimal patient data sets. The company works with pharmaceuticals, hospitals, and payer customers to maximize the insights from data supplier relationships.

- Getting better R&E to use in a validated way can have many benefits, from outcomes research to health equity studies. However, data being inconsistently being collected over sources, across sources and over time, as well as the increased risk of re-identification real-world datasets can create methodological and data management challenges.
- What HealthVerity does is work with about 75 large data partners to receive all their RWD, de-identify that data, and link individuals uniquely but anonymously, both over time and across data sets. One of the benefits of this approach is that HealthVerity can assemble features about an individual patient across different data sets. This ensemble approach is used with R&E as well. HealthVerity collects all the features around race and ethnicity for a specific individual patient, then uses methodology to deal with the potential differences in the data. There are many challenges with the collection of race and ethnicity in RWD, outlined below.
    - One challenge with race and ethnicity in RWD is how it is being collected. Depending on the source of the RWD, for example, EMR or claims data, there are a variety of ways that that data is being collected at the source. It's not always clear what modality is being used for a particular patient at any point in time. The data can be self-reported, which is the gold-standard. There is also observed data, where race and ethnicity are entered by a staff member observing a patient, and imputed data where race and ethnicity are calculated based on names or social security data.
    - Another challenge is ontology. Not all parties use the same definitions for race and ethnicity. For example, some sources have more granular representations with categories or distant categories, such as Chinese or Korean. Additionally, the way these race and ethnicity identifiers are flagged can be very different across different sources, or there can be changing values over time within a single source.
    - The last challenge is privacy and how to prevent the risk of re-identification. The risk becomes significantly higher when very specific race and ethnicity definitions for an individual are tied to age, gender, and geography. To mitigate, HealthVerity recodes certain categories up to higher levels, such that the risk of re-dentification based on the size of the cohorts is reduced.

USE CASE

Feature	Healthcare Suppliers			Consumer Suppliers		
	Insurance P520	EMR (Multiple)	Provider National Lab	Consumer Data 1	Consumer Data 2	Consumer Data 3
Ontology	Race: White Black Asian Hispanic Other Unknown	Race: White Asian Black Other Unknown Ethnicity: Hispanic or Latino Not Hispanic or Latino	Af. American Not Af. Am.	Race: White/Other Black Asian Hispanic Ethnicity: Asian Af. American Chinese Hispanic American Indian Japanese Portuguese Caucasian/White Country of origin (50+ values) Language Religion	Race: NA Country of origin (50+ values) Language Religion	Race: NA Country of origin (50+ values) Language Religion

- The slide above demonstrates the varying ontology across data sources. For Health care data suppliers, insurance records tend to be more structured to the OMB definitions, with EMR using a similar approach. Lab data, on the other hand, is incomplete in terms of race and ethnicity as a standalone

data source. On the consumer data side, there are different characteristics that companies will identify in similar, but not the same, ways.

USE CASE

Feature	Healthcare Suppliers			Consumer Suppliers		
	Insurance PS20	**EMR (Multiple)**	**Provider National Lab**	**Consumer Data 1**	**Consumer Data 2**	**Consumer Data 3**
Reason Captured	Completed by patient at time of enrollment	Completed by patient at time of initial visit Staff may ask for patient to review info periodically	Asked during collection for some tests (eg COVID)			
Accuracy	Self-reported	Self-reported Observed	Self-reported Observed	Mixture of self-reported and modeled values (Source 1 is ~50/50 actual vs modeled for example)		
Temporality	Captured	Captured	Captured	Not easily available		
Coverage	Variable by plan type		Low coverage for race			

healthverity

- Many of the differences between data suppliers are due to where the R&E are being captured in the workflow, as illustrated by the slide above. For insurance claims, we surmise this data is being captured as part of the patient enrollment process. For the EMR vendor, R&E is likely being captured on patient intake or by the staff. And then the consumer data suppliers are using a combination of capture methods to be able to try to get to different forms of granularity around the individual and household level demographics. There are two things to flag here. The first is that consumer data is a mixture of both self-reported and modeled values, but we don't know which record is modeled specifically, and which record is self-reported. The second is that for temporality, or considering the individual changing values over time, that change may not be discernible if not captured discreetly.
- There are some challenges we face when we look at individual data sets. Payer data alone contains approximately 181 million patients with race indicators on around 85 million (47%) of those patients. Almost 99% of those 85 million have a single value that's recorded over time and a small percentage have multiple values reported over time.
- To combat the issues of using a single data set, HealthVerity applies more of an ensemble model that uses internal logic to choose data sources based on non-null values and hierarchical values. For example, adding EMR data to payer data, the coverage of race and ethnicity values is improved as the missing data from one source is filled in from the second source. The combinatorial data yields 237 million people. Race and ethnicity are present in 72% of those individuals, an increase from the 40.8% when using payer data alone.
- A slightly different example uses payer data, which is presumably self-reported, and consumer data, which is a mix of self-reported and imputed data. The logic used here maps the ontology from the consumer data, which is the least common denominator. This combination yielded race or ethnicity information on roughly 93.4% of individuals.
- In summary, HealthVerity is using a state-of-the-art combination approach with RWD and continuing to try to tease out where the most information can be obtained across data sets.

Moderated Discussion

Moderator: Oscar Benavidez, MD, MBA, MPP
Discussants:

- *Workshop Champion: Allen Hsiao MD, FAAP, FAMIA (CHIO, Yale/New Haven Health Systems)*
- *Carmela Couderc (Branch Chief Technology, Content and Delivery, Office of the National Coordinator for Health Information Technology (ONC)*
- *Andrew Kress*
- *Meredith Welsh*

The moderated discussion took questions from the question-and-answer chat as well as those posed by our moderator to further expand on the workshop's presentations, and to discuss the infrastructure needed to collect and share R&E. The moderated discussion reinforced that collecting race and ethnicity isn't a straight-forward process. Selecting racial and ethnic categories, the granularity of this data and storing the data for later use is complex and can vary greatly depending on the end use. Highlights from the discussion:

- ONC supports engaging patients across health systems by promoting interoperability. Software/ EMR developers can participate in ONC's voluntary certification program that ensures that any organization that's using certified health information technology (HIT) has the capability to collect, store, and exchange specific data values that might apply to a population served by the organization.
- A participant choosing not to give data is very different than not having the ability to collect a value (for example, individuals who identify as Middle Eastern or North African). Certified HIT must use the race and ethnicity code system that is stewarded by the CDC. Someone may see the word white or Latino, but in a certified HIT system, the capability to store that code is independent of the text string. The meaning is enshrined in a specific code.
- ONC has a data standard called the United States Core Data for interoperability. It considers tribal affiliation to be a separate data element from race and ethnicity.
- Hospitals are struggling with which race and ethnicity choices (as there are so many) to allow patients to select from and how often to ask patients race and ethnicity questions.
- It's very important to collect race and ethnicity information at a granular level, even if just for storage, so that when things are updated due to changed assumptions or mappings, the information is available.
- Always have transparency to the pedigree of the data, how it was collected, and what manipulations have been done to get to an analytic data point. Regardless of the data inputs, always maintain transparency on how the data was manipulated to get to what is used at the output.
- The level of R&E necessary for an analytic data point depends on the type and scope of research project. Aggregate detail can be sufficient at times, but when examining specific groups of individuals, data must be parsed more substantially.

Please join us for future RAISE Workshops:

RAISE

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