Enhancing Public Health Data Quality

ADDRESSING MISSINGNESS AND STANDARDS CONFORMANCE

Sarangan Ravichandran and Aryan Paul
Leidos Health Group
September 27–28, 2023

EC #23-Leidos-0915-26913

AGENDA

01 Electronic Health Records (EHRs)

Relationships between RWD, EHR, demographic data & RWE

02 OMB data standard, EHR Dataset used, and Analysis Assumptions

03 Exploratory Data Analysis (EDA)

Focus on Data Quality, conformance to OMB data collection standards

04 Propose solutions to address the gaps

05 Future Directions

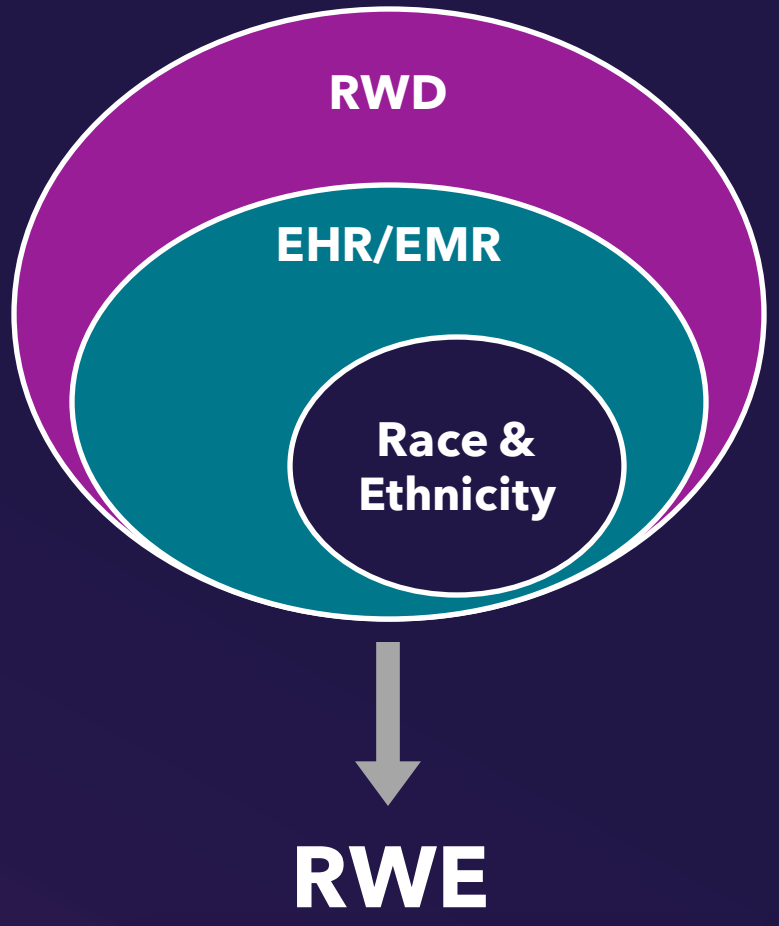
RWD: Real-World Data
RWE: Real-World Evidence

Office of Management and Budget (OMB)

Agency that oversees the implementation of various policies and the management of federal budget

DEMOGRAPHIC DATA: RELATIONSHIP TO EHR AND RWE

- Demographic data:
  - An integral part of both RWD and EHR and provides context to health information
  - Public Health Agencies such as CDC use it to make informed policy decisions, allocate resources effectively, influence treatment responses, assess disease prevalence/surveillance and healthcare utilization
  - Contributes to the development of personalized medicine, tailor interventions/treatments to specific patient profiles to predict overall healthcare efficacy
  - Facilitates the identification of healthcare disparities



EMR: Electronic Medical Record

COMMON PROBLEMS ASSOCIATED WITH DEMOGRAPHIC DATA

- **Incomplete or inaccurate data**
  - "Race/ethnicity and other intersectional data is missing nationally"
- **Lack of standardization**
- **Data Silos**:
  - Data access is not easy
- **Heterogenous**:
  - Diverse sources; difficult for harmonization and analysis

CDC Foundation
Modernizing Health Data Analytics and Forecasting. Forecasting and Modeling Listening Sessions – Final Report with Technical Notes.
August 27, 2021.

https://precision.fda.gov/challenges/30

RACE & ETHNICITY DATA COLLECTION STANDARD

- **Two-question wording**
  - Ethnicity Question:
    Are you Hispanic or Latino?
  - Hispanic or Latino
  - Not Hispanic or Latino

- **Race Question: What is your Race?**
  - American Indian or Alaska Native
  - Asian
  - Black or African American
  - Native Hawaiian or Other Pacific Islander
  - White

- **One-question (combined) wording**
  - American Indian or Alaska Native
  - Asian
  - Black or African American
  - Hispanic or Latino
  - Native Hawaiian or Other Pacific Islander
  - White

Office of Management and Budget (OMB)

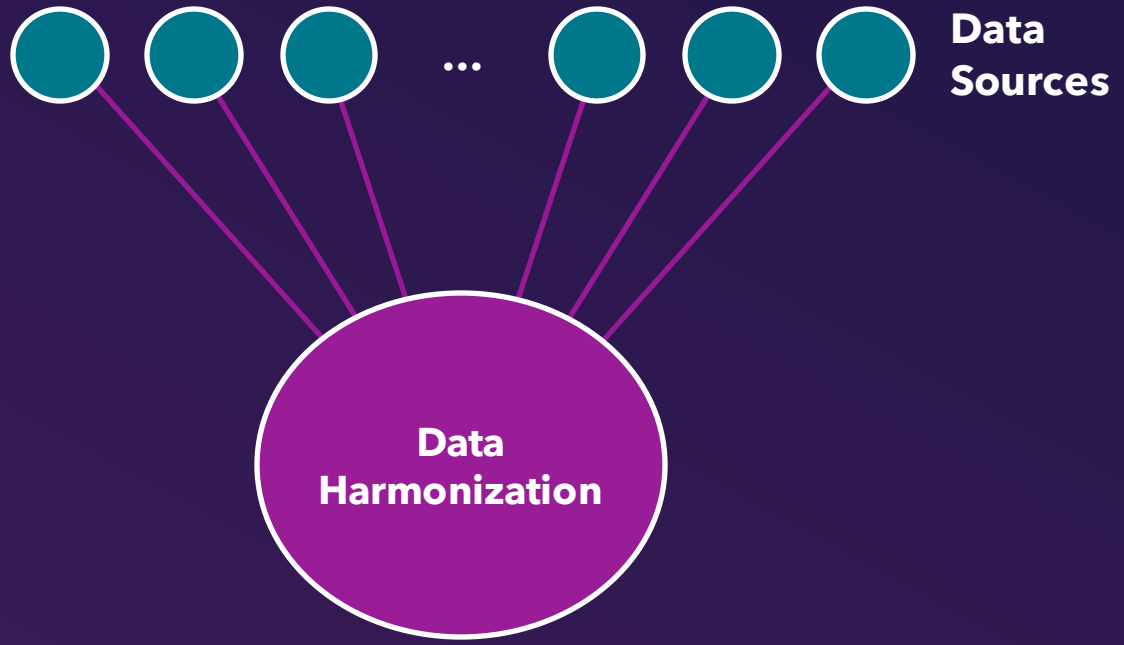
Agency that oversees the implementation of various policies and the management of federal budget

https://orwh.od.nih.gov/toolkit/other-relevant-federal-policies/OMB-standards

https://www.cdc.gov/nchs/nhis/rhoi/rhoi_history.htm

EHR LONGITUDINAL DATASET USED FOR DATA MODELING

- Diverse ambulatory healthcare provider practices across the US
- De-identified, random sample of 1 M patients'
- Jan. 01, 2017 – Dec. 31, 2021



EHR DATASET CHARACTERISTICS AND OUR ANALYSIS ASSUMPTIONS

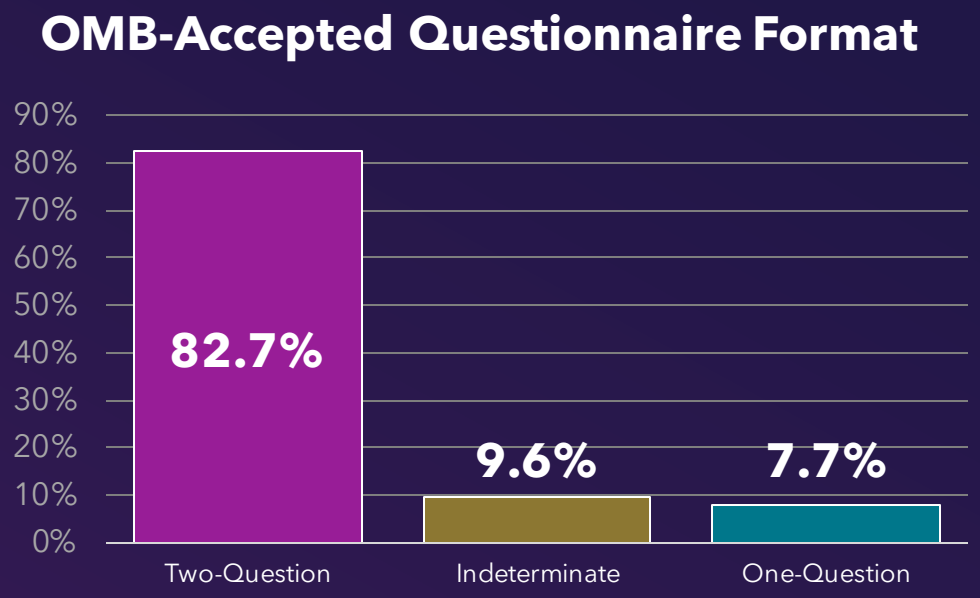
- Dataset contained Race & Ethnicity columns, no information about the questionnaire wording format
  - Non-availability of data dictionary and meta data (data source or transformation, etc.)
- We allowed for variations in the Race and Ethnicity values that conform to OMB standards, as long they could be mapped back to acceptable keywords
- We assumed that patients would default to answering all questions when presented with a prompt
- We did not find convincing evidence in this dataset that could reliably enhance or impute race/ethnicity by linking/enriching other information within this dataset

QUESTIONS AND OBJECTIVES

- Drawing from both existing research and our own practical insights, we will delve in the following questions:
  - What is the distribution of the one/two-question OMB formats?
    - Benefits include: Granularity, accurate representation of the population, etc.
  - What proportion of the responses are OMB compliant?
    - Benefits include: Consistency, harmonization, Adoption of standardization, regulatory compliance
  - What percentage of Race and Ethnicity data are missing?
    - Benefits include: Is the dataset representative? Fit-for-use data?
  - What is the distribution of the minimum OMB race and ethnicity categories?
    - Benefits include: Inequalities and disparities, study design issues

https://precision.fda.gov/challenges/30

CAN WE DISTINGUISH BETWEEN THE TWO OMB QUESTION WORDING FORMATS?



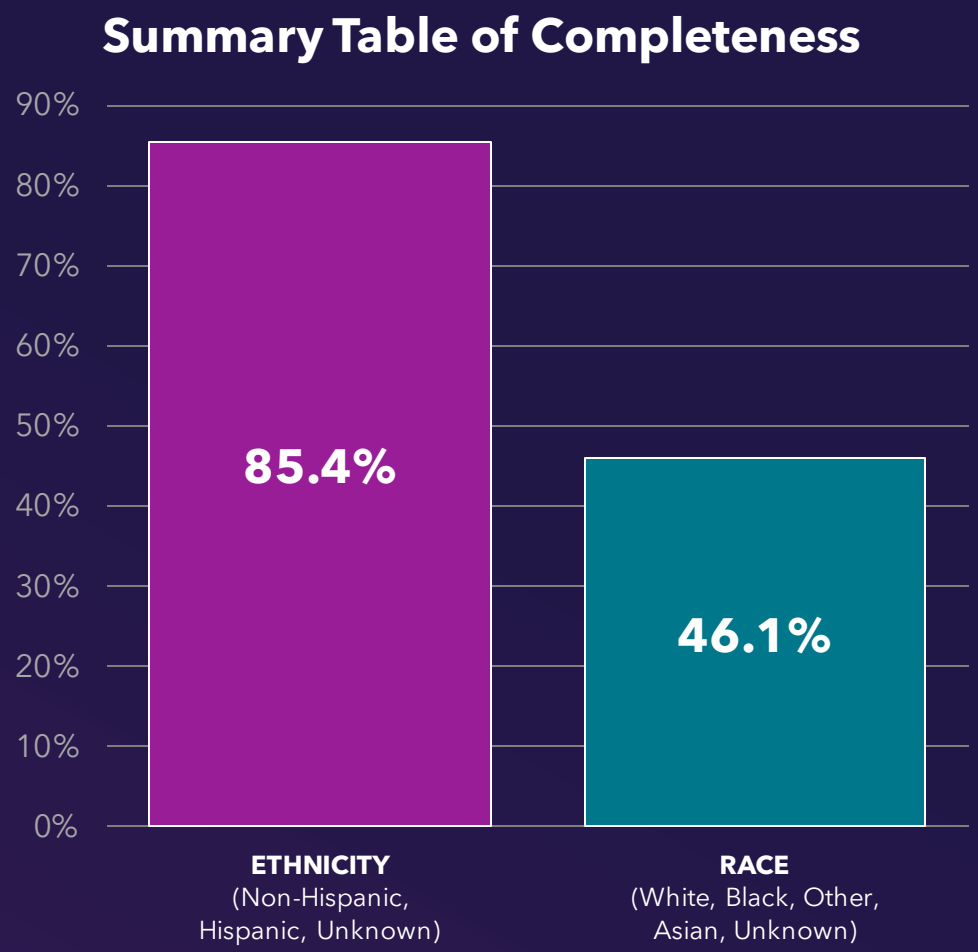
- **Limited context**: Incomplete information limits our ability to infer partial responses, reducing them to certain Race/Ethnicity combinations; Imputation requires linking relevant datasets (internal/external)

	Ethnicity	Race
2-Question Wordings	Hispanic or Latino	American Indian or Alaska Native
	Hispanic or Latino	Asian
	Hispanic or Latino	Black or African-American
	Hispanic or Latino	Native Hawaiian or Other Pacific Islander
	Hispanic or Latino	White
	Not Hispanic or Latino	American Indian or Alaska Native
	Not Hispanic or Latino	Asian
	Not Hispanic or Latino	Black or African-American
	Not Hispanic or Latino	Native Hawaiian or Other Pacific Islander
	Not Hispanic or Latino	White
	Not Hispanic or Latino	
1 or 2-Question Wordings	Hispanic or Latino	
		American Indian or Alaska Native
		Asian
		Black or African-American
		Native Hawaiian or Other Pacific Islander
		White

RACE & ETHNICITY COMPLETENESS

- Presenting responses from 1-question prompts as a 2-column format will necessarily omit some information

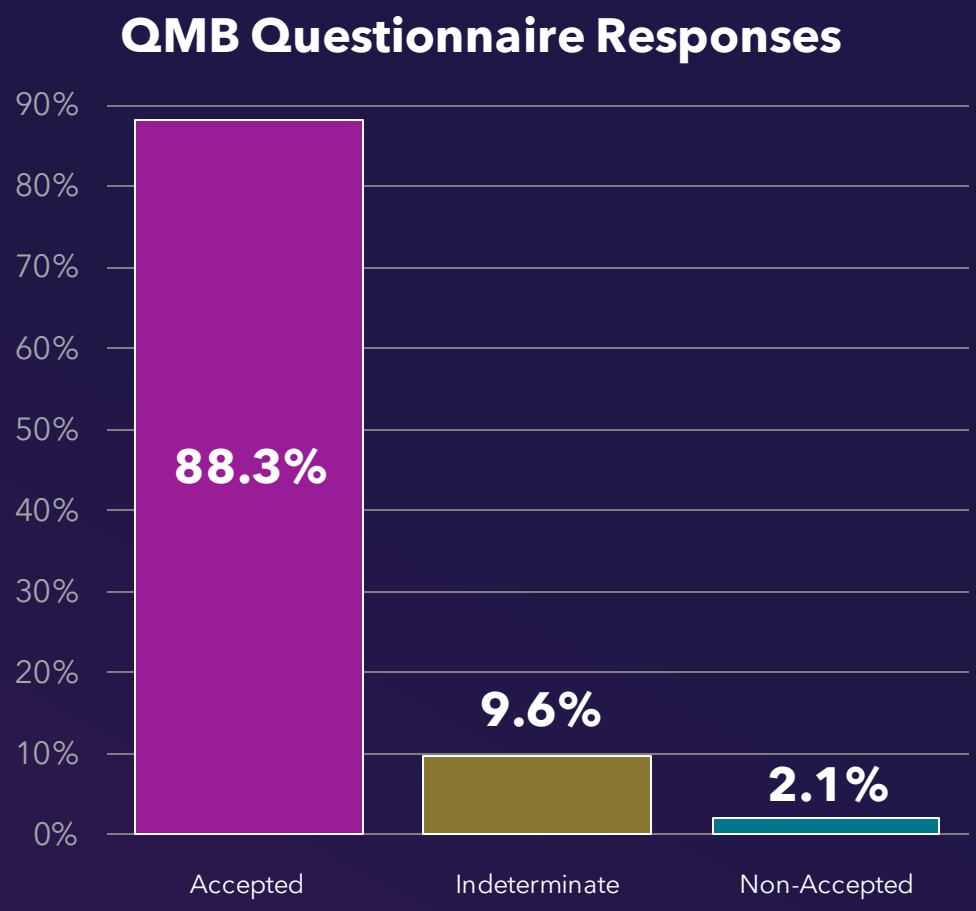
- The summary includes both literal nulls – of which there are none – and null-equivalent values ("Unknown" in this dataset) to indicate absence of information.



SUMMARY OF OMB ACCEPTED/NON-ACCEPTED RESPONSE WORDINGS

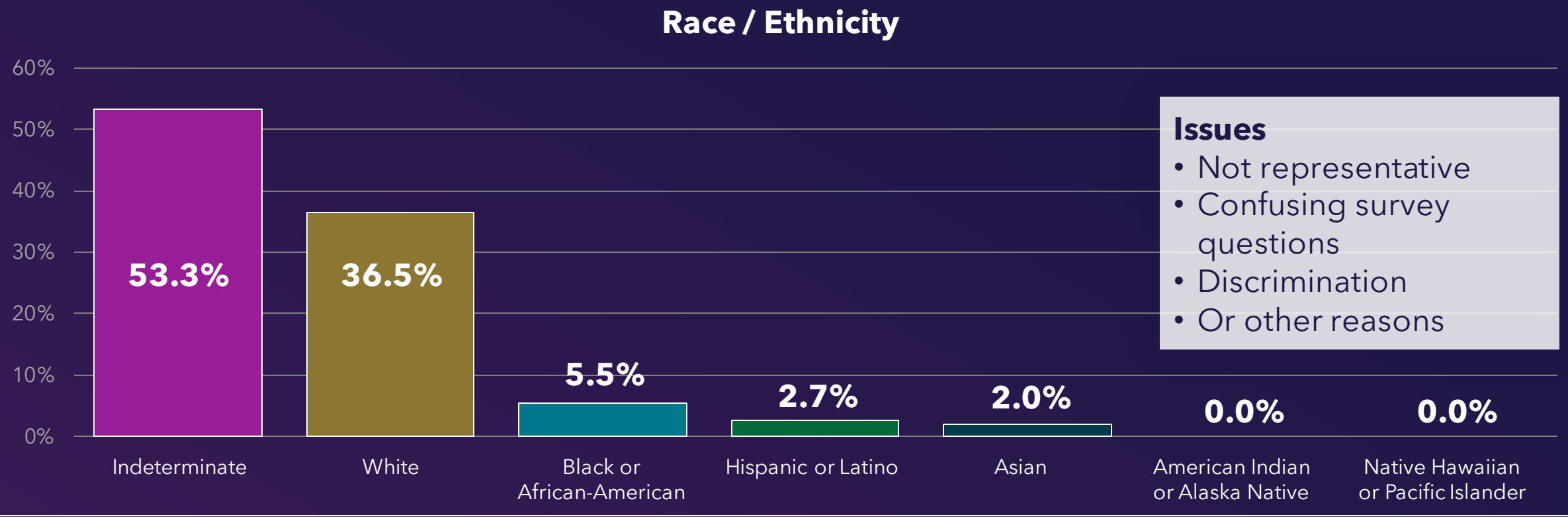
- Observations with OMB-accepted wordings have at least one non-null value for either ethnicity or race.
  - Responses to single-question prompts will only have a value in one of the columns.

- Noncompliant prompts would necessarily contain at least one answer with unacceptable wording and any prompts with no response at all cannot be assessed.



MINIMUM OMB RACE & ETHNICITY CATEGORY DISTRIBUTION

- Determined "Hispanic or Latino" by counting entries with "Hispanic" as the ethnicity value and "Unknown" as the race value. Every other value was determined by mapping the race values in the data back to the OMB minimum race categories.



NON-OMB CONFORMANT VALUES OF RACE AND ETHNICITY

	Conformant Mappable back to OMB or missing (“ ” or Unknown, a proxy for “ ”)	**Non-Conformant** “Other”; doesn’t map back to OMB
Race	97.91%	2.09%
Ethnicity	100.00%	0.00%

- Mapping responses from one-question prompts to a two-column format leads to missing information. Both literal nulls and null-equivalent values ('Unknown' in this dataset) indicate an equivalent absence of information.
- As patients are not generally required to complete the prompt and one-question responses produce missing values, they are considered conformant responses.
- The revised OMB guidance allows for additional values beyond the minimal set if they can be mapped back to the minimum set.

SOLUTION TO GENERATE SUMMARIES OF GRANULAR RACE/ ETHNICITY CATEGORIES THAT ARE OMB STANDARD COMPLIANT

- Incorporating a more-granular race categorization approach, such as **HL7 / CDC FHIR** Race and Ethnicity Standard, in addition to the OMB standard categories"
- This approach would involve modifying DB schema to include separate columns in the patients' table for tracking more-granular race/ethnicity categories
- Help reduce Social desirability bias

Unique Identifier	Hierarchical Code	Concept
2133-7	E	**ETHNICITY**
2135-2	E1	**Hispanic or Latino**
2137-8	E1.01	Spaniard
2138-6	E1.01.001	Andalusian
2139-4	E1.01.002	Asturian
2140-2	E1.01.003	Castillian
2141-0	E1.01.004	Catalonian
…	…	…

Unique Identifier	Hierarchical Code	Concept
1000-9	R	**Race**
2028-9	R2	**Asian**
2029-7	R2.01	Asian Indian
2130-5	R2.02	Bangladeshi
2131-3	R2.03	Bhutanese
2032-1	R2.04	Burmese
2033-9	R2.05	Cambodian
…	…	…

https://www.cdc.gov/nchs/data/dvs/race_ethnicity_codeset.pdf

ADDRESSING DATA QUALITY CHALLENGES: POTENTIAL SOLUTIONS

- Data Provenance
  - Source attribution, Interoperability, and Integration
- Data preprocessing steps
  - Documentation availability
- Data Quality measures
  - How missing data was handled; any imputation methods applied

ADDRESSING DATA QUALITY CHALLENGES: POTENTIAL SOLUTIONS

- Standardization
  - Clear definitions; Coordination with EHR vendors and curators; Documentation; drop-down options; user education/training;
  - Race and ethnicity are frequently used interchangeably, despite having distinct meanings.
- Metadata, Version control, and Ethical and legal considerations
- Improve Data Quality focusing on incomplete data
  - Encourage self-reporting; training for data gatherers

SUMMARY

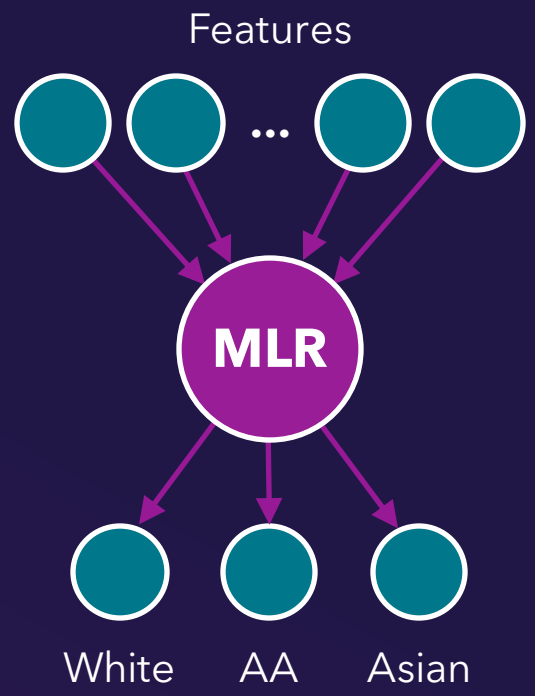
- As this sample dataset shows, EHR Race & Ethnicity data **often lacks completeness, compliance and context**. Secondary use of clinical data requires dealing with suboptimal data

- Our approach aims to mitigate this by **focusing on the available data**, making reasonable assumptions about patient behavior & accounting for minor deviations (business/technical decisions).
  - CDC can employ comparable methods and assumptions when handling RWE/EHR datasets with the above-mentioned characteristics

- **Absent strong links** between where and when the race and ethnicity **data are collected** and lacking both history and **context** for the values themselves, it is **difficult to impute** missing values.

CHALLENGES AND FUTURE DIRECTIONS

- More complete data may be amenable to more complex analysis, but some data sets, like the one used for this study, are **too limited** to readily support such techniques
- **Imputing Race and Ethnicity is complex**. If possible, collect better quality data. In the recent years, several methods have been proposed:
  - Multinomial Logistic Regression (MLR) doi: 10.1111/1475-6773.13171
  - Bayesian, doi: 10.1007/s10742-009-0047-1
  - AI/ML
  - Multiple Imputation models (https://pubmed.ncbi.nlm.nih.gov/35368775/)
  - NLP (https://pubmed.ncbi.nlm.nih.gov/31329882/)
  - Random Forest-based methods
- These approaches have shown to **reduce bias** for this purpose



Multiple Imputation of Missing Race and Ethnicity in CDC COVID-19 Case-Level Surveillance Data

Guangyu Zhang[1,*], Charles E. Rose[1], Yujia Zhang[1], Rui Li[2], Florence C. Lee[1], Greta Massetti[1], Laura E. Adams[1]

[1]CDC COVID-19 Response Team, Centers for Disease Control and Prevention, Atlanta, Georgia

[2]Health Resources and Services Administration, Rockville, Maryland, USA

Thank you!

Any questions?

– SUPPLEMENTAL SLIDES –

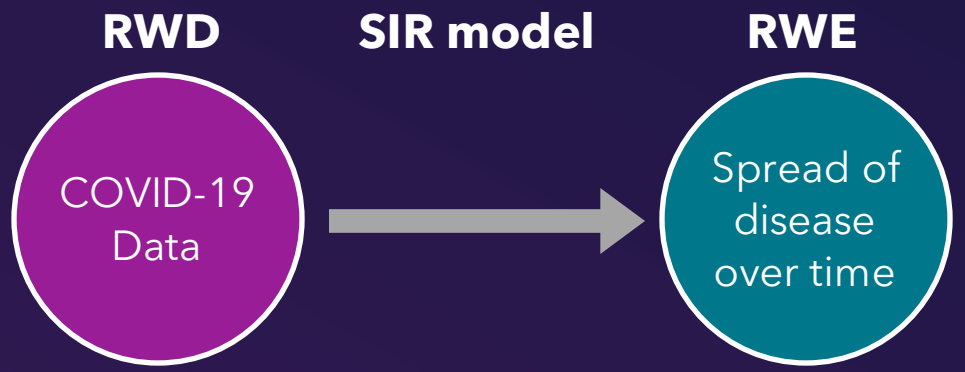
"RACE AND ETHNICITY ARE COMPLEX TERMS AND OFTEN USED INTERCHANGEABLY" GLEANED FROM STANFORD.EDU

Characteristic	Race	Ethnicity
Definition	A social construct that refers to a group of people who share physical characteristics (White, Black, Asian, Hispanic)	People who share a common culture (Irish, Italian, German, Chinese, Japanese)
Identification	Physical characteristics	Culture
Inherited or learned?	Inherited	Learned
Socially constructed	Yes (?)	No (?)
Immutable	No (?)	Yes (?)

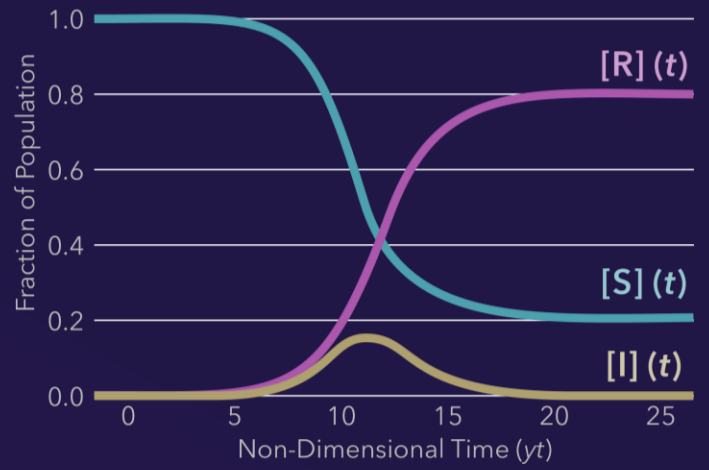
WHAT IS REAL-WORLD DATA (RWD) & REAL-WORLD EVIDENCE (RWE)?

- **RWD**: Data collected from sources other than controlled setting (or lab).
- **RWE**: Clinical evidence regarding the usage and potential benefits or risks of a product or interventions derived from analysis of RWD.

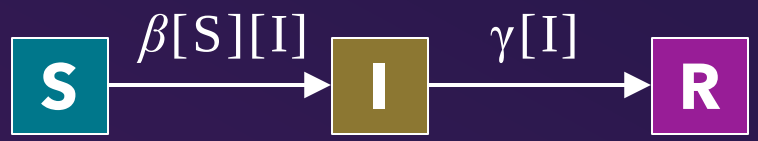
t	# of Infected
1	3
2	8
3	26
4	76
5	226
…	…
20	3



SIR Model Dynamics



Susceptible-Infectious-Recovered (SIR)



RWE (additional ex.):

- Infection Rate
- Mortality Rate
- Hospitalization Rate
- Vaccination Rate, etc.

EXAMPLES OF RACE/ETHNICITY IN RWE

- "Specifically, throughout the pandemic, reports indicated that Hispanic or Latino, Black, AI/AN, and Asian individuals had higher rates of infection, hospitalizations, and deaths compared to their white counterparts …"
  - OEI-05-20-00540, HHS-OIG report on CDC Overcomes Data Challenges to Tackle COVID-19 Health Disparities
- "Long believed to be a disease primarily of White people, we recently showed that multiple sclerosis (MS) incidence is highest in Black followed by White women and significantly lower in Hispanic and Asian individuals residing in Southern California"
  - Langer-Gould et al, https://pubmed.ncbi.nlm.nih.gov/35501161/
- New hackathons and crowdsourcing for Race/Ethnicity Data collection/usage
  - https://precision.fda.gov/challenges/30