Sentinel

Improving Data Capture of Race and Ethnicity

Monica Ter-Minassian, ScD
Mid-Atlantic Permanente Research Institute,
Kaiser Permanente Mid-Atlantic States

February 23, 2023

Introduction

- We conducted a narrative literature review to find methods to improve data capture of race and ethnicity for the FDA Sentinel distributed database.

- **Incentive:** Race and ethnicity (R/E) data from Sentinel Data Partners (Administrative claims-based and claims-linked-to-EHR health care systems) are incomplete or unable to be shared externally.

Methods

- We used a snowball approach to search the literature with PubMed, Google and Google Scholar (examined citations and papers that cited the original article)

- What could be done at the Data Partner level vs. Sentinel Common Data Model (SCDM) level?
  - Acknowledging the limitations of the SCDM, which does not include names or complete addresses (for patient privacy), **only ZIP code**, we focused on literature that discussed methods that might be most relevant to Sentinel but did not limit the review to this

Inclusion Criteria

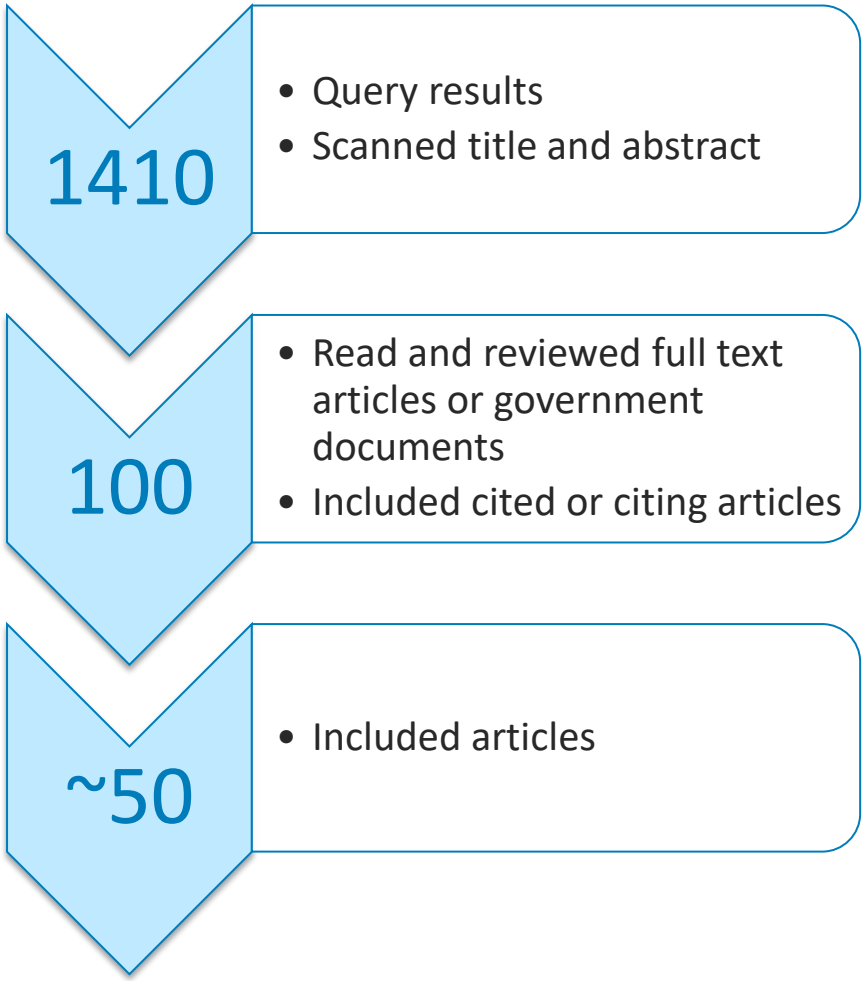
- Query restricted to full text English language articles pertaining to U.S. healthcare system data
- Limited to 2000 to current date

Exclusion Criteria

- Did not refer to race and ethnicity missingness or data quality improvement
- Used data or databases that would be inaccessible to most researchers

Methods, cont.

Final Query between years 2000 to current
Missing Race and ethnicity and
data improvement, or data quality
data augmentation, or surrogate measures
US Census 2010 or 2020
imputation
zip code, or ZCTA
claims
language preference or interpreter
birth certificates or natality data
first name lists or surname lists
pharmacoepidemiology, or pharmacovigilance, drug safety
American Indian or Alaskan Native
health system race data quality
health data collection and race
race health construct
OMB and changes and race
birth certificates and infant and race and ethnicity
self-report and race and ethnicity and clinics



Results

We focused on 3 main themes:

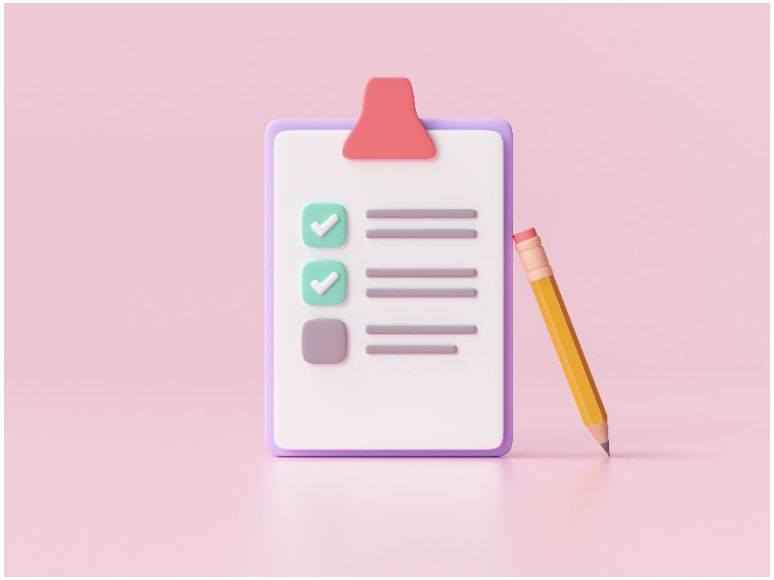
1. **Methods to collect self-reported** or administratively assigned R/E values at the healthcare system level
2. **Data augmentation** with external sources that calculate R/E distributions for locations or names
   - *Focused on ZIP code level linkage*
3. **Data imputation and validation** using internal and external sources
   - *Compared 5 large studies*

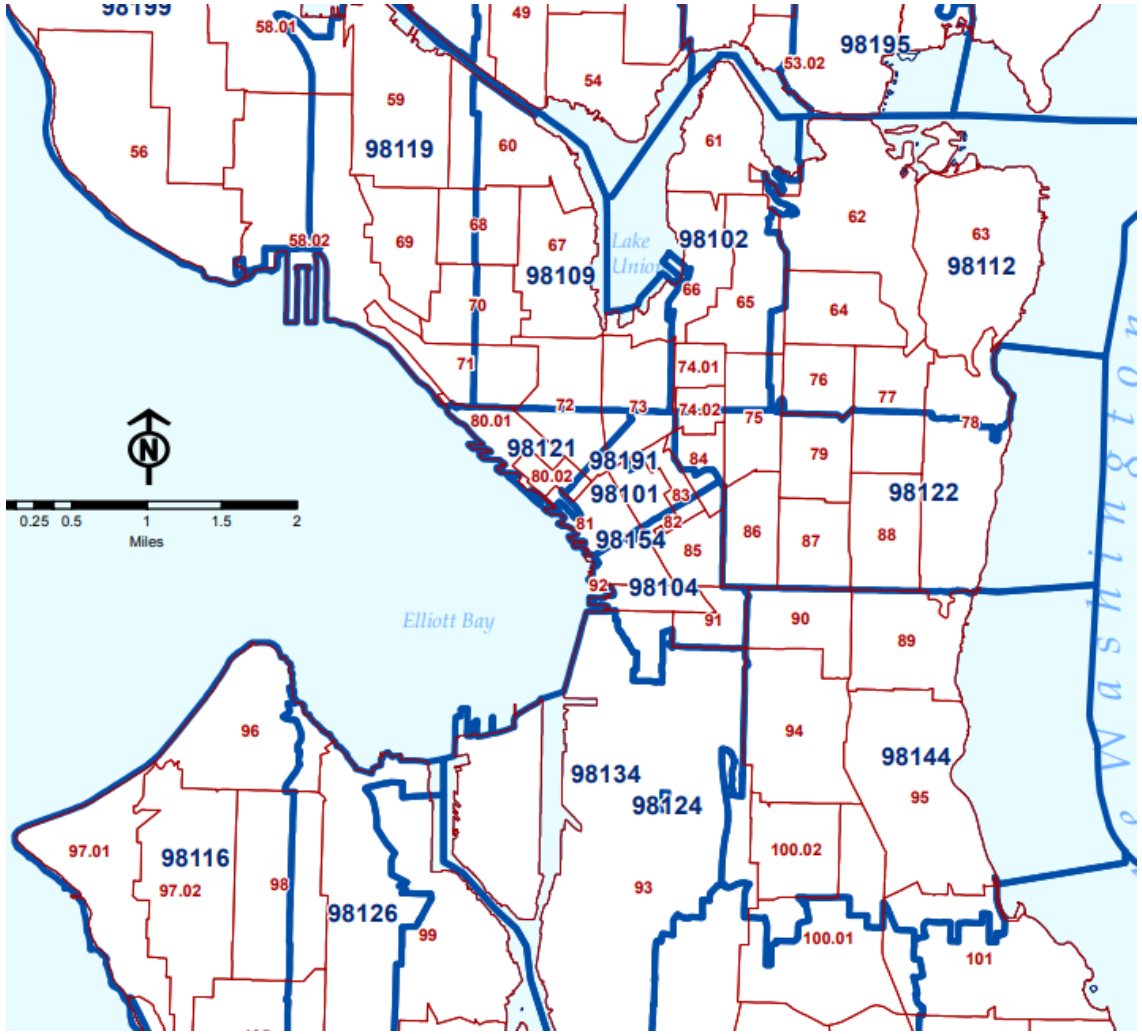
Self-Reported/Administrative R/E

Data collected at the healthcare system level varies between EHR-based and claims-based healthcare systems

- Categories generally include Office of Management and Budget (OMB) standards (Race: American Indian or Alaskan Native, Black, Asian Native Hawaiian/Pacific Islander, White; Ethnicity: Hispanic)

- EHR systems more likely to collect R/E at member registration or in the provider's office
  - Frequently, questions asked verbally to the patient and the provider or clerk then enters responses directly into EHR
- Claims more likely to administer a written form at time of member enrollment to be completed by the patient





Zip code in dark blue, census tracts in red www.seattle.gov

Data Augmentation

Most R/E data comes from the **U.S. Census Bureau**, from publicly available decennial census or 5-year American Community Survey (ACS) data

- **Geocoding** can be used to create a vector of probabilities for R/E categories associated with a geocode
  - Linking to U.S. Census block group, tract or ZIP Code Tabulation Area (ZCTA) can provide proportions of R/E at those levels, at a 5-year update frequency
  - ZIP codes change frequently, so it is important to include time frames when using ZIP codes to ZCTA or Census tract mapping
- Linking **ZIP code to Census tracts** is preferred because it is a more spatiotemporally stable and accurate geographical unit;
  - Allows linkage to ACS prior to 2013

Data Augmentation, cont.

Surname or first name lists

- Can be used to create a vector of probabilities for R/E categories associated with the name
  - U.S. Census Bureau has publicly available **surname** lists specific for Asian or Pacific Islander names and for Spanish origin names
  - Other groups have developed **first name** lists that assign probabilities to individuals based on OMB standard categories
    - Some first name lists were developed from mortgage data

Frequently Occurring Surnames in the 2010 Census: Top 1,000 Surnames (excerpt)

SURNAME	RANK	FREQUENCY (COUNT)	PROPORTION PER 100,000 POPULATION	CUMULATIVE PROPORTION	PERCENT NON-HISPANIC OR LATINO WHITE ALONE	PERCENT NON-HISPANIC OR LATINO BLACK OR AFRICAN AMERICAN ALONE	PERCENT NON-HISPANIC OR LATINO ASIAN AND NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE	PERCENT NON-HISPANIC OR LATINO AMERICAN INDIAN AND ALASKA NATIVE ALONE	PERCENT NON-HISPANIC OR LATINO TWO OR MORE RACES	PERCENT HISPANIC OR LATINO ORIGIN
SMITH	1	2,442,977	828.2	828.2	70.9	23.1	0.5	0.9	2.2	2.4
JOHNSON	2	1,932,812	655.2	1,483.4	59.0	34.6	0.5	0.9	2.6	2.4
WILLIAMS	3	1,625,252	551.0	2,034.4	45.8	47.7	0.5	0.8	2.8	2.5
BROWN	4	1,437,026	487.2	2,521.6	58.0	35.6	0.5	0.9	2.6	2.5
JONES	5	1,425,470	483.2	3,004.8	55.2	38.5	0.4	1.0	2.6	2.3
GARCIA	6	1,166,120	395.3	3,400.1	5.4	0.5	1.4	0.5	0.3	92.0

Data Augmentation, cont.

Supplemental data sources of self reported race

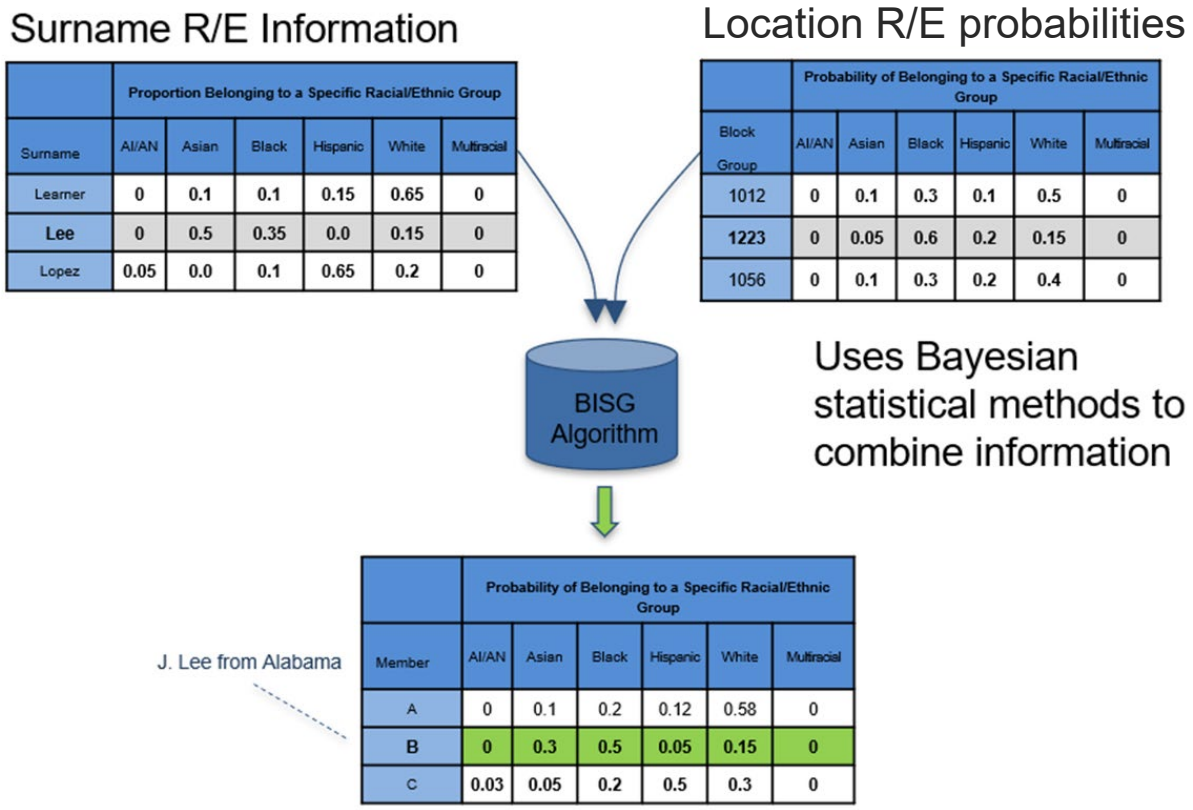
- State or hospital birth registries containing maternal and paternal self-reported R/E
- Cancer registries
- Centers for Medicare and Medicare Services data containing self-reported R/E from annual surveys
- Social Security Administration's Master Beneficiary Record
- Indian Health Service
- The Healthcare Cost and Utilization Project State Inpatient Databases
  - Race available for some states

- Data may be available with a Data Use Agreement

Data Imputation

Bayesian imputation

- The most commonly used and cited method was the Bayesian Improved Surname and Geocoding algorithm (BISG) developed by Elliot et al. (2009) at RAND Corporation
  - This method uses Bayesian imputation to combine probability information from surname and geocode
- A more recent version MBISG2.0 (Haas, 2019) included geocode, surname and first name probabilities, and adjusts for:
  - Compound last names and Spanish language preference, residents of Puerto Rico, and age
  - CMS coverage type, gender, and low-income indicators using multinomial logistic regression modeling



	Proportion Belonging to a Specific Racial/Ethnic Group					
Surname	AI/AN	Asian	Black	Hispanic	White	Multiracial
Learner	0	0.1	0.1	0.15	0.65	0
Lee	0	0.5	0.35	0.0	0.15	0
Lopez	0.05	0.0	0.1	0.65	0.2	0

	Probability of Belonging to a Specific Racial/Ethnic Group					
Block Group	AI/AN	Asian	Black	Hispanic	White	Multiracial
1012	0	0.1	0.3	0.1	0.5	0
1223	0	0.05	0.6	0.2	0.15	0
1056	0	0.1	0.3	0.2	0.4	0

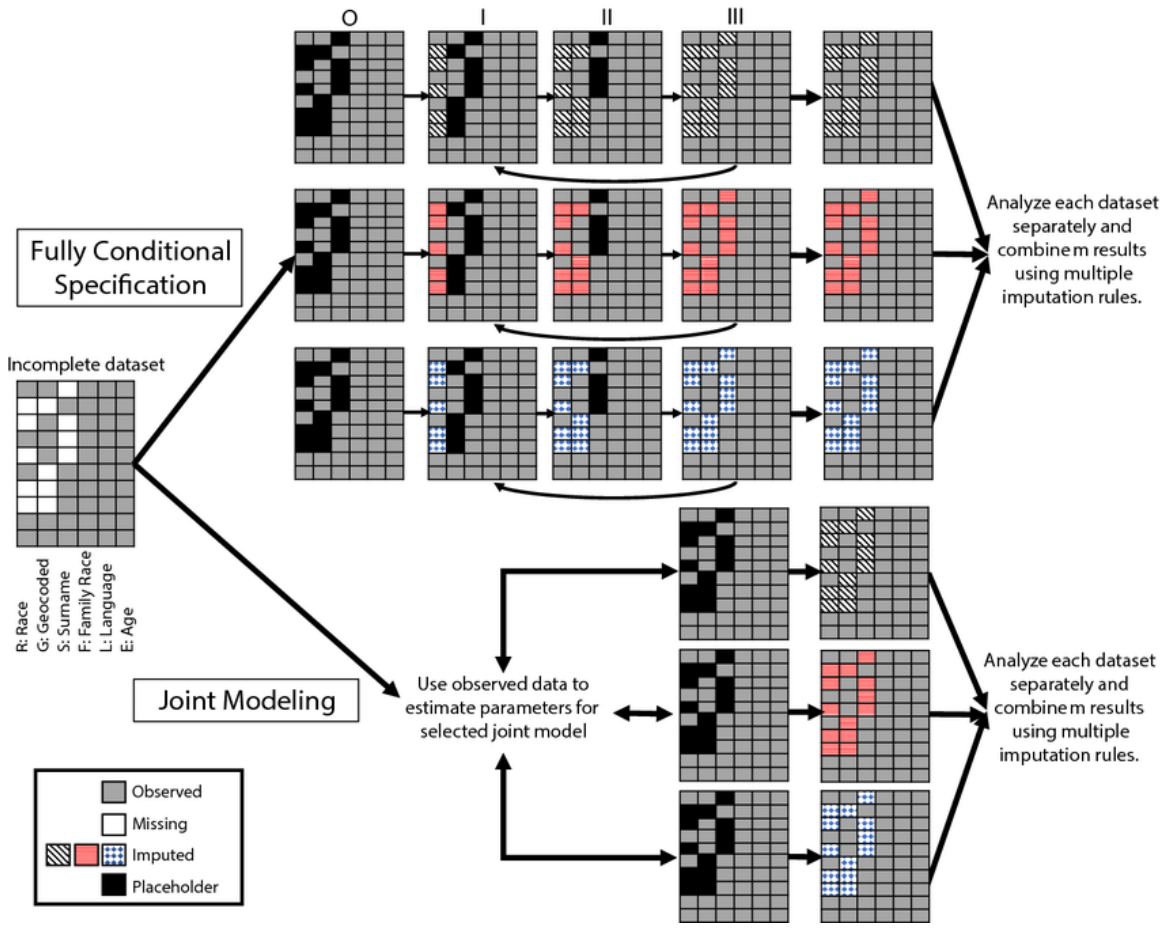
	Probability of Belonging to a Specific Racial/Ethnic Group					
Member	AI/AN	Asian	Black	Hispanic	White	Multiracial
A	0	0.1	0.2	0.12	0.58	0
B	0	0.3	0.5	0.05	0.15	0
C	0.03	0.05	0.2	0.5	0.3	0

https://www.cms.gov/files/document/new-method-improve-measurement-race-and-ethnicity-cms-data-and-applications-inequalities-quality.pptx

Data Imputation, cont.

Multiple imputation (MI)

- Replaces each missing value with a set of plausible values, producing a set of imputed datasets
- **Conditional MI:** allows variables of different types to be modeled separately
  - Multivariate imputations by chained equations (using R *mice*) or chained equations in a Bayesian framework (using R *mi*)
- **Joint MI**: assumes that data follow a joint distribution, typically a multivariate normal (MVN), and draws imputed values from this distribution (*Amelia* package in R)
  - Theoretically better, but not good for categorical variables
- **Ma et al. (2018) and Silva (2019) found that conditional MI was the best method**



Silva, Gabriella & Trivedi, Amal & Gutman, Roee. (2019). Developing and evaluating methods to impute race/ethnicity in an incomplete dataset. Health Services and Outcomes Research Methodology. 19. 1-21. 10.1007/s10742-019-00200-9.

Data Imputation Methods Comparison

After validating imputed datasets with known self-reported race and ethnicity data and assessing performance with measures such as weighted correlation, ROC and Brier score (squared errors of prediction), researchers found:

Improving R/E groups

- Most methods imputed White and Black with high performance measures
- Imputing Hispanic ethnicity improved by including Spanish language preference, residence in Puerto Rico and using the Spanish surname lists
- Imputing Asian or Native Hawaiian/Pacific Islander races improved using the Asian/Pacific Islander surname lists
- All imputation methods did not impute American Indian/ Alaskan Native well
- Only BISG and MBISG2.0 attempted to impute multiracial, but did not do so successfully

Race and Ethnicity	Improvement
American Indian/Alaskan Native	-
Asian	✓
Black	✓
Native Hawaiian/Pacific Islander	✓
Hispanic	✓
Multiracial	-
White	✓

Strengths and Limitations

Imputation relies on:

- **Self-reported** data which may be biased due to sampling methods
  - May miss less common groups such as American Indian and multiracial groups
  - **May not capture the pediatric population R/E well which relies on accurate maternal and paternal self-report**
  - Intake form questions vary in different health care systems
  - Individual interpretation when self-reporting

- **Linkage on name**
  - Surname or first name that is not unique to a R/E group may not have informative probability
  - **Smaller populations of first name datasets sampled may not be representative of the population covered by the Sentinel Distributed Database**

- **Linkage on Geocoding Level**
  - Ability to link full address or zip code only data
  - Census tract, ZCTA represent different sampling and population density
  - ZIP code to ZCTA is not a 1:1 link and may change over time
  - **Geocoding is dependent on segregated or homogeneous neighborhoods; may be less useful in time as neighborhoods continue to desegregate and diversify**

Recommendations for Sentinel

At the **Sentinel Operations level** with only ZIP code available, imputation based on zip code to census tract mapping may be the most straightforward

- Consider requesting geocoded data at the U.S. Census block group level (or geocoding with spatial disaggregation).
- Consideration: if Sentinel adopts imputed R/E values, an indication of whether R/E data is imputed should be added to the Sentinel Common Data Model to distinguish from self-reported R/E data if both are included
- Other data sources, especially CMS, could be consulted to obtain self-reported data and MBISG2.0 imputed values with a DUA

If imputation can be performed at the **Data Partner level**, the BISG method is recommended, supplemented by language preference and first name probabilities

Team

FDA:

José Hernández-Muñoz

Aloka Chakravarty

Sentinel Operations Center:

Ryan Schoeplein

Stacey Moisuk

KPMAS:

Monica Ter-Minassian

Anna DiNucci

Issmatu Barrie

Acknowledgements: Sonia Kim, Biostatistician for statistical help. and Jacqueline Puigbo, Epidemiologist for statistical guidance.

Funding: This project was supported by Task Order 7540119F19001 under Master Agreement 75F40119D10037 from the U.S. Food and Drug Administration (FDA).