National Center for Health Statistics

# Race and Ethnicity Modeling Applied to Linked Health Data

Lisa B. Mirel, Dean Resnick, Jessie L. Parker, Cindy Zhang, and Christine Cox

FCSM Session D-2: *Dying for Better Data: Using Administrative Records to Improve the Quality and Utility of Health Survey Data*

October 26, 2022

# Background

- Linked health data enable researchers to explore social determinants of health (SDOH) and health equity research
- However, some survey and administrative data lack key variables for SDOH and health equity research
- NCHS has a data linkage program that links NCHS survey data with administrative records
- To further enhance SDOH and health equity research our team explored imputing race and ethnicity information that could be applied to linked data

# Background: NCHS Data Linkage Program

- Create linked data files that support high quality research and program evaluation

- Utilize state of the art linkage methodologies and provide documentation and support for analyzing linked data files

- Explore innovative methods for providing researcher access to linked data

# NCHS surveys used in linkages

## National Health Interview Survey (NHIS)

A nationally representative, cross-sectional sample of the US civilian noninstitutionalized population, which includes a household interview survey that serves as an important source of information on the nation’s health

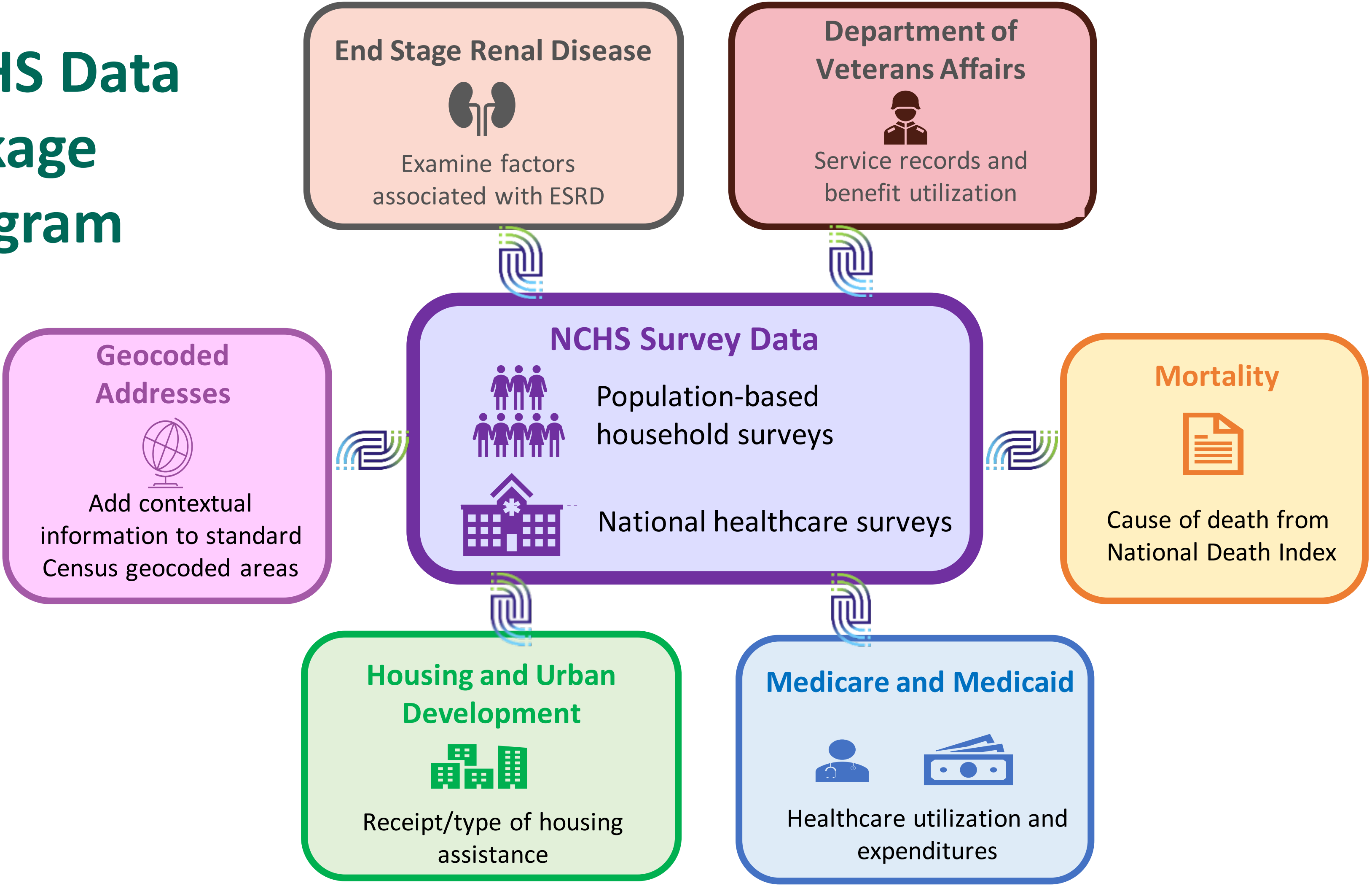
## National Health and Nutrition Examination Survey (NHANES)

A nationally representative, cross-sectional sample of the US civilian noninstitutionalized population, which includes a household interview followed by an examination in a mobile examination center that serves as an important source of information on the health and nutritional status of adults and children

## National Hospital Care Survey (NHCS)

NHCS collects data on patient care in hospital-based settings (inpatient, emergency, and outpatient departments) to describe patterns of health care delivery and utilization in the US

# NCHS Data Linkage Program



**NCHS Survey Data**

Population-based household surveys

National healthcare surveys

**End Stage Renal Disease**

Examine factors associated with ESRD

**Department of Veterans Affairs**

Service records and benefit utilization

**Geocoded Addresses**

Add contextual information to standard Census geocoded areas

**Mortality**

Cause of death from National Death Index

**Housing and Urban Development**

Receipt/type of housing assistance

**Medicare and Medicaid**

Healthcare utilization and expenditures

# Motivation: adding race and ethnicity data to linked files

- NHCS includes information on conditions and treatments of patients from sampled hospitals but data on race and ethnicity is limited
  - In the 2016 NHCS, nearly 70% of patient records were missing race and ethnicity information

- The linked NHCS-NDI data provide race and ethnicity information for decedents, but it remains missing for those assumed alive
  - This limits the analyses that can be conducted by race and ethnicity

- Analysis: assess mortality rates by race and ethnicity; the denominator should include both those who are alive and have died

# Methods

- Use model for imputing race and ethnicity

- Apply model to NHIS which includes self-reported race and ethnicity
  - Assess model overall and at precision thresholds, using positive/negative predictive value and kappa statistics

- Apply model to linked NHCS-NDI data and calculate mortality rates by race and ethnicity, overall and a precision threshold

# Model for race and ethnicity imputation

- Model builds on work by Marc Elliott, et al., described as **Bayesian Surname Geocoding** (BSG) method
- Model leverages race and ethnicity proportions (priors) derived from Census block
- First Name – used in analysis (in addition to last name)
- Name proportions among race and ethnicity used
  - e.g., $P$(Last Name = 'Clemente' | Hispanic),

    $P$(First Name = 'Anna' | Hispanic)

# Modeling strategy

- Posterior distribution computation:

$$P \sim P([Race{\scriptstyle\square}Eth] = R \mid Census\ Block) \\ \cdot P(FN \mid [Race{\scriptstyle\square}Eth] = R) \cdot P(LN \mid [Race{\scriptstyle\square}Eth] = R)$$

*P: Probability, R:* Race and Ethnicity, *FN:* First Name, *LN:* Last Name

- Imputation category assigned to group with highest probability

# Imputation categories

- Ethnicity
  - Hispanic (*takes precedence over race, e.g., persons described as Hispanic are not assigned a race group*)
- Race
  - White (non-Hispanic)
  - Black (non-Hispanic)
  - Asian or Pacific Islander (API, non-Hispanic)
  - American Indian or Alaskan Native (AIAN, non-Hispanic)

# Evaluation with 2018 NHIS data

- Imputed race and ethnicity was compared to the 2018 NHIS respondent reported race and ethnicity (“Gold Standard”)

- Assessed Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

- Compared respondent reported to imputed race and ethnicity using Cohen’s Kappa
  - overall and by sex and age

- Initial evaluation uses all records

- Refined evaluation uses records with precision ($P(Imputed\ Race = R)$> 80%)

# Respondent-reported versus imputed: PPV and NPV

| | PPV | NPV |
|---|---|---|
| **Hispanic** | | |
| Overall | 89.1 | 95.4 |
| Precision>80% | 94.2 | 96.6 |
| **Non-Hispanic Black** | | |
| Overall | 72.1 | 95.8 |
| Precision>80% | 87.8 | 97.5 |
| **Non-Hispanic White** | | |
| Overall | 87.3 | 87.3 |
| Precision>80% | 90.8 | 94.7 |

# Respondent-reported versus imputed: PPV and NPV (cont.)

| | PPV | NPV |
|---|---|---|
| **Non-Hispanic Asian*** | | |
| Overall | 70.5 | 97.8 |
| Precision>80% | 84.4 | 98.4 |
| **Non-Hispanic Other*** | | |
| Overall | 56.8 | 98.6 |
| Precision>80% | 82.5 | 99.0 |

* Comparison was made using NHIS public use categories for race and ethnicity (e.g., respondent reported non-Hispanic Asian was compared to imputed non-Hispanic API and respondent reported non-Hispanic Other was compared to imputed non-Hispanic AIAN)

# Kappa statistic: all records and precision >80%

| | Hispanic | Non-Hispanic Black | Non-Hispanic White | Non-Hispanic Asian | Non-Hispanic Other |
|---|---|---|---|---|---|
| **All** | **0.80** | *0.66* | *0.72* | *0.66* | 0.25 |
| All precision (p) >80% | **0.86** | *0.80* | **0.81** | *0.77* | 0.30 |
| **Female** | *0.78* | *0.68* | *0.71* | *0.64* | 0.25 |
| Female (p>80%) | **0.85** | **0.82** | **0.81** | *0.76* | 0.32 |
| **Male** | **0.82** | *0.64* | *0.73* | *0.68* | 0.24 |
| Male (p>80%) | **0.88** | *0.77* | **0.82** | *0.79* | 0.28 |
| **Age 65+** | **0.82** | *0.71* | *0.76* | *0.69* | 0.26 |
| Age 65+ (p>80%) | **0.90** | **0.86** | **0.87** | *0.79* | 0.29 |

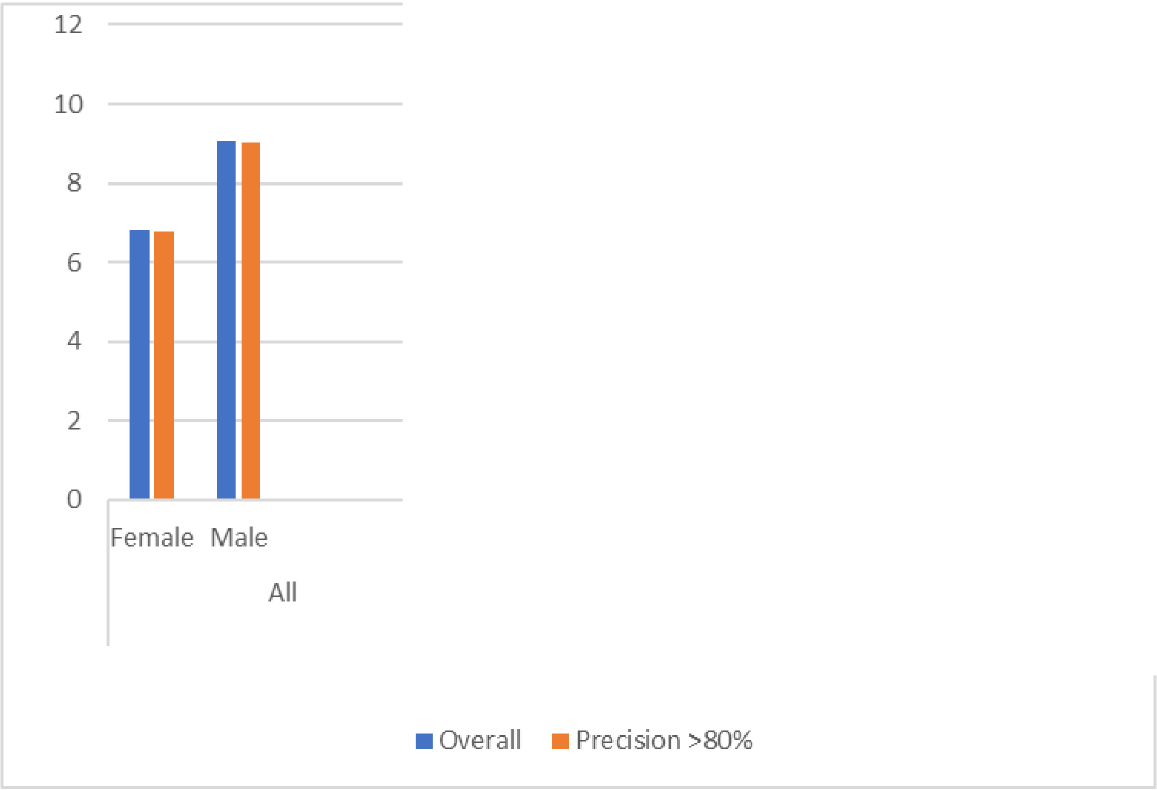
# Implementation: race and ethnicity imputation model applied to linked NHCS-NDI data

- 70% of NHCS patients are missing race and ethnicity

- Analysis: assess mortality rates by race and ethnicity; the denominator should include both those who are alive and have died

- Post-hospitalization mortality rates calculated by time after discharge (0-30 days), age (65 and over), sex and imputed race and ethnicity

# Mortality rates 0-30 days post hospital discharge for 65 and older



NOTE: small number of 2016 NHCS patients imputed to non-Hispanic, AIAN. They are not included in this tabulation.

# Summary

- This research demonstrates that it is possible to reliably impute such information using Bayesian techniques applied to data obtained from other sources
  - Precision estimates >80% seem to increase concordance

- Imputation strategy employed here is relatively straightforward and uses publicly available sources to develop the race and ethnicity distributions

- Applying statistical techniques to impute critically important health information can enable further study of the role of race and ethnicity in health outcomes

# References

- Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., & Lurie, N. (2008). A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health services research*, *43*(5p1), 1722-1736.
- Tzioumis, K. (2018). Demographic aspects of first names. *Scientific data*, *5*(1), 1-9.
- Data:
  - https://www2.census.gov/census_2010/redistricting_file--pl_94-171/
  - https://www.census.gov/topics/population/genealogy/data/2010_surnames.html
  - https://www.nature.com/articles/sdata201825

# NCHS Data Linkage Program

Contact: Lisa Mirel Lmirel@cdc.gov

**Subscribe to the NCHS Data Linkage Program LISTSERV** to receive updates! Email a message to list@cdc.gov Leave the subject line blank. In the body of the message, type or paste:

SUBSCRIBE NCHS-DATA-LINKAGE-PROGRAM lastname, firstname

where ‘lastname, firstname’ is your last and first name.

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.