An Assessment of Data Disclosure Risk for the Proposed NHANES 2017-2020 Pre-Pandemic Public Use File

Board of Scientific Counselors Meeting

January 27, 2021

Tom Krenzke, Jane Li, Lin Li, Westat

nhanes
The Nation's Mobile Health Survey

Background

- NHANES 2017-2020 pre-pandemic public use file (PUF)
  - In-scope are: Demographics, Body Measures, Blood Pressure, Cholesterol, Diabetes, Oral Health
- 30 Primary Sampling Units (PSUs) in 2017-2018 and 18 PSUs in 2019-2020
- Assume the intruder can…
  - Identify the set of PSUs and the set of respondents in 2019-2020 by differencing the 2017-2020 PUF with the 2017-2018 PUF
  - Know the names of the 18 counties involved in 2019-2020 due to outreach activities, and the physical presence of the MEC

nhanes
The Nation's Mobile Health Survey

Risk Assessment Process and Risk Reduction Factors

- Process
  - Assess PSU-level risk -- Identifying a county and associating it with a cluster of records
  - Assess individual-level risk
    - Combining categorical indirect identifying variables together
    - Outlying values of continuous variables
- Some risk reduction factors
  - Lowest geography – Variance estimation codes
  - Sampling fraction – 0.005% of nation, about 0.039% of county on average
  - Recodes
  - Variable suppression
  - Imputation
  - Controlled random treatments

nhanes
The Nation's Mobile Health Survey

PSU-level Re-identification Risk

- Goal of intruder: To identify and associate county names with individual records
- Assume the intruder knows the set of 18 counties – conservative
- Assume the intruder knows the variance estimation codes can be used to determine a set of records that can potentially be associated with a specific county
- Use NHANES data and weights to estimate 11 county-level proportions (e.g., 65+, Hispanic, Asian, Born outside US)
- Gather estimates from the American Community Survey (ACS) for the 18 counties

nhanes
The Nation's Mobile Health Survey

Probabilistic Record Linkage

- Use probabilistic record linkage (Fellegi and Sunter, 1969) to quantify the likelihood of successfully linking a county in the NHANES file to the ACS file of estimates; each file is subset to the 18 counties
  - Form pairs of records -- one record from each file
  - Scores each pair using a likelihood-ratio match weight
  - Check to see if the highest scoring pair is a correct match
- Results (assuming the intruder knows which 18 counties are in the sample)
  - 8 counties can be easily identified
  - 6 counties can be logically re-identified once the above 8 counties have been identified
  - 4 have a lower chance of re-identification

nhanes
The Nation's Mobile Health Survey

Individual-level Re-identification Risk

- Combinations of indirect identifier variables
- Estimate the re-identification risk of the file as:

$$GlobalRisk = \sum_{SU} P\left(F_k = 1 \mid f_k = 1\right)$$

  - *SU* is the set of sample uniques
  - f_k is the sample frequency in cell *k*
  - F_k is the population frequency in cell *k*

nhanes
The Nation's Mobile Health Survey

Log-linear Modeling Approach

- F_k needs to be estimated in practice
  - Skinner and Shlomo (2008) log-linear model approach is used
  - Uses weights calibrated to the county population
- Assume the intruder…
  - knows 10 indirect identifying variables accurately, including the identity of 8 or 14 counties
  - does not know who is in the sample
  - will identify sample uniques and attempt to match them to the population
- Goodness of fit measure allows to determine underfit (overestimate of risk) and overfit (underestimate of risk)
  - Usually an all-two-way interaction model is sufficient

nhanes
The Nation's Mobile Health Survey

Variables Used in Model

- PSU – County ID, where the counties that cannot be re-identified are grouped together (10 or 4)
- Gender
- Age
- Race/Ethnicity
- Country of birth
- Education attainment
- Marital status
- Ever served in armed forces
- Number of children 5 years or younger
- HH income

nhanes
The Nation's Mobile Health Survey

Results

- Six runs conducted while varying the set of identifying variables (first five assume 8 identifiable counties, last one assumes 14)

Run	Action	Risk
1	All variables	Low-to-moderate
2	Dropped # of children in HH	Low
3	Dropped HH income	Very low-to-low
4	Dropped ever served in armed forces	Very low
5	Added HH income	Low
6	Assume 14 identified counties	Low-to-moderate

nhanes
The Nation's Mobile Health Survey

Other Risk Assessments Conducted

- Relative risk
  - Exhaustive tabulations of 4-way tables from 13 indirect identifying variables
  - Record the violations of 3-anonymity
  - Identify the categories of records that cause the most violations

- Continuous variables
  - Reviewed distributions for income-to-poverty ratio, height, weight
  - Income-to-poverty ratio is currently top-coded
  - Top coding is not applied to extreme height or weight

Recommended Confidentiality Edits

- Suppress (RDC release only)
  - Education level children/youth 6-19
  - Served in military
  - Age in months at exam 0–19 (release BMI category for children/adolescents)
  - Household income and Family income (release Income-to-poverty ratio)
- Recode
  - Marital Status as 1 = Married/Living with partner, 2 = Widowed/Divorced/Separated, and 3 = Never married
  - Length of time in US -- TBD
- Re-run risk assessment analysis with above changes and then re-evaluate need to suppress
  - Age in years at screening
  - Pregnancy status at exam

Recommended Confidentiality Edits

- Current approach is to mask the variance estimation codes through controlled random swapping (Park, 2008)
- Propose to do the following:
  - Increase swapping rate for re-identified PSUs
  - Target swapping for individuals with high risk (from log-linear model or extreme height or weight)

nhanes
The Nation's Mobile Health Survey

References

- Fellegi, I., and Sunter, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

- Park, I. (2008). PSU masking and variance estimation in complex surveys. *Survey Methodology*, Vol 34, No. 2, pp. 183-194.

- Skinner, C.J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-linear Models. *Journal of American Statistical Association*, 103, 989–1001.

nhanes
The Nation's Mobile Health Survey