Procedures To Reduce the Risk of Respondent Disclosure in a Public-Use Data File: The National Immunization Survey

Meena Khare (NCHS),
Michael P. Battaglia (Abt Associates),
David C. Hoaglin (Abt Associates),
and Robert A. Wright (NCHS)

Confidentiality, Disclosure , and Data Access

- *Most Public-Use Data Files (PUFs) are at person level*
- *Potential conflict between data users' need for detailed information and protecting confidentiality*
- *Detailed information collected: Demographic, socioeconomic, geographic data and other characteristics*
- *Unique and rare characteristics of respondents increase risk of disclosure*
- *Explosion in exogenous data files and rapidly growing information technology (e.g., Growth in Birth Certificate information, editors:Doyle et al., 2001, pp45-51)*
- *Section 308(d) of the Public Health Services Act and the Privacy Act of 1974 promise confidentiality of information and protecting identity of respondents*

National Immunization Survey (NIS)

- ***Large ongoing list-assisted RDD survey, conducted by the CDC since April 1994***

- ***Measures vaccination coverage rates among children aged 19-35 months at national, state, and urban area levels (78 IAP areas)***

- ***Monitors Healthy People 2000 and 2010 Goals***
  - *>90% Coverage: 4DTP, 3Polio, 1MCV, 3Hib, 3HepB, and 4:3:1:3 series*

- ***Approximately 4% of households in the U.S.A. contain child aged 19-35 months; approximately 35,000 children with completed household interview***

Contents of the NIS PUF

- ***Household CATI Interview***
  - ***Demographic Data:*** *Age, gender, race/ethnicity of the child, mother's age and race/ethnicity*
  - ***Socioeconomic data:*** *Family income, mother's education*
  - ***Geographic Identifiers****: City, State, County*
  - ***Subject-matter data:*** *Medical conditions, Child's immunization status, vaccination dates*

- ***Provider Record Check Study, PRCS (mailed IHQ)***
  - *Vaccine-specific shot dates*
  - *Provider's information (e.g., facility type, VFC participation)*

The NIS PUF

- ❖ *Public-use data files (1995-2000): Child-level records with 78 IAP area (state and urban) identifiers*

- ❖ *Approximately 35,000 age-eligible children with household interview data*

- ❖ *Approximately 23,000 children with household interview and 'adequate' provider data; on average 295 children per IAP area*

- ❖ *PUFs released on the Internet and CD-ROMs*

NCHS Review and Clearance Process

- ***Disclosure Review Board (DRB)***

- ***Extensive review of 2-, 3-, and 4- way tables within each IAP area***
  - **Population size >100,000 in each area**
  - **Unique Cells**
    - ***Demographic, socioeconomic, unique identifiers***
  - **Small Cells (<5 cases)**
  - **Time lag between data collection and data release**

- ***Potential availability of exogenous files and list of common variables***

- ***Warning to data users; penalty for misuse of data***

- ***DRB Checklist***

Techniques Used for Release of Microdata

- ***Micro-aggregation***
- ***Deletion of data items***
- ***Deletion of sensitive records***
- ***Data swapping***
- ***Recoding of variables into broad categories***
- ***Top- and bottom-coding***
- ***Sampling***
- ***Population thresholds by selected categories or geography***
- ***Imputation and collapsing of categories***

Example: Population Thresholds

- *Each of the 78 areas identified in the PUF has total population greater than 100,000*

Example: Data Recoding

- ❖ *Race/ethnicity of child is recoded into 4 categories*
  - ◆ *Hispanic*
  - ◆ *White, non-Hispanic*
  - ◆ *Black, non-Hispanic*
  - ◆ *All other races, non-Hispanic*

- ❖ *Age of child collected in months but is recoded to:*
  - ◆ *19-23 months*
  - ◆ *24-29 months*
  - ◆ *30-35 months*

Example: Top-coding

- *Household size top-coded to 8+*
- *Ratio of family income to poverty threshold is capped at 3.0*
- *Family income is capped at “greater than $50,000”*
- *Number of vaccination providers identified by the household respondent is capped at “3 or more”*

Example: Deletion of Variables from PUF

- *Interview dates are not included*
- *Child’s date of birth is not included*
- *Provider- and household-reported vaccination dates are not included*
- *ZIP code of residence is not included*
- *Area code and central office code from RDD sample are not included*

Exogenous Files

- ***A data intruder could covertly match the NIS PUF with exogenous population file X***

- ***How does one determine whether a cell in a cross-tabulation, using variables common to the PUF and exogenous file X, indicates that the PUF includes most or all of the children in a rare population cell?***

- ***Can examine unweighted cross-tabulations of PUF data using demographic and socioeconomic variables***

- ***A cell might contain only 3 children; if PUF sample size were 3 times larger, the same cell might contain 9 children, though population size in that cell is still the same***

Exogenous Files (cont.)

- *Can look for small weighted cell sizes*

- *With multiple weight adjustments the weight assigned to a child does not really reflect "how many children they represent in the population"*

  - **Example: two children may have the same base sampling weight; but if one child is in a household that has 3 voice-use phone lines, its adjusted base sampling weight will be one-third of the other child's weight**

Method Used for NIS PUFs

- *Coarsen data in small (rare) population cells by applying a technique that distorts data records before PUF is released*
  - **Identified variables A, B, C, and D that are common to NIS PUF and exogenous file X**
  - **Both files identified 78 geographic areas**
  - **Reviewed the 4-way table A x B x C x D within a selected area and identified cells with n<5 in the exogenous file X**

Artificial Example

AREA # 1: FIVE SMALL CELLS (n<5)

A	B	C	D	Number in PUF	Number in Population File X
1	4	2	3	2	5
1	3	1	3	1	3
2	1	2	1	3	5
3	2	1	2	2	4
3	1	1	2	1	2

Area # 1: After Recoding of Variable A in the PUF

A	B	C	D	Number in PUF	Number in Population File X
2	4	2	3	**6**	**23**
2	3	1	3	**3**	**14**
2	1	2	1	3	5
2	2	1	2	**4**	**17**
2	1	1	2	**2**	**11**

The data recodes to variables A, B, C, and D are shown in purple.

Area # 1: After Recoding Variable B in the PUF

A	B	C	D	Number in PUF	Number in Population File X
2	4	2	3	6	23
2	3	1	3	3	14
2	2	2	1	**5**	**19**
2	2	1	2	4	17
2	1	1	2	2	11

The data recodes to variables A, B, C, and D are shown in purple.

Final Result

After data coarsening:

Cross-tabulation of variables A, B, C, and D for a geographic area in the released PUF does not result in identification of any small population cells in exogenous file X for which most or all children are in PUF.

Summary

- ❖ *Potential conflict between data users' need and what can be released in the PUF in order to protect confidentiality*
- ❖ *Legal obligation to protect confidentiality of respondents and to reduce risk of disclosure*
- ❖ *Warning to users; penalty for misuse*
- ❖ *Methods used in the NIS*
  - ◆ **Population in each IAP area> 100,000**
  - ◆ **Identified rare cells; used extensive cross-tabulation**
  - ◆ **Deleted sensitive variables**
  - ◆ **Used top-coding, bottom-coding, recoding**
  - ◆ **Collapsed categories to create wider classes**
  - ◆ **Imputation**
  - ◆ **Used data coarsening**
  - ◆ **Matched with exogenous file for evaluation only**
- ❖ *DRB review and approval*

References

- Pat Doyle, Julia I. Lane, Jules J. M. Theeuwes, and Laura V. Zayatz (eds.). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: North-Holland, 2001

- Federal Committee on Statistical Methodology, Confidentiality and Data Access Committee, *Checklist on Disclosure Potential of Proposed Data Releases*, *www.fcsm.gov/docs/checklist_799.doc*, July 1999

- A.O. Zarate, (1998) *"Legal, Administrative and Statistical Aspects of Confidentiality Procedures at the National Center for Health Statistics Presentation", paper presented as expert testimony on issues of 'privacy and confidentiality', for the public Meeting on the President's Initiative on Immunization Registries*, Atlanta, 16 July 1998.

- National Center for Health Statistics (NCHS), *Staff Manual on Confidentiality*, September l984

Thank you

mxk1@cdc.gov