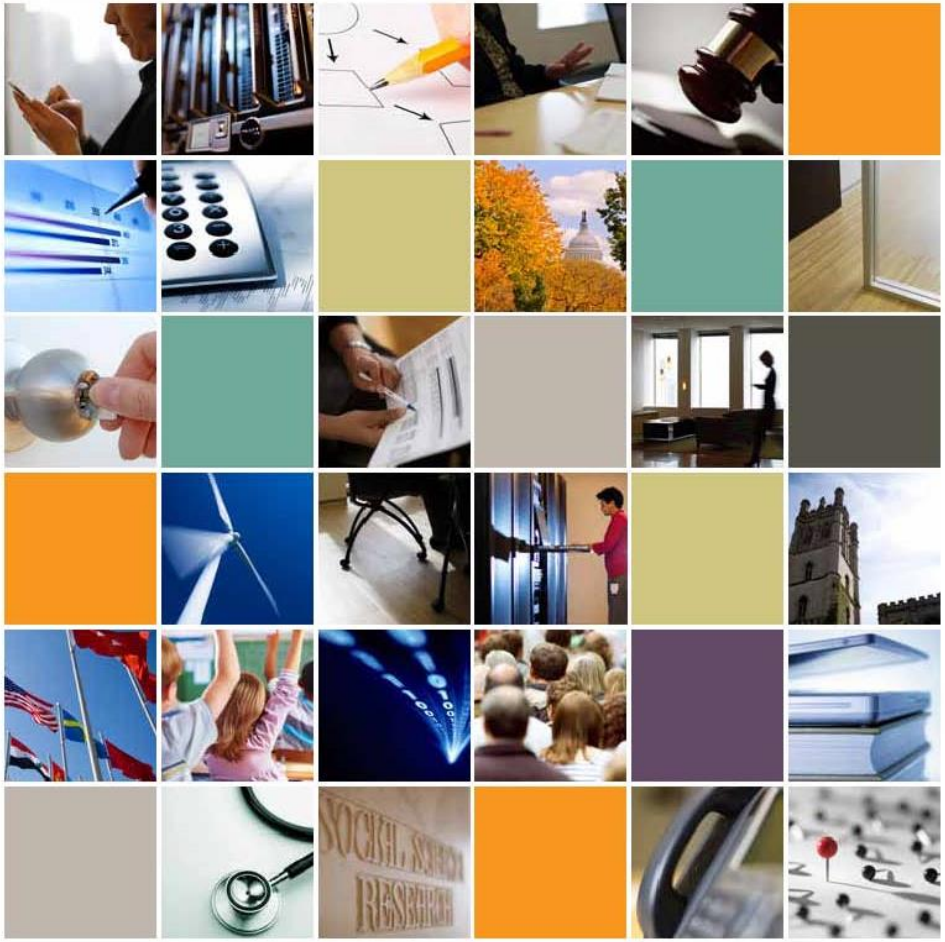
Disclosure Treatment of Sparsely Populated Geographic Areas

Josh Borton -

NORC at the University of Chicago

June 23, 2016



Common challenge

- Demand for reports at sub-state geographies, such as county
  - Presents ***disclosure issues*** if the county is sparsely populated, or has very small sub populations

Reporting Problems Caused by Sparse Geographies

- CDC EPH tracking
  - Different disclosure limitation standards and methods used by different states
  - Lack of standard rules and delivery causes problems with data harmonization

- State published data
  - Unable to report of some parts of the population
    - Geographic and/or demographic
  - Analytic utility may be limited

Best Practices for Dealing with Small Geographies

- Is the population sparse or is the data sparse?
  - Small counts for larger populations may meet disclosure standards
- Combine regions to meet disclosure standards
  - Recoding
- Suppress statistics that don’t meet disclosure standards
  - Lose ability to analyze some areas
- Consider using estimation to protect sensitive values

Options for Protecting Sparse Geographies

- De-Identify Underlying Micro-Data
  - Everything is an estimate

- De-Identify Tabular Data
  - Cells determined to be disclosive require attention
  - Frequency defined by the number of observations in the cell
- Suppress disclosive cells
  - Loss of information from suppression
- Aggregate disclosive cells
  - Information becomes more general
- Estimate disclosive cells
  - Information is less precise but remains available

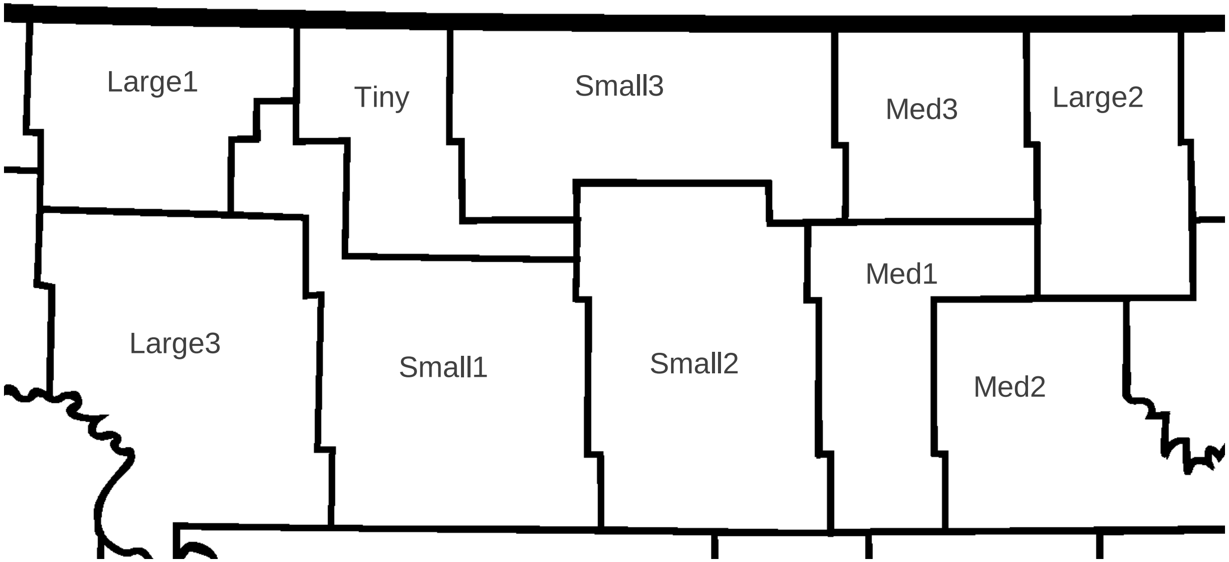
Common Tabular Methods

General Comments

- Generally use suppression and recoding
  - Leads to loss of information due to missing values
  - ‘Holes’ in a table
- Inter-table disclosure can be difficult to manage
- Traditionally risk is not measured
  - NORC has developed methods to quantify the risk of suppressed tables, including risk from inter-table dependencies

Geography for Tabular Examples

- Example of 10 fake counties
- Create tables showing counts of disease by age range



Tabular Method 1:

Suppress Disclosive Cells

- Suppression of disclosive cells allows for clear explanation of treatment to the user
- Requires the suppression of complementary, often non-disclosive cells, in order to ensure protection

Raw Data					
	Age				
County	<55	55-64	65-74	75+	TOT
large1	25	32	103	99	**259**
large2	64	50	114	116	**344**
large3	32	30	200	175	**437**
med1	**9**	16	88	82	**195**
med2	15	**9**	72	65	**161**
med3	19	25	99	41	**184**
small1	19	15	16	12	**56**
small2	13	16	14	20	**63**
small3	**7**	11	25	13	**62**
tiny	**3**	**9**	**8**	**3**	**23**
TOT	**206**	**213**	**739**	**626**	**1784**

Supression					
	Age				
County	<55	55-64	65-74	75+	TOT
large1	25	32	103	99	**259**
large2	64	50	114	116	**344**
large3	32	30	200	175	**437**
med1	~~**9**~~	~~**16**~~	88	82	**195**
med2	~~**15**~~	~~**9**~~	72	65	**161**
med3	19	25	99	41	**184**
small1	19	15	16	12	**62**
small2	13	16	14	20	**63**
small3	~~**7**~~	~~**11**~~	~~**25**~~	~~**13**~~	**56**
tiny	~~**3**~~	~~**9**~~	~~**8**~~	~~**3**~~	**23**
TOT	**206**	**213**	**739**	**626**	**1784**

Tabular Method 2:

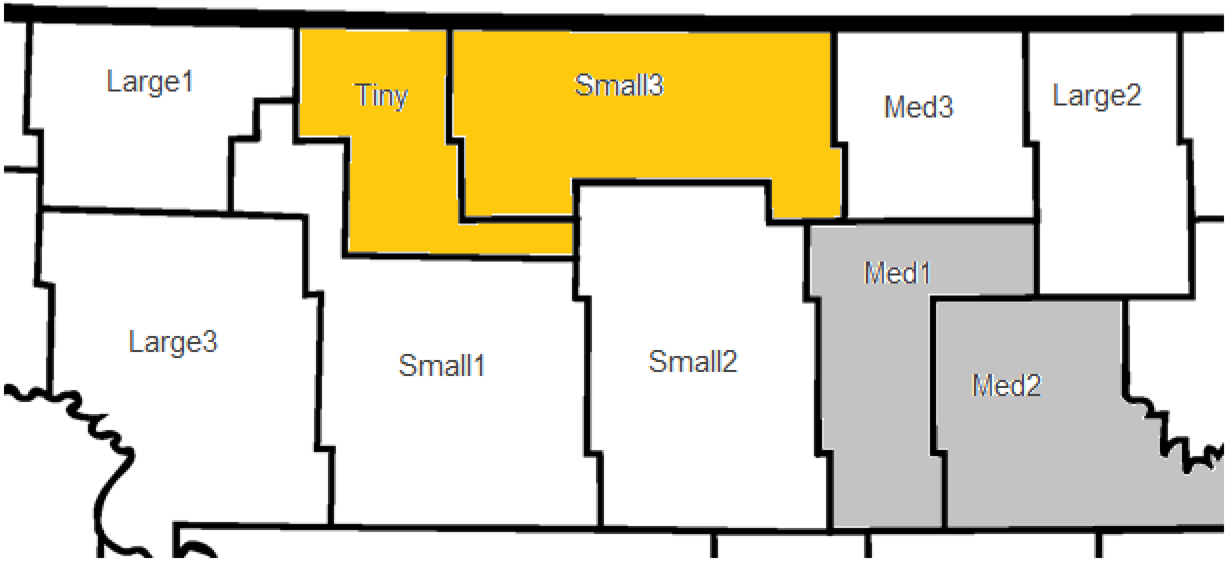
Aggregate Disclosive Cells

- Aggregation, or recoding, of cells can be used in place of suppression
- Prevents ‘holes’ in the table at the cost of less specific table dimensions
- Tables are complete, but may not be useful to those interested in values of a dimension that has been recoded

Raw Data					
	Age				
County	<55	55-64	65-74	75+	TOT
large1	25	32	103	99	**259**
large2	64	50	114	116	**344**
large3	32	30	200	175	**437**
med1	**9**	16	88	82	**195**
med2	15	**9**	72	65	**161**
med3	19	25	99	41	**184**
small1	19	15	16	12	**56**
small2	13	16	14	20	**63**
small3	**7**	11	25	13	**62**
tiny	**3**	**9**	**8**	**3**	**23**
TOT	**206**	**213**	**739**	**626**	**1784**

Aggregation					
	Age				
Race	<55	55-64	65-74	75+	TOT
large1	25	32	103	99	**259**
large2	64	50	114	116	**344**
large3	32	30	200	175	**437**
med1	**24**	**25**	88	82	**195**
med2			72	65	**161**
med3	19	25	99	41	**184**
small1	19	15	16	12	**62**
small2	13	16	14	20	**63**
small3	**10**	**20**	**33**	**16**	**56**
tiny					**23**
TOT	**206**	**213**	**739**	**626**	**1784**

Recoded Geography



Tabular Method 3:

Estimate Disclosive Cells

- It is possible to present estimates for cells that are deemed to be disclosive
  - Includes complementary suppressions
- Estimates do not increase disclosure risk of table
  - Same values a sophisticated intruder could calculate themselves
- Allows user to see a table without ‘holes’
  - User is aware which cells are true values and which are estimates

Raw Data					
	Age				
County	**<55**	**55-64**	**65-74**	**75+**	**TOT**
large1	25	32	103	99	**259**
large2	64	50	114	116	**344**
large3	32	30	200	175	**437**
med1	**9**	16	88	82	**195**
med2	15	**9**	72	65	**161**
med3	19	25	99	41	**184**
small1	19	15	16	12	**56**
small2	13	16	14	20	**63**
small3	**7**	11	25	13	**62**
tiny	**3**	**9**	**8**	**3**	**23**
TOT	**206**	**213**	**739**	**626**	**1784**

Estimation					
	Age				
Race	**<55**	**55-64**	**65-74**	**75+**	**TOT**
large1	25	32	103	99	**259**
large2	64	50	114	116	**344**
large3	32	30	200	175	**437**
med1	**9.1**	**15.9**	88	82	**195**
med2	**12.3**	**11.7**	72	65	**161**
med3	19	25	99	41	**184**
small1	19	15	16	12	**62**
small2	13	16	14	20	**63**
small3	**8.2**	**13.1**	**22.8**	**11.9**	**56**
tiny	**4.4**	**4.3**	**10.2**	**4.1**	**23**
TOT	**206**	**213**	**739**	**626**	**1784**

De-Identify Underlying Micro-Data

- Many commonly used micro-data methods introduce bias
- NORC X-ID methods avoid bias by using aggregation and sampling
  - Restructure data into small aggregates of size 10 to 20
    - Termed micro-groups
  - Produce summary statistics at the micro-group level
  - Add protection through intelligent sub-sampling of the data
- All outputs are estimates
  - Provides estimates with little error for adequately sized analysis questions
  - Provides estimates with large amount for error for individuals or very small groups

Micro Group Formation

Observation Level Data (all of these records are assigned to one microgroup)

mg_num	n3	SETTING	Age Group	New	Race	Trans Vol	Trans Amount
1	1	R	6	0	1	2	$1,576.30
1	1	R	6	0	4	55	$2,675.24
1	1	R	6	0	2	6	$638.62
1	1	R	6	0	2	9	$1,836.86
1	1	R	6	0	6	6	$1,654.05
1	1	R	6	0	1	2	$887.35
1	1	R	6	0	1	2	$1,354.50
1	1	R	6	0	1	1	$1,054.10
1	1	R	6	0	1	4	$2,885.25
1	1	R	6	0	1	1	$1,899.77
1	1	R	6	0	1	1	$717.04
1	1	R	6	0	1	2	$806.36
1	1	R	6	0	1	2	$917.66
1	1	R	6	0	3	2	$875.75

Micro Group Record

		Group Variables			Create Dummies	Micro Means		Micro Proportions					
mg_num	**n3**	**SETTING**	**Age Group**	**New**	**Race**	**Trans Vol**	**Trans Amount**	**race1**	**race2**	**race3**	**race4**	**race5**	**race6**
1	14	R	6	0	Micro Proportion	6.79	$1,412.78	0.643	0.143	0.071	0.071	0.000	0.071

Treating Micro-Data vs Tabular Data

- Treated Micro-Data
  - More versatile and allows the production of nearly any table that is desired
  - All cells are estimates
    - Large cells are good (very good) estimates
    - Small cells will have more error, for protection
- Treated Tabular Data
  - Requires that all tables be known at the time of treatment
  - Presents the real value where the data allows
    - Non-disclosive cells that are not needed for complementary suppression
  - Estimation of suppressed cells can improve user experience

Josh Borton
borton-joshua@norc.org

Thank You!

NORC at the UNIVERSITY of CHICAGO

insight for informed decisions™

Introducing Uncertainty via Sub-Sampling

Observation Level Data (all of these records are assigned to one microgroup)

mg_num	n3	SETTING	Age Group	New	Race	Trans Vol	Trans Amount
1	0.00	~~R~~	~~6~~	~~0~~	~~1~~	2	$1,576.30
1	0.00	~~R~~	~~6~~	~~0~~	~~4~~	~~55~~	~~$2,675.24~~
1	0.00	~~R~~	~~6~~	~~0~~	2	6	$638.62
1	0.00	~~R~~	~~6~~	~~0~~	2	9	$1,836.86
1	0.00	~~R~~	~~6~~	~~0~~	6	6	$1,654.05
1	4.65	R	6	0	1	2	$887.35
1	0.00	~~R~~	~~6~~	~~0~~	1	2	$1,354.50
1	0.00	~~R~~	~~6~~	~~0~~	~~1~~	1	$1,054.10
1	0.00	~~R~~	~~6~~	~~0~~	~~1~~	4	$2,885.25
1	4.66	R	6	0	1	1	$1,899.77
1	4.66	R	6	0	1	1	$717.04
1	0.00	~~R~~	~~6~~	~~0~~	~~1~~	2	$806.36
1	0.00	~~R~~	~~6~~	~~0~~	1	2	$917.66
1	0.00	~~R~~	~~6~~	~~0~~	~~3~~	2	$875.75

Micro Group Record

		Group Variables			Create Dummies	Micro Means		Micro Propotions					
mg_num	**n3**	**SETTING**	**Age Group**	**New**	**Race**	**Trans Vol**	**Trans Amount**	**race1**	**race2**	**race3**	**race4**	**race5**	**race6**
1	13.98	R	6	0	Micro Proportion	3.07	$1,315.21	0.628	0.246	0.000	0.000	0.000	0.127