National Cancer Institute

Evaluating Population Uniqueness for Population-based Cancer Registry data Using U.S. Census Microdata Samples

Mandi Yu, Dave Stinchcomb, and Kathy Cronin
National Cancer Institute

Louisville, Kentucky
2011 NAACCR

1

National Cancer Institute

Overview

- Background
- Study Objectives
- Data Sources
  - SEER Microdata
  - Census 2000 5% PUMS and 100% Summary File
- Methods
- Results
- Conclusion and Limitations

2

National Cancer Institute

The NCI Surveillance, Epidemiology, and End Results Program (SEER)

- Collects and disseminates population-based cancer data
  - Covers ~28% of the U.S. population
  - Includes demographics, geographic locations, and cancer characteristics
  - Contains > 6 million tumors from 1973-2008
  - SEER research data file is released under a data use agreement
- A primary source to identify and produce cancer trends, geographic variations, and ethnic disparities in cancer outcome

3

National Cancer Institute

Confidentiality Issues with SEER Data

- Increasing demand for small geographic cancer data
- However, cancer patients are more likely to be identified when one combines detailed geography with seemly innocuous demographics
- Risk is ever increasing due to new developments
  - Making in-house variables available
  - Disseminating linked variables: SEER-Medicare, SEER-NLMS, and SEER-MHOS
  - Being matched with records in external databases (public or commercial) on common variables
- The agency is obligated to ensure the risk of re-identifying patients through released SEER data is low.

4

National Cancer Institute

Types of Disclosure

- Definitions: Duncan, et al. (1993)
  - Identity disclosure
  - Attribute disclosure
  - Inferential disclosure: less of concern for microdata
- Revealing identity of an unknown cancer patient is potentially more harmful than revealing attributes of an already known cancer patient

5

National Cancer Institute

Measures of Identification Risk

- Internal risk: # of unique records (RU) in a data file
  - NAACCR Record Uniqueness Program
  - Assume the intruder knows an individual has cancer
  - Not account for population multiplicity
- External risk: # of unique records that are also unique in the population (PU&RU)
  - Requires external population data
    - Estimated from the sample, or
    - Directly obtained from an external source
  - Assume the intruder is more interested to know whether an individual has cancer

6

National Cancer Institute

Methods for Estimating Population Uniqueness (PU)

- Most existing methods are developed for representative random survey samples
  - PU is estimated from the sample
  - Non-parametric approaches such as Subsampling and Equivalent Classes (Zayatz, 1991; Greenberg and Zayatz, 1992)
  - Model-based approaches (Bethlehem, Keller, and Pannekoek 1990; Skinner 1994; Chen and Keller-McNulty 1998; Samuels 1998; Benedetti and Franconi 1998; Skinner and Elliot 2002; etc.)
- However, SEER is not a random sample of the general population because cancer patients usually have distinct characteristics.

7

National Cancer Institute

Study Objectives

- Develop methods to routinely assess the proportions of PU&RUs given specifications of SEER data files
- Evaluate the feasibility of estimating # PUs from Census Microdata Samples
  - Follow CDAC's suggestion to "assure that the risk of disclosure from released data when combined with other relevant publicly available data is very low"
  - Publically accessible, authoritative, and high-quality
  - Contains detailed demographic and socioeconomic data that allows flexibility in selecting key variables
  - Methods can be used in future assessments based on yearly updated microdata samples from the American Community Survey

8

National Cancer Institute

Data Sources

- SEER Microdata
- External data set
  - Census 2000 5% PUMS
  - Census 2000 100% Bridged Single-Race County Population Estimates Summary File (SF)
  - SF is treated as the gold standard

9

National Cancer Institute

SEER Microdata

- Mainly based on SEER 17 Registries Database
- Includes 16 cancer registries: Los Angeles, San Francisco-Oakland, San Jose-Monterey, Great California, Connecticut, Detroit, Atlanta, Rural Georgia, Hawaii, Iowa, Kentucky, Louisiana, New Jersey, New Mexico, Seattle and Utah (Alaska Native was excluded)
- Cases diagnosed in 2000
  - # of tumors: 355,236; # of patients: 346,955
- Includes all cancer sites
- County-level data

10

National Cancer Institute

Key Variables

- State and County: 476 counties
- Age (86): Single year age from 0 to 85 (top-coded)
- Gender (2): Male and Female
- Race (5): White, Black, American Indian or Alaska Native (AIAN), Asian, Native Hawaiian, and Pacific Islander (API); and Others
- Hispanic Origin (2): Hispanic and Non-Hispanic
- Total # of possible combination cells: ~818,720

11

National Cancer Institute

Census 2000 5% PUMS

- Contains a random sample of data collected on census 2000 long form
- Detailed demographic and socioeconomic information
- Geographic threshold: Public Use Microdata Area (PUMA) with min pop size of 100,000
- One person per row with a person weight, which can be interpreted as # of individuals with the same characteristics that this record represents
- Average weight is 18.79 and the values range from 0 to 256

12

National Cancer Institute

Census 2000 100% SF

- Summary tables collected on census short form survey
- Bridged Single-Race County Population Estimates Summary File (SF)
- Limited to basic demographic information
- Issues may complicate our study
  - Data swapping procedures, *such that "user should not assume that tables with cells having a value of one or two reveal information about specific individuals"- Census Bureau Documentation*
  - Measurement discrepancy in key variables (e.g. Measurement error or Imputation error)

13

National Cancer Institute

Estimating RU and PU&RU

- A "nonparametric" approach
- SEER patients were matched with pop records on key variables
- For each combination cell, k, estimate # of SEER records (f_k) and # of Pop units (F_k)
- RU={k: f_k=1}, f_k is directly calculated from SEER data
- RU&PU={k: f_k=1, F_k=1}
- Two methods to estimate F_k from census PUMS
  - Sum of person weights in PUMS
  - Sum of # of persons in a pseudo pop generated from PUMS

14

National Cancer Institute

Estimating Population Totals

- Pop totals for counties <100,000 is not estimable
- One solution is to impute for county geocodes using PUMA-County relationship multiple times (M=5)
- Estimate the pop total in each combination cell using M imputed data separately as it were the actual data, then, the multiple imputed estimates can be obtained by taking the average of M sets of point estimates

15

National Cancer Institute

PUMA-County Relationships

- Three types of relationships
  - Pattern 1: 1 or Multiple PUMA-to-1 County
  - Pattern 2: 1 PUMA-to-Multiple Counties
  - Pattern 3: Mixed matched PUMA and County
- Imputation procedures were applied to PUMS subjects who reside in a pattern-2 or -3 PUMA
- Imputation Assumptions
  - Pattern 2: Race is distributed proportionally across all counties nested within a PUMA
  - Pattern 3: Race is distributed homogenously within a PUMA

16

National Cancer Institute

A Problem with this Approach

- 99.96% of person weight in a PUMS is greater than 1
- Therefore, estimated pop totals are almost always greater than 1
- Also true for imputed data since a PUMS data record with weight>1 can only be assigned to one county;
- Consequently,
  - F_k is almost always >1
  - # of PU is zero

17

National Cancer Institute

An Alternative Imputation Method

- Generate a pseudo pop by replicating each PUMS record w times (w is the person weight);
- Imputation for county is then carried out based on this pseudo population
- This method allows assign w records to different counties, thus permits PU;
- This approach only helps with imputed portion of data.

18

National Cancer Institute

Data Matching Results

- SEER patients are classified into two groups:
  - Match group: Patients matched to at least one population unit
  - Zero Match group: Patients that were not matched to any population units
  - Possible reasons for zero match
    - Measurement discrepancy
    - Census under count
    - Low sampling fraction (applied to PUMS only)

19

National Cancer Institute

Results from SF (Gold Standard)

	TOTAL	COVERED					NOT COVERED		Combined* %
	N	N	Cov. Rate	RU (%)	PU	PU&RU (%)	N	RU (%)	
Entire SEER File	346,643	346,281	99.90	25,093 (7.25)	233	232 (.07)	362	350 (96.69)	582 (0.17)
CA	144,315	144,225	99.94	5,509 (3.82)	40	39 (.03)	90	88 (97.78)	127 (0.09)
CT	20,272	20,242	99.85	705 (3.48)	4	4 (.02)	30	28 (93.33)	32 (0.16)
GA	11,202	11,197	99.96	924 (8.25)	4	4 (.04)	5	5 (100.00)	9 (0.08)
HI	5,572	5,558	99.75	372 (6.69)	7	7 (.13)	14	14 (100.00)	21 (0.38)
IA	16,269	16,243	99.84	3,305 (20.35)	39	39 (.24)	26	26 (100.00)	65 (0.40)
KY	22,140	22,070	99.68	4,425 (20.05)	62	62 (.28)	70	68 (97.14)	130 (0.59)
LA	21,186	21,168	99.92	3,473 (16.41)	19	19 (.09)	18	18 (100.00)	37 (0.17)
MI	22,588	22,584	99.98	432 (1.91)	3	3 (.01)	4	4 (100.00)	7 (0.03)
NJ	48,208	48,125	99.83	2,333 (4.85)	16	16 (.03)	83	77 (92.77)	93 (0.19)
NM	7,593	7,585	99.89	1,746 (23.02)	12	12 (.16)	8	8 (100.00)	20 (0.26)
UT	6,999	6,992	99.90	977 (13.97)	15	15 (.21)	7	7 (100.00)	22 (0.31)
WA	20,299	20,292	99.97	892 (4.40)	12	12 (.06)	7	7 (100.00)	19 (0.09)

* Combined # PU&RU is estimated assuming all RU records that are not covered by the population data are population unique.

20

National Cancer Institute

Results from PUMS (based on data with survey weights)

	COVERED				NOT COVERED		Combined %
	N	Coverage Rate	RU (%)	PU	N	RU (%)	
Entire SEER File	342,176.4	98.71	22,203.0 (6.49)	-	4,466.6	3240 (72.54)	0.93
CA	143,575.2	99.49	4,987.2 (3.47)	-	739.8	609.8 (82.43)	0.42
CT	20,123.0	99.26	612.0 (3.04)	-	149.0	121 (81.21)	0.60
GA	11,017.6	98.35	802.8 (7.29)	-	184.4	126.2 (68.44)	1.13
HI	5,492.2	98.57	323.0 (5.88)	-	79.8	63 (78.95)	1.13
IA	15,822.6	97.26	3,042.4 (19.23)	-	446.4	288.6 (64.65)	1.77
KY	21,156.0	95.56	3,835.8 (18.13)	-	984.0	657.2 (66.79)	2.97
LA	20,494.2	96.73	3,019.0 (14.73)	-	691.8	472 (68.23)	2.23
MI	22,502.0	99.62	382.0 (1.70)	-	86.0	54 (62.79)	0.24
NJ	47,752.4	99.05	2,072.6 (434)	-	455.6	337.4 (74.06)	0.70
NM	7,289.0	96.00	1,524.2 (20.91)	-	304.0	229.8 (75.59)	3.03
UT	6,842.8	97.77	849.0 (12.41)	-	156.2	135 (86.43)	1.93
WA	20,109.4	99.07	753.0 (3.74)	-	189.6	146 (77.00)	0.72

21

National Cancer Institute

Results from PUMS (based on pseudo population data)

	COVERED					NOT COVERED		Combined %
	N	Coverage Rate	RU (%)	PU	PU|RU (%)	N	RU (%)	
Entire SEER File	344,072	99.26	23,352 (6.79)	72	67 (.01)	2,567	2,087 (81.30)	2,154 (0.62)
CA	143,683	99.56	5,073 (3.53)	14	13 (.01)	632	524 (82.92)	537 (0.37)
CT	20,123	99.26	612 (3.04)	-	- (.00)	149	121 (81.21)	121 (0.60)
GA	11,100	99.09	860 (7.75)	3	3 (.03)	102	69 (67.39)	72 (0.64)
HI	5,499	98.69	328 (5.96)	-	- (.00)	73	58 (79.45)	58 (1.04)
IA	16,095	98.93	3,174 (19.72)	1	1 (.00)	174	157 (90.25)	158 (0.97)
KY	21,828	98.59	4,209 (19.28)	23	22 (.10)	312	284 (90.96)	306 (1.38)
LA	20,947	98.87	3,301 (15.76)	16	13 (.06)	239	190 (79.60)	203 (0.96)
MI	22,502	99.62	382 (1.70)	-	- (.00)	86	54 (62.79)	54 (0.24)
NJ	47,773	99.10	2,086 (4.37)	1	1 (.00)	435	324 (74.49)	325 (0.67)
NM	7,481	98.52	1,654 (22.11)	9	9 (.11)	108	96 (88.89)	105 (1.38)
UT	6,912	98.76	906 (13.11)	5	5 (.07)	87	78 (89.63)	83 (1.19)
WA	20,129	99.16	767 (3.81)	1	1 (.00)	170	132 (77.65)	133 (0.66)

22

National Cancer Institute

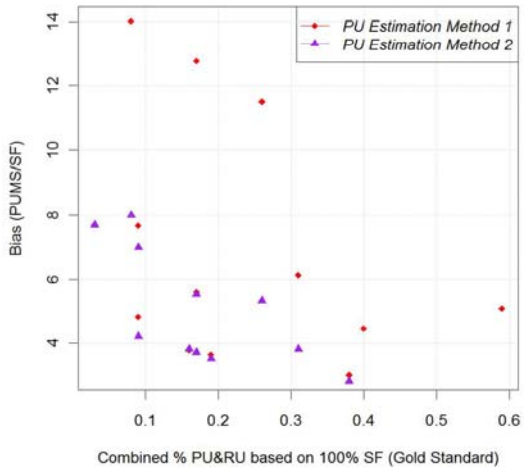
Who are the Zero Match Patients?

- Compared with the population data, zero match SEER patients tend to be
  - Older (60 years of age and older)
  - Male
  - Hispanic
  - Not-married
  - Non-white
  - Residing in a smaller counties

23

National Cancer Institute

Relationship: Bias and Risk Size



24

National Cancer Institute

Conclusion

- PUMS has great potentials to be used in the evaluation of population uniqueness
- Expanded weight method performs better than original weight method
- Both methods produce conservative risk estimates
- The magnitude of the upward bias is in the neighborhood of 3-4.
- When more variables are used in as keys, for example, Marital status, census-tract SES attributes, etc. bias may decrease

25

National Cancer Institute

Limitations and Remedies

- Limitations
  - Upward bias due to the small sampling fraction of PUMS
  - Estimation uncertainty due to imputation for county
  - Measurement discrepancy in key variables between SEER and Census data, especially Race and Hispanic
  - Deterministic matching doesn't consider such errors
- Remedies
  - Access more data (with larger sampling fraction) with lower geographic thresholds
  - Build relationship between Sampling fraction and estimation bias on area population composition and area attributes and extrapolate

26