Disclosure Risk Assessment for Population-based Cancer Microdata

Mandi Yu[1], Kathleen Cronin[2], and David Stinchcomb[3]

[1]Contractor, Surveillance Research Program, Division of Cancer Control and Population Science, National Cancer Institute, National Institutes of Health, 6116 Executive Blvd., Suite 504, Rockville, MD 20852
[2] Surveillance Research Program, Division of Cancer Control and Population Science, National Cancer Institute, National Institutes of Health, 6116 Executive Blvd., Suite 504, Rockville, MD 20852
[3]Westat Inc. 1600 Research Blvd., Rockville, MD 20850

Abstract

The authors developed and tested a non-parametric method of estimating the risk of disclosing information about whether known population individuals have cancers. This method matches cancer patients diagnosed in 2000 in a research data file from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program with individuals in Census 2000 Public Use Microdata Sample (PUMS) by county of residence and several common demographic key variables and estimates the proportion of patients who are unique in both files. To overcome the lack of direct estimates of population totals for counties with less than 100,000 residents, the authors developed two methods to impute the incomplete county codes in PUMS based the relationships among Public Use Microdata Area (PUMA), County, and Race. The uniqueness estimates were then validated against the gold standard obtained by matching SEER data with 100% Census 2000 Summary File 1 (SF1.) Older racial and ethnic minorities residing in less populous areas are at higher risks of being identified. Both imputation methods produce conservative risk estimate and the magnitudes of upwards bias are 3-4 times. The bias tends to be greater for areas with larger risks. This research is the first attempt to systematically evaluate the risk of disclosing the attribute of whether an individual has cancer, and it builds substantial foundation for establishing routine procedures for assessing such risks using yearly updated microdata from the American Community Survey (ACS) in the future.

Key Words: Disclosure Risk, SEER, Cancer Registry, Population Uniqueness, Census PUMS, and ACS

1. Introduction

The National Cancer Institute's Surveillance, Epidemiology, and End Results Program (SEER) routinely collects and publishes data on cancer patient demographics, geographic locations, tumor characteristics, and treatment information from population-based cancer registries. It has been the most authoritative source of data for describing cancer incidence and survival at national, regional, and local levels. There has been an increasing need for small geographic area data to identify areas with elevated cancer rates and to plan and monitor the impact of cancer control and prevention activities at local levels. However, the identification of cancer patients is likely when one combines detailed geography with basic demographics. Disclosure not only discredits the agency but also presents a significant impact on to a patient, as it tends to have great potential for tangible harms given the sensitive nature of individual medical conditions. In order for the agency to make informed decisions in balancing data utility and the risk of disclosure, it is important to quantify the risk of disclosure systematically.

Record Uniqueness (RU) in a cancer surveillance data file means that there is only a single patient with the same basic socio-demographic characteristics such as gender, age, race, and geography. While the risk of disclosing new information about a known cancer patient presented by record uniqueness in a cancer surveillance data file has been well addressed [1]. However, significant gaps exist for evaluating the disclosure risk when matching the cancer surveillance data to another general population data file based these same basic demographic variables. Because this second file is not a file of cancer patients, but drawn from the general population (e.g. census data, credit card files, voter registry, etc,), if a person is unique based on the same set of attributes we call this person Population Unique (PU). If an intruder knows that a person has cancer and basic demographics, he can use RU to reveal additional details about the cancer patient (e.g. tumor characteristics, treatment, number of prior cancers). However, if an intruder has matched registry data to a population file, and knows that a person is both RU and PU, then they can use this information to reveal that a person has cancer based just on basic demographic information (potentially a much more sensitive revelation). This precisely follows the definition of disclosure provided by various authors, in which a patient is identified, rather than just disclosing additional information about a known patient [2, 3]. One of the methodological difficulties that hinder the development in this area is that the assessments require population data, which has to either be estimated from the released data [4-13] or be acquired from another source. The estimation approach is not feasible because cancer patients usually have different characteristics than the general population, inferring the population distribution from the cancer patients would be problematic without acquiring addition information about the difference between these two. Therefore, population information can only be acquired externally.

This article regards the number of RU patients who are also unique in the population, denoted as PU, as a measure of disclosure risk and develops a nonparametric model to assess such risks. This measure of risk, denoted as $PU|RU$, has been widely used and considered in the context of releasing census or survey sample data [5, 7, 12, 14-16]. This study is the first to apply this measure in the case of estimating risk for a public health surveillance data. It particularly concerns the threat of disclosure arising from the possibility that an *intruder* might successfully identify a patient through matching his/her released identifying information to known individuals in the population.

In this study, we first evaluate RU and PU from SEER data linked to US Census 2000 100% Summary File 1 (SF1) based on gender, single year of age, race (White, Black, American Indians and Alaksa Natives, and Asian & Pacific Islander)) and county. Because the SF1 is census rather than a sample, we can identify with relative certainty (with the exception of people missed) if a person is PU. However, the SF1 is a summary (rather than individual record) file which provides counts based on four basic demographic characteristics, and intruders may want to use additional identifying characteristics besides the 4 basic demographic variables. In this instance, linking a file such as the 5% Public Use Microdata Sample (PUMS) from the U.S. Census long form survey may be attractive for the intruder, since it contains a rich source of additional characteristic of the individual (e.g. marital status, and nativity). However, since data sources such as PUMS are a sample rather than a census, and counties with populations under 100,000 are combined to reduce the chance of identification, the intruder can never be absolutely certain that a person is population unique, and sophisticated statistical techniques must be employed to infer a high probability of PU.

In this paper we first evaluate the extent of PU|RU using SEER and the SF1. Secondly we use a basic algorithm (including imputation of county) to estimate PU|RU using SEER and the PUMS data. We chose the PUMS data because it allowed one consider various disclosure risk scenarios by choosing different sets of identifying information given the data richness. This test also prepares us for future use of yearly updated American Community Survey (ACS) sample data in annual assessments of SEER new releases given the similarities between the two survey data. The issue of measure change over time, for example, patients' moving out their initial diagnosis areas, can be better coped by using timely ACS estimates. For validation purpose, the risk estimates were compared with those obtained by using the SF1. Despite its high accuracy, since it is based on complete population enumeration, we did not consider use it in routine assessments because the data is limited to four basic demographics and only updated once every 10 years. Only considering limited sets of disclosure risk scenarios and assuming constant identifying information across 10 years could be problematic and limiting.

In the remaining sections, we first introduce data attributes type and their relations with the risk of disclosure. We then define two disclosure measures, justify our choices, discuss properties of those measures, show evaluation results, and discuss areas for future research.

2. Materials and Methods

2.1. Data attributes

A cancer registry microdata set like SEER's research data file can be viewed as a data file with each row representing a cancer event for one patient, and each column representing an attribute of this event. From the standpoint of confidentiality, the attributes can be classified into four categories: direct identifiers, quasi-identifiers, sensitive attributes, and non-sensitive attributes. The direct identifiers include names, addresses, unique identification numbers (for example, Social Security Number), etc. and they are routinely removed from a released data file to prevent direct identity disclosure. Quasi-identifiers, also called key identifiers or key variables, are variables that can be used to indirectly identify patients, for example, age, gender, date of birth, race, ethnicity, and small area geographic locations. This set of variables is the background information an intruder might easily have about the known target individuals. A sensitive attribute is the information an intruder does not have but attempts to obtain and the disclosure of this attribute might result in discrimination embarrassment or economic harm to an identified patient. Data on cancer diagnoses, test results, cause of death, etc. are usually in this category. The rest of data items are non-sensitive attributes, and the disclosure of which generally do not cause harm.

Establishing the distinction between these four types of variables usually involves difficult and complex judgments, because the categorization may not necessarily be mutually exclusive and static over time. With more data made publicly available, sensitive attributes such as diagnosis and treatment information may become available for identification. Attributes that are not sensitive to one party may appear highly sensitive to another party.

There are two general types of disclosure that are of concern: identity disclosure and attribute disclosure [17]. Identity disclosure refers to the identification of an entity (such as a cancer patient or a health care organization) and attribute disclosure refers to an intruder finding out something new about the target entity. For a microdata with detailed information attached to each record, identity disclosure usually leads to the revelation of attributes, and therefore, is of primary importance.

2.2. Measures of disclosure risk

The identity disclosure risk measures in the literature are mainly based on either the number/percentage of unique records or the probability of identification. Because unique records can be identified with high certainty and cancer surveillance data usually contains many unique records since cancer is rare, measures based on uniqueness are more useful. Two pieces of knowledge are important for a successful identity disclosure attempt: response knowledge [4] and population uniqueness. Response knowledge is a term originally used in surveys and it refers to the knowledge that a person has been interviewed for a particular survey. Translated into cancer surveillance, it is the knowledge that a person has cancer (physicians and hospitals are required by law to report all cancer diagnoses). If an intruder has reasons to believe that a particular person has cancer, and consequently his data must be in the surveillance database, identity disclosure can be easily accomplished if this person is unique in the dataset. Response knowledge significantly increases the risk of disclosure and oftentimes this information is what an intruder would like to know most. The knowledge of population uniqueness usually requires complete enumeration of a population. If a record is both record unique and population unique, then the disclosure becomes much more likely since a one-to-one relationship between a data record and a target individual can be established, and the knowledge that this person has cancer will be disclosed together with other attribute information that is attached to his record.

For simplicity, in the rest of article, we used the term sample to refer to SEER data and the term population to refer to the general population data. F_k and f_k denote the population size and the sample size in cell k of a cross-classification of key variables with a total of K cells respectively. $\sum_{k=1}^{K} F_k = N$ and $\sum_{k=1}^{K} f_k = n$, where N and n are the total population size and total sample size. We defined the set of record uniques as $RU = \{k: f_k = 1\}$, the set of population uniques as $PU = \{k: F_k = 1\}$, and the set of record uniques that are also population unique as $RU|PU = \{k: f_k = 1, F_k = 1\}$. A global disclosure risk measure for the entire sample data file is $\tau = \sum_{k=1}^{K} I(f_k = 1, F_k = 1)$, where $I(\cdot)$ is the indicator function.

2.3. Data sources and key variables

This study involves three datasets: the SEER research data file for which we seek risk estimates; 5% Census 2000 PUMS file, the population data that intruders have access to for matching SEER patients; and finally 100% Census 2000 Summary File 1, a population data based on complete enumeration, which provides population statistics for calculating 'gold standard' risk estimates for validation.

This SEER data includes 346,955 cancer patients diagnosed in 2000, the same year as the U.S. 2000 Census, from SEER 17 Registries[1]. We excluded the data from Alaska Native Tumor Registry because all patients are American Indians and Alaska Natives, and additional confidentiality constraints are in place for these data. We obtained both PUMS and SF1 from U.S. Census Bureau. PUMS contains rich data representing 5 percent of U.S. population. Each record represents w_i population individuals, where w_i is the person weight attached to record i. The smallest geographic units in PUMS are Public Use Microdata Areas (PUMAs) with a minimum population of 100,000. This threshold was set by the Census Bureau to prevent the disclosure of individual information from released data. SF1 contains information collected on Census short-form questionnaires. Three merits make it a perfect validation source: it is based on complete enumeration of U.S. population; statistics are available at much lower geographic levels with high precision, such as county and census tract; and single-race population estimates at county level are available. However, it cannot be used routinely because the contents are limited to four basic demographic variables[2] and risk assessments are not possible in situations where the intruders have access to more background information.

Five common key variables among all three datasets are selected: single year of age (A) in 85 categories (top-coded at 85), sex (S) in 2 categories, race (R) in 5 categories (White, black, American Indian or Alaska Native (AIAN), Asian, Native Hawaiian, and Pacific Islander (API), and Others), Hispanic origin (H) in 2 categories, and State-county geocodes (SC) in 468 categories[3]. The total possible number of combination cells is about 795,600.

2.4. Methods

2.4.1. Record matching and risk estimation

For both population data of PUMA and SF1, we matched SEER patients with population records on key variables. Based on the matching results, SEER patients are classified into two groups: matched group and not-matched group. For matched group, the number of RU and the number of $RU|PU$ patients are estimated. When PUMS is used, $\widehat{F_k}$ is obtained by summing the survey weights in cell k: $\widehat{F_k} = \sum_{i\in k} w_i$, where i denote individuals. Then $\hat{\tau}$ is estimated by $\hat{\tau}_1 = \sum_{k=1}^{K} I\left(f_k = 1, \widehat{F_k} = 1\right)$. When SF is used, τ can be directly calculated since both f_k and F_k are known. For not-matched group, we first calculated $\hat{\theta} = \sum_{k=1}^{K} I(f_k = 1)$, then we predicted how many of them are also population unique, $\hat{\tau}_2$. The combined estimate is $\widehat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

[1] The SEER's 17 Registries are Los Angeles, San Francisco-Oakland, San Jose-Monterey, Great California, Connecticut, Detroit, Atlanta, Rural Georgia, Hawaii, Iowa, Kentucky, Louisiana, New Jersey, New Mexico, Seattle, Utah, and Alaska Native Tumor Registry.

[2] The entire questionnaire includes seven questions: name, household relationship, sex, age, Hispanic or Latino origin, race, and home ownership (whether home is owned or rented).

[3] There are 486 counties in SEER areas.

It is vital to point out that there are two possible reasons for non-matching: measurement errors [18] or perturbation errors [19, 20] in key variables and incomplete coverage of the population data due to sampling or census undercount. A true population match may appear to have a different combination of key variables from the SEER patient due to measurement discrepancies; such leads to a non-match. Individuals with rare characteristics may not appear in a sample representing a portion of the population due to sampling. The smaller the sampling fraction is the more under-represented individuals. Some individuals were not counted by the census and the undercount rates vary by age, sex, race, geographic area population density, and economic status [21]. These non-matching factors are in fact protective to the confidentiality of the release data because they make record matching difficult [22].

2.4.2. Imputing county codes in PUMS

Because county codes are not available in PUMS for those with population less than 100,000, we created imputations, using PUMA-county-race relationships extracted from SF1, multiple times and estimated the imputed population totals in the same way as if we were using the original data. Imputation uncertainty was taken into account by taking the average of multiple estimates [23].

There are mainly four types of PUMA-county relations. Counties with large populations are PUMAs by themselves (referred to have a relation of “1 to 1”) or subdivided into one or multiple PUMAs (“M to 1”). For these counties, population totals can be estimated using the sum of the weights over all PUMAs nested within a county. One PUMA can also be formed by a group of small adjacent counties (“1 to M”) and occasionally, and PUMA boundaries can cross county boundaries and be made of parts of several counties (“Mixed”). For respondents residing in these counties, we imputed their county codes assuming race is distributed homogenously within a PUMA. We could have based the imputations on just the PUMA-county relation or a relation between PUMA, county and other variables. Goodman-Kruskal’s lambda association tests suggest that Race has the strongest association with the geography.

We developed two imputation methods. In both methods, we first created a set of race-PUMA strata with population size of n_s, where s denotes stratum. Within each stratum, we randomly allocated records into one of the nested counties with probabilities proportional to their sample weights w_i. We repeated the second step multiple times (M=5) to create multiply imputed data sets. The only difference between the two methods is how many population individuals we assumed each record represents. In the first method, we assumed each respondent represents w_i population individuals. Assigning one record to a county is equivalent to allocating w_i population individuals with the same characteristics to that county, therefore, the estimated population size for that cross-classification cell, $\widehat{F_k}$, is guaranteed to be no less than w_i. When $w_i > 1$, then $\widehat{F_k} > 1$ and cell k is not a population unique cell. The distribution of weights suggests that only a few records represent themselves and most records are associated with weights much greater than one. Therefore, we expected large underestimation bias in $\hat{\tau}$ since it mostly equals zero. Alternatively, we considered an expanded-weights approach. We created a pseudo population data by expanding each record to w_i records with the same characteristics, thus each respondent only represents himself. We then randomly assigned each respondent to a county with equal probability of $1/\sum_i^{n_s} w_i$. This approach allows the assigning of pseudo respondents who are generated from one respondent to different counties, thus it is possible to have combination cells with frequencies of one and consequently produce less bias in $\hat{\tau}$.

3. Results

We first presented the gold standard risk estimates obtained using SF 1. We then compared them with two sets of risk estimates obtained using PUMS, where county codes are imputed using imputation method 1 and 2 respectively.

Table 1 shows the estimated numbers and proportions of RU, PU, and $PU|RU$ by coverage status using SF1. Overall, for 99.9% of SEER patients, at least one match can be found in the census population. Among these matched patients, the proportion of RU is very high with an overall value of 7.25% across all cancer sites and all SEER areas. The proportions are higher for regions with low population density, such as New Mexico, Iowa, Kentucky, Utah, and Louisiana, which are around 20%. In contrast, RU is relatively rare in populous regions, such as Connecticut, California, Detroit[4], and Seattle[5], and the proportions range from 3% to 4%. The proportions of PU follow the same pattern as those of RU but at much smaller magnitudes (approximately 100 times smaller.) Almost all PU are also RU, which suggests that the overall impact of measurement errors on record match is negligible.

Almost all of the 362 SEER patients without population matches are record uniques. Because the measurements in key variables are comparable based on previous results, it is highly possible that the impact of measurement errors on matching rates is also negligible, and the main contributing factor for non-matching is census undercount. It is desirable to have conservative disclosure risk estimates, therefore, we treated not-covered RU as PU. We then derived a combined estimate of $PU|RU$ for the entire SEER data file (presented in the last column of Table 3) by summing the number of $PU|RU$ among covered SEER patients and the number of RU among not covered SEER patients. On the average, the proportion of $PU|RU$ is 0.17%. The data with the highest risk is from the Kentucky registry (0.59%) and the data with the lowest risk is from Detroit SEER registry (0.03%). Compared with the general population, non-match SEER patients tend to be 60 years of age and older, male, non-white, Hispanic, not married, and residing in smaller counties (data not shown and available upon request.)

Table 2 shows the numbers and proportions of SEER records and counties by imputation status and PUMA-county relationship. On average, 17.2 % SEER patients residing in 78.6% counties have missing data on county codes in PUMS. This proportion varies by area population density. For SEER areas with low population densities, the proportion of missing data is as high as 71.8% for Kentucky and 68.3% for Iowa. In contrast, for populous regions, this proportion reduces to 0% for Connecticut, 3.6% for New Jersey and 4.6% for California.

[4] Metropolitan Detroit Cancer Surveillance System covers Macomb, Oakland, and Wayne counties.

[5] FHCRC Cancer Surveillance System covers Clallam, Grays Harbor, Island, Jefferson, King, Kitsap, Mason, Pierce, San Juan, Skagit, Snohomish, Thurston, and Whatcom counties.

Table 1 : Estimated Number of SEER Records (2000) that are RU, PU, and $PU|RU$ using 100% Census SF 1

	Overall	Matched				Non-matched	Combined
	N	Match Rate (%)	RU (%)	PU	PU/RU (%)	RU (%)	PU/RU (%)
Entire SEER File	346,643	99.90	25,093 (7.25)	233	232 (.07)	350 (96.69)	582 (0.17)
CA	144,315	99.94	5,509 (3.82)	40	39 (.03)	88 (97.78)	127 (0.09)
CT	20,272	99.85	705 (3.48)	4	4 (.02)	28 (93.33)	32 (0.16)
GA*	11,202	99.96	924 (8.25)	4	4 (.04)	5 (100.00)	9 (0.08)
HI	5,572	99.75	372 (6.69)	7	7 (.13)	14 (100.00)	21 (0.38)
IA	16,269	99.84	3,305 (20.35)	39	39 (.24)	26 (100.00)	65 (0.40)
KY	22,140	99.68	4,425 (20.05)	62	62 (.28)	68 (97.14)	130 (0.59)
LA	21,186	99.92	3,473 (16.41)	19	19 (.09)	18 (100.00)	37 (0.17)
MI*	22,588	99.98	432 (1.91)	3	3 (.01)	4 (100.00)	7 (0.03)
NJ	48,208	99.83	2,333 (4.85)	16	16 (.03)	77 (92.77)	93 (0.19)
NM	7,593	99.89	1,746 (23.02)	12	12 (.16)	8 (100.00)	20 (0.26)
UT	6,999	99.90	977 (13.97)	15	15 (.21)	7 (100.00)	22 (0.31)
WA*	20,299	99.97	892 (4.40)	12	12 (.06)	7 (100.00)	19 (0.09)

Notes: * Three SEER regions do not cover complete states. In Georgia, SEER covers metro Atlanta and several additional rural counties. In Michigan, SEER covers metro Detroit. In Washington, SEER covers metro Seattle. See http://seer.cancer.gov/registries/ for details.

Table 2: The Numbers and Proportions of SEER 2000 Persons (Ps) and Counties (Cs) by Imputation Status and PUMA-County Relation for SEER 17 Registries

PUMA-County Relationship*	Direct Estimation				Indirect Estimation Via Imputation			
	M to 1		1 to 1		1 to M		Mixed	
	# Ps (%)	# Cs (%)	# Ps (%)	# Cs (%)	# Ps (%)	# Cs (%)	# Ps (%)	# Cs (%)
Total	248,93 (7.2)	35 (7.5)	261,855 (75.5)	65 (13.9)	51,066 (14.7)	361 (77.3)	8,825 (2.5)	6 (1.3)
CA	7,456 (5.2)	10 (17.2)	130,330 (90.3)	24 (41.4)	4,993 (3.5)	23 (39.7)	1,536 (1.1)	1 (1.7)
CT	3,324 (16.4)	4 (50.0)	16,948 (83.6)	4 (50.0)	-	-	-	-
GA**	-	-	10,615(94.8)	5 (33.3)	587 (5.2)	10 (66.7)	-	-
HI	6,82 (12.2)	1 (25.0)	4,056 (72.8)	1 (25.0)	834 (15.0)	2 (50.0)	-	-
IA	3,477 (21.4)	5 (5.1)	1,686 (10.4)	1 (1.0)	11,106 (68.3)	93 (93.9)	-	-
KY	736 (3.3)	1 (0.8)	5,515 (24.9)	2 (1.7)	15,889 (71.8)	117 (97.5)	-	-
LA	3,132 (14.8)	5 (7.8)	3,533 (16.7)	2 (3.1)	8,598 (40.6)	53 (82.8)	5,923 (28.0)	4 (6.3)
MI**	-	-	22,588 (100.0)	3 (100.0)	-	-	-	-
NJ	3,559 (7.4)	5 (23.8)	42,877 (88.9)	14 (66.7)	406 (0.8)	1 (4.8)	1,366 (2.8)	1 (4.8)
NM***	1,010 (13.3)	2 (6.1)	2,514 (33.1)	1 (3.0)	4,065 (53.6)	30 (90.9)	-	-
UT	684 (9.8)	1 (3.4)	4,423 (63.2)	3 (10.3)	1,892 (27.0)	25 (86.2)	-	-
WA**	833 (4.1)	1 (7.7)	16,770 (82.6)	5 (38.5)	2,696 (13.3)	7 (53.8)	-	-

Notes: * M to 1: Multiple PUMAs correspond to 1 county
1 to 1: 1 PUMA corresponds to 1 county
1 to M: 1 PUMA corresponds to multiple counties
Mixed: 1 PUMA is comprised of multiple small counties and part(s) of a large county.

** Three SEER regions do not cover complete states. In Georgia, SEER covers metro Atlanta and several additional rural counties. In Michigan, SEER covers metro Detroit. In Washington, SEER covers metro Seattle. See http://seer.cancer.gov/registries/ for details.

*** Four cases with missing information on county are excluded.

Table 3 shows the results obtained using PUMS with county codes imputed using method 1. Compared with the gold standards, the match rates decrease slightly for all SEER regions from an average of 99.90% to 98.71% with a difference of 1.19%. The declines are smaller for populous regions. Three regions with the smallest declines are Michigan (0.36%, Metropolitan Detroit), California (0.45%), and Connecticut (0.59%). Three regions with the largest declines are Kentucky (4.12%), New Mexico (3.89%), and Louisiana (3.19%). As we expected, the estimated proportions of PU are zero for all regions. Among not-covered SEER patients, approximately 73% are RU on average, with 65% to 85% across regions. These proportions are much smaller than those gold standard estimates. Under the same assumption about non-matched RU, we calculated the combined estimates of the proportion of $PU|RU$ which are shown in the last column of Table 4. The average proportion across all regions is nearly 1%. The largest proportion is associated with data from New Mexico registry (3.03%), followed by Kentucky registry (2.97%). The smallest proportion is associated with Detroit registry (0.24%) and the second smallest is with California (0.42%). The overestimation bias, on average, is 5.6 times, and ranges from 3.0 for Hawaii to 14.0 for Georgia. Table 4 shows the results from PUMS with county codes imputed using method 2. Similar to the results in Table 3, the coverage rates decreased but at smaller magnitudes. The proportions of $PU|RU$ are less than 0.1% for covered SEER patients for all regions except for Kentucky, and New Mexico. For Connecticut and Detroit, imputation procedures are not needed because all counties have 100,000 or more population. Therefore, the results are the same in both imputation methods. However, the magnitudes of the proportions of $PU|RU$ are small compared with the gold standard. Among not-covered SEER patients, the proportions of RU slightly increased over Imputation Method 1, but are also smaller than the gold standard. The accuracy of combined $PU|RU$ estimates improved over that of method 1, but still exhibits certain upward discrepancies from the gold standards. The overestimation bias, on average, is 3.7 times, and ranges from 2.4 for both Iowa and Kentucky to 8.0 for Georgia.

Table 3: Estimated Number of SEER Records (2000) that are RU and PU using Census 2000 5% PUMS by Coverage Status, Imputation Method 1

	Matched			Not-matched	Combined
	Match Rate (%)	RU (%)	PU	RU (%)	PU/RU (%)
Entire SEER File	98.71	22,203.0 (6.49)	0	3240 (72.54)	0.93
CA	99.49	4,987.2 (3.47)	0	609.8 (82.43)	0.42
CT	99.26	612.0 (3.04)	0	121 (81.21)	0.60
GA*	98.35	802.8 (7.29)	0	126.2 (68.44)	1.13
HI	98.57	323.0 (5.88)	0	63 (78.95)	1.13
IA	97.26	3,042.4 (19.23)	0	288.6 (64.65)	1.77
KY	95.56	3,835.8 (18.13)	0	657.2 (66.79)	2.97
LA	96.73	3,019.0 (14.73)	0	472 (68.23)	2.23
MI*	99.62	382.0 (1.70)	0	54 (62.79)	0.24
NJ	99.05	2,072.6 (434)	0	337.4 (74.06)	0.70
NM	96.00	1,524.2 (20.91)	0	229.8 (75.59)	3.03
UT	97.77	849.0 (12.41)	0	135 (86.43)	1.93
WA*	99.07	753.0 (3.74)	0	146 (77.00)	0.72

Notes: * Three SEER regions do not cover complete states. In Georgia, SEER covers metro Atlanta and several additional rural counties. In Michigan, SEER covers metro Detroit. In Washington, SEER covers metro Seattle. See http://seer.cancer.gov/registries/ for details.

Table 4: Estimated Number of SEER Records (2000) that are RU and PU using Census 2000 5% PUMS by Coverage Status, Imputation Method 2.

	Matched				Not-matched	Combined
	Match Rate (%)	RU (%)	PU	PU/RU (%)	RU (%)	PU/RU (%)
Entire SEER File	99.26	23,352 (6.79)	72	67 (.01)	2,087 (81.30)	2,154 (0.62)
CA	99.56	5,073 (3.53)	14	13 (.01)	524 (82.92)	537 (0.37)
CT	99.26	612 (3.04)	-	- (.00)	121 (81.21)	121 (0.60)
GA*	99.09	860 (7.75)	3	3 (.03)	69 (67.39)	72 (0.64)
HI	98.69	328 (5.96)	-	- (.00)	58 (79.45)	58 (1.04)
IA	98.93	3,174 (19.72)	1	1 (.00)	157 (90.25)	158 (0.97)
KY	98.59	4,209 (19.28)	23	22 (.10)	284 (90.96)	306 (1.38)
LA	98.87	3,301 (15.76)	16	13 (.06)	190 (79.60)	203 (0.96)
MI*	99.62	382 (1.70)	-	- (.00)	54 (62.79)	54 (0.24)
NJ	99.10	2,086 (4.37)	1	1 (.00)	324 (74.49)	325 (0.67)
NM	98.52	1,654 (22.11)	9	9 (.11)	96 (88.89)	105 (1.38)
UT	98.76	906 (13.11)	5	5 (.07)	78 (89.63)	83 (1.19)
WA*	99.16	767 (3.81)	1	1 (.00)	132 (77.65)	133 (0.66)

Notes: * Three SEER regions do not cover complete states. In Georgia, SEER covers metro Atlanta and several additional rural counties. In Michigan, SEER covers metro Detroit. In Washington, SEER covers metro Seattle. See http://seer.cancer.gov/registries/ for details.

Figure 1 shows the relative bias of $\hat{\tau}$ by imputation method for each SEER area. For each method, relative bias was calculated as the difference in $\hat{\tau}$ between PUMS and SF1 divided by SF1. A much larger bias occurred for areas with lower risks of disclosure for both methods. Because person weights are distributed similarly across low- and high-risk areas (see Table 5), differential sampling fractions are less likely to be the main reason for bias as otherwise discussed in the literature of estimating population uniqueness from random sample data [5, 10, 13, 24]. Rather, this finding suggests that the inflation of bias is closely related to the large number of population individuals with infrequent combinations of demographics within a geographic area.

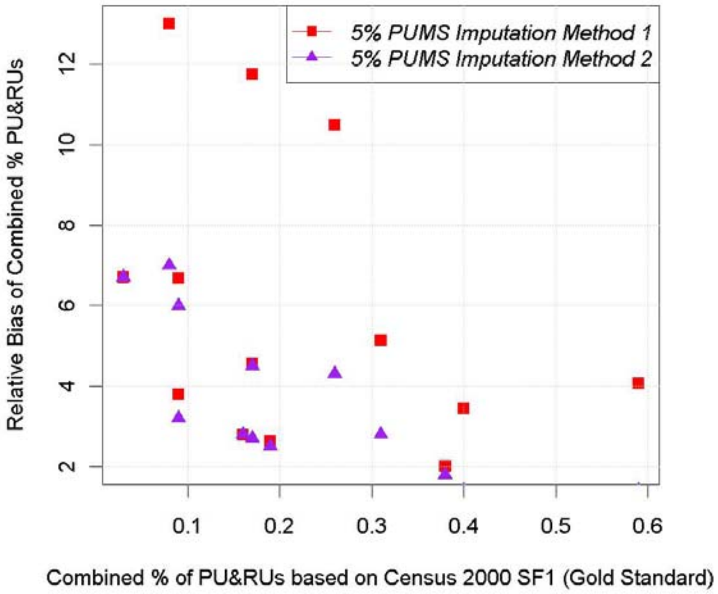

Figure 1: Relative Bias of Combined Proportion of PU/RU by Imputation Method

4. Discussion

The overall proportion of SEER patients who can be uniquely identified from the U.S. general population is small, which is less than 1 percent. However, given the size of the ever-growing SEER data system, this small percentage translates to hundreds of patients who are at great risk of having their private health information disclosed. The sustainability or even the survival of SEER data system is highly dependent upon the agency's ability to keep data anonymous. This study is one of the first steps to ensure SEER data confidentiality through statistical approaches.

Despite the upward bias, results from this study suggest that the census microdata sample file has great utility for assessing the proportion of record unique patients in a population based cancer surveillance data who are also unique in the population. The conservative estimation can lead to data being overly withheld from legitimate researchers and other data users. The bias can be due to several reasons: (1) low sampling fraction in PUMS; (2) the invalidity of assuming all not-covered SEER record uniques are population unique; (3) perturbation errors in county codes due to imputation; and (4) finally the underlying population distribution in an area. The first and second reasons oftentimes go hand in hand such that the smaller the sampling fraction, the more individuals, including those with less rare combinations of characteristics that are becoming unrepresented in PUMS due to sampling. For future study, an approach that allows relaxing the assumption about the population frequencies of non-matched unique SEER patients should be considered. A model should be developed to predict how many of these SEER uniques are also unique in the population. However, solving this problem is no less challenging than providing a solution to the original research question in which we sought estimates of population uniqueness for all SEER unique patients.

Alternatively, since a population unique individual will appear to be record unique if this person has cancer and resides in a SEER covered area, the proportion of PUs among SEER areas provides a upper bound for τ. The size of overestimation could be small as suggested by the results for matched SEER records shown in Table 1, although a formal evaluation of such bias is warranted. Then the future research question becomes how to estimate the proportion of PUs from a representative microdata sample of SEER covered regions, such as the Census or ACS PUMS, using probabilistic models [4, 5, 9, 24]. Methodological challenges such as low sampling fraction as well as complex sampling design features in PUMS should be addressed. By comparing the estimates obtained using different set of key variables, one could infer which variables or variable grouping schemes contribute most to the uniqueness, thus statistical procedures should be applied to control disclosure.

References

1. Howe, H.L., A.J. Lake, and T. Shen, *Method to Assess Identifiability in Electronic Data Files.* American Journal of Epidemiology, 2006.
2. Fellegi, I.P., *On the Question of Statistical Con®dentiality.* Journal of the American Statistical Association, 1972. **67**: p. 7-18.
3. Stephen E. Fienberg, U.E.M., and Ashish P. Sanil, *A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data.* Journal of Official Statistics, 1997. **13**(1): p. 75-89.
4. Bethlehem, J.G., W.J. Keller, and J. Pannekoek, *Disclosure Control of Microdata.* Journal of the American Statistical Association, 1990. **85**(409): p. 38-45.
5. Chen, G. and S. Keller-McNulty, *Estimation of identification disclosure risk in microdata.* Journal of Official Statistics, 1998. **14**(79-95).
6. Greenberg, B. and L.V. Zayatz, *Strategies for Measuring Risk in Public Use Microdata Files.* Statistica Neerlandica, 1992. **46**(1): p. 33-48.
7. Samuel, S.M., *A Bayesian, species-sampling-inspired approach to the uniques problems in microdata disclosure risk.* Journal of Official Statistics, 1998. **14**: p. 373-383.
8. Shlomo, N., *Release Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility*. Journal of Privacy and Confidentiality, 2010. **2**(1): p. 73-91.
9. Skinner, C.J. and R.G. Carter, *Estimation of a measure of disclosure risk for survey microdata under unequal probability sampling*. Vol. 29. 2003. 197-201.
10. Skinner, C.J. and M.J. Elliot, *A measure of disclosure risk for microdata.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 2002. **64**: p. 855-867.
11. Skinner, C.J. and D.J. Holmes, *Estimating the re-identification risk per record in microdata*. Vol. 14. 1998. 361-372.
12. Skinner, C.J., et al., *Disclosure control for census microdata.* Journal of Official Statistics, 1994. **10**(31-51).
13. Zayatz, L.V., *Estimation Of The Percent of Unique Population Elements On A Microdata File Using The Sample*, in *Statistical Research Division Report Series*. 1991, Bureau of The Census: Washington, DC.
14. Carter, R.B., J.-R, Briggs, M., *Analysis of the risk of disclosure for census microdata*. 1991, Social survey methods division, Statistics Canada, Ottawa: Ottawa.
15. Elliot, M.J., Skinner, C.J. and Dale, A. , *Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk.* Research in Official Statistics, 1998. **1**(2): p. 53-67.

16. Fienberg, S.E. and U.E. Makov, *Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data.* Journal of Official Statistics, 1998. **14**(385-397).
17. Lambert, D., *Measures of disclosure risk and harm.* Journal of Official Statistics, 1993. **9**(2): p. 313-331.
18. Clegg, L., et al., *Quality of race, Hispanic ethnicity, and immigrant status in population-based cancer registry data: implications for health disparity studies.* Cancer Causes Control, 2007. **18**(2): p. 177-187.
19. Zayatz, L., *Disclosure Limitation for Census 2000 Tabular Data*, in *Joint ECE/Eurostat work session on statistical data confidentiality*. 2003: Luxembourg.
20. U.S. Census Bureau, *2000 Census of Popualtion and Housing, Public Use Microdata Sample, United States: Technical Documentation*. 2008.
21. U.S. Census Monitoring Board Presidential Members, *Final Report to Congress, Section 5, Effect of Census 2000 Undercount on Federal Funding to States and Selected Counties, 2002-2012*. 2001, U.S. Census Monitoring Board.
22. Yu, M., *Disclosure risk assessments and control*, in *Program in Survey Methodology*. 2008, The University of Michigan: Ann Arbor. p. 165.
23. Rubin, D.B., *Multiple Imputation for Nonresponse in Surveys*. 1987, New York:: John Wiley & Sons, Inc. .
24. Skinner, C. and N. Shlomo, *Assessing Identification Risk in Survey Microdata Using Log-Linear Models.* Journal of the American Statistical Association, 2008. **103**(483): p. 989-1001.