

An Identifiability Assessment of *CINA Deluxe* with Area-based SES Measures

Holly L. Howe, PhD

North American Association of Central Cancer Registries, Inc.

Andrew Lake

Information Management Services, Inc.

1 Background

It is well known that health inequalities exist among various segments of the population, with their elimination an overarching goal of *Health People 2010*. [1] Health inequalities are evident among various geographic locations, race and ethnic groups, social and economic categories, and by personal characteristics related to disability, gender, and sexual orientation. [2] Public health surveillance systems routinely include information on gender, race, ethnicity, other demographic characteristics and some area-based measures of geography; however, inclusion of economic information is a general omission. [3] One consequence of the absence of these data is descriptions of inequalities are formalized in terms of the variables that are collected – like race and ethnicity—rather than the potentially stronger predictors of economic deprivation.

Cancer incidence data are heavily reliant on information documented in the medical record. Personal economic information is generally not included in a medical record, making it difficult to measure cancer inequalities in economic terms. However, through electronic linkage, area-based economic measures can be obtained, thus permitting exploration of the impact of area economic deprivation on cancer outcomes. The social and economic environment is a powerful construct in measuring deprivation or inequality. [4-7] Further, Krieger and colleagues [2,4-6] have identified that the size of the geographic area affects the predictability of area measures of economic deprivation, with census tract and block group areas being sufficiently sensitive to inequality gradients in a variety of health outcomes. Singh and colleagues [7] examined disparities in cancer mortality and found that differences in economic indicators among census tract areas were greater than they were among counties. Nonetheless, the analysis using county data, in the absence of census tract data for all decedent cases, demonstrated differential cross-sectional patterns in mortality among counties of low and high poverty rates. [7]

The *CINA Deluxe* research file, that NAACCR makes available to its member researchers, attempts to balance patient confidentiality concerns with file usefulness and member registries that contribute their data to the file determine this balance. In 2004, registries believed that the balance for a file like *CINA Deluxe* should not include a geographic identifier smaller than a county. The decision was based on several considerations:

- (1) Individual registry data are part of a multi-registry aggregated data file that is accessible from one source, but one that is overseen by NAACCR, not the registry; and

- (2) Census tract information is excluded from NAACCR data submissions due, in part, to known incomplete data in all registries and unknown accuracy of census tract assignment by commercial geocoding software.

Many registries have recently become more comfortable with the inclusion of a county identifier on the NAACCR file submissions due to several NAACCR policies related to data use:

- (3) Registries are informed by NAACCR of all research projects that propose to use their data and obtain their consent for that use;
- (4) NAACCR has all research projects conducted under the auspices of NAACCR reviewed by the NAACCR Institutional Review Board (IRB); and
- (5) All projects that plan to use or publish state or county identifiers employ extra precautions in the consenting process.

However, a small number of registries still do not submit the county variable or do not give permission for a county variable to be included on a research file due to state law or local policy that govern release of these data. Thus, the county variable is not a standard data item as it would necessitate the omission from all NAACCR research projects, the data submitted from quality registries with this county restriction.

A variety of socioeconomic indicators have been used as area-based, or ecologic, measures to identify disparities and inequalities in health outcomes, including cancer incidence [See e.g., 2,4-7]. Researchers and advocacy groups have asked NAACCR about the availability of socioeconomic measures in cancer registries and on the NAACCR research analytic file, CINA Deluxe. Since no consensus on the most appropriate or best socio-economic measure has yet been identified in the scientific literature, we relied on the work of Krieger [2] and Singh [7] to identify a meaningful measure that would be simple, useful, and meaningful across many geographic areas and over time. [2]

2 Purpose

A balance between patient privacy (record confidentiality) and information release must be maintained. Confidentiality concerns include not only the capability to identify a patient from a data file, but also the potential to gain new information about a known patient, or to re-identify a patient through linkage of the registry file with other electronic files. This balance can be accomplished in several ways: omission of personal identifiers to eliminate direct identification, omission of certain variables in released information, or aggregation of variable values to diminish indirect disclosure through unique combinations of data values.

The purpose of this project was to assess empirically the increased identifiability or uniqueness of records included in NAACCR's research file, CINA Deluxe, that would result from the addition of county-based socioeconomic indicators obtained from the U.S. Bureau of the Census. The empirical assessment was conducted to identify the number and proportion of records that would be identifiable as single unique records or within sets of five or fewer records with the addition of one or more county-based socioeconomic measures (CBSM). Further, beyond merely assessing uniqueness, we would also ascertain how, if any, modifications to the data variables or the file, might appropriately balance the inclusion of CBSM information and also maintain an appropriate protection of confidentiality.

3 Method

3.1 Data Source

The NAACCR data submission for 1995 - 2001 was comprised of 35 population-based cancer registries (see Table 1 for complete list) that included a specific county identifier (totaling 1,499 counties) on their file. This file had more than 4.67 million cancer cases that were included in this assessment of case uniqueness.

Alabama	Hawaii	Massachusetts	Oklahoma
Arizona	Idaho	MI: Detroit Metro	Oregon
CA: Greater Bay Area	Illinois	Missouri	Rhode Island
CA: Los Angeles	Indiana	Montana	South Carolina
Delaware	Iowa	Nebraska	Utah
District of Columbia	Kentucky	New Hampshire	Washington
Florida	Louisiana	New Jersey	West Virginia
GA: Atlanta	Maine	New Mexico	Wisconsin
		New York	Wyoming

3.2 Record Uniqueness Program

The North American Association of Central Cancer Registries, Inc. (NAACCR), in collaboration with Information Management Services, Inc., developed the Record Uniqueness software as a tool to test data files for potential patient identifiability related to small numbers, or small cell sizes. [8] Record Uniqueness (RU) was based on a SAS program originally developed by the Illinois State Cancer Registry to evaluate their public use files for risk of confidentiality breach. [8] This tool evaluates the data file for identifiability and potential re-identifiability, since most often, confidential data items, e.g., name, address, or social security number, are not released to researchers or available on public use files. The program identifies when any combination of designated variables results in a unique record or in a set of five or fewer unique records.

Record Uniqueness allows the user to define the variables that will be assessed for uniqueness. It also allows the user to input variables or recodes of variables (e.g., SEER site recodes rather than the default primary site). A Default Variable Set is pre-selected as it contains the basic variables that should be included in any analysis for unique records. These are age, sex, race, year of diagnosis, and primary cancer site. If data are requested for more than one geographic area (e.g., states within the United States or counties within a state), then the relevant geographic variable should also be added to the Default Variable Set for an analysis of unique records.

Once the data have been processed, Record Uniqueness outputs descriptive information on the number of records in each analysis and the number and proportion of unique records. It also lists the number and proportion of combinations that produces sets of five or fewer unique records. In addition, the program provides the relative contribution (or weight) of each variable

in the Variable Set to the total number and proportion of unique records on the data file. RU guides the user to decrease the number of unique records, by identifying the variable with the greatest contribution to unique records. By re-categorizing this variable into a variable with fewer values, one can achieve the greatest reduction in the number of unique records. Through this aggregation of values into larger groups of values and an iterative process of testing for unique records, a user can create a data file that meets acceptable thresholds of unique records and achieve a balance between protection of patient (record) confidentiality and data file information. When the re-aggregation of variable values into fewer categories is insufficient in reducing the magnitude of unique records, or becomes meaningless with regard to data utility, then one should consider omitting the variable from the data file to achieve the desired uniqueness threshold.

The NAACCR CINA Deluxe Advisory Group, a committee that reviews development and release of NAACCR data files for researchers (CINA Deluxe) and the public (CINA+ Online), defined two record uniqueness thresholds for data files. As one would expect, the number and proportion of unique records increases as the number of variables involved in the combinations increases. Also the variables with the largest number of values (e.g., primary site) also are more likely to contribute to record uniqueness than variables with few categories (e.g., sex), particularly when the distribution of cases are not evenly spread among the values. The suggested threshold [8] for files created for researchers is that no more than 20 percent of the variable combinations should identify categories of five (5) or fewer patients based on the key variables requested or the Default Variable Set (including geography). Similarly, for public use files, fewer than 5 percent of the variable combinations should identify sets of five (5) or fewer patients based on the Default Variable Set (including geography if it is available on the data file).

3.3 Approach

The first step of the analysis assessed whether inclusion of multiple CBSMs on the file had a differential impact on uniqueness than inclusion of only one CBSM. In this phase of analysis, we examined three CBSMs including either one or two combinations of the measures. The county identifier for each case was linked with files of the U.S. Bureau of the Census to obtain three area-based socioeconomic indicators for the county. These were:

- The percent of the county population below the poverty level
- The median household income for the county, and
- The percent of county residents 25 years and older that had a high school diploma.

The three CBSMs were appended to each record for the purposes of this assessment. Based on the recommendation of Krieger [4,5] for a useful area based socioeconomic measure, the proportion of persons living below the poverty level would be the preferred single measure if only one had to be selected, although we were restricted to the county as the smallest available geographic area.

Following this, an iterative approach to assessing unique records was conducted. After each iteration, the resulting file was examined as to whether it achieved the threshold standards for unique records (sets of five or fewer records at lower than 20 percent). If the threshold was not achieved, the relative contribution (weight) of each variable was used to revise the variable(s) that had the greatest uniqueness impact. These were re-aggregated with fewer categories for the next iteration. The exception to this was that when the CBSM had the highest

weight it was overlooked in the early iterations in order to achieve the greatest precision (information) from this variable. After all meaningful iterations involving re-aggregation of the variables were attempted, then the CBSM values were grouped by rounding the percents to the nearest integer to reduce the number of categories.

The analysis started with SEER site groups, rather than primary cancer sites. The re-aggregation of SEER site codes included a set when only major site groups were identified and also when only minor sites were identified. Age was recoded in several ways for the iterations involving re-aggregation of five-year age groups. In one scenario, five-year age groups were classified into four age groups (see table 2) and a second modification grouped 0-19 year olds into one age category and all other age groups through 85+ years into five-year age groups (15 categories).

The following table summarizes the variables included in the record uniqueness assessment for this study. For each variable, the table provides the number of values included for each variable. The table includes all permutations that were tested for age, site, and the CBSM (percent of the county population below the poverty level).

Table 2. Variables in the Record Uniqueness Analysis with the Count of Values in the Variable	
Variable	Values Count
Registry	35
Race (W-B-Other- Unknown)	4
Age	
Five-year age groups	18
20-Year age groups (0-19, 20-39, 40-64, 65+)	4
Modified five-year groups (0-19, 20-24, 25-29...85+)	15
Sex	2
Year of Diagnosis	7
Primary Site	
SEER Site Recode	78
Major Sites	16
Minor Sites	45
% County Poverty	
Actual	283
Rounded to Integers	40

The median and percent CBSMs are available to the tenth-decimal place and the median household income is available to the nearest whole dollar. In the final iterations to achieve the unique record thresholds without eliminating variables, the CBSM, percent county poverty, was rounded to the nearest whole number. At this level of precision, the impact of omitting variables was assessed on record uniqueness and the extent to which the desired thresholds could be achieved.

4 Results

4.1 Single or Multiple CBSMs

Table 3 presents the results of the first set of analyses on the impact on the number and proportion of adding one or multiple CBSMs to CINA Deluxe. Adding one CBSM to the file, and regardless of which CBSM, increased the proportion of unique case sets of five or fewer cases to about one-half of the records on the file. Of the three CBSMs, percent poverty had the lowest estimate at 49.2 percent. In each CBSM analysis, the CBSM was the variable contributing most to uniqueness, followed by cancer site groups and five-year age groups.

The last section of table 3 relates the impact of adding two CBSMs to the file, where it was found that the uniqueness was only marginally affected by the addition of a second CBSM. In this case, the addition of both median county household income and percent county poverty yielded the same number (2.3 million) and proportion (51.0 percent) of unique records as median county household income alone. However the weights and the order of variables contributing to uniqueness changed. When both CBSMs were on the file, site recode contributed the most to uniqueness, followed by five-year age groups, then median household income and percent county poverty.

At the conclusion of these analyses, it was decided to focus on the addition of just one CBSM variable to attempt to reduce the sets of unique records from about 50 percent to the desired threshold of less than 20 percent. The CBSM, percent population in the county below the poverty level, was selected for the remaining analyses.

4.2 Variable Recodes and Re-aggregation

The weights listed in Table 3 for the percent county poverty suggest that the contribution to uniqueness is, in order, percent poverty, site recode, five-year age groups, followed by registry, race recode, year of diagnosis, and sex. Following the interest in maintaining the greatest information for the CBSM, the values of site and age were re-grouped into fewer categories using several approaches to determine whether the threshold for record uniqueness could be achieved. These results are presented in Table 4.

The first two iterations focused on cancer site groups, one using only SEER major site groups and the other, SEER minor site groups. In the first case, using SEER major site groups, the record uniqueness of cases sets of five or fewer cases dropped to 33.1 percent. The minor site group iteration did not fare as well, achieving a small reduction from 49.2 percent to 44.7 percent. Neither of these achieved the desired threshold and substantial information was lost in both cases for site-specific information.

The next iteration collapsed age categories into four major groups (roughly 20 years each). In this situation, record uniqueness dropped to 26.7 percent, not yet achieving the desired threshold. In this scenario, percent county poverty and site recode variables were contributing most to uniqueness. Having already explored meaningful recodes for site, we determined that the CBSM needed to be grouped. Thus, the precision of the CBSM was categorized by rounding the decimal place to the nearest full number. When integer-level poverty data were substituted in the revised

(20-year) age recode model, record uniqueness dropped to 17.6 percent, a value that fell below our desired threshold.

Because 15-20 year age categories are broad, and would dramatically affect the values of age-adjusted rates, several additional modifications were tried to see whether percent poverty at the integer level could achieve the threshold while five-year age groups were retained and if not, if only childhood cases were re-grouped into one category of 0-19 years, while all other ages remained in five-year categories. As shown at the bottom of table 4, neither of these options resulted in a satisfactory proportion of unique records, and in fact the result for both was similar, about 37 percent of the cases resulted in unique case sets of five or fewer records.

4.3 Variable Omission

Based on all the iterations, it appeared that to achieve the desired threshold of unique records and to maintain meaningful precision levels of key variables, we would need to consider omitting variables from the data set. Again this decision was guided by the desire to maintain the maximum precision in the CBSM. The options and results of omitting each of the default variables are summarized in Table 5. As shown, either omitting race (31.8 percent unique record sets) or sex (29.8 percent unique case sets) does not help achieve the threshold of 20 percent or fewer unique case sets. Eliminating the site recode or five-year age groups has the greatest impact on unique case sets, but the loss of information from either of these variables would greatly diminish the utility of the data file. Eliminating either year of diagnosis or registry code is effective in reducing unique record sets to below the uniqueness threshold. Should one omit year of diagnosis from the data file, there would be 14.9 percent unique case sets of five or less; if one omitted the registry variable, there would be 12.3 percent unique case sets.

5 Discussion

This assessment has empirically demonstrated that a county-based socioeconomic measure can be added to the CINA Deluxe file through linkage with information from the US Bureau of the Census, while maintaining an appropriate level of record confidentiality. However, some modifications are necessary to achieve thresholds that give an appropriate balance for patient confidentiality and file usefulness. This balance may change in the future. New suggestions or new permutations of a file can be retested with the Record Uniqueness program.

Further, gradients in economic deprivation and cancer incidence inequalities may be less pronounced using county-level measures. As Krieger has found, the size of the geographic area does matter. Results using county measures may show fewer inequalities than exist in truth, or may not detect inequalities in some areas. When possible, perhaps within one cancer registry, research on cancer inequalities as related to economic deprivation and disparities should use census tract areas as a more robust measure. [4]

The recommendation of the NAACCR Data Evaluation and Publication Committee is that the default variables from RU should include the following definitions: SEER site recode; five-year age groups, race (white/black/other), sex, and the CBSM, percent county population below

poverty (rounded to the integer). Further, NAACCR should produce two independent CINA Deluxe-CBSM files: one that includes year of diagnosis, but omits registry code, and the second that includes registry code, but omits year of diagnosis. As this cumulative file grows, further evaluation will be undertaken to determine whether the file that includes registry code can also include a year of diagnosis variable that is aggregated into multi-year, rather than single year, intervals and still reach the desired record uniqueness threshold.

6 Conclusion

The analyses yielded several options for a CINA Deluxe-CBSM data file. In 2005, NAACCR will release a CINA Deluxe-CBSM file in two formats as described above. This will be available to NAACCR researchers through the same discretionary release process that is used to review research applications and provide access to the data set.

Researchers have also requested CINA Deluxe files with a county identifier that would enable them to link and append data from other electronic data files that would be used to address myriad research questions. A county identifier is available to researchers, but only after special request in their research proposals and written consent by all participating registries.

In the future and depending on the research questions and uses, a different configuration for the file may be released (for example, after an assessment, it might include the CBSM as a categorical variable with only 4 or five categories, and both registry name and year of diagnosis, if this combination achieves the desired thresholds for record uniqueness.

7 References

1. US Department of Health and Human Services. Healthy People 2010 (Conference edition, two volumes). Washington DC: US Govt Printing Office, 2000. (URL: <http://www.health.gov/healthypeople/about/goals.htm>, Last accessed December 21, 2004).
2. The Public Health Disparities Geocoding Project Monograph. Geocoding and Monitoring US Socioeconomic Inequalities in Health: An introduction to using area-based socioeconomic measures, version 7.04. President and Fellows of Harvard College, 2004.
3. Kreiger N, Chen JT, Ebel G. Can we monitor socioeconomic inequalities in health? A survey of US Health Departments' data collection and reporting practices. *Public Health Rep* 1997; 112:481-91.
4. Krieger N, Chen JT, Waterman PD, Soobader M, Subramanian SV, Carson R. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? *Am J Epidemiol* 2002; 156:471-82.
5. Krieger N, Chen Jt, Water,am PD, Rehkopf DH, Subramanian SV. Race/ethnicity, gender, and monitoring area-based socioeconomic measures – the public health disparities geocoding project. *Am J Public Health* 2003; 93(10):1655-71.

6. Krieger N, Waterman P, Lemeieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *Am J Public Health* 2001; 91:1114-6.
7. Singh GK, Miller BA, Hankey BF, Edwards BK. Area Socioeconomic Variations in U.S. Cancer Incidence, Mortality, Stage, Treatment, and Survival, 1975-1999. NCI cancer Surveillance Monograph Series, Number 4. Bethesda, MD: National Cancer Institute, 2003. NIH Publication No. 03-5417.
8. NAACCR. Record Uniqueness Software. [URL: http://www.naacr.org/index.asp?Col_SectionKey=7&Col_ContentID=312, Last Accessed March 1, 2005].

Table 3. Record Uniqueness Assessment of CINA Deluxe, 1995-2001, with Various County-based Socioeconomic Measures, CINA Deluxe File 1995-2001, 4.67 million cases

Variable List	Unique Cases	Percent Unique Cases	Unique Case Sets of 5 or Less	Percent Unique Case Sets of 5 or Less	RU Weight
Registry, Race Recode, 5-yr Age Recode, Site Recode, Sex, Year DX AND:					
Pct HS Diploma	1,060,575	22.6983	2,306,844	49.3708	
Pct HS Diploma					5.2484
Site Recode					4.95924
5-yr Age Recode					3.42442
Registry					2.06537
Race Recode					1.82819
Year DX					1.67485
Sex					0.73275
Pct Poverty	1,049,929	22.4704	2,299,113	49.2053	
Pct Poverty					5.15467
Site Recode					4.98515
5-yr Age Recode					3.44328
Registry					2.20534
Race Recode					1.82314
Year DX					1.72377
Sex					0.74528
Median Household Income	1,122,478	24.0231	2,381,489	50.9683	
Median Household Income					5.7223
Site Recode					4.60011
5-yr Age Recode					3.1433
Race Recode					1.60432
Year DX					1.49811
Registry					1.2074
Sex					0.6936
Percent Poverty and					
Median Household Income	1122478	24.0231	2381489	50.9683	
Site Recode					4.30944
5-yr Age Recode					2.90144
Median Household Income					2.89992
Pct Poverty					2.01056
Race Recode					1.34966
Year DX					1.32568
Registry					0.68078
Sex					0.67395

Table 4. Record Uniqueness Assessment of CINA Deluxe, 1995-2001, with a County-based Socio-economic Measure and Various Variable Recodes , CINA Deluxe File 1995-2001, 4.67 million cases

Variable List	Unique Cases	Percent Unique Cases	Unique Case Sets of 5 or Less	Percent Unique Case Sets of 5 or Less	RU Weight
Registry, Race Recode, 5-yr Age Recode, Major Site , Sex, Year DX, Pct Poverty	540,694	11.5719	1,548,756	33.1463	
Pct Poverty					5.50175
5-yr Age Recode					3.58988
Major Site					3.51871
Registry					2.4479
Race Recode					2.14181
Year DX					1.83828
Sex					0.74634
Registry, Race Recode, 5-yr Age Recode, Minor Site , Sex, Year DX, Pct Poverty	895,096	19.1567	2,090,596	44.7427	
Pct Poverty					5.37685
Minor Site					4.61168
5-yr Age Recode					3.59623
Registry					2.307
Race Recode					1.90846
Year DX					1.73937
Sex					0.73337
Registry, Race Recode, 20-yr Age Recode , Site Recode, Sex, Year DX, Pct Poverty	459,128	9.8262	1,247,831	26.7059	
Pct Poverty					5.38193
Site Recode					5.2924
Registry					2.32655
Race Recode					2.08401
Year DX					2.03222
20-yr Age Recode					1.86331
Sex					0.86993
Registry, Race Recode, 20-yr Age Recode , Site Recode, Sex, Year DX, Pct Poverty (integer)	264,235	5.65512	822,184	17.5963	
Site Recode					5.66202
Pct Poverty					4.01818
Registry					3.14883
Year DX					2.32034
Race Recode					2.16178
20-yr Age Recode					1.94556
Sex					0.88635
Registry, Site Recode, Sex, Race Recode, Year DX, Pct Poverty (Integer), 5-yr Age Recode	693,494	14.8421	1,754,815	37.5563	
Site Recode					5.29957
Pct Poverty					3.90265
5-yr Age Recode					3.65113
Registry					3.01285
Year DX					2.04606
Race Recode					1.88676
Sex					0.75015
Registry, Site Recode, Sex, Race Recode, Year DX, Pct Poverty (Integer), 0-19 Age Grp and other five-year age groups	683,736	14.6332	1,749,624	37.4452	
Site Recode					5.4314
Pct Poverty					3.9595
0-19 then 5-yr age recode					3.39498
Registry					3.14115
Year DX					2.2936
Race Recode					2.09476
Sex					0.85295

Table 5. Record Uniqueness Assessment of CINA Deluxe, 1995-2001, Options that meet Uniqueness Thresholds with a County-based Socioeconomic Measure
 CINA Deluxe File 1995-2001, 4.67 million cases

Variable Omitted	Unique Cases	Percent Unique Cases	Unique Case Sets of 5 or Less	Percent Unique Case Sets of 5 or Less	Variable List
None	693,494	14.8421	1,754,815	37.5563	Registry, Site Recode, Sex, Race Recode, Year DX, Pct Poverty, 5-yr Age Recode
Race	522,308	11.1784	1,487,792	31.8415	Registry, Site Recode, Sex, Year DX, Pct Poverty, 5-yr Age Recode
Sex	493,796	10.5682	1,390,263	29.7542	Registry, Site Recode, Race Recode, Year DX, Pct Poverty, 5-yr Age Recode
Site recode	39,237	0.8397	168,028	3.5961	Registry, Sex, Race Recode, Year DX, Pct Poverty, 5-yr Age Recode
Age Recode	120,560	2.58021	464,425	9.9396	Registry, Site Recode, Sex, Race Recode, Year DX, Pct Poverty
Year of Diagnosis	215,902	4.62071	696,925	14.9155	Registry, Site Recode, Sex, Race Recode, Pct Poverty, 5-yr Age Recode
Registry	169,003	3.617	576,790	12.3444	Site Recode, Sex, Race Recode, Year DX, Pct Poverty, 5-yr Age Recode