RHODE ISLAND

Rhode Island Cancer Registry (RICR); Junhie Oh, RICR Administrator

Address Quality Review and Validation to Improve Geocode Matching

NATIONAL PROGRAM OF CANCER REGISTRIES SUCCESS STORY

SUMMARY: Demands of geographic identifiers for small areas, such as city/town, ZIP or census tract levels, have increased in cancer researches and programs. Central cancer registries are challenged by technique-, time- and labor-intensive geocoding processes to provide the national cancer surveillance system with quality and standardized geocoding output.

To yield more effective and quality geocoding output, and identify less-burdensome staff time/resource investment, RICR selected 2011-2015 reportable cases and implemented a sequential and systemic address quality review and validation, but not in an exhaustive way.

"Standardization" of address information, prior to running the Geocoder, improved the geocoding matching status in a certain level. Although, we learned that "complete" street-level addresses do not necessarily return with accurate or valid census tracts in the Geocoder. There is no single "automated" geocoding procedure or tool that would match 100% of quality and valid data. RICR explores and tests the third-party geocoding software or web-based search engines that would supplement the Geocoder, and effectively invest staff time and resources.

CHALLENGE: Demands of geographic identifiers for small areas, such as city/town, ZIP or census tract levels, have increased in cancer researches and programs. Such research/study outcomes undeniably rely on accurate, complete and valid address information from patient's medical records. Beyond the local registries' efforts to strive for quality information, central cancer registries assume unique responsibilities to provide the national cancer surveillance system with geocoding output produced and tested by a standardized geocoding method, thereafter, that can be aggregated and comparable with other registries' corresponding data.

Like most of other states, Rhode Island Cancer Registry (RICR) is benefited by a state-of-the-art and easy-to-use geocoding platform, the NAACCR Geocoder (developed by Texas A&M University). RICR can conduct geocoding fast and effectively at a minimum level cost and skill requirement, and incorporate the standardized and required geocoding data elements into the annual final datasets. Although, NAACCR has recommended registries carefully and thoroughly review the Geocoder output and quality, given that the tool's sensitivity and specificity would not equally apply to all states' unique environment.

To yield more effective and quality geocoding output, and identify less-burdensome staff time/resource investment, RICR selected 2011-2015 reportable cases and implemented a sequential and systemic address quality review and validation, but not in an exhaustive way.

SOLUTION:

Step 1 -- Prior to the Geocoder run with the reportable cancer records diagnosed in 2011-2015 (N=32,473), "quick" and "basic" address clean-ups were conducted:

- All city/town names were up-cased
- 172 records' (0.5%) 5+4 ZIP codes were simplified to 5-digit only ZIP; Previous year's RICR geocoding experience did not show an enhanced functionality by ZIP code precision, and 99.5% of the patient address were reported with 5-digit ZIP
- 65 records' (0.2%) misspelled city/town names were corrected (e.g. "North Kingstown" spelled wrong as "North Kingston" or "North Kingtown")
- 47 records' (0.1%) shorten city/town names were fully spelled out (e.g. "N Scituate" or "NO Scituate" to "North Scituate")

Step 2 -- Geocoder run to yield a baseline geocoding output file before conducting the 2nd level of more extensive address clean-ups. The output file was named as "Before QA", and saved to compare with "After QA" later.

Step 3 -- 2nd level address clean-ups were conducted:

- "Before QA" output file was read in the SAS 9.4, and city/town and ZIP code were listed and reviewed; (1) City/town names used in the hospital reports were mixtures of the official and non-official municipal names, the census designated place (CDP) names within city/town, and conventional small area (villages) names that are not recognized by the census designation. (2) Numbers of the ZIP codes were found to be "not clean" with not-existing ZIPs (e.g. 02803, 02843, and etc), or out of the valid range (02801-02904) in Rhode Island.
- City/town and ZIP cross-tabulation was created, and reviewed against the USPS database; 1,417 records (4.4%) did not have correct combinations of city/town and ZIP. USPS database was downloaded from the NAACCR Geocoding Resource page: Address Validation Reference Data (https://www.naaccr.org/gis-resources/#Address). Wisconsin Cancer Registry first created, in 2011, the Access databases derived from the USPS database, and updated in 2014.
- Prior to the 2nd Geocoder run, city/town and ZIP information were cleaned up to obtain the best possible matches:
  - Approximately 900 records' city/town names or ZIP codes were re-assigned. Two-thirds of these, SAS program replaced mismatched cities with the UPSP-"preferred" cities by ZIP codes (e.g., "Providence" is preferred to "North Providence" for the ZIP "02904"; "Providence" is preferred to "Cranston" for the ZIP "02905". One third of the street level addresses were manually checked in the USPS address search engine (https://tools.usps.com/zip-code-lookup.htm?byaddress), and the recommended "correct" city/town names and ZIP codes were used for re-coding.
  - 350 records' city/town and ZIP combinations were found to be somewhat "correct". However, small places/villages within city/town - some are census-designated and some are not – did not correspond with the USPS reference table. (e.g., Greenville is a census-designated place within the City of Smithfield. The USPS-preferred ZIP for Greenville is not same with the ZIP for Smithfield. Similarly, ZIP 02879 is preferred for "Wakefield", a village in town of South Kingstown, to "Peace Dale", "Narragansett" or "South Kingstown", although these areas' geopolitical or census designation boundaries are overlapped.
  - A small number of the addresses (~70) showed some random errors, and these city-ZIP combinations were not explained with physical proximity. Manual searches with street level addresses were the only possible way to determine true values of city or ZIP.
  - Another small set of ZIP codes' (~40) had single-digit diversions that were considered most likely from data entry errors. Manual address search confirmed these errors and corrected.

Step 4 -- Geocoder was run again with cleaned-up addresses. Geocoding quality was compared between the "After QA" and "Before QA" output files.

RESULTS:

- Overall, about 3% of the 2010 Census Tract output was based on "incomplete" address information, such as ZIP only, PO Box, or City only, indicating that a majority of the 2011-2015 cases were reported with complete street-level addresses (Table 1).

Table 1. NAACCR Certainty Type Summary	Record count	%
Residence Street Address	31,529	97.1%
Residence ZIP	590	1.8%
PO Box ZIP	335	1.0%
Residence City/ZIP with only 1 Census Tract	13	<0.1%
Blank (geocoding not assigned)	6	<0.1%
Total	32,473	100%

- A certain level of "standardization" process, prior to running the Geocoder, improved the address "Match Type" outcomes, increasing "exact" matches, and reducing "relaxed" matches (Table 2). This level of improvement did not exactly correspond with the number of the records of which addresses were manually or programmatically corrected (appx. 1,400 addresses). Collectively evaluating Tables 1 and 2, we can tell a "complete" street-level address does not necessarily lead to find an accurate or valid census tract.

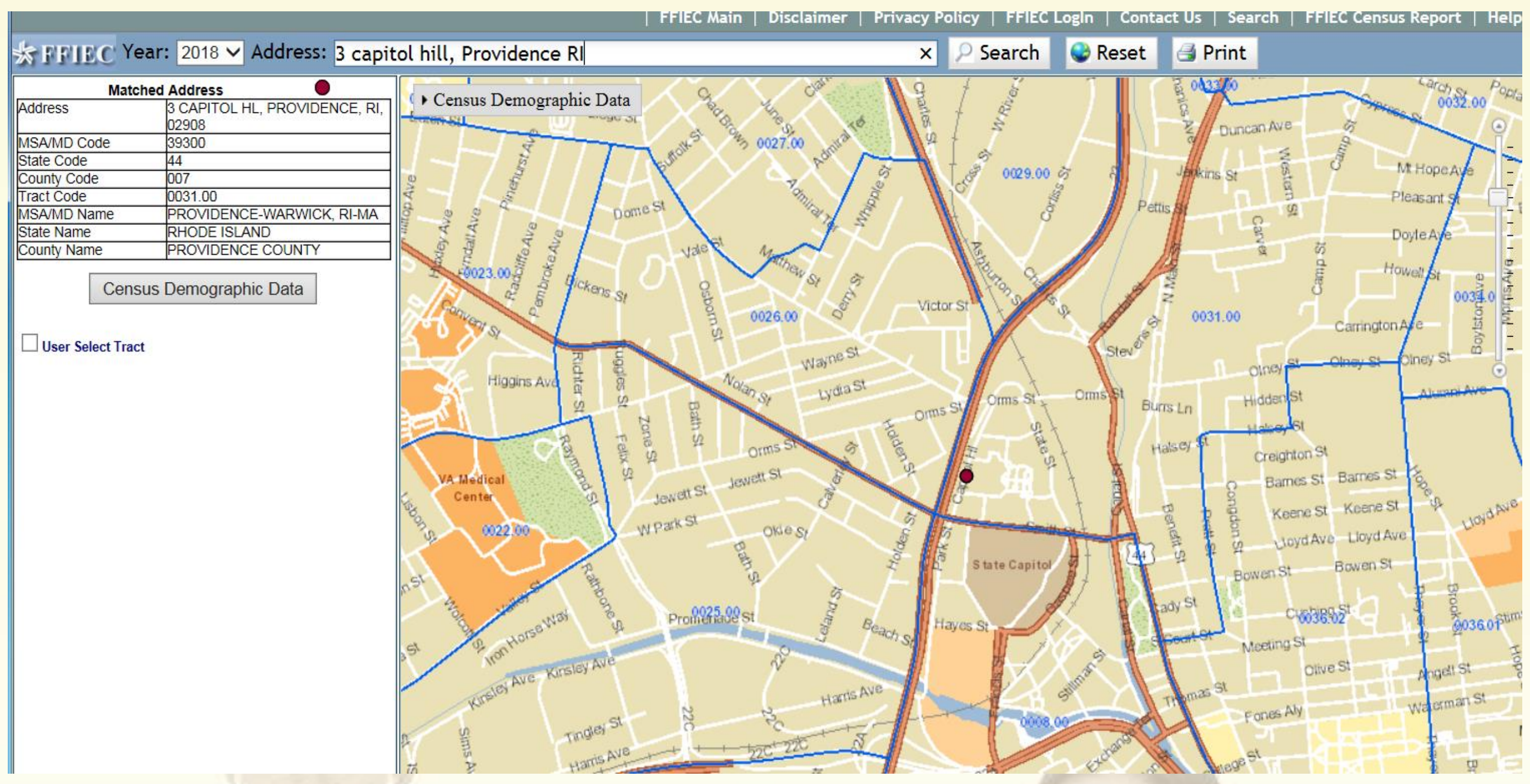
Table 2. Match Type Summary	Before QA		After QA: address clean-ups			
					Quality improvement	
	Record count (a)	% (b)	Record count (a)	% (b)	Count difference (c-a)	% difference (d-b)
Exact	28,971	89.2%	29,357	90.4%	386	1.2%
Relaxed	1,726	5.3%	1,354	4.2%	-372	-1.1%
Relaxed; Soundex	1,762	5.4%	1,756	5.4%		
Blank	14	<0.1%	6	<0.1%		
Total	32,473	100%	32473	100%		

SUSTAINING SUCCESS:

Reviewing address quality and geocoding output with the reportable cases diagnosed in 2011-2015, we found the current USPS reference dataset useful to validate the best city-ZIP match in addresses. Even with a limited success, "standardization" of address information, prior to running the Geocoder, improved the geocoding matching status.

A big lesson learned by RICR is: there is no single "automated" geocoding procedure or tool that would return with 100% of quality and validity. The process inevitably engages with a certain level of labor- and time-intensive manual reviews and validations. Underlying Geocoder dataset and algorithm cannot always assign "correct" geocodes for all Rhode Island cases. In addition to the USPS address search engine we used for this QA activity, third-party geocoding software or web-based search engines can be tested to supplement the Geocoder. An ideal system would have multi-faceted usability for registrars to use their time and resources efficiently. The Federal Financial Institutions Examination Council's (FFIEC) Geocoding System can be one of the systems. Figure 1 demonstrate the RI Dept of Heath address and how the geocoding and mapping results are displayed in the FFIEC Geocoding System.

Figure 1. Federal Financial Institutions Examination Council's (FFIEC) Geocoding System



Year: 2018 Address: 3 capitol hill, Providence RI

Matched Address	
Address	3 CAPITOL HL, PROVIDENCE, RI, 02908
MSA/MD Code	39300
State Code	44
County Code	007
Tract Code	0031.00
MSA/MD Name	PROVIDENCE-WARWICK, RI-MA
State Name	RHODE ISLAND
County Name	PROVIDENCE COUNTY

Census Demographic Data

□ User Select Tract

NPCR
NATIONAL PROGRAM OF CANCER REGISTRIES

Centers for Disease Control and Prevention
National Center for Chronic Disease Prevention and Health Promotion

CS 305996-A