*Geospatial Health 6(2), 2012, pp. 273-284*

# A multi-stage approach to maximizing geocoding success in a large population-based cohort study through automated and interactive processes

Jennifer S. Sonderman[1], Michael T. Mumma[1], Sarah S. Cohen[1], Elizabeth L. Cope[1], William J. Blot[1,2], Lisa B. Signorello[1,2]

[1]*International Epidemiology Institute, Rockville, MD 20850, USA;* [2]*Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt University, and Vanderbilt-Ingram Cancer Center, Nashville, TN, 37232, USA*

**Abstract.** To enable spatial analyses within a large, prospective cohort study of nearly 86,000 adults enrolled in a 12-state area in the southeastern United States of America from 2002-2009, a multi-stage geocoding protocol was developed to efficiently maximize the proportion of participants assigned an address level geographic coordinate. Addresses were parsed, cleaned and standardized before applying a combination of automated and interactive geocoding tools. Our full protocol increased the non-Post Office (PO) Box match rate from 74.5% to 97.6%. Overall, we geocoded 99.96% of participant addresses, with only 5.2% at the ZIP code centroid level (2.8% PO Box and 2.3% non-PO Box addresses). One key to reducing the need for interactive geocoding was the use of multiple base maps. Still, addresses in areas with population density <44 persons/km$^2$ were much more likely to require resource-intensive interactive geocoding than those in areas with >920 persons/km$^2$ (odds ratio (OR) = 5.24; 95% confidence interval (CI) = 4.23, 6.49), as were addresses collected from participants during in-person interviews compared with mailed questionnaires (OR = 1.83; 95% CI = 1.59, 2.11). This study demonstrates that population density and address ascertainment method can influence automated geocoding results and that high success in address level geocoding is achievable for large-scale studies covering wide geographical areas.

**Keywords:** epidemiologic methods, geographical information systems, prospective studies, residence characteristics, United States of America.

## Introduction

The Southern Community Cohort Study (SCCS) is a prospective cohort study of approximately 86,000 adults designed to investigate health disparities in understudied populations, primarily low-income African Americans and whites in the southeastern United States of America (USA) (Signorello et al., 2005, 2010). Enrollment largely took place at community health centers (CHCs) in a 12-state area in the southeastern USA. The personal information and biological specimens collected at enrollment and through follow-up surveys has created an invaluable resource for gaining an understanding of the reasons for racial health disparities. The examination of spatial and contextual determinants of disease is a growing area of research that has major potential to be utilized within the SCCS.

Corresponding author:
Jennifer S. Sonderman
International Epidemiology Institute
1455 Research Blvd., Suite 550
Rockville, MD 20850, USA
Tel.+1 301 279 4273; Fax +1 301 424 1053
Email: Jennifer@iei.us

The first step in such analyses is typically to geocode, or assign geographic coordinates (i.e. latitude and longitude) to, each study participant's address within a geographical information system (GIS) so that it may be related to spatially referenced environmental data or area (e.g. census tract) characteristics. There are many geocoding methods available, including address range interpolation, ZIP code centroid geocoding, parcel matching and exact measurement with a global positioning system (GPS) device (Rushton et al., 2006; Armstrong and Tiwari, 2008). The method chosen depends largely upon study size, resources and positional accuracy requirements.

Despite the ever-increasing number of spatial analyses in the literature, detailed reports of geocoding methods from epidemiological studies are still relatively sparse (McElroy et al., 2003; Rose et al., 2004; Gilboa et al., 2006; Zhan et al., 2006; Lovasi et al., 2007; Goldberg et al., 2008; Robinson et al., 2010; Vieira et al., 2010; Duncan et al., 2011). Methods reports are important as they enable peer evaluation of the strengths and limitations of the resulting, future spatial analyses and benefit new and established studies considering a similar pursuit. An abundance of lit-

erature does exist evaluating potential bias resulting from address interpolation, such as subject loss (Oliver et al., 2005; Zhan et al., 2006; Kravets and Hadden, 2007; Zimmerman et al., 2008; Vieira et al., 2010) and positional accuracy compared with a known location (Krieger et al., 2001; Bonner et al., 2003; Cayo and Talbot, 2003; Whitsel et al., 2004, 2006; Ward et al., 2005; Zhan et al., 2006; Kravets and Hadden, 2007; Schootman et al., 2007; Zandbergen, 2007; Zimmerman et al., 2007; Mazumdar et al., 2008; Vieira et al., 2010; Zimmerman and Li, 2010), that may help investigators choose GIS software or geocoding vendors. However, these analyses were primarily conducted on small populations over a limited geographical area and ignore other real-world issues such as the effect of data collection methods and participant characteristics that may limit the ability to geocode a large study.

The SCCS presents several unique challenges to geocoding, namely the substantial inclusion of rural participants, whose addresses generally do not geocode with the same success as urban participants (Vine et al., 1997; Gregorio et al., 1999; Boscoe et al., 2002; Cayo and Talbot, 2003; McElroy et al., 2003; Oliver et al., 2005; Rushton et al., 2006; Whitsel et al., 2006; Kravets and Hadden, 2007; Boscoe, 2008; Wey et al., 2009; Lin et al., 2010), and a large population across an entire region of the USA. The time- and resource-intensive methods often employed to accurately troubleshoot addresses not matching automatically, such as manual (i.e. "interactive") geocoding, re-contacting participants, or obtaining local maps (McElroy et al., 2003; Goldberg et al., 2008; Robinson et al., 2010), would for practical purposes need to be limited due to the large number of addresses across diverse geographical areas in the SCCS. Herein we describe the multi-stage geocoding process developed in anticipation of these difficulties after an extensive review of the existing literature and the relative contribution of various methods to our overall success. We also examine how final geocoding match rates differ by selected geographic and participant characteristics as well as how these factors influenced the need for costly, and often prohibitive (Boscoe et al., 2002; Goldberg et al., 2008; Lin et al., 2010), manual remediation for the benefit of future studies facing similar challenges.

## Materials and methods

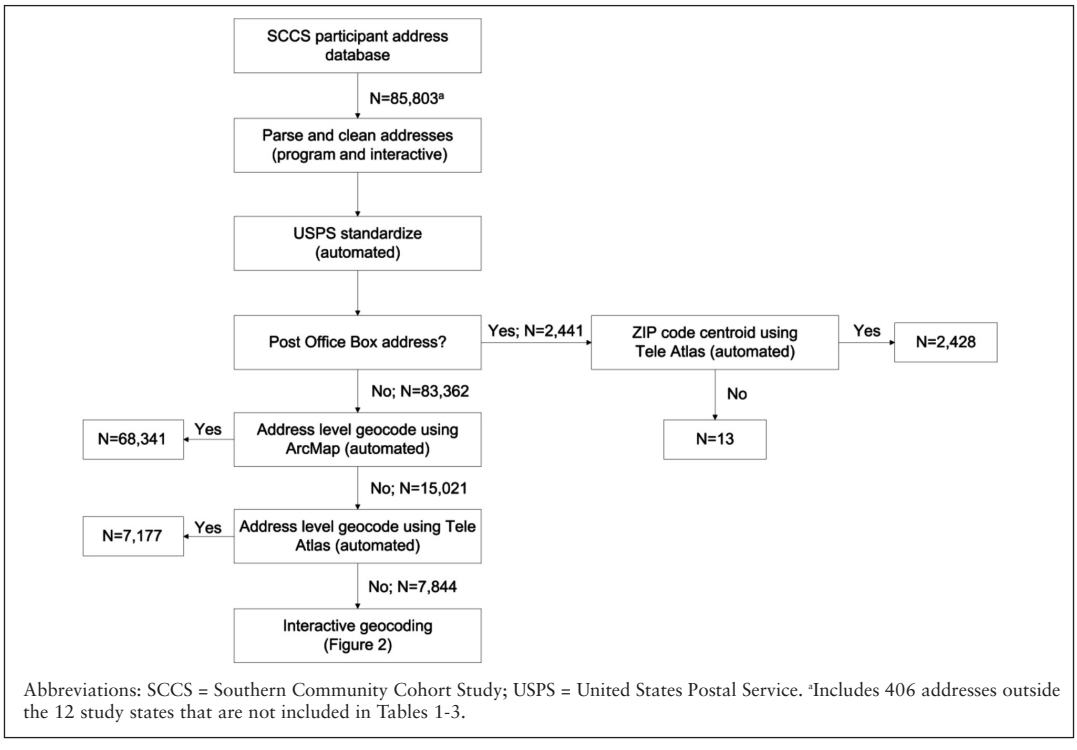
The SCCS was approved by the Institutional Review Boards at Vanderbilt University and Meharry Medical College. All study subjects provided written informed consent.

### *Participant address collection*

Details of SCCS participant enrollment have been described elsewhere (Signorello et al., 2005, 2010). Briefly, eligible participants were aged 40-79 years and were enrolled in 12 states: Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia and West Virginia. From 2002-2009, the SCCS enrolled 73,495 participants in-person at 71 CHCs (CHC participants) and an additional 12,308 via mail (general population participants). For CHC participants, trained interviewers administered a baseline computer-aided interview that collected their physical address and mailing address (if different) into four fields each: street address, city, state and ZIP code. General population participants wrote their address into the same four fields on a self-administered paper questionnaire. Baseline address collection was thus completed for the 85,803 study participants that form the basis of this geocoding work. In addition, the baseline interview elicited socio-demographic information including age, sex, race (White, Black/African American, Hispanic/Latino, Asian or Pacific Islander, American Indian or Alaska Native, other), household income (<US$ 15,000, US$ 15,000-24,999, US$ 25,000-49,999, US$ 50,000-99,999, ≥US$ 100,000), and education (<9 years, 9-11 years, high school or GED, vocational, technical, or business training, some college or junior college, college graduate, graduate school to a master's degree, graduate school beyond a master's degree).

### *Parsing and cleaning of addresses*

Misspellings or missing information in the address and a lack of address standardization may prohibit automated matching to an address base map (Vine et al., 1997; Gregorio et al., 1999; Boscoe et al., 2002; Rushton et al., 2006; Zhan et al., 2006; Kravets and Hadden, 2007; Boscoe, 2008; Zimmerman et al., 2008); therefore, we first standardized and evaluated the raw address data. Using version 9.2 of the SAS System for Windows (SAS Institute Inc., Cary, NC, USA), we programmatically parsed the self-reported street address into usable components: street number, pre-directional (e.g. "N"), street name, street type and post-directional (e.g. "SW") (Fig. 1). Though the interview included a separate question for mailing




Abbreviations: SCCS = Southern Community Cohort Study; USPS = United States Postal Service. [a]Includes 406 addresses outside the 12 study states that are not included in Tables 1-3.

Fig. 1. Geocoding procedure for Southern Community Cohort Study participant addresses.

address, interviewers occasionally entered "double" addresses into one field, such as "PO Box 789, 123 Apple Street" requiring removal of the Post Office (PO) Box information. We then fixed common entry errors (e.g. "RAOD") and regional nicknames (e.g. "JAX" for Jacksonville), and removed apartment numbers, "care of" information and other secondary identifiers such as the building complex (e.g. "MEADOW APARTMENTS") that are superfluous from a geocoding perspective. Additional errors arising during data collection included phonetic spellings of the street or city, missing information such as direction (e.g. "NW") or street type (e.g. "AVE"), and random data entry errors (e.g. a state abbreviation "AL", indicating Alabama, rather than the correct "LA", indicating Louisiana).

We then used Microsoft Access to manually continue address cleaning. Sorting alternately by each component helped identify additional common misspellings, unusual address components, and typographical errors.

*USPS standardization*

An in-house, United States Postal Service (USPS) certified address correcting software programme, QAS Batch version 4.57 (Experian QAS, Boston, MA, USA) further edited the SCCS addresses to facilitate automated address matching. If the software was able to match the address to its reference file containing all valid USPS delivery points, it corrected and standardized the address according to USPS standards, added the four digit extension to the ZIP code, and qualitatively reported the degree of uncertainty in the changes. Prior to finalizing the address standardization, we closely examined all changes applied to the input addresses to accept an uncertainty level where major changes (e.g. a change in more than two digits of the ZIP code or a change in state) were rare.

*Geocoding with ArcMap*

We then attempted to geocode the 83,362 non-PO Box addresses to the address level through address interpolation using in-house ESRI ArcMap 9.3.1 software (ESRI, Redlands, CA, USA). GIS software and commercial vendors can typically geocode 70-80% of addresses automatically with a reasonable degree of spatial accuracy using this method (Gregorio et al., 1999; McElroy et al., 2003; Oliver et al., 2005; Gilboa et al., 2006; Zimmerman et al., 2008; Lin et al., 2010; Vieira et al., 2010), in which coordinates are assigned to a street address by matching to a base map containing street line segments, an accompanying address number range and geographic coordinates. If the street name and number match within the specified "zone"

(e.g. city, state and ZIP code), a coordinate is proportionally assigned along the street segment.

To increase the probability of a match, both the 2006 StreetMap USA (ESRI, 2006) and the 2008 TIGER/Line® shapefiles (US Census Bureau, 2008) were utilized as base maps. Addresses were matched with a spelling sensitivity of 80, a minimum match score of 71 and a side offset of 11.1 m. Ties, or addresses with multiple, equally likely matches in the base map, were not considered matched. We closely examined addresses that matched with scores as low as 65 (some studies have used as low as 60 (Ward et al., 2005; Gilboa et al., 2006) or as high as 80 (McElroy et al., 2003; Robinson et al., 2010; Duncan et al., 2011), and a score of 71 was the lowest at which we felt confident that the coordinate was a reasonable representation of the input address. At this score and spelling sensitivity, the matched street name and type were often identical, but the street number could be different by as much as 10; in contrast, for addresses matching with a score of 70, the difference in street number was typically greater than 50.

Geocoded locations from both StreetMap USA and TIGER were closely compared for congruency. Coordinate locations farther than 3.2 km apart (N = 46) were manually reviewed and compared with the location obtained separately using Google Earth (Google, Mountain View, CA, USA), used previously in lieu of a true gold standard (Lovasi et al., 2007), for potential systematic errors giving preference to one base map over the other. As none were identified, and slightly more SCCS participant addresses matched in StreetMap USA, these coordinates were uniformly given preference over those from TIGER.

*Geocoding using online vendor*

Addresses failing to geocode in ArcMap (N = 15,021; or 18.0% of the 83,362 addresses attempted), as well as the 2,441 PO Box addresses (see Fig. 1), were submitted to the online-based EZ-Locate™ Client version 2.47 available from Tele Atlas at www.geocode.com (Tele Atlas, Lebanon, NH, USA) using their USA_Geo_002 base map (Tele Atlas, 2006). This vendor was chosen for the relatively low cost, high degree of detail regarding the precision of the match (i.e. address match, ZIP code centroid match, etc.), provision of both batch and interactive geocoding, and preservation of input address confidentiality. For non-PO Box addresses unable to be geocoded to the address level (N = 7,844), Tele Atlas provided the more precise of the delivery-weighted ZIP + 4, ZIP + 2, or 5-digit ZIP code centroid, which, as opposed to the geographical center of the ZIP code, was weighted towards the highest concentration of valid delivery points within the ZIP code (Tele Atlas, 2006); these coordinates were retained as a fallback if an address level geocode could not subsequently be obtained through interactive methods (Fig. 2). The best possible geocode for a PO Box, the delivery-weighted 5-digit ZIP code centroid, was also obtained from Tele Atlas for 2,428 out of 2,441 PO Box records (99.5%).

To assess the consistency of geocoding provided by Tele Atlas with that from ArcMap, we submitted a sample of 2,000 addresses that had successfully geocoded using both StreetMap USA and TIGER to obtain the analogous coordinate from Tele Atlas.

*Interactive geocoding*

After the largely automated procedures described above and in Fig. 1, 7,844 (9.4%) non-PO Box addresses had either not geocoded or had only a ZIP code centroid coordinate and were thus slated for individual evaluation and processing (Fig. 2).

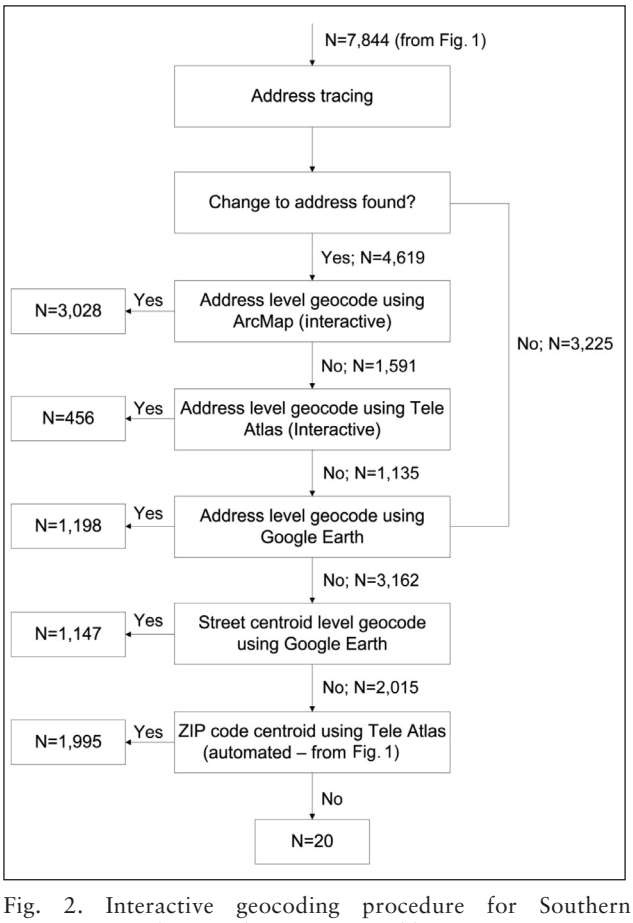



Fig. 2. Interactive geocoding procedure for Southern Community Cohort Study participant addresses.

Time for extensive online research was allowed for each address, and a human decision-making element was necessary. Due to their low cost and frequent updates, Internet resources have been used often in studies to manually geocode addresses (McElroy et al., 2003; Goldberg et al., 2008; Lin et al., 2010; Robinson et al., 2010; Duncan et al., 2011) and we utilized several in our protocol. The participant's address history was first traced on LexisNexis (www.lexisnexis.com) to help determine whether the address initially failed to geocode due to a typing error. Geocoding staff then manually searched for the address in Google Earth, which typically provided a suggestion if the exact input address was not found, followed by MapQuest (www.mapquest.com), Melissa Data (www.melissadata.com), or a generic web search, which might find the address, provide different suggestions or produce other identifying information. If a correction was found, for consistency we returned (although now fully manually) to our sequential process of ArcMap (using StreetMap USA), then Tele Atlas, then, if necessary, Google Earth (Fig. 2) to obtain coordinates.

At this stage we defined an additional geocode level, the street centroid (coordinates representing the midpoint of the street centerline), accepted only when the street could be found in Google Earth with no address number range and was less than 2 km in length, as this would generally provide a more precise coordinate than a ZIP code centroid. If the address still could not be geocoded to the address or street centroid level (N = 1,995), we attempted to improve the precision of the coordinate (e.g. to a ZIP + 4 centroid from a 5-digit ZIP code centroid) by correcting the ZIP code, if possible (N = 13). Otherwise, the ZIP code centroid provided by the initial batch submission to Tele Atlas was retained (N = 1,982).

*Additional quality control*

In addition to the procedures described above, we implemented extensive consistency checks throughout the automated and interactive processes to verify that the resulting coordinate was a valid representation of the input address. For example, all substantial edits to an address (e.g. a change in state or more than two digits of the ZIP code) were reviewed, and we ensured that identical input addresses (e.g. for spouses) received an identical geocode. We also re-reviewed a 5% random sample of those sent for interactive geocoding to ensure that errors were virtually absent.

*Statistical analysis*

Our systematic geocoding procedures allowed for a determination of the performance of each procedural stage at improving geocoding match rates. Geocoding with ArcMap was performed on raw, unparsed and unstandardized address data to determine a baseline (i.e. "unprocessed") level of geocoding success. To isolate the contribution of parsing/cleaning and USPS standardization, we also conducted this test on parsed and cleaned address data, before applying USPS standardization.

The distributions of geocoding match rates in the hierarchy of address level, street centroid level and ZIP code centroid level were calculated and evaluated in relation to state, population density, participant enrollment source (CHC or general population) and participant sex, race, annual household income and education. Unprocessed and final geocoding success (address or street centroid level, using the full procedures shown in Fig. 1 and Fig. 2) were calculated within quartiles of population density, in relation to these same factors. Population density was calculated as the Census 2000 population per square land kilometer for the ZIP Code Tabulation Area (ZCTA), available from the US Census Bureau 2000 Summary File 3 (US Census Bureau, 2000). Quartiles for ZCTA population density were based on the distribution of ZIP codes within the SCCS address database and rounded to the nearest whole number (Q1: <44 persons/km$^2$, Q2: 44-236 persons/km$^2$, Q3: 237-920 persons/km$^2$, Q4: >920 persons/km$^2$). *P* values for trend across quartiles were calculated using the Cochran-Armitage trend test.

To identify factors associated with the most resource-intensive stage in our protocol, a multivariate logistic regression model with robust standard errors to account for correlation within ZIP code was used to calculate odds ratios (OR) and 95% confidence intervals (CI) for the outcome defined as "requiring interactive geocoding" (i.e. necessitating procedures in Fig. 2); this model was restricted to the 83,362 non-PO Box addresses and included covariates for state, ZCTA population density quartiles, enrollment source (CHC *vs.* general population), participant sex (male *vs.* female), race (white, other *vs.* African American), household income categories (US$ 15,000-24,999, US$ 25,000-49,999, ≥US$ 50,000 *vs.* <US$ 15,000), and education categories (<9 years, 9-<12 years, ≥16 years *vs.* 12-<16 years). The GENMOD procedure in SAS/STAT version 9.2 of (SAS Institute Inc., Cary, NC, USA) was used for modeling. All P values are two-sided.

## Results

Coordinates provided by the three base maps (StreetMap USA, TIGER and Tele Atlas) in our batch processes (Fig. 1) were found to be highly consistent. The median distance among the 62,644 pairs geocoding in both StreetMap USA and TIGER was 26.9 m (range 0.0-17,129.7 m). Nearly all (99.3%) of the sample addresses we submitted to Tele Atlas for our consistency test that had been successful in both StreetMap USA and TIGER were geocoded to the address level by Tele Atlas; the median distance between coordinates from Tele Atlas and StreetMap USA was 37.2 m (range = 1.2-10,962.9 m) and the median distance between coordinates from Tele Atlas and TIGER was 26.7 m (range = 2.8-10,937.7 m).

At the completion of our full geocoding protocol, 406 (0.5%) enrollment addresses were located outside the 12 study states and were excluded from the tables presented here. A mapped distribution of the remaining non-PO Box addresses is provided in Fig. 3. Geocoding match rates at the address (our gold standard), street centroid and ZIP code centroid levels for the total 85,397 addresses are shown in Table 1. Geocoding success at the address level exceeded 95% for six states and was lowest in states with the highest relative proportion of PO Box addresses including Mississippi (89.9%), South Carolina (89.3%), Arkansas (87.9%) and West Virginia (78.1%). For most of the other factors in Table 1, crude address level success rates were also predictably correlated with the frequency of PO Box addresses. Overall, our methods resulted in 5.2% of geocoded coordinates falling to the ZIP code centroid level (2.8% were PO Box addresses and 2.3% were non-PO Box addresses), and only 0.04% of addresses completely failing to geocode.

A total of 74.5% of our non-PO Box addresses would have geocoded to their final, correct location using only ArcMap, without parsing, cleaning or standardization. This "unprocessed" success rate is presented in Table 2 along with final success rates (defined as address or street centroid level geocoding following the full procedure of Fig. 1 and Fig. 2), in relation to both population density and other demographic factors. The gain in geocoding success attributed to our full procedure compared with ArcMap batch processing of raw addresses was 23.1% overall, and was significantly more beneficial in areas of lowest population density (31.4%, P <0.001). Final success rates across all states and strata of participant demographics varied significantly by population density, with the exception of addresses in North Carolina, and virtually all addresses in the three highest density quartiles geocoded at either the address or street centroid level.

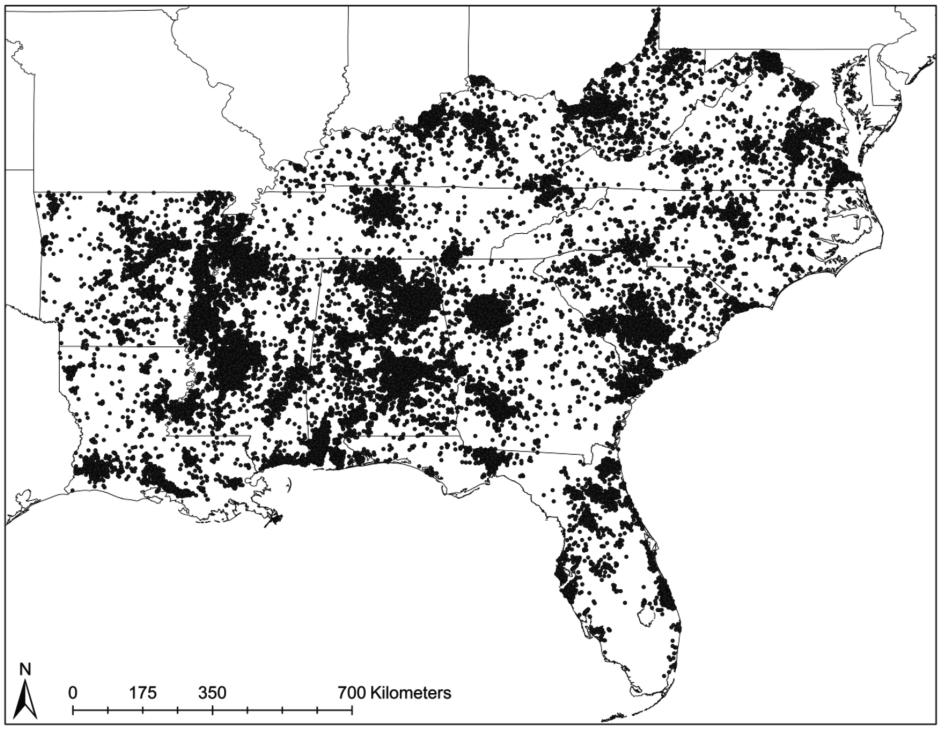



Fig. 3. Locations for 82,960 non-PO Box addresses across the 12 study enrollment states, Southern Community Cohort Study, USA 2002-2009.

Table 1. Geocoding match rates at each level for 85,397 non-PO Box and PO Box addresses by selected characteristics, Southern Community Cohort Study, USA 2002-2009.

| | Geocode level | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Street address | | Street centroid | | ZIP code centroid (non-PO Box address) | | ZIP code centroid (PO Box address) | | Did not geocode | |
| Characteristic | N | % | N | % | N | % | N | % | N | % |
| Total population | 79,811 | 93.46 | 1,148 | 1.34 | 1,981[a] | 2.32 | 2,424 | 2.84 | 33 | 0.04 |
| State | | | | | | | | | | |
| Alabama | 17,048 | 94.88 | 275 | 1.53 | 322 | 1.79 | 315 | 1.75 | 8 | 0.04 |
| Arkansas | 4,556 | 87.90 | 160 | 3.09 | 209 | 4.03 | 258 | 4.98 | 0 | 0.00 |
| Florida | 5,395 | 96.56 | 26 | 0.47 | 29 | 0.52 | 134 | 2.40 | 3 | 0.05 |
| Georgia | 10,153 | 96.29 | 100 | 0.95 | 138 | 1.31 | 148 | 1.40 | 5 | 0.05 |
| Kentucky | 6,870 | 97.09 | 48 | 0.68 | 66 | 0.93 | 92 | 1.30 | 0 | 0.00 |
| Louisiana | 3,273 | 97.58 | 29 | 0.86 | 29 | 0.86 | 23 | 0.69 | 0 | 0.00 |
| Mississippi | 12,046 | 89.90 | 321 | 2.40 | 416 | 3.10 | 608 | 4.54 | 9 | 0.07 |
| North Carolina | 2,041 | 97.66 | 2 | 0.10 | 8 | 0.38 | 39 | 1.87 | 0 | 0.00 |
| South Carolina | 4,573 | 89.28 | 78 | 1.52 | 130 | 2.54 | 339 | 6.62 | 2 | 0.04 |
| Tennessee | 7,052 | 98.44 | 33 | 0.46 | 30 | 0.42 | 48 | 0.67 | 1 | 0.01 |
| Virginia | 3,873 | 93.17 | 28 | 0.67 | 184 | 4.43 | 70 | 1.68 | 2 | 0.05 |
| West Virginia | 2,931 | 78.12 | 48 | 1.28 | 420 | 11.19 | 350 | 9.33 | 3 | 0.08 |
| Population density quartiles[b] | | | | | | | | | | |
| Q1 | 18,128 | 83.12 | 702 | 3.22 | 1,482 | 6.80 | 1,497 | 6.86 | 0 | 0.00 |
| Q2 | 20,328 | 96.52 | 176 | 0.84 | 268 | 1.27 | 289 | 1.37 | 0 | 0.00 |
| Q3 | 20,331 | 98.55 | 110 | 0.53 | 75 | 0.36 | 115 | 0.56 | 0 | 0.00 |
| Q4 | 20,437 | 98.74 | 122 | 0.59 | 97 | 0.47 | 42 | 0.20 | 0 | 0.00 |
| Enrollment source | | | | | | | | | | |
| CHC | 68,769 | 94.02 | 1,062 | 1.45 | 1,644 | 2.25 | 1,636 | 2.24 | 30 | 0.04 |
| General population | 11,042 | 90.09 | 86 | 0.70 | 337 | 2.75 | 788 | 6.43 | 3 | 0.02 |
| Sex | | | | | | | | | | |
| Male | 32,380 | 93.47 | 501 | 1.45 | 813 | 2.35 | 926 | 2.67 | 21 | 0.06 |
| Female | 47,431 | 93.45 | 647 | 1.27 | 1,168 | 2.30 | 1,498 | 2.95 | 12 | 0.02 |
| Race | | | | | | | | | | |
| African American | 52,872 | 94.45 | 735 | 1.31 | 888 | 1.59 | 1,464 | 2.62 | 21 | 0.04 |
| White | 23,110 | 91.46 | 359 | 1.42 | 963 | 3.81 | 826 | 3.27 | 9 | 0.04 |
| Other | 3,829 | 92.27 | 54 | 1.30 | 130 | 3.13 | 134 | 3.23 | 3 | 0.07 |
| Annual household income | | | | | | | | | | |
| <US$ 15,000 | 42,786 | 93.74 | 649 | 1.42 | 997 | 2.18 | 1,190 | 2.61 | 20 | 0.04 |
| US$ 15,000-24,999 | 16,335 | 93.64 | 241 | 1.38 | 412 | 2.36 | 450 | 2.58 | 7 | 0.04 |
| US$ 25,000-49,999 | 10,892 | 93.21 | 145 | 1.24 | 282 | 2.41 | 364 | 3.11 | 3 | 0.03 |
| ≥US$ 50,000 | 7,366 | 92.51 | 66 | 0.83 | 212 | 2.66 | 316 | 3.97 | 2 | 0.03 |
| Education (years) | | | | | | | | | | |
| <9 | 6,012 | 90.50 | 131 | 1.97 | 238 | 3.58 | 256 | 3.85 | 6 | 0.09 |
| 9-11 | 16,217 | 93.55 | 271 | 1.56 | 403 | 2.32 | 437 | 2.52 | 7 | 0.04 |
| 12-15 | 45,665 | 93.90 | 615 | 1.26 | 1,079 | 2.22 | 1,261 | 2.59 | 14 | 0.03 |
| ≥16 | 10,187 | 93.53 | 94 | 0.86 | 202 | 1.85 | 404 | 3.71 | 5 | 0.05 |

Abbreviations: CHC = Community Health Center; PO = Post Office. [a]Of these, all were delivery-weighted. 7 were ZIP + 4 centroids, 595 (30.0%) were ZIP + 2 centroids, and 1,379 (69.6%) were 5-digit ZIP code centroids. [b]Q1: <44 persons/km$^2$, Q2: 44-236 persons/km$^2$, Q3: 237-920 persons/km$^2$, Q4: >920 persons/km$^2$.

Fig. 4 shows the sequential gains in geocoding success at the address or street centroid level at each procedural stage and the general disparity in success across quartiles of population density. Among these non-PO Box addresses, programmatic and interactive parsing and cleaning efforts resulted in a small (2.0%) gain over ArcMap geocoding of unprocessed addresses, and use of USPS address standardization software further increased success by 3.7%. After ArcMap geocoding, use of Tele Atlas was preferentially beneficial for the lowest population density group (17.1% gain) and, importantly, resulted in a substantial shrinking in the success disparity across these geographic groups. Of the 7,844 addresses sent for interactive geocoding, 5,829 (74.3%) were improved to an address or street centroid level geocode. This resource-intensive process was beneficial for all groups, resulting in an important, final

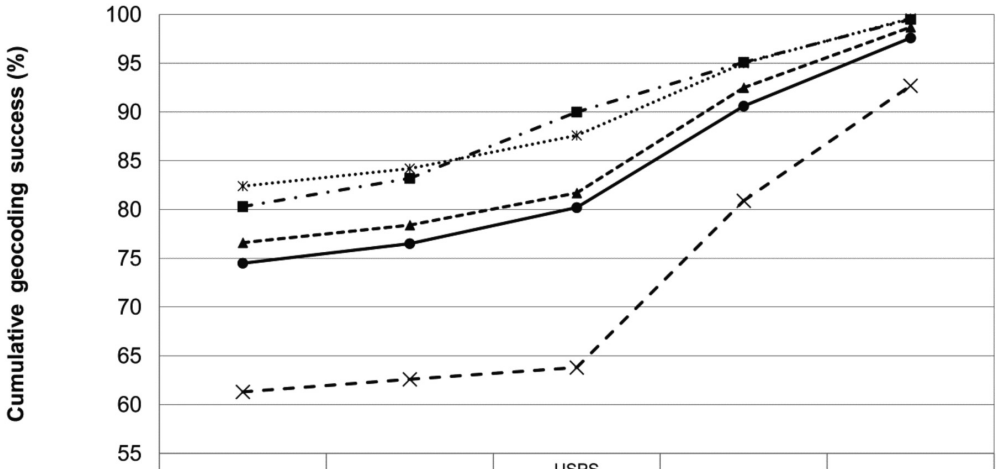
Table 2. Unprocessed[a] and final[b] geocoding match rates for 82,960 Non-PO Box addresses by selected characteristics, Southern Community Cohort Study, USA 2002-2009.

| | | Population density quartiles[c,d] | | | |
|---|---|---|---|---|---|
| | Overall | Q1 | Q2 | Q3 | Q4 |
| Characteristic | Unprocessed (%) / Final (%) | | | | |
| Total Population | 74.5 / 97.6 | 61.3 / 92.7 | 76.6 / 98.7 | 82.4 / 99.6 | 80.3 / 99.5 |
| State | | | | | |
| Alabama | 74.2 / 98.1 | 58.1 / 93.3 | 77.0 / 99.1 | 78.8 / 99.6 | 81.6 / 99.6 |
| Arkansas | 70.7 / 95.8 | 60.5 / 92.7 | 77.2 / 99.0 | 88.7 / 99.7 | 91.1 / 100.0 |
| Florida | 79.1 / 99.4 | 64.8 / 97.9 | 77.9 / 99.7 | 82.2 / 99.9 | 85.3 / 99.5 |
| Georgia | 75.7 / 98.7 | 59.8 / 90.5 | 82.2 / 99.6 | 81.4 / 99.7 | 74.4 / 99.3 |
| Kentucky | 82.9 / 99.1 | 61.9 / 94.3 | 78.6 / 98.9 | 84.7 / 99.7 | 85.9 / 99.5 |
| Louisiana | 81.5 / 99.1 | 70.4 / 97.7 | 78.7 / 99.5 | 91.8 / 99.7 | 94.4 / 100.0 |
| Mississippi | 70.5 / 96.7 | 66.6 / 95.4 | 77.9 / 99.0 | 84.5 / 99.6 | 85.3 / 100.0 |
| North Carolina | 85.8 / 99.6 | 76.9 / 98.6 | 86.2 / 100.0 | 87.8 / 99.7 | 88.6 / 99.7 |
| South Carolina | 67.0 / 97.3 | 52.9 / 94.5 | 73.3 / 99.3 | 78.5 / 99.2 | 83.1 / 98.5 |
| Tennessee | 80.1 / 99.6 | 74.1 / 97.8 | 81.7 / 99.5 | 83.0 / 99.8 | 78.5 / 99.6 |
| Virginia | 75.1 / 95.4 | 49.7 / 75.0 | 87.7 / 99.7 | 87.4 / 100.0 | 77.4 / 99.9 |
| West Virginia | 53.1 / 87.6 | 25.2 / 58.2 | 56.0 / 91.5 | 80.8 / 98.6 | 80.0 / 100.0 |
| Enrollment source | | | | | |
| CHC | 73.3 / 97.7 | 60.2 / 93.0 | 75.0 / 98.8 | 81.1 / 99.6 | 79.2 / 99.5 |
| General population | 82.2 / 97.1 | 67.5 / 91.1 | 83.3 / 98.2 | 90.9 / 100.0 | 92.1 / 100.0 |
| Sex | | | | | |
| Male | 73.1 / 97.5 | 59.4 / 92.2 | 75.3 / 98.5 | 80.8 / 99.5 | 77.2 / 99.3 |
| Female | 75.4 / 97.6 | 62.3 / 93.0 | 77.3 / 98.8 | 83.6 / 99.7 | 83.2 / 99.8 |
| Race | | | | | |
| African American | 75.3 / 98.3 | 62.7 / 94.8 | 77.9 / 99.2 | 82.2 / 99.6 | 78.8 / 99.5 |
| White | 72.9 / 96.0 | 59.6 / 89.6 | 75.0 / 97.9 | 83.2 / 99.6 | 85.5 / 99.5 |
| Other | 72.9 / 96.7 | 55.0 / 88.3 | 73.9 / 98.6 | 82.7 / 99.7 | 84.3 / 99.5 |
| Annual household income | | | | | |
| <US$ 15,000 | 73.8 / 97.7 | 60.5 / 93.0 | 76.3 / 98.7 | 81.0 / 99.6 | 78.9 / 99.5 |
| US$ 15,000-24,999 | 74.6 / 97.6 | 61.6 / 92.9 | 75.8 / 98.5 | 82.5 / 99.5 | 81.3 / 99.6 |
| US$ 25,000-49,999 | 76.1 / 97.5 | 62.5 / 92.7 | 78.3 / 98.7 | 85.5 / 99.9 | 83.1 / 99.4 |
| ≥US$ 50,000 | 77.8 / 97.2 | 63.7 / 90.4 | 77.7 / 98.9 | 87.2 / 99.8 | 88.2 / 99.5 |
| Education (years) | | | | | |
| <9 | 68.6 / 96.2 | 57.4 / 91.7 | 71.7 / 98.3 | 79.8 / 99.1 | 77.5 / 99.4 |
| 9-11 | 72.3 / 97.6 | 60.0 / 93.0 | 75.2 / 98.6 | 80.6 / 99.6 | 76.8 / 99.5 |
| 12-15 | 75.5 / 97.7 | 62.1 / 92.8 | 77.4 / 98.7 | 82.7 / 99.7 | 81.4 / 99.5 |
| ≥16 | 78.1 / 98.1 | 64.4 / 92.8 | 78.5 / 99.1 | 85.6 / 99.9 | 84.0 / 99.7 |

Abbreviations: CHC = Community Health Center; PO = Post Office. [a]Unparsed, uncleaned, and unstandardized addresses that geocoded correctly at the address level using only ArcMap. [b]Parsed, cleaned, and standardized addresses that geocoded at the address or street centroid level following the full procedures shown in Fig. 1 and Fig. 2. [c]Q1: <44 persons/km$^2$, Q2: 44-236 persons/km$^2$, Q3: 237-920 persons/km$^2$, Q4: >920 persons/km$^2$. [d]All final, two-sided Cochran-Armitage P values for trend across population density quartiles were <0.001 with the exception of Florida ($P_{trend}$ = 0.02) and North Carolina ($P_{trend}$ = 0.25).

7.0% gain to bring the address and street centroid level geocoding of non-PO Box addresses to 97.6%. Nearly all remaining non-PO Box addresses geocoded to either the ZIP + 4 (N = 7; 0.01%), ZIP + 2 (N = 595; 0.7%), or 5-digit ZIP code centroid (N = 1,379; 1.7%). Including the 2,424 ZIP code centroids obtained for PO Box addresses within the 12 study states, our methods achieved a geographic coordinate for 85,364 (99.96%) of the 85,397 participant addresses.

Table 3 presents the results of a multivariate logistic regression model evaluating the independent contribution of each demographic characteristic in Table 1 to whether the (non-PO Box) address required interactive geocoding. Several of these factors were significant determinants of this resource-intensive process. Among the 12 states, addresses in West Virginia were most likely to require interactive geocoding. Compared with addresses in the highest quartile of population density, those in the lowest quartile were more than five times as likely (OR = 5.24; 95% CI = 4.23, 6.49) to require interactive geocoding. Other significant predictors of interactive geocoding were enrollment at a CHC (*vs.* general population, OR =




| | Unprocessed addresses (ArcMap) | Addresses parsed and cleaned (ArcMap) | USPS standardization of addresses (ArcMap) | Tele Atlas | Interactive geocoding |
|---|---|---|---|---|---|
| Overall SCCS | 74.5 | 76.5 | 80.2 | 90.6 | 97.6 |
| Density Q1 | 61.3 | 62.6 | 63.8 | 80.9 | 92.7 |
| Density Q2 | 76.6 | 78.4 | 81.7 | 92.5 | 98.7 |
| Density Q3 | 82.4 | 84.2 | 87.6 | 95.0 | 99.6 |
| Density Q4 | 80.3 | 83.2 | 90.0 | 95.1 | 99.5 |

Abbreviations: SCCS = Southern Community Cohort Study; USPS = United States Postal Service. [a]Q1: <44 persons/km$^2$, Q2: 44-236 persons/km$^2$, Q3: 237-920 persons/km$^2$, Q4: >920 persons/km$^2$.

Fig. 4. Cumulative geocoding success (defined as address- or street centroid-level coordinates) at each procedural stage for 82,960 non-PO Box addresses within the 12 Southern Community Cohort Study states, overall and by population density quartile[a].

1.83; 95% CI = 1.59, 2.11), being male (*vs.* female, OR = 1.35; 95% CI = 1.28, 1.42), and having less than 9 years of education (*vs.* 12-15 years, OR = 1.37; 95% CI: 1.25, 1.51).

## Discussion

Geocoding SCCS participant residences presented several challenges, chief among them the sheer volume of addresses. Our first goal was to clean and standardize the nearly 86,000 addresses to maximize success using in-house and largely automated processes and minimize the number of addresses requiring use of an outside vendor or interactive geocoding. Drawing upon prior reports in smaller, more geographically concentrated populations, we undertook intensive and multi-step address processing and used two reference base maps in pursuit of this goal. Studies attaining the highest level of geocoding success have often matched against more than one base map prior to interactive geocoding (McElroy et al., 2003; Zhan et al., 2006; Lovasi et al., 2007). Still, 18% of our non-PO Box addresses required outsourcing to an online vendor. This step, as it turns out, was very beneficial to our overall match rates, particularly for addresses in the least populated areas, and demonstrates the utility of involving a third base map when geocoding. Because of this relatively simple and inexpensive step, only 9.4% of our addresses required interactive geocoding, half the proportion we originally expected based on other reports (Gregorio et al., 1999; McElroy et al., 2003; Oliver et al., 2005; Gilboa et al., 2006; Lin et al., 2010).

Although we were able to limit the need for interactive geocoding to 9.4% of our cohort, this still amounted to thousands of addresses that required careful and time-consuming hand processing using a variety of methods. We found that a number of factors influenced the need for this process, some of which are amenable to managing in future studies at the data collection stage. For example, although it would seem that the in-person interview at the CHC would result in more accurate data collection, CHC participant addresses were 83% more likely to require interactive geocoding than the general population participants who completed a paper questionnaire, after adjustment for differences in socioeconomic characteristics between these populations such as education and household income. Relative to addresses from general population participants making less than US$ 25,000 per year, an income level more comparable with the CHC participants, addresses from CHC participants were still more likely to require interactive geocoding (OR = 1.66; 95% CI = 1.42, 1.94). Further stratifica-

Table 3. Multivariate logistic regression-derived odds ratios[a] for requiring interactive geocoding among 82,960 non-PO Box addresses, Southern Community Cohort Study, USA 2002-2009[a].

| Characteristic | OR | 95% CI |
|---|---|---|
| State at enrollment | | |
| Alabama | 1.00 | Referent |
| Arkansas | 0.91 | 0.61, 1.35 |
| Florida | 0.84 | 0.67, 1.07 |
| Georgia | 1.01 | 0.81, 1.25 |
| Kentucky | 0.83 | 0.69, 1.01 |
| Louisiana | 0.60 | 0.42, 0.85 |
| Mississippi | 0.73 | 0.52, 1.03 |
| North Carolina | 0.58 | 0.42, 0.81 |
| South Carolina | 1.06 | 0.85, 1.32 |
| Tennessee | 0.82 | 0.65, 1.03 |
| Virginia | 1.22 | 0.78, 1.89 |
| West Virginia | 2.61 | 1.94, 3.51 |
| Population density quartiles[b] | | |
| Q1 | 5.24 | 4.23, 6.49 |
| Q2 | 1.67 | 1.37, 2.04 |
| Q3 | 1.09 | 0.90, 1.32 |
| Q4 | 1.00 | Referent |
| Enrollment source | | |
| CHC | 1.83 | 1.59, 2.11 |
| General population | 1.00 | Referent |
| Gender | | |
| Male | 1.35 | 1.28, 1.42 |
| Female | 1.00 | Referent |
| Race | | |
| African American | 1.00 | Referent |
| White | 1.05 | 0.94, 1.18 |
| Other | 1.06 | 0.90, 1.26 |
| Annual household income | | |
| <US$ 15,000 | 1.00 | Referent |
| US$ 15,000-24,999 | 0.98 | 0.92, 1.05 |
| US$ 25,000-49,999 | 0.91 | 0.82, 1.00 |
| ≥US$ 50,000 | 0.94 | 0.82, 1.08 |
| Education (years) | | |
| <9 | 1.37 | 1.25, 1.51 |
| 9-11 | 1.10 | 1.03, 1.18 |
| 12-15 | 1.00 | Referent |
| ≥16 | 0.82 | 0.74, 0.90 |

Abbreviations: CHC = Community Health Center; CI = confidence interval; OR = odds ratio; PO = Post Office. [a]Adjusted for all characteristics listed in the Table. [b]Q1: <44 persons/km$^2$, Q2: 44-236 persons/km$^2$, Q3: 237-920 persons/km$^2$, Q4: >920 persons/km$^2$.

tion by population density and education (not shown) suggested that this finding was not the result of confounding by these factors. Thus, socioeconomic status does not appear to be the driving factor for this observed difference in automated geocoding success. A possible explanation is the method by which the participant orally stated their address, with the interviewer typing from what was heard, and an increased possibility of gross misspellings compared with data entry from the participant's own spelling from a paper questionnaire. This is supported by the address cleaning step of our interactive process; of the addresses routed to interactive geocoding, a "fix" for the address was found for 62% of the CHC *vs.* 31% of the general population participants. It may also be advisable to pay close attention to address collection for study populations with less than a high school education to lessen the chance of automated geocoding failure. Other general lessons learned during this effort include first, that the collection of nearest cross street information would assist with locating the address during interactive geocoding or would enable geocoding the intersection as an alternative to the street or ZIP code centroids, and second, that data entry forms should be designed to capture address components as separate input fields (e.g. street number, street name, directionals), reducing the time and effort needed to retroactively parse the addresses.

An issue over which study investigators have little to no control is the relatively poor coverage of base maps for rural areas. Low population density was the strongest predictor of the need for interactive geocoding, the intensity of the interactive geocoding, and ultimately the acceptance of ZIP code centroid coordinates. Rural route addresses in our study population were rare (1%), even within areas with <44 persons/km$^2$ (3%). More common were PO Boxes, which constituted 7% of the addresses in the lowest density quartile, compared with only 0.2% of those in the highest. To allow the inclusion of all SCCS participants in future spatial analyses, ZIP code centroid coordinates were obtained for nearly 100% of the PO Box addresses and should be of sufficient quality for analyses with large-scale exposures, such as average exposure to ultraviolet radiation or ambient temperature. For analyses requiring a higher level of detail, we may consider further address correction methods, such as re-contacting participants (Vine et al., 1997; McElroy et al., 2003; Goldberg et al., 2008) or contacting local postmasters for assistance with the address (Boice et al., 2003; Hurley et al., 2003).

Address interpolation, which was used to geocode the vast majority of our addresses, does involve the potential for positional error and is thus one limitation of our overall approach. Given the large geographical extent of the SCCS, an evaluation of the positional accuracy of our ultimate coordinates was not feasible; however, like match rates in general, positional error is typically larger in rural, less densely populated areas. This may stem from greater error within the reference base maps and longer street segments, both of which may produce interpolated coordinates farther from the actual residence (Bonner et al., 2003; Cayo

and Talbot, 2003; Ward et al., 2005; Whitsel et al., 2006; Lovasi et al., 2007; Zimmerman and Li, 2010) and potentially introduce differential exposure misclassification by population density in analyses requiring extremely accurate coordinates for exposure assignment (Bonner et al., 2003; Cayo and Talbot, 2003; Ward et al., 2005; Zandbergen, 2007; Mazumdar et al., 2008). The main alternative, however, is to geocode using local property parcel maps (Dearwent et al., 2001; Cayo and Talbot, 2003; Rushton et al., 2006; Whitsel et al., 2006; Zandbergen, 2008), which is not practical for most studies as large as the SCCS. For the common "neighbourhood"-level analyses requiring linkage to an enumeration area (e.g. US census tract), the effect of positional error is likely to be reduced as the area size is typically inversely related to population density.

In summary, we demonstrated that extremely high success in address level geocoding is attainable for very large scale studies using a combination of readily accessible geocoding tools in both an automated and interactive fashion. Overall, we geocoded 99.96% of SCCS participant addresses, with only 5.2% at the ZIP code centroid level. The multi-stage protocol we developed also substantially reduced, though did not completely eliminate, differences in geocoding success by population density and some demographic characteristics. This report represents one of the most detailed descriptions of the utility of various geocoding methods in a large population-based study to date, and may benefit large and small studies alike in maximizing geocoding success for spatial analyses.

## Acknowledgements

This work was supported by an American Reinvestment and Recovery Act grant from the National Cancer Institute at the National Institutes of Health (3R01 CA092447-08S1). The authors would like to thank Mr. Brendan Williams for data cleaning and ArcMap geocoding assistance, Mr. Mark Steinwandel for the preparation of address datasets, and Mr. Benjamin Sonderman, Ms. Victoria Rainbolt and Ms. Cristina Ortiz for interactive geocoding. The authors declare that we have no conflicts of interest to report.

### References

Armstrong MP, Tiwari C, 2008. Geocoding methods, Materials, and first steps toward a geocoding error budget. In: Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman D (Eds). Geocoding health data: the use of geocographic codes in cancer prevention and control, research, and practice. CRC Press, Boca Raton, FL, USA, pp. 11-35.

Boice JD, Bigbee WL, Mumma MT, Blot WJ, 2003. Cancer incidence in municipalities near two former nuclear materials processing facilities in Pennsylvania. Health Phys 85, 678-690.

Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL, 2003. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology 14, 408-412.

Boscoe FP, 2008. The science and art of geocoding: tips for improving match rates and handling unmatched cases in analysis. In: Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman D (Eds). Geocoding health data: the use of geocographic codes in cancer prevention and control, research, and practice. CRC Press, Boca Raton, FL, USA, pp. 95-109.

Boscoe FP, Kielb CL, Schymura MJ, Bolani TM, 2002. Assessing and improving census tract completeness. J Registry Manag 29, 117-120.

Cayo MR, Talbot TO, 2003. Positional error in automated geocoding of residential addresses. Int J Health Geogr 2, 10.

Dearwent SM, Jacobs RR, Halbert JB, 2001. Locational uncertainty in georeferencing public health datasets. J Expo Anal Environ Epidemiol 11, 329-334.

Duncan DT, Castro MC, Blossom JC, Bennett GG, Gortmaker SL, 2011. Evaluation of the positional difference between two common geocoding methods. Geospat Health 5, 265-273.

ESRI, 2006. Streetmap USA in ESRI data and maps (machine-readable data files).

Gilboa SM, Mendola P, Olshan AF, Harness C, Loomis D, Langlois PH, Savitz DA, Herring AH, 2006. Comparison of residential geocoding methods in population-based study of air quality and birth defects. Environ Res 101, 256-262.

Goldberg DW, Wilson JP, Knoblock CA, Ritz B, Cockburn MG, 2008. An effective and efficient approach for manually improving geocoded data. Int J Health Geogr 7, 60.

Gregorio DI, Cromley E, Mrozinski R, Walsh SJ, 1999. Subject loss in spatial analysis of breast cancer. Health Place 5, 173-177.

Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P, 2003. Post office box addresses: a challenge for geographic information system-based studies. Epidemiology 14, 386-391.

Kravets N, Hadden WC, 2007. The accuracy of address coding and the effects of coding errors. Health Place 13, 293-298.

Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW, 2001. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 91, 1114-1116.

Lin G, Gray J, Qu M, 2010. Improving geocoding outcomes for the Nebraska Cancer Registry: learning from proven practices. J Registry Manag 37, 49-56.

Lovasi G, Weiss J, Hoskins R, Whitsel E, Rice K, Erickson C, Psaty B, 2007. Comparing a single-stage geocoding method to a multi-stage geocoding method: how much and where do they

disagree? Int J Health Geogr 6, 12.

Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ, 2008. Geocoding accuracy and the recovery of relationships between environmental exposures and health. Int J Health Geogr 7, 13.

McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA, 2003. Geocoding addresses from a large population-based study: lessons learned. Epidemiology 14, 399-407.

Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW, 2005. Geographic bias related to geocoding in epidemiologic studies. Int J Health Geogr 4, 29.

Robinson JC, Wyatt SB, Hickson D, Gwinn D, Faruque F, Sims M, Sarpong D, Taylor HA, 2010. Methods for retrospective geocoding in population studies: the Jackson Heart Study. J Urban Health 87, 136-150.

Rose KM, Wood JL, Knowles S, Pollitt RA, Whitsel EA, Diez Roux AV, Yoon D, Heiss G, 2004. Historical measures of social context in life course studies: retrospective linkage of addresses to decennial censuses. Int J Health Geogr 3, 27.

Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL, 2006. Geocoding in cancer research: a review. Am J Prev Med 30, S16-S24.

Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, Higgs G, 2007. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. Ann Epidemiol 17, 464-470.

Signorello LB, Hargreaves MK, Blot WJ, 2010. The Southern Community Cohort Study: investigating health disparities. J Health Care Poor Underserved 21, 26-37.

Signorello LB, Hargreaves MK, Steinwandel MD, Zheng W, Cai Q, Schlundt DG, Buchowski MS, Arnold CW, McLaughlin JK, Blot WJ, 2005. Southern community cohort study: establishing a cohort to investigate health disparities. J Natl Med Assoc 97, 972-979.

Tele Atlas, 2006. USA_Geo_002 (Service Description Document).

US Census Bureau, 2000. Census 2000 Summary File 3 (SF3) - Sample Data (machine-readable data files).

US Census Bureau, 2008. 2008 TIGER/Line® Shapefiles (machine-readable data files).

Vieira VM, Howard GJ, Gallagher LG, Fletcher T, 2010. Geocoding rural addresses in a community contaminated by PFOA: a comparison of methods. Environ Health 9, 18.

Vine MF, Degnan D, Hanchette C, 1997. Geographic information systems: their use in environmental epidemiologic research. Environ Health Perspect 105, 598-605.

Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, Mix W, Colt JS, Hartge P, 2005. Positional accuracy of two methods of geocoding. Epidemiology 16, 542-547.

Wey CL, Griesse J, Kightlinger L, Wimberly MC, 2009. Geographic variability in geocoding success for West Nile virus cases in South Dakota. Health Place 15, 1108-1114.

Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, Heiss G, 2006. Accuracy of commercial geocoding: assessment and implications. Epidemiol Perspect Innov 3, 8.

Whitsel EA, Rose KM, Wood JL, Henley AC, Liao D, Heiss G, 2004. Accuracy and repeatability of commercial geocoding. Am J Epidemiol 160, 1023-1029.

Zandbergen PA, 2007. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. BMC Public Health 7, 37.

Zandbergen PA, 2008. A comparison of address point, parcel and street geocoding techniques. Comput Environ Urban Syst 32, 214-232.

Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH, 2006. Match rate and positional accuracy of two geocoding methods for epidemiologic research. Ann Epidemiol 16, 842-849.

Zimmerman D, Li J, 2010. The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. Int J Health Geogr 9, 10.

Zimmerman DL, Fang X, Mazumdar S, Rushton G, 2007. Modeling the probability distribution of positional errors incurred by residential address geocoding. Int J Health Geogr 6, 1.

Zimmerman DL, Fang X, Mazumdar S, 2008. Spatial clustering of the failure to geocode and its implications for the detection of disease clustering. Stat Med 27, 4254-4266.