SSM - Population Health 15 (2021) 100887

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

Matching participant address with public records database in a US national longitudinal cohort study

Marquita S. Brooks [a,*], Aleena Bennett [a], Gina S. Lovasi [b], Philip M. Hurvitz [c], Natalie Colabianchi [d], Virginia J. Howard [e], Jennifer Manly [f], Suzanne E. Judd [a]

[a] Department of Biostatistics, University of Alabama at Birmingham (UAB), Birmingham, AL, USA
[b] Urban Health Collaborative, Dornsife School of Public Health, Drexel University, Philadelphia, PA, USA
[c] Center for Studies in Demography and Ecology, University of Washington, Seattle, WA, USA
[d] Environment and Policy Lab, University of Michigan School of Kinesiology, Ann Arbor, MI, USA
[e] Department of Epidemiology, UAB, Birmingham, AL, USA
[f] Neurology at Gertrude H. Sergievsky Center and the Taub Institute for Research in Aging and Alzheimer's Disease, Columbia University, New York, NY, USA

ARTICLE INFO

Keywords:
Residential address
Residential history
Environmental exposure

ABSTRACT

Background: Epidemiological studies utilize residential histories to assess environmental exposure risk. The validity from using commercially-sourced residential histories within national longitudinal studies remains unclear. Our study assessed predictors of non-agreement between baseline addresses from the commercially-sourced LexisNexis database and participants in the national longitudinal study, REasons for Geographic and Racial Differences in Stroke (REGARDS). Additionally, we assessed differences in stroke risk by neighborhood socioeconomic score (nSES) based on participant reported address compared to nSES from LexisNexis/REGARDS matched baseline address.

Methods: From January 2003–October 2007, REGARDS enrolled 30,239 black and white adults aged 45 and older within the continental United States and collected their baseline address. ArcGIS Desktop 10.5.1 with ESRI 2016 Business Analyst Data was used to geocode baseline addresses from LexisNexis and REGARDS. Logistic regression was used to estimate the likelihood that LexisNexis address matched REGARDS baseline address for each participant. Survival analysis was used to estimate association between nSES and incident stroke.

Results: Approximately 91% of REGARDS participants had a LexisNexis address. Of these geocoded addresses, 93% of REGARDS baseline addresses matched LexisNexis addresses. Odds of agreement between LexisNexis and REGARDS was higher for older-aged participants (OR = 1.02 per year, 95% CI: 1.01, 1.02), blacks compared to whites (OR = 1.16, 95% CI: 1.05, 1.29), females compared to males (OR = 1.15, 95% CI: 1.04, 1.26), participants with an income of $34k-74k compared to an income less than $20k (OR = 1.62, 95% CI: 1.39, 1.89). Odds of agreement were lower for residents in Midwest compared to residents in the south (OR = 0.82, 95% CI: 0.73, 0.94). No significant differences in nSES-stroke associations were observed between REGARDS only and LexisNexis/REGARDS matched addresses; however, differences in interactions were observed.

Conclusion: Agreement between LexisNexis and REGARDS addresses varied by sociodemographic groups, potentially introducing bias in studies reliant on LexisNexis alone for residential address data.

1. Introduction

Epidemiological studies interested in retrospective assessment of environmental exposures often use historical residential addresses from either self-report or other data sources to establish prior location of residence for participants (Hertz et al., 2017; Ling et al., 2019; Wheeler & Wang, 2015). LexisNexis is a commercially available data provider that several studies have utilized to obtain residential histories (Hertz et al., 2017; Jacquez et al., 2011; Ling et al., 2019; Medgyesi et al., 2021).

Two prior studies using LexisNexis compared commercially sourced addresses with those collected directly from participants. A bladder cancer case-control study that included 11 counties of southeastern Michigan, found that 71.5% of LexisNexis residential addresses matched

* Corresponding author.1665 University Blvd., RPHB 514F, Birmingham, AL, 35233, USA.
E-mail address: msgray92@uab.edu (M.S. Brooks).

https://doi.org/10.1016/j.ssmph.2021.100887
Received 19 February 2021; Received in revised form 9 July 2021; Accepted 2 August 2021
Available online 3 August 2021

case-control study residential addresses (Jacquez et al., 2011). The California Teachers Study (CTS)—a prospective cohort study of female professional employees in California public schools—found that 85% of addresses matched between LexisNexis and CTS (Hertz et al., 2017). The CTS examined demographic predictors of an address match and found lower match rates among black women and women younger than 40 years of age (Hertz et al., 2017). A third study on early childhood residential mobility utilized LexisNexis geocoded addresses to assess location-based pesticide exposure and found that LexisNexis address histories may be useable for assessing environmental exposures, but this study did not compare quality of commercially available residential histories to an alternative source of address data (Ling et al., 2019). Although these studies each utilized LexisNexis addresses, none were national or a longitudinal with tracking of residential mobility. It is not clear to what degree bias may exist in addresses obtained by LexisNexis arising from systematically greater error in strata such as those with less residential stability.

An integrated review of environmental and genetic risk factors that impact risk of multiple sclerosis (MS) revealed that sun exposure in childhood, adolescence, adulthood and over the life course is associated with higher risk of MS (Abbasi et al., 2017; Abdollahpour et al., 2018; Bäärnhielm et al., 2012; Bjørnevik et al., 2014; Dalmay et al., 2010; Islam et al., 2007; Kampman et al., 2007; Lucas et al., 2011). Although these findings are similar, they are not consistent, especially for observational studies (Espinosa-Ramírez et al., 2014). Other studies have identified social and economic neighborhood characteristics as factors that are important to health, including cardiovascular disease, stroke, and mortality (Brown et al., 2011; Diez-Roux et al., 2001; Freedman et al., 2011; Pujades-Rodriguez et al., 2014; Smith et al., 1998; Steenland et al., 2004; Sundquist et al., 2004; Winkleby & Cubbin, 2003). Some studies revealed that residence in a socioeconomically disadvantaged neighborhood can increase risk of incident stroke (Brown et al., 2011; Kim et al., 2021; Kleindorfer et al., 2006; Lisabeth et al., 2007), which varied by sex or race (Brown et al., 2011; Grimaud et al., 2011; Hart et al., 2000; Lisabeth et al., 2007; Pujades-Rodriguez et al., 2014). Another study found that stroke incidence increased with a decrease in neighborhood socioeconomic score (nSES) (Howard et al., 2016). Although these studies revealed that residence in low nSES can impact risk of stroke, these studies did not address the robustness of such associations to using two different residential history sources for exposure assessment. One study shed light on the importance of complete residential histories to accurately assess environmental exposures (Nikkilä et al., 2018). This study found that susceptibility to a disease could vary with residential mobility—the movement of participants from one address to another—which could be misclassified if the exposure of interest depends on a participant's residence (Nikkilä et al., 2018).

These aforementioned findings motivate the present study aim to examine the implications of using residential history data from LexisNexis as compared with baseline residential address of participants from the REasons for Geographic and Racial Differences in Stroke (REGARDS) cohort, a national longitudinal study. Our study assessed disagreement between the two sources of residential history and the robustness of an association between nSES and stroke risk between these residential history data sources.

2. Materials and methods

2.1. Study population

The study used data from the REGARDS cohort; detailed descriptions of the study have been provided elsewhere (Howard et al., 2005; Thacker et al., 2016). Briefly, REGARDS enrolled 30,239 community-dwelling participants across the United States from January 2003 to October 2007 to investigate racial and geographical differences in stroke incidence. The study sampled adults who were at least 45 years old from the stroke belt (North Carolina, South Carolina, Georgia, Tennessee, Mississippi, Alabama, Louisiana and Arkansas), stroke buckle (coastal plain of North Carolina, South Carolina and Georgia), and other regions in the contiguous USA, with an oversampling of white and black individuals, and persons living in the stroke belt. The final overall sample was comprised of 21% from the stroke buckle, 35% from the rest of the stroke belt, and 44% from the other 40 contiguous states, with 42% black participants and 55% women. Participants were selected from commercially available lists purchased from Genesys Inc. That included that included telephone numbers and addresses. Participants were then contacted by mail followed by telephone. Trained interviewers first obtained verbal consent, and then using a computer-assisted telephone interview (CATI), obtained measures of demographics and an array of potential risk factors. A brief physical exam including blood pressure measurements, blood and urine samples, anthropometry, and an electrocardiogram (ECG) was conducted at the participants' residence locations 3–4 weeks after the CATI. Follow-up surveys are conducted by telephone interview every 6 months. All participating institutions' institutional review boards approved the study, and all participants provided verbal and written consent.

2.2. Source of address data

Participants' residential histories were obtained in 2018 from LexisNexis (http://www.LexisNexis.com), a provider of commercially available personal profiles. A finder file was sent to LexisNexis, including participant SSN, first name, last name, birth date, and most current address. LexisNexis provided in return up to 10 "best" addresses that were available for each participant. The addresses purchased from LexisNexis included address, city, state, ZIP code, date first seen, and date last seen. Addresses were included in this study if REGARDS enrollment year was between date first seen and date last seen of the LexisNexis address. Each address in the LexisNexis data meeting this criterion was compared to the REGARDS baseline address, with potential for multiple addresses from the LexisNexis data source. For purposes of this analysis, we only included LexisNexis addresses during the enrollment year for each participant because the REGARDS team physically went to the participants' homes during that year, providing the most valid available address. The number of addresses was defined as the total count of addresses collected during the enrollment year for addresses available through LexisNexis.

2.3. Address geocoding

The purchased telephone numbers and Genesys addresses were utilized to recruit participants via CATI phone calls and mailings. During the CATI call, addresses were obtained directly from the participant for the baseline in-home visit. Therefore, we used the location of the in-home visit as the baseline address for REGARDS participants.

ArcGIS Desktop 10.5.1 with ESRI 2016 Business Analyst data was used for geocoding both the LexisNexis and REGARDS address data. Using the address locators available in the Business Analyst data, a composite addresses locator was created to match initially to the point address locator ("rooftop" level accuracy), then to match against street address (using address range interpolation) for those addresses failing to match at the rooftop level. The minimum match score was set to 90. The REGARDS addresses were extensively cleaned before and during the geocoding process. This process included the removal of apartment numbers, spelling corrections, and the replacement of P.O. Box information with physical addresses. After cleaning, addresses were geocoded. If addresses did not return a matched geocoded location, then the address were further investigated to correct any issues that may have been due to incorrect city, state, or ZIP codes for those participants. We also attempted to geocode all addresses that were received from LexisNexis using the same software settings.

2.4. Stroke outcome

Methods for incident stroke have been described in detail elsewhere (Howard et al., 2011). During follow-up calls, participants or a proxy were asked about possible history of stroke, transient ischemic attack, hospitalizations, emergency department visits, or death. Report of these generated a request to retrieve medical records for that participant. A senior stroke nurse was responsible for initial reviews of medical records to exclude any that did not exhibit signs of stroke. Afterwards, the records for participants with a suspected stroke were adjudicated by physicians. For deaths with no medical record, proxy interview or death certificates were used in adjudication. The WHO definition was used to identify stroke events (WHO, 1989).

2.5. Primary exposure: neighborhood socioeconomic score

Based on previous published methods, a summary nSES variable was created using six variables from 2000 decennial census data that represent education (percentage of adults aged 25 years and older who completed college and adults aged 25 years and older who completed high school as well), occupation (Percentage of individuals aged 16 and older who are employed in professional, managerial, or executive occupations), and wealth/income (log of housing units median value, log of median household income, and percentage of those households that receive dividend, net rental income, and interest) (Diez-Roux et al., 2001; Diez-Roux et al., 2001). The residential census block from the 2000 US census was identified using the REGARDS geocoded addresses to obtain neighborhood characteristics at an individual level. Since census block groups have small population and spatial size, they can be used a proxy for neighborhoods, whose characteristics which have shown to be strong predictors of health (Diez-Roux et al., 2001; Krieger, 1992). For these variables, z scores from each census block were estimated by subtracting the overall mean for the different census block groups and dividing this by the standard deviation (Diez-Roux et al., 2001). The summing Z scores for the aforementioned variables constructed the summary nSES index score, which ranged from −11.8 to 29.0, with a higher value indicating higher nSES advantage (Diez-Roux et al., 2001; Howard et al., 2016).

2.6. Sociodemographic characteristics

During the baseline telephone interview, REGARDS collected data on age, race, sex, education (less than high school, high school graduate, college graduate or above), and income (<$20k, $20-34k, $35-74k, ≥ $75k). Region was coded based on the location of the baseline address according to US Census regions (Northeast, Midwest, South, and West). Diabetes was defined using a fasting glucose level of >125 mL/dL, or if a participant was non-fasting, > 200 mL/dL, or self-reported medication use to control glucose levels. Self-reported history physician diagnosis or ECG was used to define atrial fibrillation.

2.7. Statistical analysis

For descriptive statistics, the means, standard deviations, and frequencies were assessed between geocoded LexisNexis and REGARDS data. To address the potential for participant selection bias, we compared socio-demographic information between people included versus those who were excluded from analyses. Longitudes and latitudes to six decimal places were compared between the geocoding results for each address source. Where these values were not an exact match, distances were calculated between locations. In cases where LexisNexis provided multiple addresses covering REGARDS participant enrollment year, only exact matches or the nearest address were kept for analysis. Locations less than 1 m apart were also considered to be matches.

We used logistic regression to estimate the likelihood that the LexisNexis address during enrollment year matched the REGARDS baseline address for each participant. Models included the number of moves during enrollment year, age, race, sex, income, education level, and census region (Midwest, Northeast, South, and West) as exposures. We also tested for interaction between sex and race. We used survival analysis to assess the potential differences in risk of stroke based on nSES score between LexisNexis/REGARDS matched baseline addresses and REGARDS only baseline addresses. We also tested for interactions of nSES with race, age, education, and sex. All analysis were performed using SAS 9.4 (SAS Institute, Cary, NC).

While for our main analyses, only LexisNexis and REGARDS geocoded locations within 1 m were considered to indicate a matched address, we performed sensitivity analyses using alternative cut points informed by the observed distribution of distances between LexisNexis and REGARDS addresses (Supplemental Table 1) and a tolerance of 5 m used in a previous study (Goldberg, 2008) (Supplemental Table 2).

3. Theory

The reason environmental exposure and residential address histories have received great attention is largely due to research that has revealed variations in impact of environmental exposures on health over a life course (Abbasi et al., 2017; Abdollahpour et al., 2018; Bäärnhielm et al., 2012; Bjørnevik et al., 2014; Boothe et al., 2014; Dalmay et al., 2010; Espinosa-Ramírez et al., 2014; Islam et al., 2007; Kampman et al., 2007; Lucas et al., 2011). For outcomes that depend on residential histories, it is important to utilize accurate and complete residential histories to better estimate risk of environmental exposure (Nikkilä et al., 2018). This study theorizes that accurate baseline residential addresses in REGARDS could provide a better assessment of initial exposures, which could improve risk assessment of a disease, which is why this study seek to compare and contrast LexisNexis and REGARDS baseline residential addresses.

4. Results

Of 30,239 enrolled participants, there were 56 without useable data and 287 whose baseline addresses from REGARDS were not able to be geocoded, resulting in a sample size of 29,896 (99%) participants (Fig. 1). Of these, 26,877 (91%) participants had a LexisNexis address history available for the enrollment year. Table 1 shows that the subset with LexisNexis data was similar to the superset in age, sex, race, income, education levels, and residence in the different census regions. Supplemental Table 2 compared sociodemographics between participants included versus participants excluded from our analysis and no significant differences were observed between groups.

Of the 26,877 participants, there were 26,794 participant addresses that could be geocoded at baseline using both REGARDS and LexisNexis data. We removed 22 participants with missing education information, which lead to an analytical sample of 26,772 participants. Of these, 24,811 (92.7%) addresses matched between REGARDS and LexisNexis and 1961 (7.3%) addresses did not match.

A larger count of LexisNexis addresses during enrollment year was associated with lower odds of agreement (OR = 0.57, 95% CI: 0.54, 0.59; Table 2). Table 2 reveals that odds of agreement was greater for older participants (OR = 1.02 per year of age, 95% CI: 1.01, 1.02), for black race vs. white race (OR = 1.16, 95% CI: 1.05, 1.29), and for females than for males (OR = 1.15, 95% CI: 1.04, 1.26). Furthermore, in comparison to a reference income of < $20 k per year, odds of agreement were higher for an income of $20k34 k (OR = 1.31, 95% CI: 1.13), an income of $35 k-74 k (OR = 1.62, 95% CI: 1.39, 1.89), and an income of more than $75 k (OR = 1.36, 95% CI: 1.14, 1.63) (see Table 2). Compared to residents in the south, odds of agreement were lower for residents in the Midwest (OR = 0.82, 95% CI: 0.73, 0.94; Table 2). Of all sociodemographic variables, education was the only factor that was not found to be an independent predictor of agreement between LexisNexis and REGARDS geocoded addresses (p = 0.377; Table 2). Our interaction

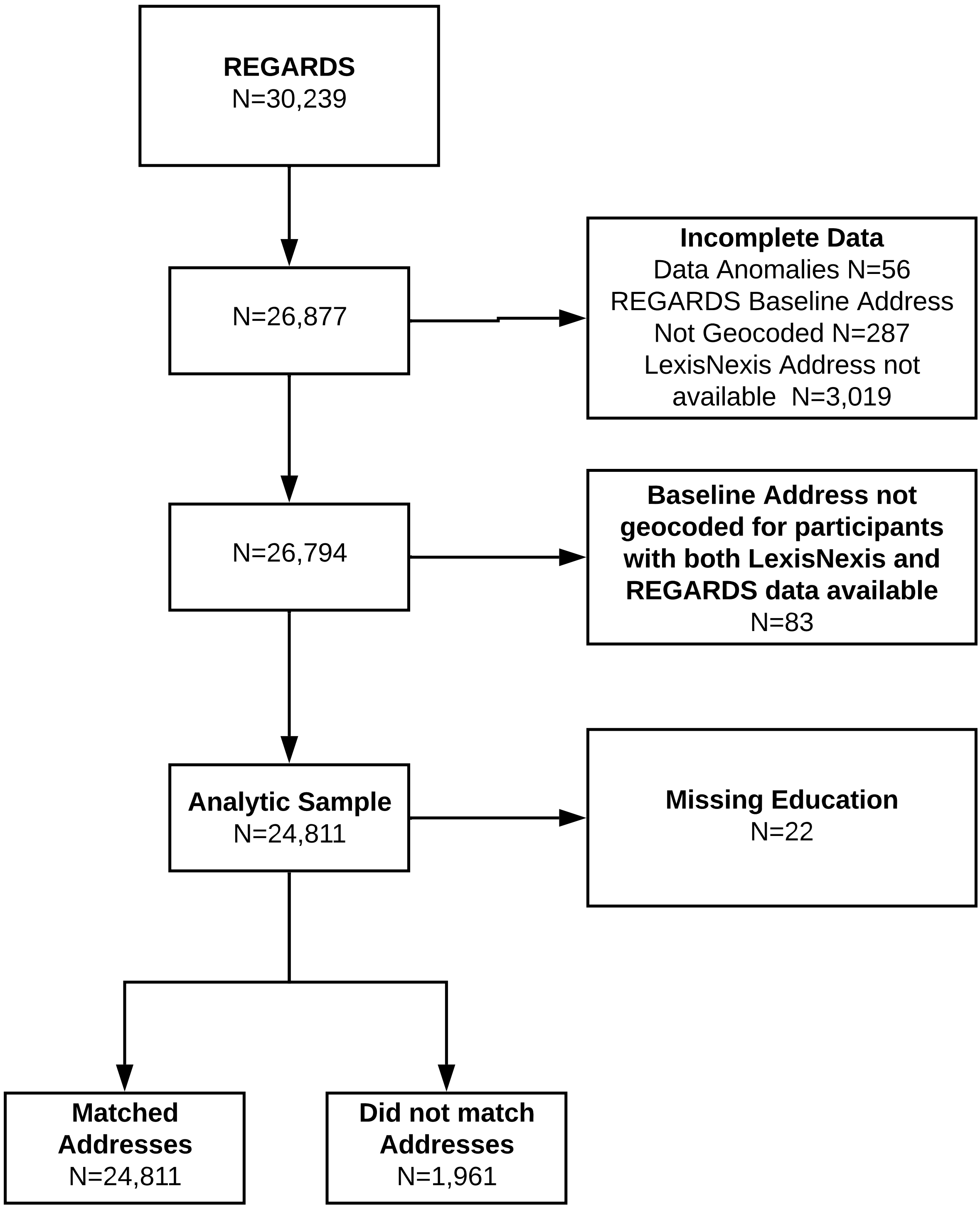

Fig. 1. Participant selection flowchart.

testing revealed no significant interaction between sex and race in predicting agreement (p = 0.786). Supplemental Table 3 shows sociodemographic differences for baseline characteristics across these four census regions.

For participants for which none of the returned LexisNexis addresses corresponded to the REGARDS baseline address (n = 1961), the nearest LexisNexis address was a mean distance of 166 km (SD = 523 km) from the REGARDS baseline address. In our sensitivity analysis, 116 (5.9%) had a distance of ≥1 m but < 5 m between addresses. After considering a <5 m distance to be a match, findings remained unchanged (Supplemental Table 4).

Baseline characteristics for REGARDS geocoded addresses and LexisNexis/REGARDS matched addresses. Findings reveal no significant differences in baseline characteristics (Supplemental Table 5). Risk of stroke did not differ by nSES score for REGARDS only baseline addresses when compared to LexisNexis/REGARDS matched baseline addresses (HR: 0.994,CI:0.983–1.006; HR: 0.998, CI: 0.985–1.010; respectively; Table 3). After testing interactions for age, race, sex, and education, a significant interaction was observed for education and nSES in the REGARDS only geocoded addresses (p = 0.039; Table 3). After stratification by education for REGARDS only geocoded addresses, there was a lower risk of stroke among people will a college graduate education or

Table 1

Baseline characteristics of geocoded REGARDS and LexisNexis residential histories.[a].

	LN Geocoded (N = 26,877; 91.0%)	REGARDS geocoded (N = 29,896; 99.0%)
Number of participants		
Number of moves during enrollment year, Mean (SD)	1.41 (0.76)	1.02 (0.13)
Age (y), Mean (SD)	64.84 (9.37)	64.85 (9.43)
45–54	3348 (12.5)	3748 (12.5)
55–64	10257 (38.2)	11426 (38.2)
65–74	8750 (32.6)	9605 (32.1)
75+	4522 (16.8)	5117 (17.1)
Female	14798 (55.1)	16490 (55.2)
Black	11238 (41.8)	12444 (41.6)
Income ($US/y)		
< $20 k	4679 (17.4)	5423 (18.1)
$20 k-$34 k	6424 (23.9)	7240 (24.2)
$35 k-$74 k	8094 (30.1)	8817 (29.5)
≥ $75 k	4377 (16.3)	4715 (15.8)
Refused	3303 (12.3)	3701 (12.4)
Education		
Less than HS	3249 (12.1)	3754 (12.6)
High school	6934 (25.8)	7715 (25.8)
Some college	7184 (26.8)	8009 (26.8)
College +	9488 (35.3)	10393 (34.8)
Census Region	14900 (55.4)	16574 (55.4)
Midwest	4136 (15.4)	4656 (15.6)
North East	1875 (7.0)	2140 (7.2)
South	18187 (67.7)	20182 (67.5)
West	2679 (10.0)	2918 (9.8)

[a] Unless otherwise noted, all numbers are expressed as n (%).

Table 2

Predictors of non-agreement between LexisNexis geocoded addresses during enrollment year and REGARDS baseline geocoded address.

Predictors	Odds Ratio (95% CI)	p-value
Number of addresses during year of enrollment	0.57 (0.54, 0.59)	<.0001
Age	1.02 (1.01, 1.02)	<.0001
Race (black vs. white[a])	1.16 (1.05, 1.29)	0.0036
Sex (female vs. male[a])	1.15 (1.04, 1.26)	0.0058
Income ($US/y)		<.0001
< $20 k	Ref[a]	
$20 k-$34 k	1.31 (1.13, 1.51)	
$35 k-$74 k	1.62 (1.39, 1.89)	
≥ $75 k	1.36 (1.14, 1.63)	
Refused	1.34 (1.12, 1.60)	
Education		0.3769
Less than high school	Ref[a]	
High school graduate	1.11 (0.94, 1.31)	
Some college	1.16 (0.98, 1.37)	
College graduate and above	1.14 (0.96, 1.36)	
Region		0.0299
South	Ref[a]	
West	0.94 (0.80, 1.11)	
Midwest	0.82 (0.73, 0.94)	
Northeast	0.97 (0.80, 1.17)	

[a] Reference category.

more when compared to people with a high school degree only (HR: 0.974, CI: 0.956–0.992; Table 4).

5. Discussion

In our study, we were interested in comparing the residential information obtained from LexisNexis to residential information provided by participants in REGARDS, a national longitudinal study that has collected data on > 30,000 participants. Our findings showed that the odds of agreement between LexisNexis and REGARDS geocoded addresses was lower for participants with less stable residential histories.

Table 3

Hazard ratios and interactions for risk of stroke in REGARDS (N = 25,544) vs. LexisNexis/REGARDS (N = 21,400) matched baseline addresses.

Model[a]	Hazard Ratio	Confidence Limits	
Neighborhood SES in REGARDS	0.994	0.983	1.006
Neighborhood SES in LexisNexis	0.998	0.985	1.010
Interactions	**REGARDS geocoded baseline addresses only**	**LexisNexis and REGARDS matched baseline addresses[b]**	
Age*nSES	0.849	0.553	
Race*nSES	0.118	0.403	
Sex*nSES	0.861	0.938	
Education*nSES	0.039	0.158	

[a] Model adjusted for age, sex, race, age*race, income, education, Hypertension, Diabetes, and Atrial Fibrillation.

[b] These include people who LexisNexis address matched the REGARDS address.

Table 4

Education Stratified Hazard ratios for risk of stroke and nSES in REGARDS only geocoded addresses.

Model[a]	Hazard Ratio	Confidence Limits	
Less than HS	0.980	0.937	1.026
HS grad	1.008	0.983	1.033
Some college	1.018	0.995	1.042
College graduate or above	0.974	0.956	0.992

[a] Model adjusted for age, sex, race, age*race, income, education, Hypertension, Diabetes, and Atrial Fibrillation.

Agreement between LexisNexis and REGARDS geocoded addresses varied by age, race, sex, income, and census region. This variability could lead to misclassification of participants, which can bias the conclusions drawn in environmental exposure analyses that use residential histories from LexisNexis. Agreement did not vary by education, which could be attributed to the enrollment age of participants being 45 years or older, who may have been further along in their careers at time of enrollment. Furthermore, 9% of REGARDS participants did not have LexisNexis address histories available during enrollment year; bias could result if there were systematic reasons for this discrepancy. Among participants who had both REGARDS and LexisNexis address histories available, approximately 93 percent matched, even when using a strict criterion of being geocoded to points <1 m apart. Based on our initial hypothesis, no significant results were observed when comparing risk of stroke in REGARDS only to LexisNexis/REGARDS matched baseline addresses; however, a significant interaction for education and nSES was observed in the REGARDS only geocoded addresses, which suggests that REGARDS cohort addresses may be valuable in detecting interactions contributing to stroke disparities.

Several previous studies have utilized LexisNexis residential histories to assess environmental exposures (Hertz et al., 2017; Jacquez et al., 2011; Ling et al., 2019). However, issues in residential history collection or validity of residential histories have led to potential exposure misclassification (Hughes & Pruitt, 2017). One study used electronic medical record to collect residential histories; however, these histories are usually incomplete and suffer from bias in the collection process (Hughes & Pruitt, 2017). Another study interested in assessing residential traffic exposure in childhood leukemia found an association during the postnatal period and suggested that epidemiologic studies should utilize complete residential histories when estimating exposure to residential traffic (Boothe et al., 2014). When tracking exposure over a life course, it is important that studies utilize complete residential histories that accurately reflect mobility for participants with greater residential instability (Nikkilä et al., 2018). Additionally, because there are known, observable differences in risk of stroke based on neighborhood socioeconomic characteristics that differ by race and sex (Brown et al., 2011; Grimaud et al., 2011; Hart et al., 2000; Howard et al., 2016; Lisabeth et al., 2007; Pujades-Rodriguez et al., 2014), it is important to have unbiased and comprehensive residential location data.

Our study had several notable strengths. Because a REGARDS team was physically present at the baseline visit, the baseline addresses collected for these participants can be considered a gold standard. Some limitations of our study include relying on the date of enrollment as the time for evaluating agreement; whereas residential stability would ideally be assessed using moves prior to enrollment date, address changes were most validly assessed following enrollment.

6. Conclusions

REGARDS prospective data collection and strict process for cleaning residential address data could lower the misclassification of exposures for cohort addresses versus using LexisNexis as the source for residential addresses. It is vital that studies consider the quality of residential histories when assessing environmental exposure based on geocoded residential addresses. Additionally, future work should consider the residential histories that are utilized to assess neighborhood-level effects or other time and position-dependent effects to decrease the magnitude of bias present in studies that utilize LexisNexis for residential histories. This could help to decrease bias and systematic errors, especially for participants with a higher degree of residential instability.

Authors' statement

Marquita S. Brooks contributed substantially contributed to the conceptualization of this work, interpretation of data, and major contributor in the writing of this manuscript. Aleena Bennett substantially contributed to the design of this work, analyzed and interpreted the data, and was a major contributor in the writing of this manuscript. Gina S. Lovasi, Phil M. Hurvitz, Natalie Colabianchi, Virginia J. Howard, Jennifer Manly, and Suzanne E. Judd substantially contributed to the design of this work, interpretation of data, and revision as well as editing of manuscript. Suzanne E. Judd was also responsible for supervision.

Ethical statement

All authors read and approved the final manuscript. An authorship statement has been included in this submission. All authors ensure that this work has been written entirely as an original work. A statement regarding data access has been included in the manuscript submission. We have properly acknowledged the work of others. There are no conflicts of interest to report. Our work does not involve chemicals and therefore does not require a hazards statement. All participating institutions' institutional review boards approved the study, and all participants provided verbal and written consent.

Data statement

The datasets generated and analyzed during the current study are not publicly available because the data underlying the findings include potentially identifying participant information and cannot be made publicly available due to ethical/legal restrictions. However, data including statistical code supporting this manuscript are available to researchers who meet the criteria for access to confidential data. Data can be obtained upon request through the University of Alabama at Birmingham at regardsadmin@uab.edu.

Financial disclosure statement

This research project is supported by cooperative agreement U01 NS041588 co-funded by the National Institute of Neurological Disorders and Stroke (NINDS) and the National Institute on Aging (NIA, National Institutes of Health, Department of Health and Human Service. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NINDS or the NIA. Representatives of the NINDS were involved in the review of the manuscript but were not directly involved in the collection, management, analysis or interpretation of the data. This work was supported by the National Institute of Aging (grants 1R01AG049970, 3R01AG049970-04S1), Commonwealth Universal Research Enhancement (C·U.R.E) program funded by the Pennsylvania Department of Health - 2015 Formula award - SAP #4100072543, the Urban Health Collaborative at Drexel University, and the Built Environment and Health Research Group at Columbia University. Additional funding was provided by investigator-initiated grants from National Institutes of Health (R01 NS092706, Natalie Colabianchi and Suzanne Judd, MPIs).

Declaration of competing interest

None.

Acknowledgements

The authors thank the other investigators, the staff, and the participants of the REGARDS study for their valuable contributions. A full list of participating REGARDS investigators and institutions can be found at: https://www.uab.edu/soph/regardsstudy/.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssmph.2021.100887.

References

Abbasi, M., Nabavi, S. M., Fereshtehnejad, S. M., Jou, N. Z., Ansari, I., Shayegannejad, V., Mohammadianinejad, S. E., Farhoudi, M., Noorian, A., Razazian, N., Abedini, M., & Faraji, F. (2017). Multiple sclerosis and environmental risk factors: A case-control study in Iran. *Neurological Sciences, 38*(11), 1941–1951. https://doi.org/10.1007/s10072-017-3080-9

Abdollahpour, I., Nedjat, S., Mansournia, M. A., Sahraian, M. A., & van der Mei, I. (2018). Lifestyle factors and multiple sclerosis: A population-based incident case-control study. *Mult Scler Relat Disord, 22*, 128–133. https://doi.org/10.1016/j.msard.2018.03.022

Bäärnhielm, M., Hedström, A. K., Kockum, I., Sundqvist, E., Gustafsson, S. A., Hillert, J., Olsson, T., & Alfredsson, L. (2012). Sunlight is associated with decreased multiple sclerosis risk: No interaction with human leukocyte antigen-DRB1*15. *European Journal of Neurology, 19*(7), 955–962. https://doi.org/10.1111/j.1468-1331.2011.03650.x

Bjørnevik, K., Riise, T., Casetta, I., Drulovic, J., Granieri, E., Holmøy, T., Kampman, M. T., Landtblom, A. M., Lauer, K., Lossius, A., Magalhaes, S., Myhr, K. M., Pekmezovic, T., Wesnes, K., Wolfson, C., & Pugliatti, M. (2014). Sun exposure and multiple sclerosis risk in Norway and Italy: The EnvIMS study. *Multiple Sclerosis, 20*(8), 1042–1049. https://doi.org/10.1177/1352458513513968

Boothe, V. L., Boehmer, T. K., Wendel, A. M., & Yip, F. Y. (2014). Residential traffic exposure and childhood leukemia: A systematic review and meta-analysis. *American Journal of Preventive Medicine, 46*(4), 413–422. https://doi.org/10.1016/j.amepre.2013.11.004

Brown, A. F., Liang, L. J., Vassar, S. D., Stein-Merkin, S., Longstreth, W. T., Jr., Ovbiagele, B., Yan, T., & Escarce, J. J. (2011). Neighborhood disadvantage and ischemic stroke: The cardiovascular health study (CHS). *Stroke, 42*(12), 3363–3368. https://doi.org/10.1161/strokeaha.111.622134

Dalmay, F., Bhalla, D., Nicoletti, A., Cabrera-Gomez, J. A., Cabre, P., Ruiz, F., Druet-Cabanac, M., Dumas, M., & Preux, P. M. (2010). Multiple sclerosis and solar exposure before the age of 15 years: Case-control study in Cuba, Martinique and Sicily. *Multiple Sclerosis, 16*(8), 899–908. https://doi.org/10.1177/1352458510366856

Diez-Roux, A. V., Kiefe, C. I., Jacobs, D. R., Jr., Haan, M., Jackson, S. A., Nieto, F. J., Paton, C. C., & Schulz, R. (2001). Area characteristics and individual-level socioeconomic position indicators in three population-based epidemiologic studies. *Annals of Epidemiology, 11*(6), 395–405. https://doi.org/10.1016/s1047-2797(01)00221-6

Diez-Roux, A. V., Merkin, S. S., Arnett, D., Chambless, L., Massing, M., Nieto, F. J., … Watson, R. L. (2001). Neighborhood of residence and incidence of coronary heart disease. *New England Journal of Medicine, 345*(2), 99–106. https://doi.org/10.1056/nejm200107123450205

Espinosa-Ramírez, G., Ordoñez, G., Flores-Rivera, J., & Sotelo, J. (2014). Sunlight exposure and multiple sclerosis in a tropical country. *Neurological Research, 36*(7), 647–650. https://doi.org/10.1179/1743132813y.0000000307

Freedman, V. A., Grafova, I. B., & Rogowski, J. (2011). Neighborhoods and chronic disease onset in later life. *American Journal of Public Health, 101*(1), 79–86. https://doi.org/10.2105/ajph.2009.178640

Goldberg, D. (2008). *A geocoding best practices guide*. Springfield, IL): North Am Assoc Cent Cancer Registries.

Grimaud, O., Béjot, Y., Heritage, Z., Vallée, J., Durier, J., Cadot, E., Giroud, M., & Chauvin, P. (2011). Incidence of stroke and socioeconomic neighborhood characteristics: An ecological analysis of Dijon stroke registry. *Stroke, 42*(5), 1201–1206. https://doi.org/10.1161/strokeaha.110.596429

Hart, C. L., Hole, D. J., & Smith, G. D. (2000). The contribution of risk factors to stroke differentials, by socioeconomic position in adulthood: The Renfrew/Paisley study. *American Journal of Public Health, 90*(11), 1788–1791. https://doi.org/10.2105/ajph.90.11.1788

Hertz, A., Nelson, D. O., Deapen, D., Von Behren, J., Bernstein, L., Layefsky, M., Reynolds, P., & Hurley, S. (2017). Tracing a Path to the Past: Exploring the use of commercial credit reporting data to construct residential histories for epidemiologic studies of environmental exposures. *American Journal of Epidemiology, 185*(3), 238–246. https://doi.org/10.1093/aje/kww108

Howard, V. J., Cushman, M., Pulley, L., Gomez, C. R., Go, R. C., Prineas, R. J., Graham, A., Moy, C. S., & Howard, G. (2005). The reasons for geographic and racial differences in stroke study: Objectives and design. *Neuroepidemiology, 25*(3), 135–143. https://doi.org/10.1159/000086678

Howard, V. J., Kleindorfer, D. O., Judd, S. E., McClure, L. A., Safford, M. M., Rhodes, J. D., Cushman, M., Moy, C. S., Soliman, E. Z., Kissela, B. M., & Howard, G. (2011). Disparities in stroke incidence contributing to disparities in stroke mortality. *Annals of Neurology, 69*(4), 619–627. https://doi.org/10.1002/ana.22385

Howard, V. J., McClure, L. A., Kleindorfer, D. O., Cunningham, S. A., Thrift, A. G., Diez Roux, A. V., & Howard, G. (2016). Neighborhood socioeconomic index and stroke incidence in a national cohort of blacks and whites. *Neurology, 87*(22), 2340–2347. https://doi.org/10.1212/wnl.0000000000003299

Hughes, A. E., & Pruitt, S. L. (2017). The utility of EMR address histories for assessing neighborhood exposures. *Annals of Epidemiology, 27*(1), 20–26. https://doi.org/10.1016/j.annepidem.2016.07.016

Islam, T., Gauderman, W. J., Cozen, W., & Mack, T. M. (2007). Childhood sun exposure influences risk of multiple sclerosis in monozygotic twins. *Neurology, 69*(4), 381–388. https://doi.org/10.1212/01.wnl.0000268266.50850.48

Jacquez, G. M., Slotnick, M. J., Meliker, J. R., AvRuskin, G., Copeland, G., & Nriagu, J. (2011). Accuracy of commercially available residential histories for epidemiologic studies. *American Journal of Epidemiology, 173*(2), 236–243. https://doi.org/10.1093/aje/kwq350

Kampman, M. T., Wilsgaard, T., & Mellgren, S. I. (2007). Outdoor activities and diet in childhood and adolescence relate to MS risk above the Arctic Circle. *Journal of Neurology, 254*(4), 471–477. https://doi.org/10.1007/s00415-006-0395-5

Kim, Y., Twardzik, E., Judd, S. E., & Colabianchi, N. (2021). Neighborhood socioeconomic status and stroke incidence: A systematic review. *Neurology, 96*(19), 897–907. https://doi.org/10.1212/wnl.0000000000011892

Kleindorfer, D. O., Lindsell, C., Broderick, J., Flaherty, M. L., Woo, D., Alwell, K., Moomaw, C. J., Ewing, I., Schneider, A., & Kissela, B. M. (2006). Impact of socioeconomic status on stroke incidence: A population-based study. *Annals of Neurology, 60*(4), 480–484. https://doi.org/10.1002/ana.20974

Krieger, N. (1992). Overcoming the absence of socioeconomic data in medical records: Validation and application of a census-based methodology. *American Journal of Public Health, 82*(5), 703–710. https://doi.org/10.2105/ajph.82.5.703

Ling, C., Heck, J. E., Cockburn, M., Liew, Z., Marcotte, E., & Ritz, B. (2019). Residential mobility in early childhood and the impact on misclassification in pesticide exposures. *Environmental Research, 173*, 212–220. https://doi.org/10.1016/j.envres.2019.03.039

Lisabeth, L. D., Diez Roux, A. V., Escobar, J. D., Smith, M. A., & Morgenstern, L. B. (2007). Neighborhood environment and risk of ischemic stroke: The brain attack surveillance in corpus christi (BASIC) project. *American Journal of Epidemiology, 165*(3), 279–287. https://doi.org/10.1093/aje/kwk005

Lucas, R. M., Ponsonby, A. L., Dear, K., Valery, P. C., Pender, M. P., Taylor, B. V., Kilpatrick, T. J., Dwyer, T., Coulthard, A., Chapman, C., van der Mei, I., Williams, D., & McMichael, A. J. (2011). Sun exposure and vitamin D are independent risk factors for CNS demyelination. *Neurology, 76*(6), 540–548. https://doi.org/10.1212/WNL.0b013e31820af93d

Medgyesi, D. N., Fisher, J. A., Flory, A. R., Hayes, R. B., Thurston, G. D., Liao, L. M., Ward, M. H., Silverman, D. T., & Jones, R. R. (2021). Evaluation of a commercial database to estimate residence histories in the los angeles ultrafines study. *Environmental Research, 197*, 110986. https://doi.org/10.1016/j.envres.2021.110986

Nikkilä, A., Kendall, G., Raitanen, J., Spycher, B., Lohi, O., & Auvinen, A. (2018). Effects of incomplete residential histories on studies of environmental exposure with application to childhood leukaemia and background radiation. *Environmental Research, 166*, 466–472. https://doi.org/10.1016/j.envres.2018.06.035

Pujades-Rodriguez, M., Timmis, A., Stogiannis, D., Rapsomaniki, E., Denaxas, S., Shah, A., Feder, G., Kivimaki, M., & Hemingway, H. (2014). Socioeconomic deprivation and the incidence of 12 cardiovascular diseases in 1.9 million women and men: Implications for risk prediction and prevention. *PloS One, 9*(8), Article e104671. https://doi.org/10.1371/journal.pone.0104671

Smith, G. D., Hart, C., Watt, G., Hole, D., & Hawthorne, V. (1998). Individual social class, area-based deprivation, cardiovascular disease risk factors, and mortality: The Renfrew and Paisley study. *Journal of Epidemiology & Community Health, 52*(6), 399–405. https://doi.org/10.1136/jech.52.6.399

Steenland, K., Henley, J., Calle, E., & Thun, M. (2004). Individual- and area-level socioeconomic status variables as predictors of mortality in a cohort of 179,383 persons. *American Journal of Epidemiology, 159*(11), 1047–1056. https://doi.org/10.1093/aje/kwh129

Sundquist, K., Winkleby, M., Ahlén, H., & Johansson, S. E. (2004). Neighborhood socioeconomic environment and incidence of coronary heart disease: A follow-up study of 25,319 women and men in Sweden. *American Journal of Epidemiology, 159*(7), 655–662. https://doi.org/10.1093/aje/kwh096

Thacker, E. L., Soliman, E. Z., Pulley, L., Safford, M. M., Howard, G., & Howard, V. J. (2016). Investigation of selection bias in the association of race with prevalent atrial fibrillation in a national cohort study: REasons for geographic and racial differences in stroke (REGARDS). *Annals of Epidemiology, 26*(8), 534–539. https://doi.org/10.1016/j.annepidem.2016.06.012

Wheeler, D. C., & Wang, A. (2015). Assessment of residential history generation using a public-record database. *International Journal of Environmental Research and Public Health, 12*(9), 11670–11682. http://www.mdpi.com/1660-4601/12/9/11670.

WHO. (1989). Stroke–1989. Recommendations on stroke prevention, diagnosis, and therapy. Report of the WHO task force on stroke and other cerebrovascular Disorders. *Stroke, 20*(10), 1407–1431. https://doi.org/10.1161/01.str.20.10.1407

Winkleby, M. A., & Cubbin, C. (2003). Influence of individual and neighbourhood socioeconomic status on mortality among black, Mexican-American, and white women and men in the United States. *Journal of Epidemiology & Community Health, 57*(6), 444–452. https://doi.org/10.1136/jech.57.6.444