Geospatial Health 7(2), 2013, pp. 369-374

Assessing bias associated with geocoding of historical residence in epidemiology research

Daikwon Han[1], Matthew R. Bonner[2], Jing Nie[2], Jo L. Freudenheim[2]

[1]*Department of Epidemiology and Biostatistics, Texas A&M University, College Station, TX 77843, USA;*
[2]*Department of Social and Preventive Medicine, University at Buffalo, Buffalo, NY 14214, USA*

Abstract. The use of geocoded historical residence as proxy for retrospective assessment of exposure in early life is increasing in epidemiological studies of chronic health outcomes. Dealing with historical residence poses challenges, primarily due to higher uncertainties associated with data collection and processing. A possible source of bias is connected with the exclusion of subjects, who cannot, for various reasons, be geocoded. We evaluated the potential bias that may arise due to incomplete geocoding, using birth residence data collected as part of a population-based case-control study of breast cancer in western New York state. We found that geocoded and non-geocoded populations did not differ in the distribution of most risk factors compared, and that the geocoding status did not modify the spatial patterns of the study populations. However, the results emphasize the need for epidemiological studies to consider the potential biases that may be introduced by geocoding of historical residence when investigating retrospectively chronic disease and early-life exposure.

Keywords: geocoding, historical residence, bias, USA.

Introduction

The historical residence as proxy for retrospective assessment of exposure in early life is increasingly being used in epidemiological studies of chronic health outcomes. Given the accumulated evidence on the importance of early-life factors and conditions, such as exposures at the place of birth, recent studies concerning cancer outcomes have utilised available information about the early-life environment, which can be revealed by the lifetime residential history. Numerous studies have used such records to estimate early-life and/or cumulative exposures to environmental contaminants such as pesticides, arsenic and traffic pollutants, in relation to different cancer outcomes, e.g. cancer of the breast and the bladder (Brody et al., 2004; Jacquez et al., 2005; Nie et al., 2007; Gallagher et al., 2010; Meliker et al., 2010). Additionally, residence history data make it possible to account for residential mobility and time lags between exposure and diagnosis and this may provide less biased exposure-outcome associations than studies solely based on residence at the time of diagnosis or when the interview was done (Han et al., 2005; Hurley et al., 2005; Urayama et al., 2009; Boscoe, 2011). However, most investigations are case/control studies and the data collection was based on self-report that may involve recall bias and/or exposure misclassification. Further, dealing with historical residence poses challenges due to the higher uncertainties associated with data collection, data processing and exclusion of subjects who cannot be geocoded, which may be a source of bias.

Geocoding is the process of placing subjects' addresses on a map. Epidemiological studies have increasingly relied on the geocoded residence history in investigating the relationship between retrospective exposures and the subsequent risk of adult chronic disease. Accurate and complete geocoding of data related to residential history becomes critical when biological markers for retrospective exposure in the past are lacking. There are possibilities for geographic bias – defined as the difference in spatial patterns in characteristics of subjects by geocoding status – for example, when subjects fail to be address-matched and thus must be excluded from analysis in spatial epidemiology studies. There are numerous studies documenting the implications of positional accuracy or error due to discrepancy in geographical information systems (GIS)-based geocoded locations relative to the true position on the Earth's surface, including exposure misclassification of geocoded residency in epidemiological studies (Krieger et al., 2001; Bonner et al., 2003; Ward et al., 2005; Gilboa et al., 2006;

Corresponding author:
Daikwon Han
Department of Epidemiology and Biostatistics
Texas A&M University, College Station, TX 77843, USA
Tel. +1 979 458 0068; Fax +1 979 458 1878
E-mail: dhan@srph.tamhsc.edu

Zandbergen, 2007). However, there are relatively few report, none for historical residence, assessing geographic bias associated with completeness of geocoding in spatial and environmental epidemiology (Gregorio et al., 1999; Oliver et al., 2005).

We conducted an investigation evaluating the quality of geocoding and the magnitude of potential bias due to geocoding failure of historical residency. Using historical residential data collected as part of a western New York State Exposures and Breast Cancer (WEB) study, we compared address-matched subjects with self-reported birth residence information with those, who initially failed to be geocoded but later matched based on additional information obtained from birth certificates. The purpose of the present study was therefore twofold: (i) to identify geographic bias due to geocoding failure of historical residency; and (ii) to assess selection bias by case-control status of geocoded and non-geocoded subjects in early life.

Materials and methods

Subjects were identified from the WEB study, a population-based, case-control study covering the period 1996-2001. The study subjects were women, aged 35-79 years, who were residents of the Erie and Niagara counties in New York state. Details of the WEB study have been described elsewhere (Bonner et al., 2005; Han et al., 2005). In short, the WEB study included 1,166 primary histologically confirmed, breast cancer cases from the two counties and 2,105 controls, who were randomly selected from the county populations, constituting groups that were frequency-matched with regard to age, race and county of residence. Extensive in-person interviews and self-administered questionnaires were used to ascertain study participant risk factors that included lifetime, residential history from birth to current location. Specifically, with respect to the birth addresses, we asked for information on the birth address twice and collected information on response reliability.

When the lifetime residential history was collected for the WEB study, participants were address-matched using GIS with the aid of supplementary methods such as geocoding strategies that included historical maps and city directories. We relied on several resources to geocode addresses, especially for earlier residences. An enhanced version of a street map of the study area was used as well as a programme to find and update zip-code information. We also used historic city directories to find complete residential information such as street number, town or city name and used historic maps to identify those residences (Han et al., 2005). Because geocoding of historic residency incurs a higher likelihood of missing information than is the case with more recent addresses, we developed methods using information of maiden names and partial address information provided by the participants to search for records in city directories for the appropriate time periods. In the WEB study, the final address-matching rate ranged from 80% for birth residences to 99% for current residences. We failed to geocode some birth addresses, primarily because of missing residential information such as missing street numbers or missing street names. Due to the fact that the lowest geocoding rate regarded the birth address, we paid particular attention to subjects born in the study area as additional residence information from the birth certificate was available for them.

A total of 1,510 subjects (579 cases and 931 controls), who were all born in the study area and who had their birth addresses located, were included in the study. Subjects were divided into those who were geocoded (n = 1,309), i.e. address-matched subjects using self-reported birth residence information, and those who were not (n = 201), i.e. subjects who failed to be address-matched initially due to missing self-reported birth records but matched eventually based on an additional address recorded on their birth certificates. Such additional addresses were obtained and address-matched based on certificate addresses for subjects, who were born in the study area but had no self-reported information regarding their birth residence.

Geocoded and non-geocoded populations were compared by selected risk factors and by case-control status. Selected characteristics of study participants included age, education, race, marital status, number of births, smoking and body mass index (BMI). Odds ratios (ORs) and 95% confidence intervals (CIs) of the likelihood of being a case associated with risk factors were obtained for geocoded and non-geocoded populations by homogeneity testing. Spatial clustering analysis was performed to identify geographical areas where the proportion of non-geocoded subjects was higher than expected. We employed spatial scan statistic with the spatial Poisson model for this purpose (Kulldorff and Nagarwalla, 1995). Under the null hypothesis that the proportion of non-geocoded subjects is same across zip code areas, we performed 999 Monte Carlo replications based on the data aggregated by zip-code. We used the total number of non-geocoded subjects as the case variable and the total number of subjects as the population to identify clus-

Table 1. Selected characteristics of study participants born in Erie and Niagara counties, by geocoding status (geocoded and non-geocoded population); WEB study - 1996-2001.

Characteristic		Geocoded population (n = 1,309)		Non-geocoded population (n = 201)	
		N	%	N	%
Age (years)	≤56	703	53.7	110	54.7
	>56	606	46.3	91	45.3
Education (years)	≤12	591	45.1	83	41.3
	>12	718	54.9	118	58.7
Number of births	None	203	15.5	37	18.4
	≥1	1,106	84.5	164	81.6
Marital status[1]	Married	858	90.1	137	94.5
	Never married	94	9.9	8	5.5
Race	White	1271	97.1	196	97.5
	Non-white	38	2.9	5	2.5
BMI[2]	≤30	898	68.6	133	66.2
	>30	411	31.4	68	33.8
Smoker*	No	642	49.0	81	40.3
	Yes[3]	667	51.0	120	59.7

[1]Excludes widowed, divorced/separated and other missing/questionable data; [2]BMI 12-24 months before the diagnosis (cases) or interview (controls); [3]Includes former and current smokers; *P <0.05.

ters that had an excess of non-geocoded subjects with a higher proportion of non-geocoded subjects than would be expected.

Results

The geocoded and non-geocoded populations did not differ in the distribution for many of the risk factors investigated (Table 1). There was less than 4% of difference between the two groups except with respect to two variables: marital status and whether they smoked. While the marital status was not significantly different between the two groups (P = 0.1), the difference with regard to smoking was statically significant (P = 0.02).

The likelihood of being a case associated with risk factors in the geocoded and non-geocoded groups is presented in Table 2. Only those variables with more than 4% difference were further compared with regard to case-control status. For several risk factors, including marital status, number of births and BMI, we observed that the cancer cases were more likely than the controls to never have given birth, never having married and having a BMI greater than 30. However the ORs were not significantly different. Similarly, those with cancer were more likely to have more than 12 years of education in the geocoded populations, while less likely so in the non-geocoded populations. Again, the ORs were not significantly different (P = 0.31). On the other hand, we found that the cancer cases were more likely to be smokers than controls in the geocoded populations, while this was less likely in the non-geocoded populations. There was significant heterogeneity between these two ORs (p=0.02), however, confidence intervals for both

Table 2. Odds ratio (OR) and 95% confidence interval (CI) of the likelihood of being a cancer case associated with risk factors in geocoded and non-geocoded population; WEB study, 1996-2001.

	Geocoded population		Non-geocoded population	
	OR	95% CI	OR	95% CI
Nulliparous vs. parous[1]	1.23	0.91-1.67	2.09	1.02-4.31
>12 years vs. ≤12 education	1.10	0.88-1.37	0.68	0.38-1.21
Never married vs. married	1.62	1.05-2.38	3.64	0.83-15.94
Smoker[2] vs. non-smoker*	1.24	0.99-1.55	0.90	0.50-1.61
BMI >30 vs. ≤30[3]	1.12	0.88-1.42	1.00	0.54 -1.83

[1]Excludes widowed, divorced/separated and other missing/questionable data; [2]Includes former and current smokers; [3]BMI 12-24 months before the diagnosis (cases) or interview (controls) *P <0.05.

geocoded and non-geocoded population groups included the null (Table 2).

The geocoding status did not modify the spatial patterns of study populations; there were several clusters that had an excess of non-geocoded subjects, but none of the differences was statistically significant. About 30% of the zip code areas were identified as a geographical area that had higher-than-expected proportion of non-geocoded subjects. Clusters with an excess of non-geocoded populations, identified using spatial scan statistic, are summarised in Table 3. Two primary and secondary clusters were identified as likely, but with log likelihood ratios that were not significant (p-values of 0.38 and 0.96, respectively).

Discussion

We evaluated potential bias, including geographic bias that may arise due to the incompleteness of geocoding in a study of early-life environmental factors and breast cancer. Overall our findings indicate that there is no differential error associated with geocoding status in estimating the subsequent risk of breast cancer associated with place of birth. Indeed, with exception of the smoking status, which was significant, we found that geocoded and non-geocoded populations differed only marginally in the distribution of the risk factors investigated. Furthermore, geocoding status did not modify the spatial patterns of the study populations. Although, there were several clusters that had an excess of non-geocoded subjects, none was statistically significant.

There are several sources of uncertainty in spatial and environmental epidemiology studies of early life factors, and the geocoding process could reveal many possible sources of error, including positional error and difference (Krieger et al., 2001; Rushton et al., 2006). Similarly, geocoding failure of the historical residence may cause geographic bias in spatial analysis of the relations between exposures during early life and disease risk. Because geocoded residential locations have primarily been used as proxy for retrospective exposures in early life, subsequent exposure classification and disease risk estimates depend (in part) on the quality of geocoding with respect to the historical residence. Common problems and solutions in the geocoding steps of health data have been well documented (Rushton et al., 2006; Goldberg et al., 2008). However, the potential bias due to the incompleteness of geocoding, especially of historical records, has not been fully evaluated in the research of early-life factors. Our finding that there is no difference in spatial patterns associated with geocoding status provides some credibility to the validity of study findings, including our previous ones. In addition, it reduces the likelihood of bias being introduced by missing data related to geocoding and related geospatial methods used in epidemiological research. Given the significance of early-life and lifetime exposures in chronic disease epidemiology, more effective methods for geocoding of historical residence need to be developed, including further development of imputation methods for missing residence data and utilization of other sources to validate and complement self-reported historical residence information (Curriero et al., 2010; Jacquez et al., 2011).

There are a numbers of strengths in our study, primarily richness of unique historical exposure-related factors, including lifetime residential history, collected as part of the WEB study. We achieved relatively high rates of geocoding for historical records by combining two data sources, namely self-reported and birth certificate records. Using data from two sources (self-reported birth addresses from the WEB study vs. birth residence information from the birth certificate data), we were able to compare selected characteristics of subjects by geocoding status and by case-control status as well as the potential impact on geographical (clustering) analyses. We also evaluated potentials of geographic bias by comparing spatial clustering patterns with and without additionally obtained data from the certificate. While birth certificate combined with self-reported birth data on residence may be used complementarily to increase data availability and

Table 3. Clusters with an excess of non-geocoded populations identified using spatial scan statistic.

	Number of cases	Expected	Relative risk	LLR	Location (zip code areas within the cluster)	P-value
Primary cluster	4	0.69	5.92	3.76	14202	0.38
Secondary cluster	54	42.42	1.37	1.87	14072, 14092, 14120, 14132, 14150, 4174, 14207, 14217, 14301, 14303, 14304, 14305	0.96

LLR = log likelihood ratio

completeness of early-life environment, it is important to note that the quality of data and the number of data elements from a birth certificate may also vary. Such variation can include many factors, e.g. locality and period of birth, so critical evaluation of data elements of birth certificate is required.

Our study is not without limitations. Characteristics of the study area, while being one of the major strengths of the WEB study, may restrict generalization of our findings. Although the Buffalo-Niagara region experienced relatively slow change in population growth over the time period of the study, additional factors may have contributed to the quality of geocoding historical records. For example, population growth and urban expansion patterns in the study area could hinder the application of our results in settings with different characteristics in population growth and/or environmental change. For urban regions with more rapid population growth over time, it is more likely that street names and addressing systems change more frequently than in regions, such as the one we studied, where there has been little change; such alterations would impact the ability to completely geocode historical records. Similarly, some areas with a rural mail delivery system have been given permanent addresses to establish a uniform system for addresses and to allow locating properties for emergency responses. However, historical residencies in these rural areas may not be identifiable using the updated addressing systems. Further studies should be replicated in other settings, where studies covering time periods of rapid population change, and/or areas with changed addressing systems. Lastly, geographic bias may be of less concern in a setting where polygon-based geocoding methods (e.g. based on census unit or zip code) were employed to increase the completeness of geocoding. However, problems of spatial mismatch may arise due to possible changes in the boundaries of administrative units used in the study linking historical exposures with subsequent health outcomes. Such potential error will be validated in a future study.

This study provides evidence that there may be potential bias associated with the accuracy and completeness of geocoding of residential history in epidemiology research. In addition, it shows that assessment of the quality of geocoding ensures the validity of study findings and reduces bias due to incomplete geocoding of historical records. Epidemiological studies should consider the potential biases that may be introduced by geocoding of historical residence in the investigation of retrospective exposure in early life and chronic health outcomes.

References

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

Bonner MR, Han D, Nie J, Rogerson PA, Vena JE, Muti P, Trevisan M, Edge SB, Freudenheim JL, 2005. Breast cancer risk and exposure in early life to polycyclic aromatic hydrocarbons using total suspended particulates as a proxy measure. Cancer Epidem Biomar 14, 53-60.

Boscoe F, 2011. The use of residential history in environmental health studies. In: Geospatial Analysis of Environmental Health. Maantay J, McLafferty S (eds). Springer, 93-110 pp.

Brody JG, Aschengrau A, McKelvey W, Rudel RA, Swartz CH, Kennedy T, 2004. Breast cancer risk and historical exposure to pesticides from wide-area applications assessed with GIS. Environ Health Perspect 112, 889-897.

Curriero FC, Kulldorff M, Boscoe FP, Klassen AC, 2010. Using imputation to provide location information for nongeocoded addresses. PLoS One 5, e8998.

Gallagher LG, Webster TF, Aschengrau A, Vieira VM, 2010. Using residential history and groundwater modeling to examine drinking water exposure and breast cancer. Environ Health Perspect 118, 749-755.

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.

Han D, Rogerson PA, Bonner MR, Nie J, Vena JE, Muti P, Trevisan M, Freudenheim JL, 2005. Assessing spatio-temporal variability of risk surfaces using residential history data in a case control study of breast cancer. Int J Health Geogr 4, 9.

Hurley SE, Reynolds P, Goldberg DE, Hertz A, Anton-Culver H, Bernstein L, Deapen D, Peel D, Pinder R, Ross RK, West D, Wright WE, Ziogas A, Horn-Ross PL, 2005. Residential mobility in the California teachers study: implications for geographic differences in disease rates. Soc Sci Med 60, 1547-1555.

Jacquez GM, Kaufmann A, Meliker J, Goovaerts P, AvRuskin G, Nriagu J, 2005. Global, local and focused geographic clustering for case-control data with residential histories. Environ Health 4, 4.

Jacquez GM, Slotnick MJ, Meliker JR, AvRuskin G, Copeland G, Nriagu J, 2011. Accuracy of commercially available residential histories for epidemiologic studies. Am J Epidemiol 173, 236-243.

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.

Kulldorff M, Nagarwalla N, 1995. Spatial disease clusters: detection and inference. Stat Med 14, 799-810.

Meliker JR, Slotnick MJ, AvRuskin GA, Schottenfeld D, Jacquez GM, Wilson ML, Goovaerts P, Franzblau A, Nriagu JO, 2010. Lifetime exposure to arsenic in drinking water and bladder cancer: a population-based case-control study in Michigan, USA. Cancer Cause Control 21, 745-757.

Nie J, Beyea J, Bonner MR, Han D, Vena J, Rogerson P, Vito D, Muti P, Trevisan M, Freudenheim JL, 2007. Exposure to traffic emissions throughout life and risk of breast cancer: the Western New York Exposures and Breast Cancer (WEB) study. Cancer Causes Control 18, 947-955.

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.

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.

Urayama KY, Von BJ, Reynolds P, Hertz A, Does M, Buffler PA, 2009. Factors associated with residential mobility in children with leukemia: implications for assigning exposures. Ann Epidemiol 19, 834-840.

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.

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