Annals of Epidemiology 26 (2016) 204–211

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Potential selection bias associated with using geocoded birth records for epidemiologic research

Sandie Ha PhD, MPH [a], Hui Hu BS [a], Liang Mao PhD [b], Dikea Roussos-Ross MD [c], Jeffrey Roth PhD [d], Xiaohui Xu PhD [e,*]

[a] Department of Epidemiology, College of Public Health and Health Professions, College of Medicine, University of Florida, Gainesville
[b] Department of Geography, University of Florida, Gainesville
[c] Department of Obstetrics and Gynecology, University of Florida, Gainesville
[d] Department of Pediatrics, College of Medicine, University of Florida, Gainesville
[e] Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M Health Science Center, College Station

ARTICLE INFO

Article history:
Received 9 April 2015
Accepted 13 January 2016
Available online 4 February 2016

Keywords:
Birth certificates
Selection bias
Geocode
Environmental epidemiology

ABSTRACT

Purpose: There is an increasing use of geocoded birth registry data in environmental epidemiology research. Ungeocoded records are routinely excluded.
Methods: We used classification and regression tree analysis and logistic regression to investigate potential selection bias associated with this exclusion among all singleton Florida births in 2009 ($n = 210{,}285$).
Results: The rate of unsuccessful geocoding was 11.5% ($n = 24{,}171$). This ranged between 0% and 100% across zip codes. Living in a rural zip code was the strongest predictor of being ungeocoded. Other predictors for geocoding status varied with urbanity status. In urban areas, maternal race (adjusted odds ratio [aOR] ranging between 1.08 for Hispanic and 1.18 for black compared to white), maternal age [aOR: 1.16 (1.10–1.23) for ages 20–34 compared to <20], maternal nativity [aOR: 1.20 (1.15–1.25) for non-US versus US born], delivery at a birth center [aOR: 1.72 (1.49–2.00) compared to hospital delivery], multiparity [aOR: 0.91 (0.88–0.94)], maternal smoking [aOR: 0.82 (0.76–0.88)], and having nonprivate insurance [aOR: 1.25 (1.20–1.30) for Medicaid versus private insurance] were significantly associated with being ungeocoded. In rural areas, births delivered at birth center [aOR: 2.91 (1.80–4.73)] or home [aOR: 1.94 (1.28–2.95)] had increased odds compared to hospital births. The characteristics predictive of being ungeocoded were also significantly associated with adverse birth outcomes such as low birth weight and preterm delivery, and the association for maternal age was different when ungeocoded births were included and excluded.
Conclusions: Geocoding status is not random. Women with certain exposure-outcome characteristics may be more likely to be ungeocoded and excluded, indicating potential selection bias.

Birth certificates in state vital statistics systems are widely used in epidemiologic research [1–4]. The transition from article based to digital format has tremendously improved the timeliness and quality of birth records [5]. By applying geographic information system software to the electronic birth certificate, many state vital statistics programs are able to document maternal addresses at delivery through an automated geocoding technique. The software assigns geographic coordinates to an address based on spatial reference data, such as digital street maps. This automated technique has provided more georeferenced information for birth data and has enabled sophisticated spatial analyses in epidemiologic research, especially in environmental epidemiology [6]. As a result, there has been an increasing application of geocoded birth registry in studies of environmental risk factors for adverse birth outcomes [7–11].

The limitations of geocoded addresses including positional accuracy have been well studied in epidemiologic research [12–15]. As geocoding technique is a process through probability matching between addresses and spatial reference data, a proportion of unmatched records remains as a major problem. For example, the lack of address standardization, misspelling of street addresses, and the

* Corresponding author. Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M Health Science Center, 205A SRPH Administration Building, MS 1266 212 Adriance Lab Road, College Station, TX 77843-1266. Tel.: +1-979-436-9500; fax: +1-979-458-1877.
E-mail address: xiaohui.xu@sph.tamhsc.edu (X. Xu).

http://dx.doi.org/10.1016/j.annepidem.2016.01.002
1047-2797/

limited quality of spatial reference map (e.g., no updated street information) create conditions for misclassification or missing information. In many epidemiologic studies, addresses that fail to geocode have to be excluded from studies due to missing geographic information [8,16–18]. A study by Zimmerman et al [19] showed that approximately 10%–30% of records, even higher in some subgroups, would have to be excluded if only geocoded records were considered. This exclusion not only reduces the study sample size, but possibly introduces issues related to generalizability and selection bias.

Generalizability is the extent to which the results in a given study pertain to a broader population. If ungeocoded and geocoded populations are significantly different, then results may not be generalizable from one to the other. In addition, selection bias indicates a situation where the result in a study is not valid because of different sampling probabilities related to exposure-outcome cells. For example, if individuals with the exposure and outcome are less likely to be sampled than other cross-classified cells in a 2×2 table, then the measure of association will be biased toward the null. Without consideration of these issues, any epidemiologic study that is solely based on geocoded information may end up with unreliable conclusions. To the best of our knowledge, issues related to generalizability and potential selection bias arising from differential ungeocoding have received little attention in epidemiologic research.

In this retrospective cohort study, we used classification and regression tree (CART) analysis and logistic regression to explore generalizability related to exclusion of ungeocoded births by examining whether there were significant differences between geocoded and ungeocoded birth records. We further assessed whether there is potential selection bias by a direct analysis that involved determining (1) whether births with certain exposures (e.g., characteristics) were more likely to be ungeocoded (or excluded); (2) whether these exposures were associated with important adverse birth outcomes; and (3) whether the exposure-birth outcome relationships among the geocoded population and the entire population are different.

Methods

All singleton births in Florida in 2009 ($n = 210{,}285$) were identified from the Florida Birth Vital Statistics. Singleton births were chosen to avoid duplication of addresses. Latitude and longitude for all maternal addresses at delivery were provided by the state vital statistics program. These births were independently geocoded at street address level using North American Locator in ArcGIS 9.3 (ESRI, Redlands, CA) by the Florida Department of Health Vital Statistics. Spelling sensitivity was 80, and minimum matching score was 90 on the first round. Addresses that were not matched on the first round were screened and edited for spelling and random character issues and were rematched using the same criteria. After this round, addresses that were assigned latitude and longitude were defined as geocoded; the remaining addresses were defined as ungeocoded. We used only geocodes provided by the Florida Department of Health, which were based on street address for several reasons. First, a majority of those who were not geocoded based on home address during delivery in medical records often had missing address. Therefore, we could not obtain other geographic information for a different method of geocoding. Second, for those with only zip codes available, they may be systematically different from those with full address available. Therefore, geocoding births using two methods may introduce some information bias. Third, although using zip codes for all births may increase matching rates, this introduces another major issue involving positional accuracy. Specifically geocoding to zip code centroid can improve the match rate, but this information in some studies may not be very useful if exact location of an address is required (e.g., distance to highway calculation).

Characteristics such as demographics, behavioral factors, and adverse birth outcomes were used as potential predictors of geocoding status. For demographic factors, we assessed infant sex, maternal race, maternal age, maternal education, parental marital status, parity, maternal nativity, birth facility, and private versus public medical insurance as a proxy for socioeconomic status. For behavioral factors, we assessed tobacco and alcohol use during pregnancy, adequacy of prenatal care assessed by Kotelchuck index, and prepregnancy body mass index. As markers of adverse birth outcomes, we included low birth weight (LBW) and preterm delivery (PTD). LBW was defined as birth that was born less than 2500 grams. PTD was defined as a birth that occurred before 37 weeks of gestation. We determined the proportion of each Zip Code Tabulation Area (ZCTA) that falls within the urban areas defined by the 2010 US census [20]. We further defined urbanity of each zip code based on the following cutoff proportion: rural: $<5\%$, urban: $\geq 5\%$. We selected the cutoff of 5% because this proportion indicates the probability of the address located in the urban area within the specific ZCTA, and 5% is commonly used as a cutoff to indicate small probability events.

To examine the differences in the characteristics of geocoded and ungeocoded participants, we used CART. The details of this method have been previously described [21,22]. Briefly, CART is a nonparametric regression method that sequentially splits the data into dichotomous groups, such that each resulting group contains increasingly similar responses for the outcome. The end product of a typical CART analysis is a tree diagram illustrating the paths of dichotomous splits. Every tree starts with a root node, which contains all data from which the tree will be generated. Next, the data are split into two child nodes based on the values of an independent variable in a way that the observations within the two groups have the most similar responses for the outcome (i.e., minimizing residual sums of squares). The resulting child nodes contain a subset of the observations and are further split in the same manner until a preset stopping point is reached, in this analysis a P value <0.05 was set as statistically significant. The smallest resulting nodes are called terminal nodes. For each terminal node, the CART gives an estimation of the conditional probability of observations in each node having the given outcome (in this study, being ungeocoded). The CART offers several advantages. First, it makes no assumption about monotonic or parametric relationship between predictors and outcomes. Second, it can identify complex interactions among predictors without a priori specification. It also provides results that are easy to interpret. CART analyses were performed using the PARTY package in R.

We also used univariate and multivariable logistic regression to determine the odds ratios (OR) and 95% confidence intervals (CI) for the association between selected characteristics and geocoding status, and whether these differences persist after typical adjustment that is common in studies. We stratified our analyses by urbanity status due to the strong evidence of interaction between this variable and other predictors from the CART analyses. We also used logistic regression to determine the association between exposures predictive of geocoding status and common adverse birth outcomes including LBW and PTD. We repeated these analyses for both the geocoded group and the entire study sample. Logistic regression models were performed using SAS 9.4 (Cary, NC).

Results

Table 1 summarizes study participants' characteristics by geocoding status. During the study period, 11.5% of the study