Published in final edited form as:
Spat Spatiotemporal Epidemiol. 2012 April ; 3(1): 7–16. doi:10.1016/j.sste.2012.02.002.

A research agenda: Does geocoding positional error matter in health GIS studies?

Geoffrey Jacquez, PhD[President]
BioMedware, Inc., Adjunct Assoc. Prof. of Environmental Health Sciences, University of Michigan, P: 734-913-1098 x203, C: 734-223-3784, www.biomedware.com

Geoffrey Jacquez: jacquez@biomedware.com

Abstract

Until recently, little attention has been paid to geocoding positional accuracy and its impacts on accessibility measures; estimates of disease rates; findings of disease clustering; spatial prediction and modeling of health outcomes; and estimates of individual exposures based on geographic proximity to pollutant and pathogen sources. It is now clear that positional errors can result in flawed findings and poor public health decisions. Yet the current state-of-practice is to ignore geocoding positional uncertainty, primarily because of a lack of theory, methods and tools for quantifying, modeling, and adjusting for geocoding positional errors in health analysis.

This paper proposes a research agenda to address this need. It summarizes the basics of the geocoding process, its assumptions, and empirical evidence describing the magnitude of geocoding positional error. An overview of the impacts of positional error in health analysis, including accessibility, disease clustering, exposure reconstruction, and spatial weights estimation is presented. The proposed research agenda addresses five key needs: 1) A lack of standardized, open-access geocoding resources for use in health research; 2)A lack of geocoding validation datasets that will allow the evaluation of alternative geocoding engines and procedures; 3) A lack of spatially explicit geocoding positional error models; 4)A lack of resources for assessing the sensitivity of spatial analysis results to geocoding positional error; 5)A lack of demonstration studies that illustrate the sensitivity of health policy decisions to geocoding positional error.

Keywords

geocoding error; positional uncertainty; disease clustering; environmental exposures

1. Introduction

> “It is an unfortunate reality that even though a broad range of literature exists specifically geared to exposing how minor error in geocoding accuracy can affect results based on detailed spatial models, recent research initiatives continue to employ geocoded data without regard for how the accuracy can introduce possible inconsistencies or bias into the results.”

Corresponding Author: 121 W. Washington, 4th Floor – TBC, Ann Arbor, MI 48104734.913.1098 x203, 734.913.2201 fax, jacquez@biomedware.com.

- (Goldberg, Wilson et al. 2007)

Perhaps one of the foremost problems in measurement in the analysis of geographically referenced health data is that of geocoding positional error. Geographic models underpin concepts such as disease clustering, environmental exposure assessment, neighborhood context in health disparities analysis, accessibility to restaurants and parks in studies of overweight and obesity, and local availability to health care and screening facilities. But while geocoding – the process of converting text-based addresses into geographic coordinates – is a fundamental process in diverse disciplines including health (Boulos 2004; Rushton, Armstrong et al. 2006), criminal justice (Zandbergen and Hart 2009), political science (Haspel and Knotts 2005) and computer science (Hutchinson and Veenendall 2005), the sensitivity of spatial analysis results to positional error – the difference between a true location and that returned from the geocoded address – is not routinely addressed. In health analysis, recent studies have demonstrated that the strength of the odds relationship between disease exposures modeled at geocoded locations declines with decreasing geocoding accuracy, and that “estimated measures of positional accuracy must be used in the interpretation of results of analyses that investigate relationships between health outcomes and exposures measured at residential locations” (Mazumdar, Rushton et al. 2008). Yet the state-of-practice in health analysis is to ignore geocoding positional accuracy entirely.

The availability of georeferenced data in health analysis is expanding rapidly, due to several technological and policy trends. First, there is increased availability of user-generated, location-enabled health data as segments of the population become comfortable with sharing information through smart phones, web-browsers and other means; and as search engine keywords and social media are used to assess near real-time trends in health-related symptoms, medications, and outcomes (Ginsberg, Mohebbi et al. 2009; Wilson and Brownstein 2009; Seifter, Schwarzwalder et al. 2010). The confluence of crowd sourcing (e.g. “reflexive consumerism” where patients review hospitals and professionals on the web) and volunteer geographic information (VGI, where individuals report activities at their location) is enabling significant advances in disaster response, epidemiology and exposure assessment science (Goodchild and Glennona 2010; Adams 2011). For example, by coupling technologies for near real-time sensing of pollutants with location-enabled devices such as mobile phones, VGI is being used to validate model-based high spatial resolution exposure estimates. This makes possible validation of individual-level exposure estimates as a person goes about their daily activities (Jacquez and Meliker 2010; Stevens and D’Hondt 2010).

Second, the US health care system and the Department of Health and Human Services are investing heavily in interoperable electronic health records expected to revolutionize health care and disease control and surveillance. Recent national legislation such as the Health Information Technology for Economic and Clinical Health (HITECH) Act and the Affordable Care Act (ACA) include provisions requiring the collection of detailed electronic data in standardized format for insurance and care equity purposes (Weissman and Hasnain-Wynia 2011). Many of the data records for these systems include personal identifiers – names, addresses, and related health information – that can be used to construct georeferenced databases on patients, providers and health-related resources such as screening facilities.

Third, advances in spatio-temporal epidemiology facilitate reconstruction of geocoded residential histories of patients (Jacquez, Slotnick et al. 2011). The feasibility of developing reliable geospatial data retrospectively for large, epidemiological studies has been demonstrated, and revisiting completed studies using spatial epidemiological methods is now possible (Robinson, Wyatt et al. 2009). In an era of fiscal constraints expensive, large epidemiological studies are less likely to be funded. Application of spatiotemporal

epidemiology to completed case-control, cohort and longitudinal studies holds enormous promise for gaining new insights into disease causation that leverages our nation's existing investments in health research.

Despite the burgeoning of georeferenced health data cited in the preceding paragraphs, positional uncertainty is rarely accounted for in geospatial health analysis, even though it can lead to erroneous results sufficient to lead to incorrect conclusions and flawed health policy decisions, as detailed in section 3.

What is missing is a detailed understanding of empirical geocoding error distributions, theory underpinning the sources and propagation of such error through health decision making; models of positional error, and how it may be accounted for in geohealth analyses; tools for making such theory and models accessible to health practitioners; and resources such as databases for which empirical geocoding errors are known.

This paper proposes a research agenda to address these needs, and is organized as follows. The next section describes the basics of the geocoding process, the assumptions on which geocoding is based, and empirical data describing the magnitude of street geocoding positional error. This sets the stage for section 3, which presents impacts of positional error in health analysis, including accessibility evaluation, disease clustering, exposure reconstruction, and spatial weights estimation. Section 4 details important knowledge gaps and proposes a research agenda for advancing our ability to make informed and accurate decisions using uncertain geospatial health data.

2. Geocoding process, assumptions, sources and magnitude of positional error

Overview of geocoding process

Geocoding is the process of taking address information and converting it into geographic coordinates useful for health analysis. Several approaches have been proposed, including deterministic and probabilistic address matching, among others. All involve the input of an address to be geocoded, normalization and standardization of that address into an acceptable format typically comprised of an address number, street name, city or town name, state and ZIP code, and an iterative comparison of that address to a reference data set (e.g. streets database) from which the geographic coordinates can be calculated. This calculation typically is by interpolation along a street segment for which the geographic coordinates of the beginning and end points are known, and/or a real interpolation within a parcel, ZIP code, or city polygon. Further details on the geocoding process are provided elsewhere (Goldberg 2008).

When considering accuracy two aspects of the geocoding process are of interest: *completeness* (e.g. the proportion of addresses that successfully geocoded) and *positional accuracy* (e.g. how closely the geocoded coordinates correspond to the true coordinates). This paper is concerned with positional accuracy, and its impact on the results of geospatial health analyses.

Validity of geocoding assumptions and positional error

What assumptions of the geocoding process, when violated, introduce positional error? First, geocoding assumes that all of the addresses in the address range exist and can occupy space along the street segment (e.g. 600 through 649 Main Street, Figure 1). In reality, not all addresses on a street segment may be used, and one may find 600 Main Street adjacent to

610 Main Street. Addresses also may be created when urban infilling occurs, resulting in addresses such as 600 ½ Main Street.

Second, geocoding often assumes the parcels corresponding to these addresses are homogeneous in size and dimensions, and interpolation of addresses along a street segment is linear. This has been called the “parcel homogeneity assumption” (Dearwent, Jacobs et al. 2001), and may be violated when parcels are subdivided, or have been assigned different amounts of street frontage.

Third, geocoding may assume that the centroid of the parcel is an appropriate coordinate for the corresponding address, and that the projection of that centroid to the street corresponds with the placement of that address along the street segment. This assumption is invalidated when homes are not placed in parcel centers, and by irregularly shaped parcels for which the centroid falls outside the parcel confines.

Fourth, geocoding assumes that the reference road network database is an accurate representation of the real roads. This assumption is violated when street segments represented as lines are in fact curvilinear, and when the road network database is out of date or otherwise not representative of the actual street network. A noteworthy example of violations of this assumption is “Death by GPS”, in which drivers relying on GPS have been stranded in Death Valley, California (Knudson 2011), because the road network database showed roads that did not, in fact, exist. The British have coined the term “satnav mishaps” to describe accidents that occur when roads shown as navigable by GPS are inadequate for use by vehicles. In the UK truck drivers have toppled ancient stone walls and become stuck on Medieval lanes unable to move forward or turn around (Kovash 2011). In Europe, a van driver following his GPS became stuck on the path to a mountain peak and the vehicle and driver had to be extracted by helicopter (Reporter 2010).

Additional positional error may arise when the address is not well-formed and, for example, the centroid of the ZIP code, town or coordinates of the nearest road intersection is used instead of the street address. Violations of these assumptions, along with the strategies employed for handling incomplete addresses are the principal determinants of geocoding positional error.

Magnitude and characteristics of geocoding positional errors

“Cartographic confounding” (Oliver, Matthews et al. 2005) arises when positional accuracy is a function of location, and is widely acknowledged as a pervasive characteristic of geocoded data. Yet a single accuracy value for a dataset throughout the entire area of coverage is a fundamental assumption on which most spatial health analyses are founded (Goldberg et al. 2007). What are the magnitude and characteristics of geocoding positional errors?

A review (Abe and Stinchcomb 2008) of geocoding studies reported mean positional errors from 58 to 96 meters in urban areas, and from 129 to 614 in rural settings. Bonner et al. (Bonner, Han et al. 2003) reported mean geocoding errors of 96 meters in urban and 129 meters in rural settings. Cayo and Talbot (2003) found mean errors of 58 meters in urban and 614 meters in rural areas, and Ward et al. (Ward, Nuckols et al. 2005) reported mean positional errors of 77 meters in urban and 210 meters in rural settings. The maximum geocoding positional error observed in these studies was 18,742 meters, nearly 19 kilometers (Cayo and Talbot 2003). Positional errors typical of geocoded studies of cancer have been demonstrated to bias spatial regressions (Griffith, Millones et al. 2007), lead to incorrect exposure estimates (Zandbergen 2007), confute imputed health-environment relationships and underestimate odds ratios commonly used in cancer epidemiology

(Mazumdar, Rushton et al. 2008). Other researchers have reported similar empirical geocoding error distributions, but our understanding of how these errors vary from place to place and from one setting to another is limited.

Zandbergen summarized the characteristics of geocoding positional errors using four generalizations that hold in many circumstances (Zandbergen, Hart et al. 2011). *First, geocoding positional accuracy depends strongly on the extent rural*, such that geocoding is far more accurate in urban areas. In practice this likely is determined by the density of addresses in urban areas, which reduces variability in the frontage along street segments. *Second, the sampling distributions for geocoding positional errors is not always Gaussian*. In Iowa, a mixture of Gaussian and t-distributions provided a good fit to observed positional error distributions (Zimmerman, Fang et al. 2007), and positional errors for GPS locations, geocoded addresses, and LIDAR elevation data have been found to be approximated by Rayleigh, log-normal, and normal distributions (Zandbergen 2008). A simulation study reported that nearest neighbor spatial weights are not particularly sensitive to the shape of sampling distribution of geocoding positional errors (Jacquez and Rommel 2009). It is interesting to note that the US National Standard for Spatial Data Accuracy assumes the positional error of spatial data is normal, a standard that may bear revision (Zandbergen 2008). *Third, the direction of displacement of positional errors is not symmetric* and is influenced by the rectilinear conformation of the street network. When the network is oriented on cardinal directions such that streets run north-south and east-west, geocoding positional error will be largest along these axes. *Fourth, geocoding positional errors are spatially autocorrelated*, and may be related to the characteristics of the local street network (Zimmerman and Li 2010). At larger spatial scales autocorrelation in positional errors may be associated with the rural-urban gradient.

3. Impacts of geocoding positional error in health analysis

This section considers the impacts of geocoding positional error on five salient topics underpinning health analysis: measures of accessibility; estimation of local disease rates; disease cluster statistics; exposure analysis; and spatial weights. These have been selected because they provide the basis of health policy decision-making including targeting of interventions (based on accessibility, local incidence and mortality rates, as well as the identification of disease clusters); identification of causal factors leading to disease (exposure analysis), and the implementation of spatial models for interpolation and predictive modeling (spatial weights). This section begins with a brief overview of impacts of positional error on health analysis, and then summarizes the state of knowledge on how positional error affects accessibility measures; estimates of local disease rates; disease cluster statistics; exposure analysis; and spatial weight estimation.

Overview of impacts of positional error in health analysis

Positional error can have substantial impacts on the results of geohealth analyses leading to incorrect conclusions, false positives and false negatives. Estimates of accessibility to hospitals, clinics, screening facilities, and to neighborhood resources such as food outlets and exercise opportunities, depend critically on the street network database used for geocoding (Frizzelle, Evenson et al. 2009). Substantial bias may be expected whenever underlying risk factor(s) are associated with the probability of positional error (Bonner, Han et al. 2003; Ward, Nuckols et al. 2005; Zimmerman, Fang et al. 2008). The amount of bias depends on the extent rural, with bias higher in areas with small population densities (Oliver, Matthews et al. 2005; Kravets and Hadden 2007). Models have been developed for the shape of the probability distribution of geocoding errors (Zimmerman, Fang et al. 2007), and for estimating critical parameters such as spatial intensity and risk (Oliver, Matthews et al. 2005; Zimmerman 2006; Zimmerman and Sun 2006; Kravets and Hadden 2007).

Simulation studies of relationships between environmental exposures and health have demonstrated that the strength of the imputed odds relationship between exposures and health decline with decreasing geocoding accuracy (Mazumdar, Rushton et al. 2008), an effect of sufficient magnitude to lead to false findings and incorrect conclusions. Simulated positional errors in a case-control study of bladder cancer in Michigan resulted in weight matrices with more than 40% of first-order nearest neighbors incorrectly identified (Jacquez and Rommel 2009). The amount of error in the spatial weight matrices decreased as the number of nearest neighbors considered increased.

Impacts of positional error on estimates of accessibility

A frequent use of spatial analysis in health is to evaluate accessibility of food outlets (Ball, Timperio et al. 2009; Pearce, Hiscock et al. 2009; Páez, Gertes Mercado et al. 2010; Smith, Cummins et al. 2010), exercise opportunities (Davison and Lawson 2006; Oakes, Forsyth et al. 2007; Chin, Van Niel et al. 2008; Lovasi, Moudon et al. 2008), health resources (Brabyn and Skelly 2002; Luo and Wang 2003; Jordan 2008)and other features of the human landscape. This information informs health decision making (e.g. where to place a screening facility, where are local populations defined by poor access, etc.); is used in spatial models (e.g. to quantify travel distance to health-related resources); and to evaluate availability of neighborhood services and health-related resources (salubrious and otherwise). The road network used in an analysis can have dramatic impacts on estimates of access (Frizzelle, Evenson et al. 2009), and substantial positional errors have been demonstrated when using the Masterfile of the American Medical Association as a data source for identifying locations of physician offices (McLafferty, Freeman et al. 2011). Yet, overall, little is known of the impacts of geocoding positional accuracy on measures of accessibility, and on the results of analyses that uses such measures.

Impacts of positional error on estimates of local disease rates

Many analysts and researchers have assumed the impacts of positional error on estimates of local risk and disease rates (e.g. incidence, mortality, and rate ratios) are minimal. When calculating local rates, the assignment of place of residence to the appropriate a real unit is often taken to be negligible. This assumption makes even more sense when the areas used to accumulate the rates are large, such as counties. Recent results have demonstrated this is not the case. An article in this special issue evaluated how county-based cancer rates are affected by assignment of ZIP-code level incidence counts, coming from the California Cancer Registry, to California counties (Goldberg 2011). 9% of all counties in California experienced a rate change depending on how assignment was accomplished, with a maximum change in rate of 138% in Mono county, followed by a 103% change in Butte county. Overall, 7,515 cases, 10% of all cases, were miss-assigned.

Zimmerman developed measurement-error methods for intensity estimation for 2-dimensional inhomogeneous spatial point processes, and explored how positional error impacts estimates of the intensity of the Poisson process (Zimmerman and Sun 2006; Zimmerman 2008). The intensity of a Poisson process is the local density of events, for example, the local disease risk (cases per unit area). Ignoring positional errors may lead to biased estimates of local risk (usually towards under-estimation), a reduction in power to detect local elevations in risk, and to incorrect conclusions. Likelihood-based procedures for estimating the intensity and relative risk of Poisson spatial point processes permit valid inferences to be made from locations observed with positional error. But in practice disease rates calculated by researchers, as well as state and US health agencies, rarely account for positional accuracy, even though the statistical underpinnings are in place.

Impacts of positional error on disease cluster statistics

When positional error is present, reductions in the power to detect true elevations in risk have now been demonstrated for case-only and case-control spatial clustering methods such as Dale-Evans and Cuzick and Edwards test (Zimmerman 2008; Jacquez and Rommel 2009), and for space-time interaction tests including the Knox, Mantel and Jacquez's test (Jacquez and Waller 1997; Jacquez 1999). In general, positional errors affect disease clustering methods by underestimating local disease risk and by reducing statistical power, with documented impacts including inflation of standard errors of cluster parameter estimates and a reduction in power to detect spatial clusters (See (Zimmerman and Li 2010) for a review). Methods have been developed for propagating positional errors into disease cluster statistics using point, polygon, and population-based models of positional uncertainty (Jacquez 1996), but are not routinely applied in disease clustering, likely because of a lack of understanding on the part of many analysts of the magnitudes and adverse effects of geocoding errors.

Impacts of positional error on exposure estimates

That positional uncertainty can have an impact on exposure assignment has been demonstrated for air pollutants associated with proximity to roads, where distance of place-of-residence to road centerlines determines exposure assignment (Whitsel, Quilbrera et al. 2006; Zandbergen 2007); for exposure to agricultural herbicides and pesticides, where exposures are aggregated to spatially coarse resolutions such as the section level (1 square mile) and exposure is imputed at the place-of-residence (Ward, Nuckols et al. 2005; Maxwell 2011); and for drinking water contaminated by perfluorooctanoate (PFOA) as supplied to homes (Vieira, Howard et al. 2010), among others.

As a motivating example, consider the assignment of exposure to air pollutants. A recent study in Italy (Nuvolone, Maggiore et al. 2011) assessed relationships between exposure to air pollutants based on proximity to a major roadway and health outcomes, with health outcome data both self-reported and from biomarkers. They enrolled 2,841 subjects who completed a survey on personal habits, socio-demographic status, and health. A subset of these was tested for lung function and allergy response. Exposure was assigned based on distance to a main road: 0–100 meters were considered highly exposed; 100–250 meters moderately exposed, and 250–800 meters as not exposed. Statistical analyses were stratified by gender, and subjects in the high exposure class had increased adjusted risks for persistent wheeze, Chronic Obstructive Pulmonary Disease, reduced respiratory capacity, asthma, shortness of breath and positive skin prick test for allergies. However, geocoding positional error was not accounted for in the analysis, and the sensitivity of these results to positional error is unclear (Figure 2). It is likely the number of highly exposed individuals was over-estimated, based on analyses of street geocoding and exposure misclassification (Zandbergen 2007). In general, exposure mis-assignment may be expected to be higher when the buffer zones defining exposure classes (defined by distance from the exposure feature, such as the major road) are small relative to the magnitude of the positional error, as likely was the case in this example. In practice sensitivity analyses of exposure estimates to positional error are seldom accomplished, even though the potential for exposure mis-classification is substantial.

Impacts of positional error on spatial weights

Spatial weights are used in cluster analysis to quantify potential cluster membership, in exposure assignment to determine membership in buffer zones, in spatial regression to assess proximity between pairs of locations, and in geostatistics to calculate variograms that underpin the kriging system. Thus a formal understanding of how positional uncertainty

impacts spatial weights is an important prerequisite for propagating geocoding positional error into the results of cluster assessments, interpolation and predictive models.

As an example consider a simulation study that evaluated the impacts of positional error on nearest neighbor relationships. Jacquez and Rommel assessed the extent of bias that would be introduced by positional errors into nearest neighbor weights calculated for places of residence using the population densities of US counties and for a case-control study of bladder cancer in Michigan (Jacquez and Rommel 2009). The results were sobering, and indicate that, for typical geocoding positional error levels, and using population densities for US counties, the proportion of 1st nearest neighbors misidentified increases rapidly to near 100% as population density increased (Figure 3). Using residential locations from a study of bladder cancer in southeastern Michigan (from (Meliker, Slotnick et al. 2010)), Jacquez and Rommel reported the proportion of 1st nearest neighbor mismatches to be 8% at positional errors of 200m (Figure 4). Spatial weights are widely used in geographic health analyses, and further evaluation of the impacts of positional accuracy on spatial weights is urgently needed.

4. Discussion and research agenda

This section describes gaps in our knowledge and concludes with a proposed research agenda. This agenda is designed to increase our understanding of how positional accuracy impacts health policy decisions, and will provide the theory, methods and tools necessary to undertake effective health analysis using data subject to positional errors.

Knowledge gap: What are the empirical sampling distributions and spatial autocorrelation structures of geocoding positional error?

Although strides have been made in recent years towards increasing our knowledge of the sampling distributions of geocoding positional error, this has been accomplished for only a smattering of locations across the US. What are the observed distributions of positional error, and how do they vary from one place to another?

Knowledge gap: How does positional error impact statistical power?

In general, positional errors and incompleteness reduce statistical power for identifying trends, for detecting clusters, for assigning exposures, and for evaluating associations with covariates and other risk factors. Further, the extent of positional error is known to vary with covariates (e.g. extent rural), but geographic variation and reduction in power is not fully understood. The magnitude of reduction in power varies from study to study, and theory, methods and tools for routinely assessing this power reduction have yet to be developed.

Knowledge gap: How to predict error at specific locations?

The need for routine assessment of geocoding positional error in geospatial health studies is now widely acknowledged, but its measurement has proven difficult and time consuming (Rushton, Armstrong et al. 2006; Hofferkamp and Havener 2007; Goldberg 2008; Zimmerman 2008). Statistical models of geocoding positional error distributions have been developed (Zimmerman, Fang et al. 2007), but predicting the error at specific locations has yet to be addressed. This gap in our knowledge is critical, since error magnitude and variance have proven to be location-specific, with higher mean error in rural versus urban areas of sufficient magnitudes to bias the evaluation of risks and health-environment relationships (Zandbergen 2007; Zimmerman, Fang et al. 2008).

Knowledge gap: How to propagate positional error through health analyses?

Positional error propagation in environmental modeling has been an active research area for quite some time. Heuvelink (Heuvelink 1998) considered four techniques of location error propagation (pages 36–49): first and second-order Taylor expansions, Rosenblueths' method (Rosenblueth 1975) and Monte Carlo simulation. These techniques provide mechanisms for propagating positional as well as attribute error in a broad variety of GIS-based models. To date, simulation approaches are used in which a location error model is specified and the statistical power of a spatial analysis method under different error magnitudes is explored (Jacquez 1996; Jacquez and Waller 1997). This is painstaking and time-consuming, and is not routinely accomplished. And while methods exist for substantially improving positional accuracy using GPS, manual correction and other approaches (Goldberg, Wilson et al. 2008), these are labor-intensive and it makes sense to first determine whether improved accuracy is needed, and to prioritize locations that would benefit most from enhanced positional accuracy. No approaches are currently available for routinely assessing the impacts of positional error in health studies.

Research agenda

Address geocoding is now widely used to provide geographic coordinates of place of residence, care and screening facilities, environmental hazards (superfund sites, brown fields), neighborhood features (schools, restaurants, fast food outlets, playgrounds) and other places of direct relevance to human health. These coordinates are then used to evaluate spatial relationships defining measures of accessibility (e.g. to screening facilities); estimate exposures (e.g. distance to hazards; models of exposure to carcinogens); evaluate spatial patterns (e.g. disease clusters); and to construct models of health outcomes in relation to potential causal factors. *Yet despite its wide-spread use in such critical applications, standard resources for geocoding, geocoding accuracy assessment, and for evaluating the impacts of geocoding error on public health decisions are almost entirely lacking.* There are 5 critical research topics:

Research topic1: The lack of standardized, open-access geocoding resources for use in health research—Geocoding of health outcomes is now a standard procedure in many disease registries. Geocoding procedures differ from one registry to another, and from one cancer epidemiology study to another, because *there is no standard open-access geocoding resource whose positional accuracy and geocoding error rates are fully documented and under rigorous control.* Standardized resources are needed that provide better quantitative information about positional error arising from the geocoding processes. While methods for this are available, they are only infrequently used and most positional accuracy measures for health datasets are categorical (street-level, 5 digit ZIP code, etc.) and do not detail the empirical positional error distribution.

Research topic 2: The lack of geocoding validation datasets that will allow the evaluation of alternative geocoding engines and procedures—When documenting geocoding accuracy, the current state-of-the-art is to assess the percentage of addresses for which the geocoding engine provides a latitude and longitude. The positional accuracy of the geocoded data is rarely, if ever, assessed, yet is critical information. *Standardized, geocoding validation datasets are needed for representative rural and urban areas across the nation for which the true, geographic coordinates are known.* This will allow analysts to routinely quantify the positional accuracy of the geocoding procedures used in their studies.

Research topic 3: The lack of spatially explicit geocoding positional error models—Our understanding of empirical geocoding positional error distributions is

increasing but has been developed based on only a handful of settings across the US. *Spatially explicit, predictive models of geocoding positional error are required in order to estimate positional error for study areas and addresses that lack validation datasets.*

Research topic 4: The lack of tools for assessing the sensitivity of spatial analysis results to geocoding positional error—Methods now exist for quantifying the impact of positional error on spatial weights and the results of spatial analyses, but *easily used tools that enable researchers to assess the sensitivity of their results to geocoding positional error are not yet available*.

Research topic 5: The lack of demonstration studies that illustrate the sensitivity of public health decisions to geocoding positional error—Positional errors typical of geocoded studies have been demonstrated to bias spatial regressions (Griffith, Millones et al. 2007), lead to incorrect exposure estimates (Zandbergen 2007), confute imputed health-environment relationships and underestimate odds ratios commonly used in epidemiology and health analysis (Mazumdar, Rushton et al. 2008). Yet the importance of this problem is not widely recognized by public health and disease control professionals. *Incisive studies that evaluate the impacts of geocoding positional error in disease control and surveillance, and in public health decision making, are needed to increase awareness of this important problem.*

These research topics are synergistic with important interactions and positive feedbacks. For example, the routine assessment of the sensitivity of results to positional error, and of decisions based on those results, is immediately enabled once quantitative positional accuracy measures are recorded as part of a standardized geocoding process. This will motivate stakeholders such as researchers, health policy analysts and decision makers to improve their analytical and decision-making techniques that use georeferenced health data, thereby reducing positional error, ameliorating its effect on geospatial analyses, and improving the power of the data to increase our understanding of human health and its social and environmental determinants.

Acknowledgments

The manuscript has been substantially improved in response to the comments of two anonymous reviewers.

References

Abe, T.; Stinchcomb, D. Geocoding practices in cancer registries. In: Rushton, G.; Armstrong, MP.; Gittler, J., et al., editors. Geocoding Health Data. Boca Raton, FL: CRC Press; 2008. p. 111-125.

Adams SA. Sourcing the crowd for health services improvement: The reflexive patient and “share-your-experience” websites. Social Science & Medicine. 2011; 72(7):1069–1076. [PubMed: 21414701]

Ball K, Timperio A, et al. Neighbourhood socioeconomic inequalities in food access and affordability. Health & Place. 2009; 15(2):578–585. [PubMed: 19046654]

Bonner MR, Han D, et al. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology. 2003; 14(4):408–412. [PubMed: 12843763]

Boulos MNK. Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. International Journal of Health Geographics. 2004; 3(1)

Brabyn L, Skelly C. Modeling population access to New Zealand public hospitals. International Journal of Health Geographics. 2002; 1:3. [PubMed: 12459048]

Cayo M, Talbot T. Positional error in automated geocoding of residential addresses. International Journal of Health Geographics. 2003; 2(10)

Chin G, Van Niel K, et al. Accessibility and connectivity in physical activity studies: The impact of missing pedestrian data. Preventive Medicine. 2008; 46:41–45. [PubMed: 17920671]

Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations. Journal of the Royal Statistical Society Series B. 1990; (52):73–104.

Davison K, Lawson C. Do attributes in the physical environment influence children's physical activity? A review of the literature. International Journal of Behavioral Nutrition and Physical Activity. 2006; 3:19. [PubMed: 16872543]

Dearwent SM, Jacobs RR, et al. Locational uncertainty in georeferencing public health datasets. Journal of Exposure Analysis Environmental Epidemiology. 2001; 11(4):329–334. [PubMed: 11571612]

Fotheringham, AS. Geographically Weighted Regression. In: Fotheringham, AS.; Rogerson, P., editors. The Sage Handbook of Spatial Analysis. Sage Publications; 2009.

Frizzelle B, Evenson K, et al. The importance of accurate road data for spatial applications in public health: customizing a road network. International Journal of Health Geographics. 2009; 8(1):24. [PubMed: 19409088]

Ginsberg J, Mohebbi M, et al. Detecting influenza epidemics using search engine query data. Nature. 2009; (457):1012–1014. [PubMed: 19020500]

Goldberg, D. A Geocoding Best Practices Guide. Springfield, IL: North American Association of Central Cancer Registries; 2008.

Goldberg D. The effect of administrative boundaries and geocoding error on cancer rates. Spatial and SpatioTemporal Epidemiology. 2011 ((In Prep)).

Goldberg DW, Wilson JP, et al. From text to geographic coordinates: The current state of geocoding. URISA Journal. 2007; 19(1):33–46.

Goldberg DW, Wilson JP, et al. An effective and efficient approach for manually improving geocoded data. International Journal of Health Geographics. 2008; 7(60)

Goodchild MF, Glennona JA. Crowd sourcing geographic information for disaster response: a research frontier. International Journal of Digital Earth. 2010; 3(3):231–241.

Goovaerts, P. Geostatistics for Natural Resources Evaluation. New York: Oxford University Press; 1997.

Griffith DA, Millones M, et al. Impacts of Positional Error on Spatial Regression Analysis: A Case Study of Address Locations in Syracuse, New York. Transactions in GIS. 2007; 11(5):655–679.

Haspel M, Knotts HG. Location, location, location: precinct placement and the costs of voting. The Journal of Politics. 2005; 67(2):560–573.

Heuvelink, GBM. Error Propagation in Environmental Modelling. London: Taylor and Francis; 1998.

Hofferkamp, J.; Havener, L., editors. Data Standards and Data Dictionary. Vol. II. Springfield, IL: North American Association of Central Cancer Registries; 2007. Standards for Cancer Registries.

Hutchinson, M.; Veenendall, B. Towards using intelligence to move from geocoding to geolocating. Proceedings of the 7th Annual URISA GIS in Addressing Conference; Austin, TX. 2005.

Jacquez G. Spatial statistics when locations are uncertain. Geographic Information Sciences. 1999; 5(2):77–87.

Jacquez G, Meliker J, et al. In search of induction and latency periods: Space-time interaction accounting for residential mobility, risk factors and covariates. International Journal of Health Geographics. 2007; 6(1):35. [PubMed: 17716380]

Jacquez G, Rommel R. Local indicators of geocoding accuracy (LIGA): theory and application. International Journal of Health Geographics. 2009; 8(60)

Jacquez GM. Disease cluster statistics for imprecise space-time locations. Stat Med. 1996; 15(7–9): 873–885. [PubMed: 8861156]

Jacquez, GM.; Meliker, JR. Encyclopedia of Environmental Health. Elsevier; 2010. Exposure Reconstruction Using Space-Time Information Technology.

Jacquez GM, Slotnick MJ, et al. Accuracy of commercially available residential histories for epidemiologic studies. Am J Epidemiol. 2011; 173(2):236–243. [PubMed: 21084554]

Jacquez, GM.; Waller, L. The Effect of Uncertain Locations on Disease Cluster Statistics. In: Mowerer, HT.; Congalton, RG., editors. Quantifying Spatial Uncertainty in Natural Resources: Theory and Application for GIS and Remote Sensing. Chelsea MI: Arbor Press; 1997.

Jordan L. Improving livelihoods by improving access: The application of geospatial technologies to healthcare accessibility. Harvard Health Policy Review Spring. 2008; 2008:210–222.

Knox G. The detection of space-time interactions. Applied Statistics. 1964; 13:25–29.

Knudson, T. Sacramento Bee. Sacremento, CA: Bee State News; 2011. 'Death by GPS' in desert.

Kovash, J. Mountain Notebook. Moab: Mountain Gazette; 2011. Lost GPS drivers: Alarmed and dangerous.

Kravets N, Hadden W. The accuracy of address coding and the effects of coding errors. Health Place. 2007; 13:293–298. [PubMed: 16162420]

Lovasi G, Moudon A, et al. Using built environment characteristics to predict walking for exercise. (Research). International Journal of Health Geographics. 2008; 7:10. [PubMed: 18312660]

Luo W, Wang F. Measures of spatial accessibility to health care in a GIS environment: synthesis and a case study in the Chicago region. Environment and Planning B: Planning and Design. 2003; 30:865–884.

Manly, BFJ. Randomization, bootstrap and Monte Carlo methods in biology. CRC Press; 2007.

Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967; 27(2):209–220. [PubMed: 6018555]

Marshall R. A review of methods for the statistical analysis of spatial patterns of disease. Journal of the Royal Statistical Society Series A. 1991; 154:421–441.

Maxwell SK. Downscaling pesticide use data to the crop field level in California using Landsat satellite imagery: Paraquat case study. Remote Sensing. 2011; 3:1805–1816.

Mazumdar S, Rushton G, et al. Geocoding accuracy and the recovery of relationships between environmental exposures and health. International Journal of Health Geographics. 2008; 7(13)

McLafferty S, Freeman V, et al. Spatial Error in Geocoding Physician Location Data from the AMA Physician Masterfile: Implications for Spatial Accessibility Analysis. Spat Spattemporal Epidemiol. 2011 (In Press).

Meliker J, Slotnick M, et al. Lifetime Exposure to Arsenic in Drinking Water and Bladder Cancer: A Population-Based Case-Control Study in Michigan. Cancer Causes and Control. 2010; 21:745–757. [PubMed: 20084543]

Nuvolone D, Maggiore Rd, et al. Geographical information system and environmental epidemiology: a cross-sectional spatial analysis of the effects of traffic-related air pollution on population respiratory health. Environmental Health. 2011; 10(1):12. [PubMed: 21362158]

Oakes J, Forsyth A, et al. The effects of neighborhood density and street connectivity on walking behavior: the Twin Cities walking study. Epidemiologic Perspectives & Innovations. 2007; 4:16. [PubMed: 18078510]

Oliver WN, Matthews K, et al. Geographic bias relating to geocoding error in epidemiologic studies. International Journal of Health Geographics. 2005; 4(29)

Páez A, Gertes Mercado R, et al. Relative Accessibility Deprivation Indicators for Urban Settings: Definitions and Application to Food Deserts in Montreal. Urban Studies. 2010; 47(7):1415–1438.

Pearce J, Hiscock R, et al. A national study of the association between neighbourhood access to fast-food outlets and the diet and weight of local residents. Health & Place. 2009; 15(1):193–197. [PubMed: 18499502]

Reporter, D. Daily Mail. Associated Newspapers Ltd; 2010. Don't look down: White van and driver airlifted to safety after satnav error sends him to top of mountain.

Robinson JC, Wyatt SB, et al. Methods for Retrospective Geocoding in Population Studies: The Jackson Heart Study. Journal of Urban Health. 2009; 87(1):136–150. [PubMed: 20187277]

Rosenblueth E. Point estimates for probability moments. Proc Natl Acad Sci U S A. 1975; 72:3812–3814. [PubMed: 16578731]

Rushton G, Armstrong M, et al. Geocoding in cancer research -- A review. American Journal of Preventive Medicine. 2006; 30(2):S16–S24. [PubMed: 16458786]

Seifter A, Schwarzwalder A, et al. The utility of “Google Trends” for epidemiological research: Lyme disease as an example. Geospatial Health. 2010; 4(2):135–137. [PubMed: 20503183]

Smith DM, Cummins S, et al. Neighbourhood food environment and area deprivation: spatial accessibility to grocery stores selling fresh fruit and vegetables in urban and rural settings. International Journal of Epidemiology. 2010; 39(1):277–284. [PubMed: 19491142]

Stevens, M.; D’Hondt, E. UbiComp ’10. Copenhagen, Denmark: ACM; 2010. Crowd sourcing of Pollution Data using Smart phones.

Vieira V, Howard G, et al. Geocoding rural addresses in a community contaminated by PFOA: a comparison of methods. Environmental Health. 2010; 9(1):18. [PubMed: 20406495]

Ward MH, Nuckols JR, et al. Positional accuracy of two methods of geocoding. Epidemiology. 2005; 16(4):542–547. [PubMed: 15951673]

Weissman JS, Hasnain-Wynia R. Advancing Health Care Equity through Improved Data Collection. New England Journal of Medicine. 2011; (364):2276–2277. [PubMed: 21675885]

Whitsel E, Quilbrera P, et al. Accuracy of commercial geocoding: Assessment and implications. Epidemiol Perspect Innov. 2006; (3)

Wilson K, Brownstein J. Early detection of disease outbreaks using the Internet. CMAJ. 2009; (180): 829–831. [PubMed: 19364791]

Zandbergen PA. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. BMC Public Health. 2007; 16(7)

Zandbergen PA. Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy. Transactions in GIS. 2008; 12(1):103–130.

Zandbergen PA, Hart TC. Geocoding Accuracy Considerations in Determining Residency Restrictions for Sex Offenders. Criminal Justice Policy Review. 2009; 20(1):62–90.

Zandbergen PA, Hart TC, et al. Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets. Spat Spattemporal Epidemiol. 2011 (In Review).

Zimmerman, D. Estimating spatial intensity and variation in risk from locations coarsened by incomplete geocoding. Iowa City: University of Iowa; 2006.

Zimmerman, D. Statistical methods for incompletely and incorrectly geocoded cancer data. In: Rushton, G.; Armstrong, M.; Gittler, J., et al., editors. Geocoding Health Data. Boca Raton, PL: CRC Press; 2008.

Zimmerman D, Fang X, et al. Spatial Clustering of the Failure to Geocode and its Implications for the Detection of Disease Clustering. Stat Med. 2008; 27:4254–4266. [PubMed: 18407570]

Zimmerman D, Fang X, et al. Modeling the probability distribution of positional errors incurred by residential address geocoding. International Journal of Health Geographics. 2007; 6(1)

Zimmerman D, Li J. The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. International Journal of Health Geographics. 2010; 9(1):10. [PubMed: 20158886]

Zimmerman D, Li J. Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering. Statistics in Medicine. 2010; (29):1025–1036. [PubMed: 20087879]

Zimmerman, D.; Sun, P. Estimating spatial intensity and variation in risk from locations subject to geocoding errors. Iowa City: University of Iowa; 2006.

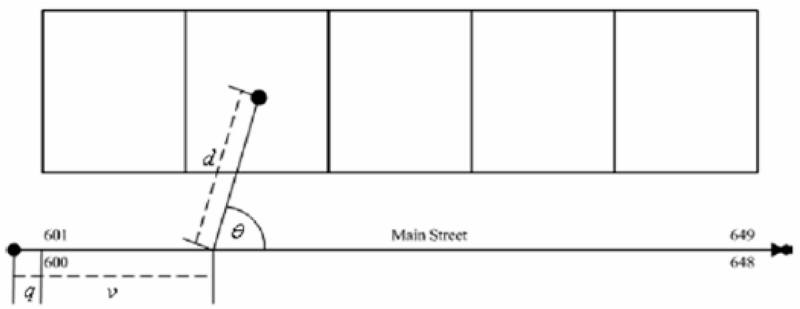

Figure 1.
Parameters of the geocoding algorithm, illustrating the parcel homogeneity assumption, a source of geocoding positional error. From Goldberg et al 2007.

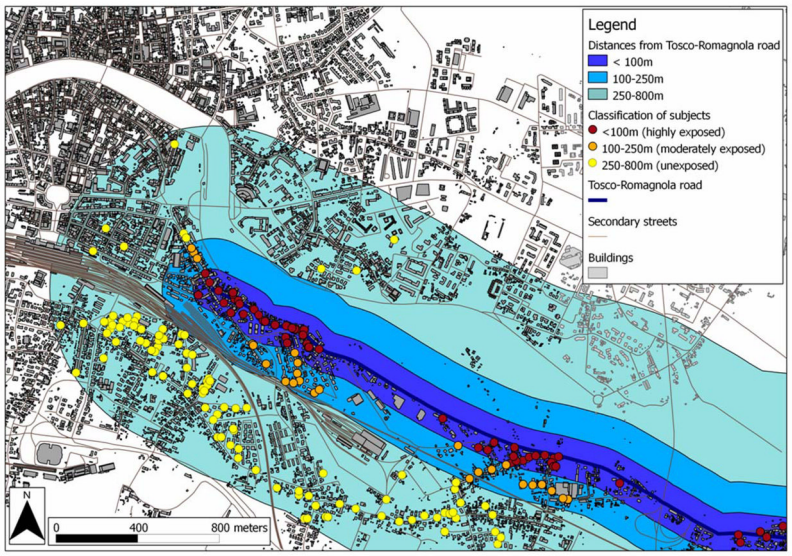

Figure 2.

Illustration of opportunity for exposure miss-assignment under geocoding positional error. This study assigned exposure for each subject based on the distance of each home from the main road. Highly exposed subjects live 0–100 meters from the road; moderately exposed live 100–250 m; and unexposed live between 250 and 800 meters from the road. A review of geocoding studies reported mean positional errors from 58 to 96 meters in urban areas (Abe and Stinchcomb 2008), sufficient to cause significant exposure assignment error when using buffers on the order of 100m. Map from Nuvolone et al. Environmental Health 2011 10:12.

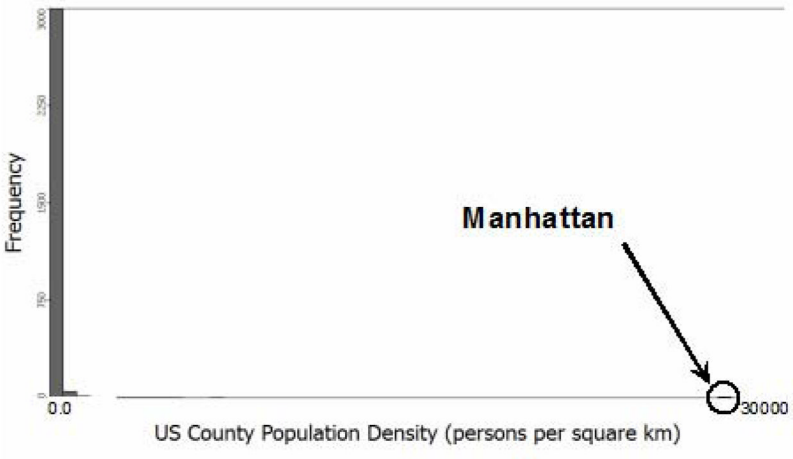

Fig 3a

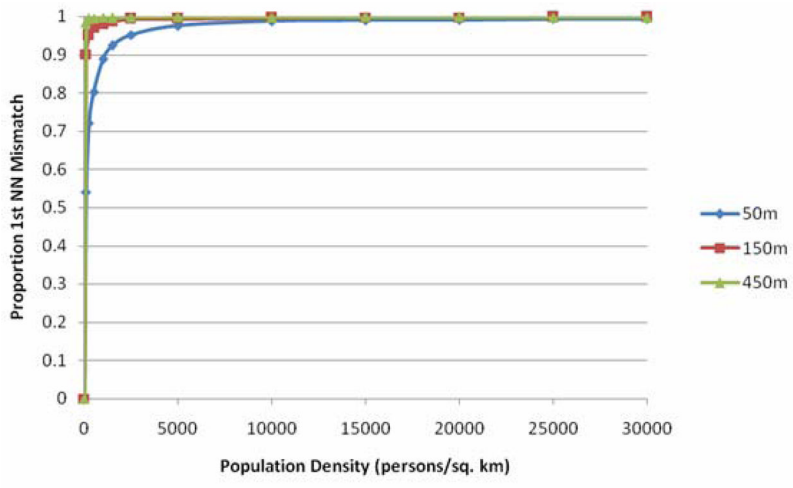

Fig 3b

Figure 3.

Impact of geocoding positional error on nearest neighbor relationships for population densities encountered across the United States. All simulations used a population of 500 individuals at the recorded densities. The different lines represent different mean geocoding error distances. Top: Population density distribution for US counties. Bottom: Proportion of 1st nearest neighbor mismatch versus representative county population densities at mean positional errors 50, 150 and 450 meters. Mean geocoding positional errors in applied studies range from 58 to 614 meters. At the population density for Manhattan virtually all first nearest neighbors are misidentified at the 3 positional error levels considered. From Jacquez and Rommel 2009.

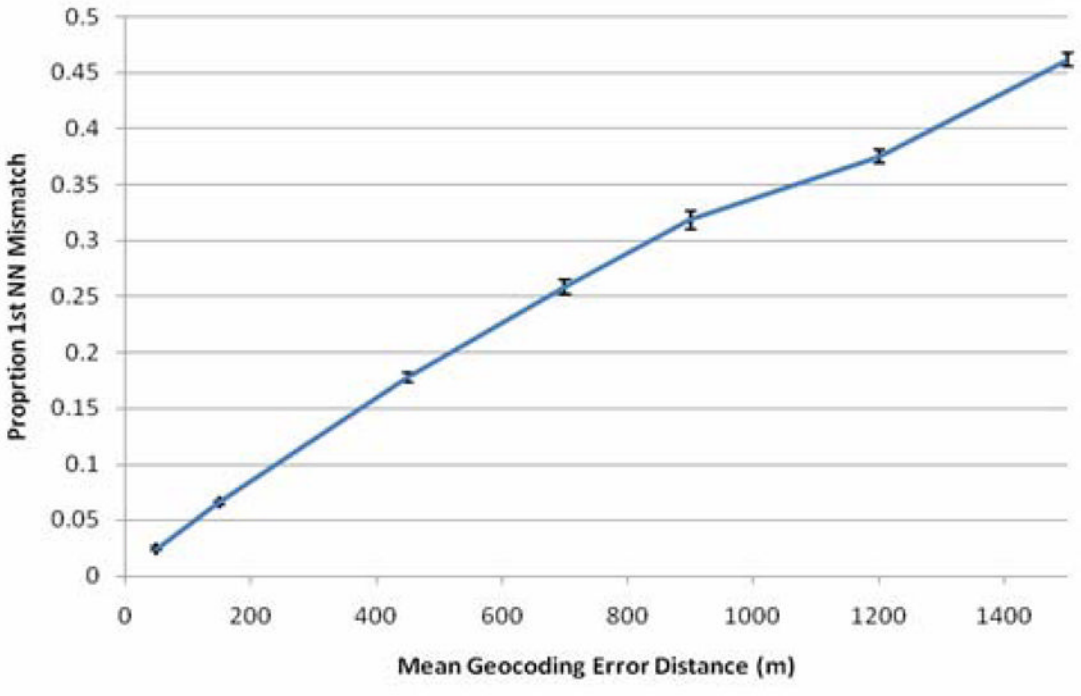

Fig 4a

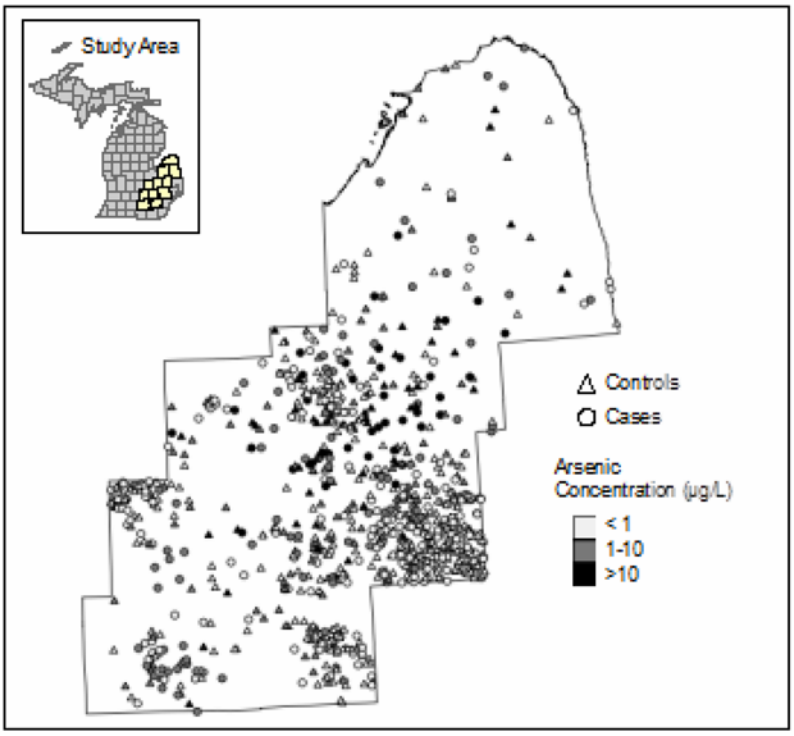

Fig 4b

Figure 4.

Impact of geocoding positional error on first nearest neighbor relationships for population densities encountered in a case-control study in Michigan versus mean geocoding error distance (meters). The confidence intervals shown are +/− one standard error from 10 simulations. From (Jacquez and Rommel 2009).