HHS Public Access

Author manuscript

Soc Sci Med. Author manuscript; available in PMC 2017 October 01.

Published in final edited form as:

Soc Sci Med. 2016 October ; 166: 49–56. doi:10.1016/j.socscimed.2016.08.009.

Confidentiality Considerations for Use of Social-Spatial Data on the Social Determinants of Health: Sexual and Reproductive Health Case Study

Danielle F. Haley[a], **Stephen A. Matthews**[b], **Hannah LF Cooper**[a], **Regine Haardörfer**[a], **Adaora A. Adimora**[c], **Gina M. Wingood**[d], and **Michael R. Kramer**[e]

[a]Department of Behavioral Sciences and Health Education, Rollins School of Public Health at Emory University, 1518 Clifton Road NE, Atlanta, GA, 30322 USA

[b]Departments of Sociology, Anthropology, and Demography, The Pennsylvania State University, 211 Oswald Tower, University Park, PA, 16802 USA

[c]Department of Epidemiology, UNC Gillings School of Global Public Health and Division of Infectious Diseases, School of Medicine, University of North Carolina at Chapel Hill, 130 Mason Farm Road, Chapel Hill, NC, 27599 USA

[d]Department of Sociomedical Sciences, Lerner Center for Public Health Promotion, Mailman School of Public Health at Columbia University, 722 West 168th Street, New York, NY, 10032 USA

[e]Department of Epidemiology, Rollins School of Public Health at Emory University, 1518 Clifton Road NE, Atlanta, GA, 30322 USA

Abstract

Understanding whether and how the places where people live, work, and play are associated with health behaviors and health is essential to understanding the social determinants of health. However, social-spatial data which link a person and their attributes to a geographic location (e.g., home address) create potential confidentiality risks. Despite the growing body of literature describing approaches to protect individual confidentiality when utilizing social-spatial data, peer-reviewed manuscripts displaying identifiable individual point data or quasi-identifiers (attributes associated with the individual or disease that narrow identification) in maps persist, suggesting that knowledge has not been effectively translated into public health research practices. Using sexual and reproductive health as a case study, we explore the extent to which maps appearing in recent peer-reviewed publications risk participant confidentiality. Our scoping review of sexual and reproductive health literature published and indexed in PubMed between January 1, 2013 and September 1, 2015 identified 45 manuscripts displaying participant data in maps as points or small-population geographic units, spanning 26 journals and representing studies conducted in 20 countries. Notably, 56% (13/23) of publications presenting point data on maps either did not

Corresponding Author Danielle F. Haley, Department of Behavioral Sciences and Health Education, Rollins School of Public Health at Emory University, 1518 Clifton Road NE, Atlanta, GA, USA, T: 919-357-1045, F: 404-727-1369, dfhaley@emory.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

describe approaches used to mask data or masked data inadequately. Furthermore, 18% (4/22) of publications displaying data using small-population geographic units included at least two quasi-identifiers. These findings highlight the need for heightened education for researchers, reviewers, and editorial teams. We aim to provide readers with a primer on key confidentiality considerations when utilizing linked social-spatial data for visualizing results. Given the widespread availability of place-based data and the ease of creating maps, it is critically important to raise awareness on when social-spatial data constitute protected health information, best practices for masking geographic identifiers, and methods of balancing disclosure risk and scientific utility. We conclude with recommendations to support the preservation of confidentiality when disseminating results.

Keywords

Social-Spatial Data; Sexual Health; Confidentiality; Quasi-identifiers; Maps

Introduction

Understanding whether and how the places where people live, work, and play are associated with health behaviors and outcomes is an essential underpinning of public health, as evidenced by the early work of John Snow (Snow, 1855) during the London cholera epidemic and the growing current interest in using geospatial data in public health research (e.g., Cooper et al., 2014; Diez-Roux, 2000; Law et al., 2004). The relative ease of using geographic information system (GIS) software (see Glossary of Key Terms, Figure 1), combined with the availability of individual-level population data ("microdata") have significantly expanded opportunities for public health researchers to explore relationships of place to health, and to visualize results using maps (Brownstein et al., 2006a; Chang et al., 2009; A. J. Curtis et al., 2006; Lozano-Fuentes et al., 2008; Palmer et al., 2013; Ruggles, 2014). The growing presence of datasets that link spatial information such as individual home address to individual attributes such as gender, race, or behaviors ("linked social-spatial data") facilitates this evolving research agenda. For people studying and intervening in sexual health, linked social-spatial data provide unique insight into etiologic patterns of disease (e.g., identifying spatial patterns of sexually-transmitted infections [STIs]), prevention planning (e.g., using space and time sampling methodologies to identify venues where high risk sexual and drug-use behaviors occur), and resource allocation (e.g., prioritization of HIV/AIDS funding to geographic areas with high prevalence of HIV infection).

While linked social-spatial data have clear benefits for public health research and interventions, their collection and use create potential risks (Wartenberg & Thompson, 2010). In 2006, a series of articles brought attention to the widespread publication of maps including unmasked individual-level point data (e.g., points representing the latitude and longitude of an individual's home), demonstrating the relative ease and troubling accuracy through which these points could be reverse coded to physical addresses (Brownstein et al., 2006a; Brownstein et al., 2006b; A. J. Curtis et al., 2006). Kounadi and Leitner demonstrated that between 2005 and 2012, the number of published articles including maps with unmasked individual point data in 19 GIScience and geography journals increased, potentially revealing more than 68,000 home addresses (Kounadi & Leitner, 2014).

Collectively, these findings underscore that despite growing literature raising the alarm about potential confidentiality breaches, as well as development of new methods for geomasking spatial data (e.g., Allshouse et al., 2010; Bader et al., 2016; Gutmann et al., 2008; Hampton et al., 2010; Kounadi et al., 2013; Kounadi & Leitner, 2015; Krumm, 2007; Seidl et al., 2015; VanWey et al., 2005), many public health researchers remain unaware of the potential risks and evolving solutions to help mitigate these risks. Our failure to effectively translate existing knowledge into practice may be due in part to the evolving intersection of two fields of inquiry (e.g., geography and public health). Discourse about geo-privacy as well as discussion of methods for effective geomasking have been concentrated largely in geography/GIScience and in highly specialized and/or technical journals (e.g., International Journal of Health Geographics, Statistics in Medicine). However, geocoding, mapping, and spatial analytic methodologies have simultaneously diffused to non-geographically trained investigators in public health. Given the rapid advances in technology, the absence of uniform guidelines for using linked social-spatial data for social and behavioral health research, and an absence of modules on linked social-spatial data in core training platforms for public health researchers (e.g., Human Subjects Protections, Good Clinical Practices), it is critically important that we raise awareness and educate investigators who may not otherwise be familiar with past work (Gutmann et al., 2008; National Research Council (U.S.). Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data., 2007; VanWey et al., 2005).

The aim of this paper is to bridge the critical gap between knowledge and practice by providing readers with a primer on key considerations for protecting participant confidentiality when disseminating study results generated from linked social-spatial data, including guidance on when geospatial data constitute protected health information (PHI) and current best practices for masking geographic identifiers. Using sexual and reproductive health, a field in which researchers routinely collect data on stigmatizing behaviors and health outcomes, as a case study, we characterize the extent to which peer-reviewed literature published and indexed in PubMed between January 1, 2013 and September 1, 2015 risks participant confidentiality by presenting maps with 1) unmasked point data or 2) small-population area-based geographic units that include additional demographic information associated with the individual or disease helping to narrow identification ("quasi-identifiers"). Geospatial data can be uniquely identifying when combined with quasi-identifiers (El Emam et al., 2010; Kounadi et al., 2013; Sweeney, 2000, 2002; VanWey et al., 2005). However, to our knowledge, no previous studies have assessed the extent to which maps in peer-reviewed publications risk participant confidentiality by including quasi-identifiers when presenting results using small-population area-based geographic units. In contrast to Kounadi and Leitner's review, we did not restrict this review to journals that specialize in GIScience and did not limit our review to maps that present participant data using points or trajectories. We conclude with recommendations to support the preservation of participant confidentiality when disseminating study results generated from linked social-spatial data.

Defining Linked Social-Spatial Data

Understanding when geospatial units constitute PHI is critical to discussions of maintaining confidentiality of linked social-spatial data for public health research (Nass et al., 2009). For the purposes of our analysis, we utilize the Health Insurance Portability and Accountability Act (HIPPA) Privacy Rule definition of PHI. We have selected this definition because it provides a clear, minimum standard of what constitutes PHI and many researchers exploring the social determinants of health in the United States (U.S.) are likely to draw data or work for agencies considered covered entities (e.g., designated health care groups, organizations, or businesses). However, the definition of PHI may vary based on additional institutional and funder regulations, and state and country guidelines (e.g., Boulos et al., 2009; El Emam et al., 2015a; Lovett et al., 2008; Yarmohammadian et al., 2010). We urge researchers using linking social-spatial data to determine the specific requirements for their own research.

According to the Privacy Rule, PHI is defined as “individually identifiable health information transmitted or maintained by a covered entity or its business associates in any form or medium” (Office of Clinical Research, 2012). Common identifiers include name and birth date. Geographic subdivisions smaller than a state (e.g., county, city, precinct, postal code, census tract, street address, latitude and longitude) are considered identifiable when *linked* to individual level health information (e.g., any information related to past, present, or future physical or mental health, including behaviors and health care utilization). For example, census tracts are geographic units utilized by the U.S. Census Bureau which contain on average 4,000 persons (though some have more or many less) and are typically homogeneous with respect to population characteristics, economic status, and living conditions (Krieger et al., 2003b). A dataset that combines an individual’s census tract with behavioral data (e.g., frequency of unprotected sex in the past six months) represents identifiable linked social-spatial data, even if the individual’s name or exact address has been stripped from the dataset.

Data are considered de-identified in accordance with the HIPPA Privacy Rule if the data do not “identify an individual and if the covered entity has no reasonable basis to believe it can be used to identify an individual” (Office of Clinical Research, 2012). There are two approaches to de-identify geographic information using these guidelines: 1) removing or aggregating geographic identifiers to large-population area-based units; and 2) applying statistical or scientific principles to render information not individually identifiable (“geomasking”) by a person with “appropriate knowledge and experience” (Office of Clinical Research, 2012).

Disclosure risk and approaches to protecting individual confidentiality

Disclosure risk associated with visualizing linked social-spatial data depends on the geographic coverage of the data, whether and how geographic units and individuals were sampled, the availability of quasi-identifiers, the availability of previously generated datasets or maps, and the heterogeneity of individuals and sample clusters (A. Curtis et al., 2011; El Emam et al., 2011; El Emam et al., 2015b; VanWey et al., 2005). To date, there is no universal standard for “adequate confidentiality protection” or “acceptable risk” (VanWey et al., 2005). Selecting an approach to reduce the probability of identifying individuals, while preserving the characteristics of the geographic data for valid inference depends in part on the nature of the data, acceptable confidentiality risk, and current and future use of the data (Armstrong et al., 1999; A. Curtis et al., 2011; El Emam et al., 2011; El Emam et al., 2015b; Kounadi & Leitner, 2015; Seidl et al., 2015; VanWey et al., 2005). For example, point data are more suitable for disease surveillance and outbreak investigation, but have a high risk of compromising individual identity if released publicly. In contrast, using area-based geographic units with larger population sizes may be less likely to compromise individual confidentiality (A. Curtis et al., 2011) and can be used to explore contextual associations of place and behavioral or disease outcomes (e.g., associations between living in high poverty areas and sexual risk behaviors), but are less sensitive to cluster detection and may fail to capture relationships that occur on a smaller geographic scale (Krieger et al., 2003a; Krieger et al., 2003b; Oakes & Kaufman, 2006). Table 1 provides an overview of several approaches to simultaneously preserve the confidentiality of individual records and the geographic attributes of the data ("geomasking"). Table 1 is based on a framework first presented by Armstrong and colleagues in 1999 (Armstrong et al., 1999), but has been expanded to include more recent adaptive geomasking and simulation techniques (Allshouse et al., 2010; Hampton et al., 2010; Wieland et al., 2008). For visual depiction of geomasking approaches, please see Kounadi and Leitner (Kounadi & Leitner, 2014) and Zandbergen (Zandbergen, 2014).

Case Study: Visual presentation of unmasked point data in recent journal publications related to sexual and reproductive health

Scoping Review Methods

We conducted a scoping review of all articles published and indexed in PubMed between January 1, 2013 and September 1, 2015 using the search terms "(map OR mapping OR geographical OR geographic OR GIS OR geospatial OR spatial) AND (sexual health OR sexual behavior OR reproductive health OR HIV OR AIDS OR STI)" and filters "English" and "Humans" (Figure 2). A scoping review is designed to rapidly assess a large volume of literature in order to provide an overview of the type, extent, and quantity of research on a given topic (Arksey & O'Malley, 2005; Levac et al., 2010; Littell et al., 2008). We based key words on the methods utilized by Brownstein and colleagues (Brownstein et al., 2006a; Brownstein et al., 2006b), but included additional content-specific key words in order to limit the search to content areas related to sexual and reproductive health. We first excluded articles that were not relevant to the review based on the title and abstract (e.g., laboratory/molecular science, fields unrelated to sexual and reproductive health [e.g., brain science, audiology], reviews and commentaries). Secondly, all remaining articles were inspected to determine whether they included maps. For articles including maps, we reviewed the methods, results, and figures to determine whether the maps displayed linked social-spatial data, and if so, the unit at which data were presented. For maps presenting data as points or geographic units with average population sizes ≤30,000, we further reviewed the methods, results, and figures to determine if additional quasi-identifiers were presented, and for point data, whether and how the data were masked. For articles presenting point data, we also researched the data source in order to identify whether the authors utilized secondary datasets in which geographic identifiers were masked by the data custodian prior to release. We selected 30,000 as a cut-off because past studies have successfully identified individuals from publically-available datasets including postal code information (Sweeney, 2002), and the average size of a U.S. postal code is 30,000. We categorized articles that presented maps as follows:

1. Maps did not display linked social-spatial data: Articles in this category included maps that did not present linked social-spatial data, such as maps for reference purposes only (e.g., map of country where research was conducted) or maps that presented locations of non-human subjects (e.g., HIV testing clinics), or microdata on individuals only (e.g., data obtained from U.S. Census).

2. Data aggregated to units with average populations larger than 30,000 people: Articles in this category included maps displaying linked social-spatial data at an aggregated unit with an average population size greater than 30,000 people and were unlikely to risk individual participant confidentiality.

3. Data aggregated to units with average populations ≤30,000 people: Articles falling in this category included maps displaying linked social-spatial data at an aggregated unit with an average population size ≤30,000 people that could be used as a proxy for a neighborhood or community or could risk individual confidentiality in the presence of quasi-identifiers.

4. Point data: Articles in this category included maps that presented point data representing home addresses or individual trajectories associated with confidential individual-level information. This category was further subdivided as follows:

   a. Masked data included maps that presented point data and either included information on whether or how the points were masked or utilized datasets that were masked by the data custodian prior to release for secondary data analyses.

   b. Insufficiently masked point data: In reviewing articles in the “masked’ category, we determined that a number of manuscripts described masking data points included in maps, but subsequently included sufficient information in the methods, results, or maps for the authors to question whether participant confidentiality was adequately protected (e.g., presence of multiple quasi-identifiers, assigning points to the nearest intersection of two streets and providing street-level maps).

   c. Unmasked or no masking information included maps that presented point data and did not include information on whether or how the points were masked. This classification is consistent with past reviews (Brownstein et al., 2006b; Kounadi & Leitner, 2014).

All manuscripts were reviewed and categorized as described above by DFH. During this process, manuscripts that could not be clearly categorized were discussed and categorized based on consensus by DFH, SAM, and MRK. All manuscripts categorized as "insufficiently masked point data" were reviewed and categorized based on consensus by DFH, SAM, and MRK. As a final step, a random selection of 20 (43%) manuscripts displaying point data or data aggregated to units with average populations ≤30,000 people was reviewed and categorized by MRK. This categorization was compared with the primary coder (DFH) and discrepancies discussed. Final intercoder reliability was 95%.

Scoping Review Results

As outlined in Figure 2, our review of 1,171 manuscripts published and indexed in PubMed between January 1, 2013 and September 1, 2015 identified 151 manuscripts related to sexual and reproductive health including maps. Of these 151 manuscripts, 73 (48%) included maps that did not display linked social-spatial data and 33 (22%) displayed data at a geographic unit with an average population size greater than 30,000. Forty-five manuscripts (30%) presented linked social-spatial data in maps using geographic units with average population sizes ≤30,000 or points. These 45 studies were published in 26 journals and represent data from 20 countries.

Twenty-three manuscripts displayed maps with linked social-spatial point data. Of these 23 manuscripts, 10 (43%) manuscripts presented masked point data: six utilized secondary datasets masked by the data custodian prior to release (i.e., DHS Program Demographic and Health Surveys), two displayed data in the absence of geographic references, and two stated data were "anonymized" but did not provide sufficient information for us to further assess masking. Thirteen manuscripts (56%) displayed maps with linked social-spatial point data that either did not include details on whether or how data were masked (n=8) or were insufficiently masked based on the methods described (n=5). For example within the "insufficiently masked" category, one manuscript noted that household points had been randomly moved a distance likely to locate the point within the same neighborhood/nearby properties. Another masked residential addresses by displaying point data for individuals in a small statistical area centroid, but displayed these points in a series of maps allowing the reader to determine not only the statistical area, but also the race, ethnicity, gender, and sexual orientation of the individual. A third manuscript displayed point data within a well-demarcated community. Although the author did not include specific geographic references in the text or map, previous publications utilizing the same dataset presented aerial photos, allowing us to identify the community and even specific households identified in the figure. These 13 manuscripts, published in 9 different journals and representing data from five countries, represent potential confidentiality breaches of 14,581 study individuals, the majority of which reported highly stigmatized behaviors (e.g., men who have sex with men [MSM], "high risk" sexual behaviors, or illicit drug use) or health conditions (e.g., HIV infection, tuberculosis). Notably, 7 of these 13 studies utilized data from U.S. populations.

Twenty two manuscripts displayed spatial data using geographic units with average population sizes ≤30,000. Four (18%) of the manuscripts in this category provided at least two additional quasi-identifiers (e.g., race, ethnicity, gender, sexual orientation, etc.) on the map or in the manuscript text. Depending on the area's sociodemographic composition, these manuscripts potentially compromised the confidentiality of up to 668 individuals. Representing data from three countries, these four studies were published in four journals. Two of these four studies utilized data from U.S. populations.

Discussion

Our scoping review identified 17 manuscripts related to sexual and reproductive health published and indexed in PubMed between January 1, 2013 and September 1, 2015 presenting insufficiently masked point data or small-population geographic units with quasi-identifiers or did not include details on whether or how data were masked, potentially compromising the confidentiality of study participants. Similar to Kounadi and Leitner, we found that over half of manuscripts including maps with point data presented point data that were either unmasked or did not include details on whether or how data were masked (Kounadi & Leitner, 2014). Notably, our review identified 45 publications including maps representing point data or small-population areal units. These manuscripts spanned 26 different journals and included data from 20 countries, underscoring the 1) broad use and publication of linked social-spatial data to explore the social determinants of health and 2) urgent need to ensure that researchers utilizing these data are well-versed on confidentiality considerations associated with using linked social-spatial data and approaches to mitigate these risks. Notably, the vast majority of studies presenting unmasked or insufficiently masked point data were based on U.S. populations. It is likely that our review, which does not extend to products not subject to the rigor of peer-review (e.g., reports, presentations), *underestimates* the extent to which presented maps compromise individual confidentiality.

The potential for harm associated with confidentiality breaches is particularly salient for individuals associated with stigmatizing behaviors (e.g., injection drug use) or conditions (e.g., HIV infection). Despite the growing body of literature describing approaches to preserving individual confidentiality when utilizing linked social-spatial data —and multiple layers of review by authors, reviewers, and editorial staff required prior to publication— peer-reviewed manuscripts which display identifiable individual point data or include quasi-identifiers in maps persist. These transgressions violate a fundamental ethical obligation to protect individual confidentiality and may be due in part to a lack of uniform guidelines and rapid advances in technology (Chang et al., 2009).

Our discussion of the confidentiality considerations surrounding the use of linked social-spatial data to explore the social determinants of sexual health follows the long debated challenge of how best to balance individual interests and the health of the public (Wartenberg & Thompson, 2010). The tension between preserving confidentiality while also ensuring scientific utility is evident when considering linked social-spatial data. Approaches are needed which simultaneously protect individual confidentiality while also maintaining spatial attributes of the data. We have a commitment to individuals involved in research to protect their data. However, disseminating study results is not only critical to advancing science, but also maximizes value from public dollars spent on research and demonstrates respect for an individual's time and efforts. Good stewardship is a key factor in ensuring these data continue to be collected and made available. Several manuscripts identified in our review attempted to mask individual point data, but did so inadequately or incompletely described efforts taken. Given the confidentiality risks associated with publishing point data on maps, even when geomasked, researchers should have clear justification for the added value of presenting data at this level (Brownstein et al., 2006a; Brownstein et al., 2006b; A. J. Curtis et al., 2006; Kounadi & Leitner, 2014). In many instances it may be possible for investigators to use point data or small areal units in analysis, but present summaries of their results as cluster statistics, aggregated maps, or tabular data, thus limiting the public dissemination of the PHI used in analysis. Decisions about whether to include maps, and the spatial unit if included, should be based on confidentiality and social harm considerations (including potential stigmatization of communities) as well as scientific utility. For example, although visually appealing, presenting point data or small areal units may not provide any additional information beyond what is already provided in study tables or results (e.g., cluster statistics). Alternatively, patterns or spatial distribution of data may be presented using other formats, such as aggregating point data to large-population areal units or presenting data in the absence of all geographic context. For an example of presenting data in the absence of all geographic context, see Chamie and colleagues (Chamie et al., 2015). However, even when presenting alternative formats, unintended consequences should be considered. For example, releasing or displaying HIV prevalence rates by smaller area-based units or mapping areas where "high risk" groups congregate may help HIV service organizations identify how best to allocate limited resources (Lorway & Khan, 2014). In contrast, this same information may result in marginalization of neighborhoods and labeling of its inhabitants as "high risk". Notably, venue-mapping may risk social harms, particularly if maps display venues where congregants engage in illegal or highly stigmatized behaviors (e.g., MSM venues in countries where homosexuality is illegal). Decisions on whether and how to display spatial data visually, including whether maps should made available only through restricted access, should be made and implemented by teams with sufficient expertise in the analytic methods being applied, regulations, and the topic area.

We aim to increase awareness and inform future dialogue so that researchers, editors, and other public health professionals can make informed decisions on how best to disseminate findings. Based on the results of our scoping review and review of current best practices, we recommend the following:

1. Develop and include a module about utilizing spatial data as a standard component of Human Subjects Training for all professional in the field, including research and editorial staff. The Collaborative Institutional Training Initiative (CITI Program) provides peer-reviewed web-based Human Subjects Training for academic, government, and commercial organizations globally and is utilized by numerous organizations likely to engage in place-based human subjects research (Collaborative Institutional Training Initiative). Given the broad reach of this organization, the use of peer-review, and the established record in providing Human Subjects Training, CITI represents a likely institution to lead this effort.

2. Include modules on confidentiality and stigma considerations in GIS tutorials and in academic coursework. This manuscript, including the citations included herein, draws upon both seminal and emerging work in this field and is intended to serve as a reference on confidentiality and stigma considerations when utilizing social-spatial data. The GIS&T Body of Knowledge (http://www.aag.org/bok/), a free online reference presenting a variety of topics relevant to GIScience may also serve as an additional resource (Ahearn et al., 2013; DiBiase et al., 2007).

3. Continue funder support for research to determine levels of privacy protection and scientific utility provided by geomasking, including acceptable confidentiality protections for dissemination by user and research stage. For example, the National Institutes of Health have demonstrated a commitment in this area, as evidenced by the Big Data to Knowledge (BD2K) Initiative (Margolis et al., 2014) and the 2016 Conference on Geospatial Approaches to Cancer Control and Population Sciences (National Cancer Institute, 2016).

4. Establish uniform reporting requirements for presenting linked social-spatial data, including (a) what geographic unit(s)/population size(s) of data may be presented; (b) guidelines for descriptions of methods used to protect individual confidentiality in publications; and (c) standard editorial procedures, including reviewer evaluative criteria, for ensuring published maps do not risk individual confidentiality. At minimum, any manuscript presenting social-spatial data using maps should be evaluated for potential confidentiality and stigma considerations by a peer-reviewer(s) with sufficient expertise in the topic area and methodology presented. However, evaluation of confidentiality considerations requires that authors report masking approaches as they would other aspects of their research (e.g., research design, analytic approaches). As noted previously, a number of the manuscripts in the "Unmasked or no masking information" category presented point data in maps but did not describe whether or how points were masked. It is possible that these authors took adequate precautions to protect participant confidentiality, but did not include this information in the manuscript due to word limitations, underscoring the need for editorial teams and reviewers to emphasize the value of this information. Of note, there may be circumstances when not fully describing masking methods is preferable (e.g., detailed description of masking facilitates reverse identification). In these instances, authors should state that they intentionally left masking procedures vague in order to protect participant confidentiality and potentially provide additional details upon request.

5. Expand CONSORT (Campbell et al., 2012) and TREND (Des Jarlais et al., 2004) reporting guidelines for randomized and non-randomized designs to include reporting of methodologies specific to linked social-spatial data (e.g., geomasking). The use of standardized reporting improves the quality reporting of research in peer-reviewed publications

(Plint et al., 2006; Turner et al., 2012) and numerous high impact peer-reviewed public health journals already endorse the use of CONSORT (e.g., The Lancet) and TREND (e.g., AIDS and Behavior).

The widespread availability of place-based data and the emerging nature of publicly available spatial data which capture activity spaces (e.g., point data collected from Twitter, geolocating phones using apps) will expand our ability to explore whether and how place contributes to the health of individuals and communities (Brownstein et al., 2006a; Chang et al., 2009; A. J. Curtis et al., 2006; Duncan et al., 2014; Lozano-Fuentes et al., 2008; Palmer et al., 2013). While these advances offer exciting research opportunities to improve the public's health, these rapidly advancing technologies highlight the challenges associated with establishing guidelines for utilizing linked social-spatial data to explore the social determinants of health, and underscore the need for ongoing dialogue across key stakeholders (e.g., editors, public health professionals, data custodians, community-based organizations, communities) and leadership by professional organizations (e.g., International Committee of Medical Journal Editors, International AIDS Society, Association of American Geographers, International Union for the Scientific Study of Population) so as to preserve individual confidentiality and minimize group-level social harms while maximizing the benefit of this research for society at large.

Acknowledgments

The authors thank the Reviewers for their thoughtful critique of and valuable contributions to this manuscript. DF Haley was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number F31MH105238 and the George W. Woodruff Fellowship of the Laney Graduate School. MR Kramer was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number K01HD074726. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Ahearn S, Icke I, Datta R, DeMers M, Plewe B, Skupin A. Re-engineering the GIS&T Body of Knowledge. International Journal of Geographical Information Science. 2013; 27:2227–2245.

Allshouse WB, Fitch MK, Hampton KH, Gesink DC, Doherty IA, Leone PA, et al. Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. Geocarto Int. 2010; 25:443–452. [PubMed: 20953360]

Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International journal of social research methodology. 2005; 8:19–32.

Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. Stat Med. 1999; 18:497–525. [PubMed: 10209808]

Bader MD, Mooney SJ, Rundle AG. Protecting Personally Identifiable Information When Using Online Geographic Tools for Public Health Research. Am J Public Health. 2016; 106:206–208. [PubMed: 26794375]

Boulos MN, Curtis AJ, Abdelmalik P. Musings on privacy issues in health research involving disaggregate geographic data about individuals. Int J Health Geogr. 2009; 8:46. [PubMed: 19619311]

Brownstein JS, Cassa CA, Kohane IS, Mandl KD. An unsupervised classification method for inferring original case locations from low-resolution disease maps. Int J Health Geogr. 2006a; 5:56. [PubMed: 17156451]

Brownstein JS, Cassa CA, Mandl KD. No place to hide—reverse identification of patients from published maps. N Engl J Med. 2006b; 355:1741–1742. [PubMed: 17050904]

Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012; 345:e5661. [PubMed: 22951546]

Cassa CA, Wieland SC, Mandl KD. Re-identification of home addresses from spatial locations anonymized by Gaussian skew. Int J Health Geogr. 2008; 7:45. [PubMed: 18700031]

Chamie G, Wandera B, Marquez C, Kato-Maeda M, Kamya MR, Havlir DV, et al. Identifying locations of recent TB transmission in rural Uganda: a multidisciplinary approach. Trop Med Int Health. 2015; 20:537–545. [PubMed: 25583212]

Chang AY, Parrales ME, Jimenez J, Sobieszczyk ME, Hammer SM, Copenhaver DJ, et al. Combining Google Earth and GIS mapping technologies in a dengue surveillance system for developing countries. Int J Health Geogr. 2009; 8:49. [PubMed: 19627614]

Collaborative Institutional Training Initiative. Mission and History.

Cooper HL, Linton S, Haley DF, Kelley ME, Dauria EF, Karnes CC, et al. Changes in Exposure to Neighborhood Characteristics are Associated with Sexual Network Characteristics in a Cohort of Adults Relocating from Public Housing. AIDS Behav. 2014

Curtis A, Mills JW, Agustin L, Cockburn M. Confidentiality risks in fine scale aggregations of health data. Computers, Environment and Urban Systems. 2011; 35:8.

Curtis AJ, Mills JW, Leitner M. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. Int J Health Geogr. 2006; 5:44. [PubMed: 17032448]

Des Jarlais DC, Lyles C, Crepaz N, Group T. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Journal Information. 2004:94.

DiBiase D, DeMers M, Johnson A, Kemp K, Luck A, Plewe B, et al. Introducing the first edition of geographic information science and technology body of knowledge. Cartography and Geographic Information Science. 2007; 34:113–120.

Diez-Roux AV. Multilevel analysis in public health research. Annu Rev Public Health. 2000; 21:171–192. [PubMed: 10884951]

Duncan DT, Regan SD, Shelley D, Day K, Ruff RR, Al-Bayan M, et al. Application of global positioning system methods for the study of obesity and hypertension risk among low-income housing residents in New York City: a spatial feasibility study. Geospat Health. 2014; 9:57–70. [PubMed: 25545926]

El Emam K, Brown A, AbdelMalik P, Neisa A, Walker M, Bottomley J, et al. A method for managing re-identification risk from small geographic areas in Canada. BMC Med Inform Decis Mak. 2010; 10:18. [PubMed: 20361870]

El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One. 2011; 6:e28071. [PubMed: 22164229]

El Emam K, Jonker E, Arbuckle L, Malin B. Correction: a systematic review of re-identification attacks on health data. PLoS One. 2015a; 10:e0126772. [PubMed: 25880057]

El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ. 2015b; 350:h1139. [PubMed: 25794882]

Gutmann M, Witkowski K, Colyer C, O'Rourke JM, McNally J. Providing Spatial Data for Secondary Analysis: Issues and Current Practices relating to Confidentiality. Popul Res Policy Rev. 2008; 27:639–665. [PubMed: 19122860]

Hampton KH, Fitch MK, Allshouse WB, Doherty IA, Gesink DC, Leone PA, et al. Mapping health data: improved privacy protection with donut method geomasking. Am J Epidemiol. 2010; 172:1062–1069. [PubMed: 20817785]

Kounadi O, Lampoltshammer TJ, Leitner M, Heistracher T. Accuracy and privacy aspects in free online reverse geocoding services. Cartography and Geographic Information Science. 2013; 40:140–153.

Kounadi O, Leitner M. Why does geoprivacy matter? The scientific publication of confidential data presented on maps. J Empir Res Hum Res Ethics. 2014; 9:34–45. [PubMed: 25747295]

Kounadi O, Leitner M. Defining a Threshold Value for Maximum Spatial Information Loss of Masked Geo-Data. ISPRS International Journal of Geo-Information. 2015; 4:572–590.

Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures--the public health disparities geocoding project. Am J Public Health. 2003a; 93:1655–1671. [PubMed: 14534218]

Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas--the Public Health Disparities Geocoding Project. Am J Public Health. 2002; 92:1100–1102. [PubMed: 12084688]

Krieger N, Waterman PD, Chen JT, Soobader MJ, Subramanian SV. Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: geocoding and choice of area-based socioeconomic measures--the public health disparities geocoding project (US). Public Health Rep. 2003b; 118:240–260. [PubMed: 12766219]

Krumm, J. Pervasive 2007. Toronto, Ontario, Canada: 2007. Inference Attacks on Location Tracks.

Law DC, Serre ML, Christakos G, Leone PA, Miller WC. Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies. Sex Transm Infect. 2004; 80:294–299. [PubMed: 15295129]

Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci. 2010; 5:69. [PubMed: 20854677]

Littell, JH.; Corcoran, J.; Pillai, VK. Systematic reviews and meta-analysis. Oxford; New York: Oxford University Press; 2008.

Lorway R, Khan S. Reassembling epidemiology: mapping, monitoring and making-up people in the context of HIV prevention in India. Soc Sci Med. 2014; 112:51–62. [PubMed: 24797356]

Lovett R, Fisher J, Al-Yaman F, Dance P, Vally H. A review of Australian health privacy regulation regarding the use and disclosure of identified data to conduct data linkage. Aust N Z J Public Health. 2008; 32:282–285. [PubMed: 18578830]

Lozano-Fuentes S, Elizondo-Quiroga D, Farfan-Ale JA, Loroño-Pino MA, Garcia-Rejon J, Gomez-Carro S, et al. Use of Google Earth to strengthen public health capacity and facilitate management of vector-borne diseases in resource-poor environments. Bull World Health Organ. 2008; 86:718–725. [PubMed: 18797648]

Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014; 21:957–958. [PubMed: 25008006]

Nass, SJ.; Levit, LA.; Gostin, LO. Institute of Medicine (U.S.). Beyond the HIPAA privacy rule : enhancing privacy, improving health through research. Washington, D.C: National Academies Press; 2009. Committee on Health Research and the Privacy of Health Information the HIPAA Privacy Rule.

National Cancer Institute. Conference on Geospatial Approaches to Cancer Control and Population Sciences. Bethesda, MD: Natcher Conference Center, NIH Campus; 2016. http://epi.grants.cancer.gov/events/geospatial/

National Research Council (U.S.). Putting people on the map : protecting confidentiality with linked social-spatial data. Washington, DC: National Academies Press; 2007. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data.

Oakes, JM.; Kaufman, JS. Methods in social epidemiology. San Francisco, CA: Jossey-Bass; 2006.

Office of Clinical Research. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. United States Department of Health and Human Services. 2012

Paiva T, Chakraborty A, Reiter J, Gelfand A. Imputation of confidential data sets with spatial locations using disease mapping models. Stat Med. 2014; 33:1928–1945. [PubMed: 24395116]

Palmer JR, Espenshade TJ, Bartumeus F, Chung CY, Ozgencil NE, Li K. New approaches to human mobility: using mobile phones for demographic research. Demography. 2013; 50:1105–1128. [PubMed: 23192393]

Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust. 2006; 185:263–267. [PubMed: 16948622]

Ruggles S. Big microdata for population research. Demography. 2014; 51:287–297. [PubMed: 24014182]

Seidl D, Paulus G, Jankowski P, Regenfelder M. Spatial obfuscation methods for privacy protection of household-level data. Applied Geography. 2015; 63:253–263.

Snow, J. On the mode of communication of cholera. London: J. Churchill; 1855.

Sweeney, L. Data Privacy Working Paper 3. Pittsburgh, Pennsylvania: Carnegie Mellon University; 2000. Simple Demographics Often Identify People Uniquely.

Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002; 10:557–570.

Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. Syst Rev. 2012; 1:60. [PubMed: 23194585]

VanWey LK, Rindfuss RR, Gutmann MP, Entwisle B, Balk DL. Confidentiality and spatially explicit data: concerns and challenges. Proc Natl Acad Sci U S A. 2005; 102:15337–15342. [PubMed: 16230608]

Wang H, Reiter JP. MULTIPLE IMPUTATION FOR SHARING PRECISE GEOGRAPHIES IN PUBLIC USE DATA. Ann Appl Stat. 2012; 6:229–252. [PubMed: 23990852]

Wartenberg D, Thompson WD. Privacy versus public health: the impact of current confidentiality rules. Am J Public Health. 2010; 100:407–412. [PubMed: 20075316]

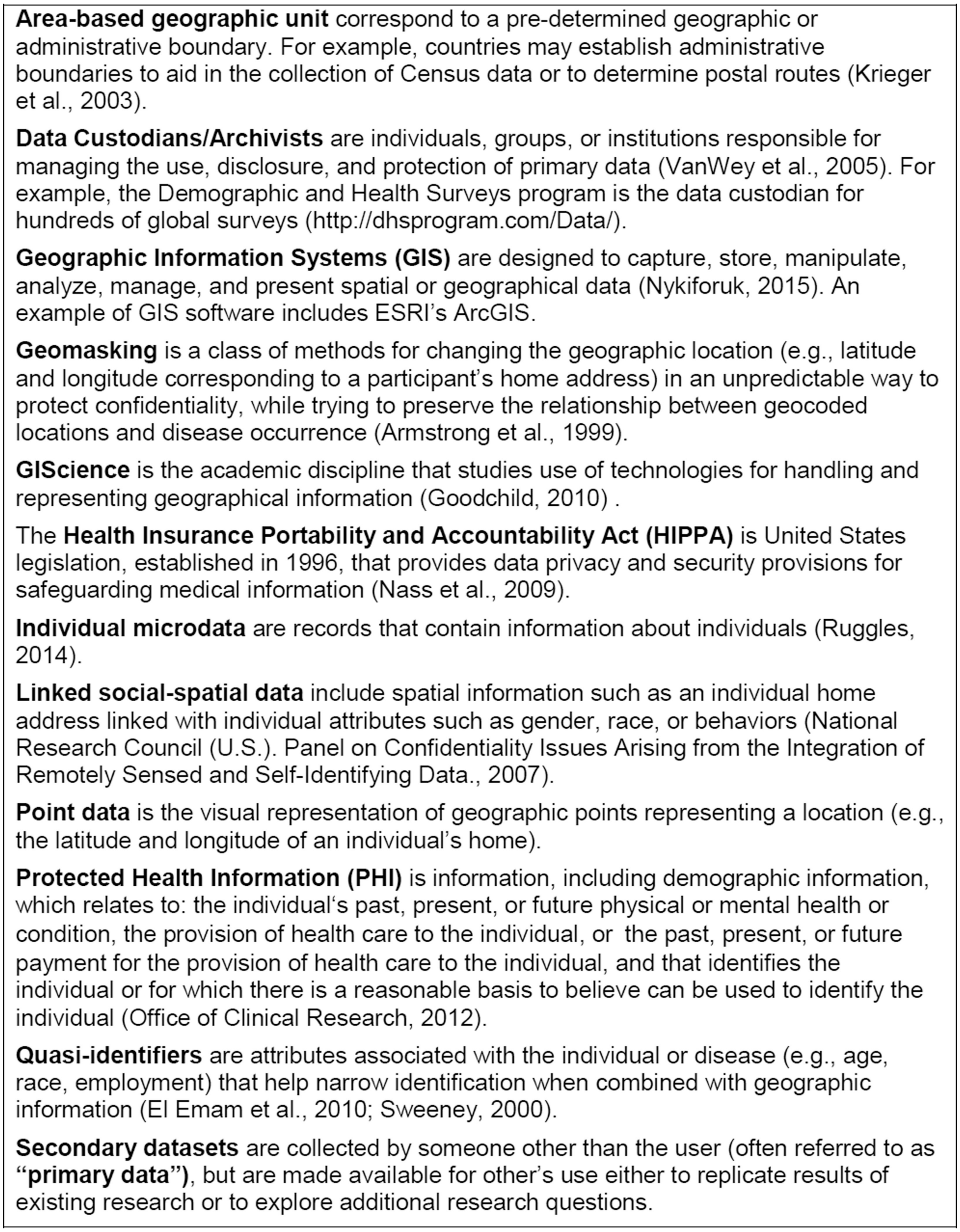
Wieland SC, Cassa CA, Mandl KD, Berger B. Revealing the spatial distribution of a disease while preserving privacy. Proc Natl Acad Sci U S A. 2008; 105:17608–17613. [PubMed: 19015533]

Yarmohammadian MH, Raeisi AR, Tavakoli N, Nansa LG. Medical record information disclosure laws and policies among selected countries; a comparative study. J Res Med Sci. 2010; 15:140–149. [PubMed: 21526073]

Zandbergen P. Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data. Advances in Medicine. 2014; 2014:14.

Research Highlights

- Using linked social-spatial data in sexual health research may risk confidentiality
- Over half of articles presenting point data on maps did not mask data adequately
- Several articles included maps with quasi-identifiers
- Publication of maps risking confidentiality occurs across a range of journals
- Findings highlight a need for heightened education for researchers and editors



Area-based geographic unit correspond to a pre-determined geographic or administrative boundary. For example, countries may establish administrative boundaries to aid in the collection of Census data or to determine postal routes (Krieger et al., 2003).

Data Custodians/Archivists are individuals, groups, or institutions responsible for managing the use, disclosure, and protection of primary data (VanWey et al., 2005). For example, the Demographic and Health Surveys program is the data custodian for hundreds of global surveys (http://dhsprogram.com/Data/).

Geographic Information Systems (GIS) are designed to capture, store, manipulate, analyze, manage, and present spatial or geographical data (Nykiforuk, 2015). An example of GIS software includes ESRI’s ArcGIS.

Geomasking is a class of methods for changing the geographic location (e.g., latitude and longitude corresponding to a participant’s home address) in an unpredictable way to protect confidentiality, while trying to preserve the relationship between geocoded locations and disease occurrence (Armstrong et al., 1999).

GIScience is the academic discipline that studies use of technologies for handling and representing geographical information (Goodchild, 2010) .

The **Health Insurance Portability and Accountability Act (HIPPA)** is United States legislation, established in 1996, that provides data privacy and security provisions for safeguarding medical information (Nass et al., 2009).

Individual microdata are records that contain information about individuals (Ruggles, 2014).

Linked social-spatial data include spatial information such as an individual home address linked with individual attributes such as gender, race, or behaviors (National Research Council (U.S.). Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data., 2007).

Point data is the visual representation of geographic points representing a location (e.g., the latitude and longitude of an individual’s home).

Protected Health Information (PHI) is information, including demographic information, which relates to: the individual‘s past, present, or future physical or mental health or condition, the provision of health care to the individual, or the past, present, or future payment for the provision of health care to the individual, and that identifies the individual or for which there is a reasonable basis to believe can be used to identify the individual (Office of Clinical Research, 2012).

Quasi-identifiers are attributes associated with the individual or disease (e.g., age, race, employment) that help narrow identification when combined with geographic information (El Emam et al., 2010; Sweeney, 2000).

Secondary datasets are collected by someone other than the user (often referred to as **“primary data”)**, but are made available for other’s use either to replicate results of existing research or to explore additional research questions.

Figure 1.

Glossary of key terms relevant to confidentiality considerations when visualizing results generated from linked social-spatial data

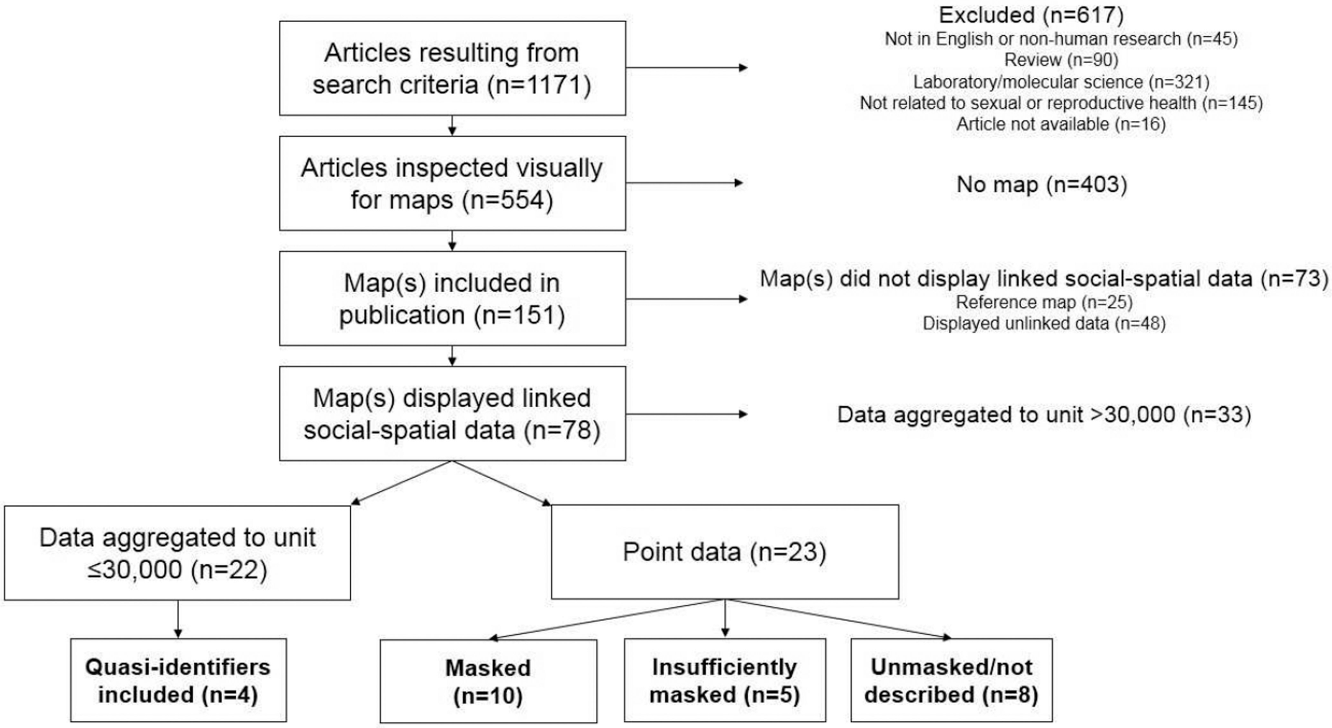

Articles resulting from search criteria (n=1171)
Excluded (n=617)
Not in English or non-human research (n=45)
Review (n=90)
Laboratory/molecular science (n=321)
Not related to sexual or reproductive health (n=145)
Article not available (n=16)
Articles inspected visually for maps (n=554)
No map (n=403)
Map(s) included in publication (n=151)
Map(s) did not display linked social-spatial data (n=73)
Reference map (n=25)
Displayed unlinked data (n=48)
Map(s) displayed linked social-spatial data (n=78)
Data aggregated to unit >30,000 (n=33)
Data aggregated to unit ≤30,000 (n=22)
Point data (n=23)
Quasi-identifiers included (n=4)
Masked (n=10)
Insufficiently masked (n=5)
Unmasked/not described (n=8)

Figure 2.

Scoping review of publications related to the social determinants of sexual or reproductive health including identifiable linked social-spatial data published and indexed in PubMed between January 1, 2013 and September 1, 2015

Table 1

Approaches to masking geographic identifiers based on the framework provided by Armstrong and Colleagues (Armstrong et al., 1999)

Approach	Description	Strengths	Challenges
Record Transformation	Records are aggregated across covariate patterns, certain records are suppressed, sampled or multiplied by random noise	Limits individual re-identification	Obscures spatial details needed for spatial analyses (e.g., cluster detection) Results in missing data
Spatial Aggregation	Data is summarized by spatial units (e.g., assigned to an areal unit polygon such as a census tract)	Limits individual reidentifcation, even at very small units (A. Curtis et al., 2011) Depending on unit, may facilitate easier data sharing/access	Obscures spatial details needed for spatial analyses (e.g., cluster detection) (Hampton et al., 2010) Units may not correspond to meaningful social or spatial divisions (e.g., modifiable unit problem) (Oakes & Kaufman, 2006) Spatial units may not perform the same for all outcomes (Krieger et al., 2002)
Point Aggregation	Points which are in geographic proximity are replaced by a composite point (e.g., points are clustered and assigned to an areal unit centroid)	May allow for analyses that require point data	Clustering techniques in and of themselves are not benign and may introduce error in spatial analyses (e.g., inaccurate cluster detection) (Hampton et al., 2010)
Affine Transformation	Points are displaced by fixed increments (translation), scaling constants (scale), rotating each point by a fixed angle around the pivot point (rotation), or a combination of the above (concatenated)	Translation preserves overall density, relative density, and directional information Techniques can be combined to introduce more uncertainty	Displacement constants cannot be shared May not provide sufficient anonymity (Wieland et al., 2008) Spatial attributes of data skewed/lost
Random Perturbation	Displaces points by a random increment and direction. Common techniques include randomized skew and Gaussian skew.	Displacement can be bounded by geographic boundaries (e.g., within census tracts) Introduction of random effects may reduce re-identification risks Gaussian skew displacement varies by population density (e.g., points in rural areas are displaced by greater distance than urban areas) (Cassa et al., 2008) Cluster detection superior to aggregation (Hampton et al., 2010)	Does not preserve relative locations and orientation of points Randomized skew does not account for underlying population density Points may be displaced a very small distance from original point Release of multiple, datasets masked using Gaussian skew may provide sufficient data for reconstruction of original data points (Cassa et al., 2008) Gaussian skew displacement parameters are user defined and requires an understanding of acceptable re-identification risk
Adaptive Techniques			
Donut Geomasking	More recent, adaptive geomasking technique that displaces points randomly by a minimum distance, but less than a maximum distance (Allshouse et al., 2010; Hampton et al., 2010)	Displacement can be bounded by geographic boundaries (e.g., within census tracts) Enhanced confidentiality protections provided by minimum displacement parameters Accounts for population density Cluster detection superior to aggregation (Hampton et al., 2010)	Displacement parameters are user defined and requires an understanding of acceptable re-identification risk Heterogeneous areas require greater displacement (Allshouse et al., 2010)
Simulation	Mathematical models (e.g., linear programming, multiple imputation) used to simulate new deidentified latitude and longitude which replace original data (Paiva et al., 2014 ;Wang & Reiter, 2012; Wieland et al., 2008)	Displacement can be bounded by geographic boundaries (e.g., within census tracts) Preserves clusters Moves point minimum specified distance Replaces participant geographic data with simulated data	Complex approaches that also require point data Displacement parameters are user defined and requires an understanding of acceptable re-identification risk Simulated data points coincide with actual data More recent approaches not

Approach	Description	Strengths	Challenges
			evaluated extensively in practice (Paiva et al., 2014; Wang & Reiter, 2012; Wieland et al., 2008)