The Journal of Infectious Diseases

SUPPLEMENT ARTICLE

Mind the Scales: Harnessing Spatial Big Data for Infectious Disease Surveillance and Inference

Elizabeth C. Lee,[1] Jason M. Asher,[3] Sandra Goldlust,[1] John D. Kraemer,[2] Andrew B. Lawson,[4] and Shweta Bansal[1,5]

[1]Department of Biology, [2]Department of Health Systems Administration, Georgetown University, and [3]Leidos, Washington D.C.; [4]Department of Public Health Sciences, Medical University of South Carolina, Charleston; and [5]Fogarty International Center, National Institutes of Health, Bethesda, Maryland

Spatial big data have the velocity, volume, and variety of big data sources and contain additional geographic information. Digital data sources, such as medical claims, mobile phone call data records, and geographically tagged tweets, have entered infectious diseases epidemiology as novel sources of data to complement traditional infectious disease surveillance. In this work, we provide examples of how spatial big data have been used thus far in epidemiological analyses and describe opportunities for these sources to improve disease-mitigation strategies and public health coordination. In addition, we consider the technical, practical, and ethical challenges with the use of spatial big data in infectious disease surveillance and inference. Finally, we discuss the implications of the rising use of spatial big data in epidemiology to health risk communication, and public health policy recommendations and coordination across scales.

Keywords. spatial big data; spatial epidemiology; disease mapping; infectious diseases; digital epidemiology; statistical bias.

During one of epidemiology's formative moments, John Snow mapped London households in which residents had cholera and succeeded in highlighting the risk of cholera associated with the Broad Street pump. Since then, spatial investigations have played a critical role in improving our understanding of the associations between risks and disease outcomes. In infectious disease epidemiology, we ask, "Which populations are at higher risk for disease?" "Where did this outbreak originate?" and "Where can we expect future disease outbreaks to arise?" Fundamentally, these are spatial questions that rely on spatial data for answers.

Traditional infectious disease epidemiology is built on the foundation of relatively high-quality and high-accuracy data on disease (eg, serological diagnostic assays) and behavior (eg, vaccination surveys). These data are usually characterized by small size, but they benefit from control groups or designed observational samples from known underlying populations, thus rendering it possible to make population-level inferences. On the other hand, digital infectious disease epidemiology typically uses existing digital traces, repurposing them to identify patterns in health-related processes. Digital data are electronic and can often be characterized as big data when they are produced in large volumes (ie, when there is a large number of subjects or a large number of measurements per subject), with high velocity (ie, when data are created in near real time), and have variety in sources and organizational structures [1]. When big data are characterized by fine spatial granularity, in which point or areal locations are identified, we refer to them here as spatial big data. Big data provide opportunities for infectious disease epidemiology and public health because they increase accessibility to populations over space and time; data on personal beliefs, behaviors, and health outcomes are now available at unprecedented breadth and depth. The trade-off to this tremendous access is the potential for loss of quality and accuracy. Streams of digital data relevant to public health may serve as proxies for a desired measure, but these data sets may not meet the assumptions for standard methods of epidemiological comparison (eg, self-reported symptoms on social media and serological diagnoses both serve as proxies for so-called true cases, but they have different biases and collection procedures and represent different populations).

The trade-off between access and accuracy and the task of separating true signal from large and varied noise characterizes the challenge and opportunity of big data for infectious disease epidemiology [2]. In this article, we focus on spatial big data and its applications to the field of spatial epidemiology. We highlight the opportunities for spatial big data to improve spatial modeling and data coverage and describe ongoing challenges as spatial big data become more pervasive in informing disease surveillance, disease control, and public health policy.

SPATIAL BIG DATA OPEN NEW DOORS IN EPIDEMIOLOGY

True to the promise of variety in big data streams, several familiar technologies produce spatial big data that can be used for infectious disease surveillance and modeling. Social media

Correspondence: S. Bansal, 408 Reiss Science Bldg, Department of Biology, Georgetown University, Washington, DC 20057 (shweta.bansal@georgetown.edu).

The Journal of Infectious Diseases® 2016;214(S4):S409–13

DOI: 10.1093/infdis/jiw344

sites like Facebook and Twitter allow users to tag individual posts with specific locations, linking geography to specific health behaviors. Mobile phones send signals with global positioning system locations, and their call data records are spatially referenced through cell tower locations, both of which enable the measurement of human activity and mobility [3, 4]. Web search data may capture user location through Internet protocol addresses, and online encyclopedia (Wikipedia) access logs may identify locations on the basis of the search language [5, 6]. Administrative medical claims and pharmacy transactions indicate the location of healthcare facilities and drugstores where patients seek care and medications [7, 8]. Restaurant reservation cancellations on sites like OpenTable may provide insight into disease incidence in specific cities [9].

Infectious disease epidemiology has already witnessed an impact from spatial big data, and the development of new methods and improvements to computational efficiency will only increase the potential of these data sources. Satellite imagery to infer climate, land use, and population density information has contributed to a better understanding of the spatial distribution of critical mosquito disease vectors and the seasonal epidemic dynamics of measles [eg, 10, 11]; and HealthMap, an automated, online news and outbreak reporting aggregator, has enabled the assimilation of disparate sources of disease occurrence data and has been used to examine spatial dynamics of cholera [12]. Mobile phone call data records have provided insights into human mobility that have informed risk maps, importation potential, and spatial dynamics of dengue and malaria [eg, 4]. Medical claims data have been used to examine spatial heterogeneity in influenza epidemic timing and severity [13, 14], while geographically referenced Twitter data have been used to identify spatial antivaccination sentiment [15].

While these studies with spatial big data have leveraged the fine spatial resolution to develop a detailed understanding of disease risk, there remain untapped opportunities with real-time surveillance, large-scale ecological inference, and adaptive disease mitigation strategies. Harnessing disease data from digital sources may enable epidemiological analyses to be performed at finer spatial scales in areas with poor coverage from traditional public health surveillance, and traditional and digital sources of spatial big data may be combined to account for the bias and gaps in each [eg, 16]. The assimilation of multiple spatial big data sources through flexible statistical modeling methods and the continuous nature of data streams could enable near real-time dynamic disease mapping and risk mapping in the near future. For example, Bayesian statistical approaches have emerged as tools for merging multiscale big data sources, incorporating explicit spatial dependencies into maps and models, and providing a framework for joining disease surveillance data across spatial scales while explicitly capturing the variation in measurement bias across locations [eg, 17]. Finally, access to multiple spatial scales of data allows one scale with missing observations to borrow information from a different scale through the addition of contextual effects in modeling inference [18].

SPATIAL BIG DATA PRESENT TECHNICAL CHALLENGES

While big data offer significant opportunities for epidemiological modeling and analysis, they also present a variety of technical and practical challenges. The measurement of incomplete and unrepresented populations, the lack of consistency and reliability in data over time, and the need for data and model validation are broad challenges with big data and statistical analysis that are discussed elsewhere [eg, 19, 20]. Here, we discuss a narrower set of challenges that arise specifically from the spatial nature of big data.

Spatial Coverage and Representation

Spatial big data may provide precise spatial information, but careful users should question the validity of available data. For example, we know that sources of spatial big data have biases in usership rates and demographic characteristics by location (Figure 1*A*) [21]. Medical claims record data only from insured and care-seeking populations, which may vary systematically according to socioeconomic and demographic characteristics. Social media sites where users volunteer spatial data tend to have more users and higher-quality information per capita in urban areas as compared to rural areas [21]. Mobile phone ownership varies by sex and literacy, and phone sharing between multiple individuals and SIM card switching complicate comparisons of these data across locations [3, 4]. As we cannot often measure the heterogeneities in user populations, these heterogeneities can translate into poor choices in sampling design (eg, how to stratify samples to get a representative population). Beyond heterogeneities in user populations, the populations captured by big data (eg, Twitter users) are not usually relevant to epidemiology; even if we could generate an unbiased sample of the population, it may not provide information important to public health. All of these issues complicate analyses that seek to compare different locations. Ultimately, issues with spatial coverage and representation cause problems for statistical inference, which often depends on assumptions of independent random variation and representative sampling for validity. Future research should compare analyses of spatial big data and analyses of designed observational data, to demonstrate the validity of spatial big data samples and to understand which features of a big data sample can produce robust statistical inference.

Spatial Uncertainty and Noise

Each source of big data provides a different type of spatial insight, despite the high spatial resolution among the sources. Users of social media volunteer their geographical locations in their profiles or posts, while Internet search engines can log spatial information automatically every time a Web search is performed.

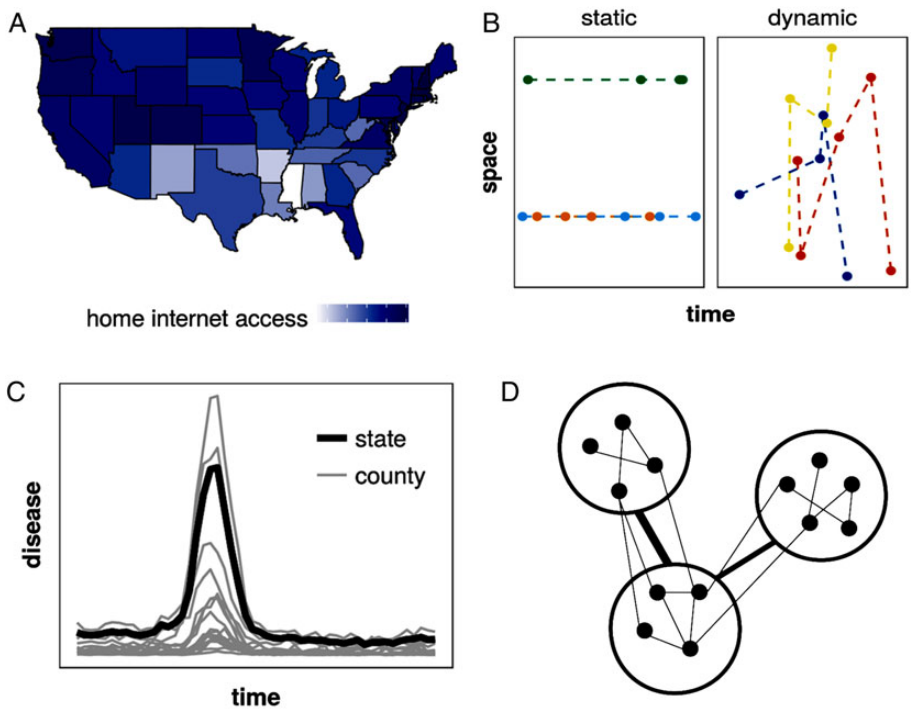

Figure 1. *A*, Spatial big data have spatial biases in the populations they represent. For instance, as reported by the 2013 American Community Survey, there is spatial variation in home Internet access across the United States, which might affect the populations generating search query data in Google Trends. *B*, With static spatial data (left), individuals (represented with different colors) report case events (points) at fixed locations. For instance, 2 individuals visited the same physician's office with symptoms multiple times (points along the time axis), so their events are recorded at the same position along the space axis (see overlapping trajectories in the lower part of the space axis), while another individual visited a different physician's office with symptoms 3 times in a similar period (upper part of space axis). Events from the same individual are connected with a dashed line. With dynamic spatial data (right), events are recorded as individuals move through space. For example, the dark blue individual (see trajectory that begins earliest on the time axis) recorded 4 events when they tweeted about symptoms at work, at the grocery store, at the pharmacy, and at home, so their case events occur at 4 different positions along the space axis. Events occur in time dynamically (as shown in this figure), but events may also be aggregated to regular intervals (eg, weekly). *C*, Data at different spatial scales may have different magnitudes and variability in time, after adjustment for population size, even if they are derived from the same data source. For instance, we observe time-varying fluctuations and variation in epidemic peak timing and magnitude in the county-level disease data (gray) that are lost in the state-level data (black). *D*, One possible method to protect privacy is to mask individual-level data by aggregating collected data to larger spatial resolutions. In reality, individuals (black circles) may be connected to other individuals through mobile phone calls (black lines). The publicly released data may be aggregated to the level of neighborhoods (green circles), and the number of calls between individuals from different neighborhoods (green lines) would be represented with different weights (here, depicted with varying thickness according to number of individual calls).

Sometimes the data are tied to a static location, as in the case of medical claims and healthcare facilities, but the cell towers associated with call data records and the locations of geographically tagged tweets vary dynamically over time (Figure 1*B*). Across the combinations of features—self-reported or automated, and static or dynamic—among these data sources, there are additional layers of uncertainty to consider in the context of epidemiology. For one, when spatial information in big data is not clearly specified, systematic biases in the results may be generated from the data-cleaning process itself (eg, addresses may be less likely to be geolocated in rural areas) [eg, 22]. Second, locations of potential transmission events will often differ from locations where disease is reported. While these components are explicitly differentiated in medical claims data (ie, transmitted in the community and reported at healthcare facilities), social media posts affiliated with dynamic movements could provide undifferentiated information about both transmission and reporting event locations. Big data provide information at unprecedented levels of spatial precision, but the spatial information fundamental to infectious disease epidemiology (eg, location and conditions that caused a disease transmission event) continues to remain obscured. As big data become more prevalent in epidemiological analysis, public health officials should take care not to conflate spatial precision with spatial accuracy in statistical inference for disease transmission and control.

Spatial Scales and Misalignment

When spatial big data are available at the level of individuals or precise spatial coordinates, practitioners may need to choose the scale of analysis and aggregate data accordingly. Analyzing data at the individual scale is prone to overfitting and the atomistic fallacy, in which we may make incorrect inferences at the group- or population-level on the basis of relationships observed in individual-level data [23]. For example, if we observe an association between body mass index (BMI) and hospitalization for influenza among individuals, it may be incorrect to assume

that populations with a high average BMI would have higher rates of influenza-associated hospitalization. On the other hand, analyzing data at aggregated scales is prone to the ecological fallacy, in which inferences about individuals are derived falsely from population-level observations [23, 24]. As an example, if we observed a negative association between average income and cholera prevalence at a national scale, it would be erroneous to assume that poor individuals have a higher risk of cholera than wealthy individuals. Similarly, statistical relationships between predictors and disease outcomes may change when analyses are performed at different spatial aggregations. For instance, Google Flu Trends attempted to estimate influenza activity across different regions of the United States by modeling the relationship between Google search terms and visits for influenza-like illness (ILI), as reported in traditional influenza surveillance systems [5]. However, the set of search terms identified as "most predictive" of ILI activity were tuned to a specific spaital scale (region-level), and may not apply to finer-resolutions (eg, zipcode-level) [5, 25]. Additionally, spatial questions often require the use of multiple data sources, and spatial misalignment arises when data are collected at different spatial scales and need to be incorporated into a single analysis. For instance, we may seek to understand the spatial distribution of cases at the state level when data were collected at the parish or county level (switching between 2 areal scales), or translate case data associated with household coordinates to cases at the county level (switching between point and areal scales; Figure 1*C*). Spatial big data have expanded the types of spatial information available for data aggregation—posts geographically tagged on social media might provide information at the level of countries, cities, neighborhoods, landmarks, and latitude-longitude coordinates—potentially engaging statistical change of support problems, even for one individual in a single day [24]. The multiplicity of highly resolved spatial scales also poses concerns for standard data checks, since traditional public health data will not necessarily be available at scales appropriate for validating comparisons to spatial big data [7, 16]. Finally, choices about how to deal with spatial misalignment have consequences for modeling results. For instance, recent studies have asked whether Zika virus–associated microcephaly was occurring at unusually high rates in different Brazilian states. Birth rate data might be collected at one spatial scale according to regular demographic surveys, but data systems tracking microcephalic live births would likely have finer spatial detail. Depending on the choice of spatial scale, the combination of these 2 data sources creates the potential for both overestimation and underestimation of microcephaly rates.

Spatial Confidentiality and Ethics

The practice of collecting data without seeking appropriate ethical approval presents some risk for digital infectious disease epidemiology, and the access to fine-grain spatial information further deepens this concern. Safeguards currently implemented for collecting and sharing spatial big data have focused on the obfuscation and aggregation of shared data to protect privacy and on the anonymization and de-identification of individuals. Many research institutions have standardized practices to protect individual privacy that follow the guidance of institutional review boards, disclosure review boards for public use data, and federal laws (eg, the Health Insurance Portability and Accountability Act of 1996, in the United States), but these organizations do not often recognize high-resolution spatial data as a source that should be covered under human subjects protection policies [26]. Several studies have provided examples in which seemingly anonymized data could be mined (or linked with other databases) for de-anonymization: de Montjoye et al [27] showed that 4 spatiotemporal position points from mobile phone records can be sufficient to uniquely identify 95% of individuals in a large de-identified data set; and Homer et al [28] showed that the sheer quantity of data collected could be sufficient to re-identify individuals in a genetic database. These issues already push the limits of existing ethical review mechanisms and our understanding of de-anonymization. In the future, guidelines to protect privacy and confidentiality may require the masking of individual-level records through the aggregation of data to coarser spatial resolutions (Figure 1*D*), the provision of synthetic data sets that attempt to mimic underlying distributions [29], or the distillation of spatial big data to parameters commonly used in epidemiological models. Investigations may consider the optimal choice of spatial scale in the context of trade-offs between the accurate representation of process heterogeneity, the protection of privacy [26], and the improvement of computational efficiency [30]. Nevertheless, public data become increasingly vulnerable to breaches of privacy as additional data are released and data-mining techniques improve over time.

IMPLICATIONS FOR PUBLIC HEALTH COMMUNICATION AND POLICY

The promise of high spatial and temporal resolutions in spatial big data opens opportunities for change in the standard practice of public health. In circumstances where adjacent or subordinate administrative units issue separate public health recommendations (eg, US federal, state, and local governments may issue independent influenza vaccination recommendations), spatial big data may enable these entities to derive their policies from analyses of a common data set and encourage coordination of preparedness activities across scales [eg, 14]. There is a growing panoply of adaptive, behavioral, and health economic modeling methods aimed at identifying the most-effective interventions for human and livestock diseases. As these methods begin to find use during ongoing outbreaks, the combination of spatial big data and adaptive models could enable the real-time adaptive management of infectious diseases and the coordination of disease control efforts across spatial scales.

In the long term, some sources of big data may become more readily available at finer spatial resolutions than the administrative regions at which policy decisions are made, even to the level of the individual. Spatial big data have already changed consumer-marketing strategies: rather than targeting geographic areas with certain socioeconomic and behavioral characteristics, marketers can now target individual users on the basis of behaviors demonstrated in their digital traces [31]. Should epidemiological modeling and design reflect these cultural changes to public health data? Perhaps an analogous scenario would see individual epidemiological data being used to inform optimal intervention strategies, ignoring the administrative boundaries that typically constrain decision making. It is difficult to imagine how such a public health infrastructure could operate—resources must still be coordinated and expended by administrative units, and policy decisions must still apply to populations (rather than individuals) to maintain feasibility. Nevertheless, epidemiological analyses with spatial big data expand the possibilities for multiscale coordination of infectious disease surveillance, response, and forecasting.

The real-time high-volume nature of spatial big data makes more epidemiological information readily available to policymakers, but it also creates challenges for the communication of public health information. Spatial big data enable small-area analyses, which are simultaneously highly precise to spatial locations and highly uncertain in modeling results about risk of disease. Similarly, the rise of epidemic forecasting technologies based on spatial big data might present predictions about risk and epidemic outcomes in precise locations even though the forecasts themselves are subject to uncertainty [16]. Consumers of analyses derived from spatial big data—clinicians, public health officials, epidemiologists, and modelers—should develop conscientious practices for communicating uncertainty about spatial results to the public.

Notes

Acknowledgments. We thank Shashank Khandelwal and 2 anonymous reviewers for their careful comments on earlier drafts of this work.

Disclaimer. The opinions, findings, and conclusions or recommendations expressed in this material are those of the author and not necessarily those of Jayne Koskinas Ted Giovanis Foundation for Health Policy and its directors, officers, or staff.

Financial support. This work was supported by the Jayne Koskinas Ted Giovanis Foundation for Health and Policy (dissertation support grant to E. C. L.); the National Cancer Institute (grant R01CA172805 to A. B. L.); and the RAPIDD Program of the Science & Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health.

Potential conflicts of interest. All authors: No reported conflicts. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Laney D. 3-D data management: controlling data volume, velocity and variety. Application Delivery Strategies. Stamford, CT: META Group Inc., **2001**.
2. Khoury MJ, Ioannidis JPA. Big data meets public health: human well-being could benefit from large-scale data if large-scale noise is minimized. Science **2014**; 346:1054–5.
3. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO. Heterogeneous mobile phone ownership and usage patterns in Kenya. PLoS One **2012**; 7:e35319.
4. Wesolowski A, Qureshi T, Boni MF, et al. Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proc Natl Acad Sci U S A **2015**; 112:11887–92.
5. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature **2009**; 457:1012–4.
6. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol **2014**; 10:e1003892.
7. Viboud C, Charu V, Olson D, et al. Demonstrating the use of high-volume electronic medical claims data to monitor local and regional influenza activity in the US. PLoS One **2014**; 9:e102429.
8. Pivette M, Mueller JE, Crépey P, Bar-Hen A. Drug sales data analysis for outbreak detection of infectious diseases: a systematic literature review. BMC Infect Dis **2014**; 14:604.
9. Nsoesie EO, Buckeridge DL, Brownstein JS. Guess who's not coming to dinner? Evaluating online restaurant reservations for disease surveillance. J Med Internet Res **2014**; 16:e22.
10. Kraemer MUG, Sinka ME, Duda KA, et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. Elife **2015**; 4:1–18.
11. Bharti N, Tatem AJ, Ferrari MJ, Grais RF, Djibo A, Grenfell BT. Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery. Science **2011**; 334:1424–7.
12. Tuite AR, Tien J, Eisenberg M, Earn DJD, Ma J, Fisman DN. Cholera epidemic in Haiti, 2010: using a transmission model to explain spatial spread of disease and identify optimal control interventions. Ann Intern Med **2011**; 154:593–601.
13. Gog JR, Ballesteros S, Viboud C, et al. Spatial transmission of 2009 pandemic influenza in the US. PLoS Comput Biol **2014**; 10:e1003635.
14. Lee EC, Viboud C, Simonsen L, Khan F, Bansal S. Detecting signals of seasonal influenza severity through age dynamics. BMC Infect Dis **2015**; 15.
15. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol **2011**; 7:e1002199.
16. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012-2013 season. Nat Commun **2013**; 4.
17. Corberán-Vallet A, Lawson AB. Prospective analysis of infectious disease surveillance data using syndromic information. Stat Methods Med Res **2014**; 23:572–90.
18. Aregay M, Lawson AB, Faes C, Kirby R. Bayesian multiscale modeling for aggregated disease mapping data. Stat Methods Med Res **2015**.
19. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science **2014**; 343:1203–5.
20. Althouse B, Scarpino S, Meyers L, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. EPJ Data Sci **2015**; 4:1–8.
21. Hecht B, Stephens M. A tale of cities: urban biases in volunteered geographic information. In: International AAAI Conference on Weblogs and Social Media (University of Michigan, Ann Arbor, Michigan, USA). Online, North America: AAAI Publications, **2014**. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8114/8120. Accessed 15 September 2016.
22. Skelly C, Black W, Hearnden M, Eyles R, Weinstein P. Disease surveillance in rural communities is compromised by address geocoding uncertainty: a case study of campylobacteriosis. Aust J Rural Health **2002**; 10:87–93.
23. Lawson AB. Statistical methods in spatial epidemiology. 2nd ed. Wiley series in probability and statistics. West Sussex, England: John Wiley & Sons, **2006**.
24. Gotway CA, Young LJ. Combining incompatible spatial data. J Am Stat Assoc **2002**; 97:632–48.
25. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol **2013**; 9: e1003256.
26. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data NRC. Putting people on the map: protecting confidentiality with linked social-spatial data. Gutmann MP, Stern PC, eds. Washington, DC: National Academies Press, **2007**.
27. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD. Unique in the crowd: the privacy bounds of human mobility. Sci Rep **2013**; 3.
28. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet **2008**; 4.
29. Kinney SK, Reiter JP, Reznek AP, Miranda J, Jarmin RS, Abowd JM. Towards unrestricted public use business microdata: the synthetic longitudinal business database. Int Stat Rev **2011**; 79:362–84.
30. Deeth LE, Deardon R. Spatial data aggregation for spatio-temporal individual-level models of infectious disease transmission. Spat Spatiotemporal Epidemiol **2016**; 17:95–104.
31. Dalton CM, Thatcher J. Inflated granularity: spatial "big data" and geodemographics. Big Data Soc **2015**; 2:1–15.