Journal of the American Medical Informatics Association, 25(10), 2018, 1402–1406
doi: 10.1093/jamia/ocy071
Advance Access Publication Date: 7 June 2018
Perspective

Perspective

Using mobile location data in biomedical research while preserving privacy

Daniel M Goldenholz,[1,2] Shira R Goldenholz,[2] Kaarkuzhali B Krishnamurthy,[2,3] John Halamka,[4] Barbara Karp,[5] Matthew Tyburski,[6] David Wendler,[7] Robert Moss,[8] Kenzie L Preston, PhD,[6] and William Theodore[1]

[1]Clinical Epilepsy Section, NINDS, NIH, [2]Epilepsy Division, Beth Israel Deaconess Medical Center, [3]Department of Neurology, Steward Health Care, [4]Department of Emergency Medicine, Beth Israel Deaconess Medical Center, [5]Combined NeuroScience IRB, Office of Clinical Director, NINDS, NIH, [6]Intramural Research Program, National Institute on Drug Abuse, NIH, [7]Section on Research Ethics, Department of Bioethics, NIH, and [8]SeizureTracker LLC

Corresponding Author: Daniel M Goldenholz, Beth Israel Deaconess Medical Center, Division of Epilepsy, 330 Brookline Ave. Baker 5, Boston, Massachusetts 02218, USA (daniel.goldenholz@bidmc.harvard.edu)

Received 2 January 2018; Revised 25 April 2018; Editorial Decision 14 May 2018; Accepted 16 May 2018

ABSTRACT

Location data are becoming easier to obtain and are now bundled with other metadata in a variety of biomedical research applications. At the same time, the level of sophistication required to protect patient privacy is also increasing. In this article, we provide guidance for institutional review boards (IRBs) to make informed decisions about privacy protections in protocols involving location data. We provide an overview of some of the major categories of technical algorithms and medical–legal tools at the disposal of investigators, as well as the shortcomings of each. Although there is no "one size fits all" approach to privacy protection, this article attempts to describe a set of practical considerations that can be used by investigators, journal editors, and IRBs.

Key words: ethics, big data, mHealth, private health information, location-based-services, geographic information systems, privacy

BACKGROUND

"Big Data"[1] is ubiquitous in today's world. Indeed, a 2015 estimate for the magnitude of global data produced daily was over 2 exabytes (ie 2.5×10^{18} bytes), with future projections accelerating beyond that.[2] Human biomedical research is commonly conducted with large-scale datasets that have identifiers removed (de-identified, such as in 45 CFR 164.512). The introduction of inexpensive location-based services (eg global positioning system, "GPS") into mainstream portable products has revolutionized the ability to track individuals. With these developments, it is inevitable[3] that more seemingly "de-identified" datasets will include location information that could be sufficient to re-identify individuals,[4] thereby threatening personal privacy. While location-based investigations could benefit many aspects of healthcare,[5, 6] there is a pressing need to identify strategies that mitigate the risks to privacy inherent in including location data in seemingly de-identified datasets.

Consider the path that data take in the context of a research study (Figure 1). First, patient data are collected and stored by the primary investigators, using a method approved by the local institutional review board (IRB). The data may be published, shared with secondary collaborators, or posted to a repository. If secondary collaborators are involved, they may have an independent IRB overseeing their activities. All of the links along this pathway require consideration in terms of data privacy concerns in a consistent, thoughtful way that various IRBs or other reviewing bodies are likely to accept.

The chance of re-identification varies by the nature of the data (ie embedded location information) and its intended usage. Home addresses, for example, being static are widely available in public

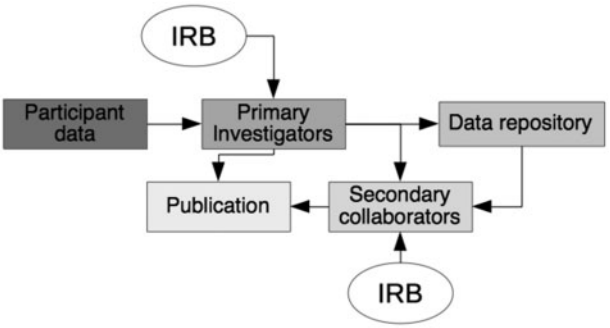

Figure 1. Path of location data. Initially, location data are produced from study participants. These data are collected by the primary investigators, modified by the requirements of the institutional review board (IRB). After analysis, the data can be published, shared with secondary collaborators, and/or sent to a data repository. If a repository is used, secondary collaborators may or may not have any connection with the primary investigators. Secondary collaborators would typically have their own independent IRB that monitors their use of the data. The secondary collaborators also have the opportunity for publication. Anywhere along these multiple pathways that the data traverse, there is a possibility for re-identification of study participants if privacy is not protected.

databases, thereby representing a higher risk possibility for re-identification. However, data involving a subject's walk through her neighborhood, if not maintained along with a specific date or time, might be more challenging to link to the subject herself, thus representing a lower risk possibility for re-identification. If a malicious hacker had access to additional information from other public or private databases, he/she could use machine learning approaches to re-identify large sets of participants. See Figure 2 and the Supplementary Figures for more examples.

The HIPAA privacy regulations[7] offer two options to maximize subject safety by protecting privacy: "Expert Determination" and "Safe Harbor." The Expert Determination method permits the use of statistical or scientific principles to protect individual patient privacy. The Safe Harbor option recommends deleting all of the 18 "identifiers," including address and localization information smaller than state. Here, we will focus on the Expert Determination method for de-identifying data because it can allow for more flexibility for location-based research.

IRBs

IRBs are tasked with not only recognizing that harm could occur, but with considering the magnitude and significance of that harm. Evolving technology allows for data capture that could, inadvertently or maliciously, allow for re-identification. Many IRB members may not be sufficiently informed about such developments, thus preventing them from recognizing potential for harm.

What follows is a set of useful questions for IRBs to consider while reviewing a protocol that include location data:

1. W*hat are the risks of collecting, sharing, and publishing individual-level location-based data?* Each stage of the protocol should be considered, including collection, use, and sharing of such data. Each time the location data move or will be accessed by new individuals should be treated as a distinct stage. An implicit assumption is that individuals should not be uniquely identifiable in published reports, although there are studies in which this is important and acceptable, provided there is explicit participant consent. Almost any study that plans to publish or share individual level location-based data carries the risk of re-identification.

2. *What types of harm (and what magnitude of harm) could subjects face if re-identification occurs?* Examples of types of harm include physical, emotional, social, legal, and financial, with the magnitude varying by situation and circumstance. For example, research participants in the witness protection program face mortal danger if re-identified; this represents physical and social harm, with the magnitude being very large. If a study reveals embarrassing personal habits, there may be emotional damage to participants who are re-identified; the magnitude is likely somewhat lower than in the previous example. Similarly, members of stigmatized religions might experience social harms such as exclusion from certain establishments if they were publicly re-identified. If health insurance companies used re-identification to stratify research participants into low- or high-risk categories, participants may suffer financially.

3. *What known parties (if any) would be interested in re-identification of these data, and how valuable would the data be to them? Do these parties have the necessary skill set to re-identify these data?* Perhaps an interested group is a little-known political party. If it is known the party has insufficient skills and financial resources to re-identify the data, this can help mitigate the concern of re-identification. Conversely, suppose a multinational corporation with vast resources is an interested group. In that case, the risk is higher, as it is known that it could employ complex statistical strategies, as well as purchase other datasets to re-identify participants.

4. *What risk mitigation strategies have been provided? Are additional strategies needed?* Some protocols may provide no risk mitigation strategies at all, and under some circumstances, that may be reasonable. In other cases, the IRB may determine that the risk mitigation strategies included in the study are insufficient, and may recommend further protections (see Table 1).

5. *Can the IRB assess whether the method of protecting against re-identification is adequate and appropriate? Is consultation with a data scientist needed?* Some methods of mitigation are simple to comprehend, such as removing data, or cutting off part of the ZIP code. Others are quite complex (such as K-anonymity or simulation), and if an IRB is unsure of the applicability and reliability of a technique, expert consultation may help. The techniques must be evaluated in the context of the sensitivity of the data, and the types of interested parties being considered.

6. *How will the risks be explained to participants so that truly informed consent is obtained?* Vetting a meaningful consent process can be challenging, particularly for an IRB whose members may not have sufficient expertise with the technology being used in a particular study. Any location data carry re-identification potential to various degrees, and should be assessed on a case-by-case basis. Occasionally, additional support for the IRB may be needed from a data scientist to act as a subject expert. See our Supplementary Table of suggestions for informed consent.

MITIGATION STRATEGIES - LEGAL

An important risk mitigation strategy is legal—incorporating a Certificate of Confidentiality (CoC) issued by the National Institutes of Health.[8] CoC now automatically covers NIH-funded research. The purpose of the CoC is to protect "investigators and institutions from being compelled to release information that could be used to identify subjects with a research project." It precludes the investigator or institution from being compelled to release data that could be used

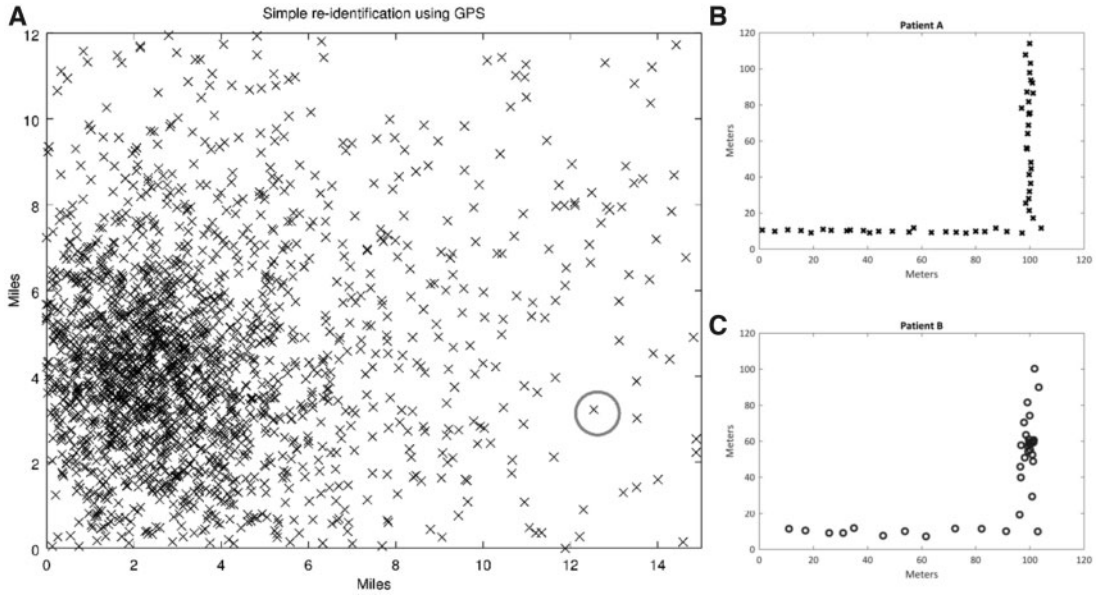

Figure 2. Simulated examples of re-identification risk. A) Simple re-identification risk. In this example, a city has decreasing population density farther from the city center. Each x represents a home location of a de-identified subject. The subject with the gray circle, although de-identified, lives far enough from the city center that even crude location data would be sufficient to uniquely identify his/her address. B and C) Spatiotemporal behavioral re-identification risk. Bob and Alice are both wearing trackers. Bob rode a bicycle and stopped at his favorite café at some point. Alice walked without stops. In this theoretical example, the tracking data without names were posted in a public repository using true GPS coordinates, along with heart-monitoring data. Although the GPS data shown here had been de-identified, Patient A (plot B) appears to be moving at a relatively constant speed throughout the path (accounting for noise), whereas Patient B (plot C) appears to have clustered some location data around $x = 100$, $y = 65$, as if he had spent additional time there. In addition, the spacing of markers for subject 2 is farther apart (except for his one stop) than 1, suggesting that subject 2 had a faster method of travel. If a malicious hacker knew that Patients A and B were either Alice or Bob, and was aware of Bob's owning a bicycle, he/she may be able to identify the individuals. This example shows that Alice is most likely Patient A, and Bob is most likely Patient B. After identification, the hacker can also use the heart data to connect medical diagnoses (eg atrial fibrillation) to identified participants. Note that timestamps are not shown in this figure intentionally—with them, the analysis would have been even easier.

Table 1. Summary of strategies for mitigating risk of re-identification of subjects when using location-based data in de-identified datasets

Strategies to mitigate risk	Examples
Certificate of confidentiality	Protection from "involuntary disclosure," such as subpoenas
Aggregating data	Summarizing by state, country, or ZIP code, k-annoymity
Obscuring data	Removal, encryption, cloaking, adding noise, decreasing resolution, simulation
Preserving relative location	Distances, spatial shift, spatial rotation, Lipschitz embedding
Location-derived data	Turning GPS data into rainfall data, etc.
Temporal manipulations	Removing times + scrambling order, relative times, adding noise, cloaking, subsampling
Attribute manipulations	Swapping attributes, decreasing variability in attributes

to re-identify subjects in "any Federal, State, or local civil, criminal, administrative, legislative, or other proceedings." The legality of the CoC thus far has been upheld.[9]

Although location data were not originally listed in the regulations establishing CoC as identifying, they are included by inference because it covers "any other item or combination of data about a research participant which could reasonably lead, directly or indirectly by reference to other information, to identification."

Some caveats exist with CoC preventing disclosures, such as certain legal situations (such as disclosure of non-identifying information within a database),[9] and cases in which re-identification would not significantly harm or damage the individual. The protections do not extend to registries or NIH-mandated data sharing plans. Also, CoC can be used only in research in which informed consent is obtained. In addition, a research subject can provide written permission for release of data under different circumstances. Under the terms of a CoC, investigators are permitted to voluntarily release data that can identify subjects, provided disclosure was discussed in the informed consent form. For instance, investigators may inform participants that they will voluntarily report on patients with communicable diseases to health authorities.

MITIGATION STRATEGIES - TECHNICAL

There are several technical strategies to minimize the ability of others to use data for identification, some of which require a data scientist. One strategy is to aggregate the data,[10] such as offering a summary table about groups of subjects without providing data points for individual subjects. This approach is limited to studies in which each group size is sufficiently large. To ensure all groups are sufficiently large, one could use "k-anonymity," making a set of k subjects that are indistinguishable when using location alone.[11]

A second strategy is to obscure location data so that they are unavailable or unreadable to anyone without formal permission to use them. All forms of data obscuring (except encryption) involve some degree of data quality degradation. Options for obscuring location data include: removal,[4] encryption,[12] cloaking,[11] adding noise,[10,13] decreasing the resolution,[14] and simulation.[15–18] Of note, modern methods for simulating synthetic data can provide nuanced data that can be minimally "lossy" while preserving privacy.[17,18] Regardless of the technique, it is important to disclose which method was used when sharing the data (eg in publication). Many of these techniques fall under the broad heading of "geographic masking." The tradeoff with obscuring data is that subsequent investigations may suffer in efficacy due to lost detail.

A third option includes preserving only a portion of the location data, such as relative distance to a fixed position, such as distance from home. Alternatively, all locations could be all be "shifted" a certain number of miles in a random direction, thus preserving relative locations but dropping absolute location. Similarly, all locations could be rotated, eg changing all North to West, West to South, etc. A general approach to preserve privacy using relative distances is called "Lipschitz embedding"[19]; this technique transforms the distances before storing them, making privacy attacks more difficult.

In addition, the location data may be relevant only because of data that correlate with location, such as elevation or rainfall. Preserving only the correlated data while dropping the location data decreases the risk of re-identification.

The temporal relationship among locations may increase the vulnerability for re-identification[20] (see Figure 2), so a variety of temporal manipulations could be performed that are analogous to spatial manipulations.

A final consideration is social-spatial linkage analysis, which uses the connection to other databases to re-identify data that are otherwise de-identified.[10,21] When location data are provided with links to other subject attributes (age, gender, ethnicity, etc.), this results in additional points of vulnerability.[22] One possible solution is to swap attributes in a random fashion when specific links between locations and attributes are not required for subsequent analysis.[22]

CONCLUSIONS

There does not appear to be a "one-size-fits-all" solution to mitigating risk. Each strategy has strengths and weaknesses, and each may be well suited to certain situations but not others. The only strategy to completely protect privacy is removal of all location data. However, complete removal is often undesirable. The alternative is to accept a degree of risk guided by the specific situation. For instance, suppose the aim of a psychological study is determining a subject's preference for soft versus loud music. If location data were re-identified, unintended parties would gain access to the relatively harmless knowledge of music amplitude preference for individuals. Conversely, if an HIV incidence study reported obscured GPS locations of subjects with AIDS, re-identification could result in serious personal harms. Thus, the former study may wish to use less cumbersome risk mitigation techniques, while the latter may require multiple strategies and a re-identification risk assessment.[15]

As location data become more commonly available in datasets, stakeholders will face more questions about assessing and mitigating the risk of lost privacy. Methods to mitigate the risk of re-identification were summarized. None of these strategies alone will be sufficient for all cases, and, with changing technology and software capabilities, these concerns will need to be re-addressed at regular intervals. Recommendations for how IRBs can approach these questions were provided as well.

FUNDING

This research was funded in part by the Intramural Program of the National Institute for Neurological Disease and Stroke and the Intramural Research Program of the National Institute on Drug Abuse, NIH.

Funding was unrestricted and played no role in the development of this manuscript.

The researchers were independent of the funding sources.

CONTRIBUTORS

DMG conceived the project, conducted simulations, interpreted results, and wrote the manuscript.

SRG, KBK, JH, BK, DW, RM, WT contributed to the interpretation of results and editing the manuscript for intellectual content.

MT and KP provided de-identified data which they analyzed and interpreted for the appendix, and they contributed to the interpretation of the results and editing the manuscript for intellectual content.

All authors provided approval of the final version of the document, and all agree to be accountable for the accuracy and integrity of their work and that of their coauthors.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank Dr Laurence Muhlbaier and Dr Kevin Weinfort from Duke University for their insights on this manuscript.

Conflict of interest statement. None declared.

REFERENCES

1. Mooney SJ, Westreich DJ, El-Sayed AM. Commentary: Epidemiology in the era of big data. *Epidemiology* 2015; 26 (3): 390–4.
2. Walker B. The Impact of Big Data on Our Everyday Lives - Infographic. VoucherCloud. 2015. https://www.vouchercloud.net/resources/big-data-infographic (Accessed January 1, 2017).
3. Kelly K. *The Inevitable*. New York: Viking; 2016.
4. Krumm J. Inference attacks on location tracks. In: *Pervasive Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007: 127–43.
5. van Rheenen S, Watson TWJ, Alexander S, *et al*. An analysis of spatial clustering of stroke types, in-hospital mortality, and reported risk factors in Alberta, Canada, using geographic information Systems. *Can J Neurol Sci* 2015; 42 (05): 299–309.
6. Epstein DH, Tyburski M, Craig IM, *et al*. Real-time tracking of neighborhood surroundings and mood in urban drug misusers: application of a new method to study behavior in its geographical context. *Drug Alcohol Depend* 2014; 134: 22–9.
7. Law P. Health insurance portability and accountability act of 1996. Public Law 104-191. *US Statut Large* 1996; 110: 1936–2103.
8. Certificates of Confidentiality (CoC). https://humansubjects.nih.gov/coc/index. Accessed February 27, 2017.

9. Wolf LE, Patel MJ, Williams Tarver BA, *et al*. Certificates of confidentiality: protecting human subject research data in law and practice. *J Law Med Ethics* 2015; 43: 594–609.
10. Zandbergen P. A. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Adv Med* 2014; 2014: 1.
11. Jia J, Zhang F. Nonexposure accurate location K-anonymity algorithm in LBS. *Sci World J* 2014; 2014: 1.
12. Xie Q, Wang L. Privacy-preserving location-based service scheme for mobile sensing data. *Sensors (Basel)* 2016; 16 (12): 1993.
13. Hampton KH, Fitch MK, Allshouse WB, *et al*. Mapping health data: improved privacy protection with donut method geomasking. *Am J Epidemiol* 2010; 172 (9): 1062–9.
14. El Emam K, Brown A, AbdelMalik P. Evaluating predictors of geographic area population size cut-offs to manage re-identification risk. *J Am Med Inform Assoc* 2009; 16 (2): 256–66.
15. Wang H, Reiter JP. Multiple imputation for sharing precise geographies in public use data. *Ann Appl Stat* 2012; 6 (1): 229–52.
16. Jung H-W, El Emam K. A linear programming model for preserving privacy when disclosing patient spatial information for secondary purposes. *Int J Health Geogr* 2014; 13 (1): 16.
17. Yu M, Reiter JP, Zhu L, *et al*. Protecting confidentiality in cancer registry data with geographic identifiers. *Am J Epidemiol* 2017; 186 (1): 83–91.
18. Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, *et al*., eds. Advances in Neural Information Processing Systems 27. New York, NY: Curran Associates, Inc.; 2014: 2672–80.
19. Kroll M, Schnell R. Anonymisation of geographical distance matrices via Lipschitz embedding. *Int J Health Geogr* 2016; 15 (1): 1.
20. Malizia N. Inaccuracy, uncertainty and the space-time permutation scan statistic. *PLoS One* 2013; 8 (2): e52034.
21. VanWey LK, Rindfuss RR, Gutmann MP, *et al*. Confidentiality and spatially explicit data: concerns and challenges. *Proc Natl Acad Sci U S A* 2005; 102 (43): 15337–42.
22. Gutmann M, Witkowski K, Colyer C, *et al*. Providing spatial data for secondary analysis: issues and current practices relating to confidentiality. *Popul Res Policy Rev* 2008; 27 (6): 639–65.