Privacy, security, and the public health researcher in the era of electronic health record research

Neal D. Goldstein[1,2], Anand D. Sarwate[3]

1. Christiana Care Health System, Department of Pediatrics, 4745 Ogletown-Stanton Road, MAP 1, Suite 116, Newark, DE 19713 USA

2. Drexel University Dornsife School of Public Health, Department of Epidemiology and Biostatistics, 3215 Market Street, Philadelphia, PA 19104 USA

3. Rutgers, The State University of New Jersey, Department of Electrical and Computer Engineering, 94 Brett Road, Piscataway, NJ 08854 USA

Abstract

Health data derived from electronic health records are increasingly utilized in large-scale population health analyses. Going hand in hand with this increase in data is an increasing number of data breaches. Ensuring privacy and security of these data is a shared responsibility between the public health researcher, collaborators, and their institutions. In this article, we review the requirements of data privacy and security and discuss epidemiologic implications of emerging technologies from the computer science community that can be used for health data. In order to ensure that our needs as researchers are captured in these technologies, we must engage in the dialogue surrounding the development of these tools.

Keywords: electronic health record; privacy; data security; analysis

Abbreviations: EHR, electronic health record; IRB, institutional review board; HIPAA, Health Insurance Portability and Accountability Act; MPC, multiparty computation; PHI, personal health identifiers

Correspondence: Neal D. Goldstein, Christiana Care Health System, Department of Pediatrics, 4745 Ogletown-Stanton Road, MAP 1, Suite 116, Newark, DE 19713, ngoldstein@christianacare.org, 1-302-733-4200

DOI: 10.5210/ojphi.v8i3.7251

Introduction

Public health research often requires maintaining the privacy and security of sensitive health information. These data may require strict safeguards to protect both the privacy of the research participant and the security of their information. Broadly speaking, privacy ensures that research subjects are not identifiable, whereas security ensures the data remain inaccessible to non-essential personnel. Despite best efforts, the scientific literature and popular press are replete with examples of data security breaches and privacy violations [1-3]. Combined with

the increase in collaborative projects between researchers and institutions, demands upon data privacy and security will only increase over time. Information technology groups may place more restrictions to protect data security and privacy, and institutional review board (IRB) applications may entail stricter scrutiny regarding protections, yet current technological solutions may be out of reach to the average epidemiologist. The goal of this commentary is to draw a distinction between the requirements of privacy and security, introduce future technologies for protecting the individual, and call for an active dialogue to articulate requirements for such technologies to be useful in epidemiological analyses for data derived from a common and timely source: the electronic health record (EHR).

A Motivating Example

Consider an epidemiologist collaborating with a hospital who has access to an EHR-derived, de-identified dataset representing a retrospective, open cohort of patients. Depending on how many years back these data go, this dataset may be quite rich in both the number of observations and variables. Even though patients may be not directly identifiable from the data, they may be readily linkable to other data sources (such as public records, social media, or online forums) that could be used for re-identification [1]. How do we ensure the privacy of these patients, while recognizing the analysis may require specific markers at the individual level? How do we ensure the security of the data acknowledging the collaborative nature of public health work, and the possible need to share the data with other investigators? In order to answer these questions, we review legislated requirements to health care data and then discuss the role of the researcher.

Privacy, Security, and Secondary Data

Depending upon the location of the researcher, there are various information privacy and security laws and regulations that promulgate the use of health information. The European Union established the Data Protection Directive to govern the use of sensitive personal data, including health and demographic markers (e.g., race) within its member states. One example of law to comply with this directive is the United Kingdom's Data Protection Act, although it has been criticized for failing to provide adequate safeguards [4]. This directive will be superseded by the General Data Protection Regulation in 2018, which explicitly includes provisions for using data for secondary purposes in research. In Canada, the Personal Information Protection and Electronic Documents Act provides the overarching guidelines for the use of personal information in business, including the health sector, and in the U.S. the Health Insurance Portability and Accountability Act (HIPAA) governs basic protections for using health information, including protected health information or PHI. Releasing health information for research purposes under these regulatory environments follows prototypical patterns: we illustrate these approaches using HIPAA as an example.

The data collection process is protected by oversight by a review process involving an IRB, human subjects review board, or ethics committee, depending on the source of the data. For example, HIPAA requires an ethics board to make determinations about use of PHI in research on human subjects. Examples of PHI include obvious identifiers, such as patient name, birth date, and social security number, but also subtler identifiers, such as admission and discharge dates, postal codes, and Internet Protocol address numbers [5].

Certain exemptions to informed consent may be possible. Through the HIPAA Privacy Rule, patient data from the EHR typically do not require informed consent so long as they are retrospective and reside at the institution. In fact, these data will likely receive an exempt IRB review provided subjects are non-identifiable [6]. Consequently, secondary analyses of EHR data are becoming quite popular in public health research. These data sets are subject to certain privacy and security requirements, and should not be viewed as onerous to the researcher nor relegated to groups external to the research process.

A data set has to be certified as shareable before it can be released. There are two main mechanisms in HIPAA for ensuring privacy in research datasets [7]. The less-commonly used "expert determination'' approach requires a trained individual to declare that there is no reasonable threat of re-identification. The Safe Harbor provision provides a list of identifiers that must be removed from data prior to use [8]. Both approaches have significant drawbacks. There have been several well-publicized works showing how to combine anonymized data with public records or other data to re-identify individuals. Examples include data sets on movie ratings [2], Internet searches [9], and genomics [3]. Although we do not have evidence of "wholesale" application of these re-identification attacks, experts should be more cautious about declaring that a particular data set is sufficiently de-identified [10]. From an epidemiological perspective, the Safe Harbor clause may remove some features of clinical care, such as admission date, potentially important for answering a pressing research question, and inclusion of these PHI will require additional IRB scrutiny.

Privacy policies also require security safeguards for storage and access to private data. For example, the HIPAA Security Rule enumerates administrative, physical, and technical safeguards required of electronic health information that the institution must implement. These safeguards may include strong passwords with two-step authentication (password plus an additional token, such as biometric fingerprint), ensuring software are up-to-date with the latest security patches, and encrypted storage solutions. To ensure compliance with the highly technical Security Rule, institutions can seek certification by organizations such as HITRUST (https://hitrustalliance.net). Ultimately it is up to the epidemiologist to abide by these solutions, and not try to circumvent security schemes by copying the research dataset to a USB stick or personal laptop. One positive from recent high profile health data breaches [11] is the movement away from bypassing security.

Bringing Informatics to Public Health

Working with secondary data is of course not new, nor are the required safeguards. A commentary from 1996 noted that privacy and security issues were more of a social and policy shortcoming than a technological hurdle [12], and now that EHRs are in use by >75% of U.S.-based providers [13], the protections needed are even greater. Making this issue particularly salient today are the hosts of recent examples describing large-scale healthcare data breaches affecting 500 or more individuals [11]. Among these publicized breaches occurring at the healthcare provider from 2009 to the time of writing (n=1131), the EHR was the source of the breach in 8% (n=86) and a network server in 14% (n=156). Meanwhile a personal computer or portable device was the source in 47% (n=528) and email in 11% (n=119), suggesting that the onus is not solely upon IRB, or information technology groups, to ensure the safeguards of the data: the researcher must take responsibility for the responsible use and management of the data. Our goal is to encourage the public health community to engage with the developers of

new privacy and security tools — by formalizing and articulating how they use data — and become stakeholders in emerging technological solutions for data sharing across institutions.

Collaborative research involving retrospective analyses of EHR data can involve many different access models. Although an onsite epidemiologist can perform the analysis using a secure computer within the institution, in many cases data must be shared with other researchers. Sharing a copy of the data (e.g., by using an encrypted USB drive or Internet service) requires the research parties enter an IRB-approved data sharing agreement. This all too common paradigm also presents the greatest risk: the institution effectively loses control over the data. The shared data can be stolen and potentially re-identified by linking to external data sources. At the other extreme, data may be available only locally. Any individual working on this research project would be sequestered in a proverbial (or literal!) silo: a windowless, locked room without network connectivity. This places a great burden upon the researchers and may not even be realistic given the global collaborations that many researchers undertake.

These two approaches for privacy and security strike a different balance between the rights of individuals and the public good. In public health, the balance of having data that are useful for epidemiological analysis while protecting the individual can be paradoxical: regression analyses are often conducted at a level that require individual level information yet the results are generalizable to a population of people. We therefore operate squarely in the gray zone between these two competing interests.

Separating the Data from the Analysis

To return to the two questions posed earlier in our example scenario, there are several new technologies under active development in the computer science community that offer significant promise for collaborative research on private and sensitive data. In such systems, data live in a secured distributed system and the questions of those data (i.e., the analysis) are asked remotely [14-17]. While this may be attractive from an institutional perspective, the statistical and analytic techniques provided in existing solutions may be insufficient to perform typical epidemiological analyses. We contend that epidemiologists should engage computer scientists to ensure such systems will provide useful functionalities for their work.

The simplest model is one in which a researcher develops a statistical method (e.g. in SAS, R, Stata, etc.) and a secure server evaluates that model on the actual data. End-to-end encryption prevents eavesdropping on the results, but such an approach also implies that the researcher is trusted to not reveal information from the queries. A more complex model may restrict the kinds of operations that can be performed on the data to a prespecified menu. Yet without input from researchers, these preprogrammed analyses may not accommodate typical epidemiologic methods.

Research consortia (for example around specific diseases [18]) exemplify a more complex scenario: several researchers, each with their own data, wish to collaborate. New technologies such as secure multiparty computation (MPC) can perform an encrypted computation such that neither the researcher nor server learns more about each other's data than the result of the computation [19]. In this setting the communication and computation are encrypted. Current MPC implementations suffer from high computational cost, but there has been rapid development of more practical approaches in the literature [20-22]. This approach

could be promising for performing more sophisticated joint studies beyond simple meta-analyses.

A different paradigm is that proposed by differential privacy, which gives a statistical privacy guarantee for privacy: the result of a computation should not reveal too much about individual data records [23]. Differentially private algorithms guarantee this by randomizing the result of the statistical computation; the simplest instance of this is addition of noise [24]. In a differentially private remote-access system, researchers may have different levels of access to the data. Again, a limited menu of analytic techniques can be made available, with more advanced epidemiologic modeling available to vetted researchers. Differential privacy involves balancing privacy concerns with the utility of the analyses: too much noise can render results meaningless but very private. Prototype systems are evaluating the practicality of differential privacy in a variety of systems from search-engine analytics [25] and mobile devices [26] to neuroimaging [16] to social science [27] and medical informatics [19].

Engaging With Emerging Technologies

MPC and differential privacy are technologies under active development today, and several research programs are attempting to bring them into mainstream usage in practical settings. This commentary is a call to the public health community to engage with developers to ensure that researchers' interests are represented. Specifically, we identify four actionable items.

First, disclose to the appropriate parties the paradigm under which epidemiological analysis occurs. Much of our research and methods are built around individual level data and in order to correlate risk factors with disease require access to these types of data. In other words, individual level analyses require individual level data. The tools that separate the data from the analysis must include these types of methods. Second, become stakeholders in the conversation. The groups that are developing these tools are likely different from the groups that will use these tools. The computer science and informatics communities develop technologies in response to known needs from applied researchers. In order for our research paradigm to be incorporated we need to be proactive and engage the appropriate groups. One way to accomplish this is by familiarizing ourselves with technological developments and engaging in joint research projects to design prototype systems, as other communities have done [27-28]. Third, lower the barriers to using these tools. At present, there is still too great a level of expertise to installing and using the systems. These modalities are not embedded in our statistical software, and therefore may be inaccessible to the researcher despite being mandated by institutional policy. Fourth, call others to action. This can be done by organizing research activities, conference workshops on data privacy and security, and writing white paper requirements for the technologies. Another avenue may be to seek funding for seed projects or collaboration on an Institutes of Medicine big challenges paper to influence the direction of the field.

In conclusion, recent data breaches indicate that privacy and security of health data derived from EHRs are continuing concerns, and public health researchers need to proactively participate in the development of new technologies that better ensure data protection. We have identified four actionable items that can help ensure that our methodological requirements are considered in the next generation of collaborative research tools.

Acknowledgements
No acknowledgements

Financial disclosures
No financial disclosures

Competing Interests
No competing interests

References
1. Sweeney L. Weaving technology and policy together to maintain confidentiality. J Law Med Ethics. 1997 Summer-Fall;25(2-3):98-110, 82.
2. Narayanan A, Shmatikov V. Robust De-anonymization of Large Sparse Datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (SSP) (pp. 111–125). 2008.
3. 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 Aug 29;4(8):e1000167.
4. Blightman K, Griffiths SE, Danbury C. Patient confidentiality: when can a breach be justified? Contin Educ Anaesth Crit Care Pain. 2014;14(2):52-56.
5. Legal Information Institute. 45 CFR 164.514 - Other requirements relating to uses and disclosures of protected health information. https://www.law.cornell.edu/cfr/text/45/164.514. Accessed August 12, 2016.
6. U.S. Department of Health and Human Services (HHS). 45 CFR 46. http://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html. Accessed August 12, 2016.
7. U.S. Department of Health and Human Services (HHS). Guidance on De-identification of Protected Health Information. Available at: http://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf. November 26, 2012.
8. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. 2010 Mar-Apr;17(2):169-77.
9. Backstrom L, Dwork C, Kleinberg, JM. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. Communications of the ACM. 2011 Dec;54(12):133-41.
10. Harvard Law Bill of Health. Ethical Concerns, Conduct and Public Policy for Re-Identification and De-identification Practice: Part 3 (Re-Identification Symposium). http://blogs.harvard.edu/billofhealth/2013/10/02/ethical-concerns-conduct-and-public-policy-for-re-identification-and-de-identification-practice-part-3-re-identification-symposium/. Accessed August 12, 2016.
11. U.S. Department of Health and Human Services (HHS) Office for Civil Rights. Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information. Breaches Affecting 500 or More Individuals. https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf. Accessed June 1, 2016.
12. Barrows RC Jr, Clayton PD. Privacy, confidentiality, and electronic medical records. J Am Med Inform Assoc. 1996 Mar-Apr;3(2):139-48.

13. Centers for Disease Control and Prevention. 2016. Adoption of Certified Electronic Health Record Systems and Electronic Information Sharing in Physician Offices: United States, 2013 and 2014. http://www.cdc.gov/nchs/data/databriefs/db236.htm. Accessed February 1, 2016.
14. Carter KW, Francis RW, Bresnahan M, et al. ViPAR: a software platform for the Virtual Pooling and Analysis of Research Data. Int J Epidemiol. 2015 Oct 8. pii: dyv193. [Epub ahead of print].
15. Gaye A, Marcon Y, Isaeva J, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014 Dec;43(6):1929-44.
16. Sarwate AD, Plis SM, Turner JA, et al. Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. Front Neuroinform. 2014 Apr 7;8:35.
17. Lindell Y, Pinkas B. Secure Multiparty Computation for Privacy-Preserving Data Mining. Journal of Privacy and Confidentiality. 2009;1(1):59–98.
18. Autism Brain Imaging Data Exchange. Introduction. http://fcon_1000.projects.nitrc.org/indi/abide/. Accessed August 12, 2106.
19. Meeker D, Jiang X, Matheny ME, et al. A system to build distributed multivariate models and manage disparate data sharing policies: implementation in the scalable national network for effectiveness research. Journal of the American Medical Informatics Association. 2015; 22(6): 1187–1195.
20. Damgård I, Geisler M, Krøigaard M, et al. Asynchronous Multiparty Computation: Theory and Implementation. In Jarecki S, Tsudik G, eds. Public Key Cryptography – PKC 2009. Berlin, Germany: Springer Berlin Heidelberg; 2009:160-79.
21. Pinkas B, Schneider T, Smart NP, et al. Secure Two-Party Computation Is Practical. In Matsui M, ed. Advances in Cryptology – ASIACRYPT 2009. Berlin, Germany: Springer Berlin Heidelberg; 2009:250–267.
22. Ben-David A, Nisan N, Pinkas B. FairplayMP: a system for secure multi-party computation. In Proceedings of the 15th ACM Conference on Computer and Communications Security (pp. 257–266). 2008.
23. Dwork C, McSherry F, Nissim K, et al. Calibrating Noise to Sensitivity in Private Data Analysis. In Halevi S, Rabin T, eds. Theory of Cryptography. Berlin, Germany: Springer Berlin Heidelberg; 2006:265–284.
24. Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science. 2014;9(3-4):211–407.
25. Erlingsson Ú, Pihur V, Korolova A. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (pp. 1054–1067). 2014.
26. Evans, J. What apple users need to know about differential privacy. Computer World. 2016 Jun.
27. Harvard School of Engineering and Applied Sciences. Privacy Tools for Sharing Research Data. http://privacytools.seas.harvard.edu. Accessed June 10, 2016.
28. Plis SM, Sarwate AD, Wood D, Dieringer C, Landis D, Reed C, Panta SR, Turner JA, Shoemaker JM, Carter KW, Thompson P, Hutchison K and Calhoun VD. COINSTAC: A Privacy Enabled Model and Prototype for Leveraging and Processing Decentralized Brain Imaging Data. Front. Neurosci. 2016;10:365.