Between Access and Privacy: Challenges in Sharing Health Data

Bradley Malin[1,2], Kenneth Goodman[3], Section Editors for the IMIA Yearbook Special Section
[1] Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA
[2] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA
[3] Institute for Bioethics and Health Policy, University of Miami, Miami, Florida, USA

Summary

Objective: To summarize notable research contributions published in 2017 on data sharing and privacy issues in medical informatics.

Methods: An extensive search of PubMed/Medline, Web of Science, ACM Digital Library, IEEE Xplore, and AAAI Digital Library was conducted to uncover the scientific contributions published in 2017 that addressed issues of biomedical data sharing, with a focus on data access and privacy. The selection process was based on three steps: (i) a selection of candidate best papers, (ii) the review of the candidate best papers by a team of international experts with respect to six predefined criteria, and (iii) the selection of the best papers by the editorial board of the Yearbook

Results: Five best papers were selected. They cover the lifecycle of biomedical data collection, use, and sharing. The papers introduce 1) consenting strategies for emerging environments, 2) software for searching and retrieving datasets in organizationally distributed environments, 3) approaches to measure the privacy risks of sharing new data increasingly utilized in research and the clinical setting (e.g., genomic), 4) new cryptographic techniques for querying clinical data for cohort discovery, and 5) novel game theoretic strategies for publishing summary information about genome-phenome studies that balance the utility of the data with potential privacy risks to the participants of such studies.

Conclusion: The papers illustrated that there is no one-size-fits-all solution to privacy while working with biomedical data. At the same time, the papers show that there are opportunities for leveraging newly emerging technologies to enable data use while minimizing privacy risks.

Keywords

Health records, genomics, data sharing, privacy, security

Yearb Med Inform 2018:55-9
http://dx.doi.org/10.1055/s-0038-1641216

Introduction

Elegant simplicity can sound naïve. For instance, a wise biomedical researcher might try to inspire students by suggesting something like "the systematic collection and analysis of data and information lead to knowledge; knowledge when applied properly leads to improved health." A wise lawyer, philosopher, or legislator might observe that "biomedical data and information are often about people; among the rights people enjoy is that of privacy." Somewhere between abstract wisdom and naïve platitudes, researchers work to foster the growth of knowledge, often under constraints that attempt to find a balance between unfettered data collection and sharing and the right of data sources to have some say in that collection and sharing.

There is something gorgeous in Pierre Charles Alexandre Louis observing, in 1834, "As to different methods of treatment, it is possible for us to assure ourselves of the superiority of one or other … by enquiring if the greater number of individuals have been cured by one means than another. Here it is necessary to count. And it is, in great part at least, because hitherto this method has not at all, or rarely, been employed, that the science of therapeutics is so uncertain." [1] This "numerical method," extolled by Osler as "simple" and "self-evident," [2] is correctly regarded as among the foundations of evidence-based medicine.

To be sure, that was in the days of Not Very Big Data. Looking for patterns in a database was often a matter of, well, counting: if one number was bigger than another, then it was reasonably safe to infer that something had been discovered. Nowadays, repositories store vast amounts of data and machines, sometimes intelligent machines, analyze it. We still seek patterns, and in doing so amplify one of the most interesting challenges in the history of empirical inquiry: how ought we seek the benefits of increased knowledge and simultaneously respect expectations that personal data and information will not be used inappropriately?

This is a non-trivial question to answer because there are many factors that can influence the answer and there is likely no one-size-fits all solution. Patients and clinical trial participants (along with their delegates and surrogates) may wish to control who gets access to information about them and under what conditions. At the same time, there is a pull to reuse data about patients to discover new knowledge, as well as to support learning healthcare systems, integrating data from disparate healthcare organizations. And while such secondary uses of health data have the potential to benefit society, healthcare organizations have legitimate concerns that the systems in which they store personal and potentially sensitive information will be hacked and exposed, leaving patients open to privacy violations and the organization susceptible to reputational loss and fines. These are only some of the players in this complex system and merely a sampling of the various viewpoints on what health information should be shared and for what purposes. As the quantity of health data continues to grow at an unprecedented rate, and the sophistication of attacks to commit intrusions grows, the medical informatics community has worked to develop new technologies and implement best practices to resolve the tension. It is in

this Big Data and Big System setting that we set out to recognize the most notable recent work in the area.

Paper Selection Methods

The paper selection process began with a search of papers in PubMed/Medline, Web of Science, ACM Digital Library, IEEE Xplore, and AAAI Digital Library. The search, which was performed by one of the section editors, was conducted in January 2018 for papers published during 2017. The search focused on papers published in the English language and related to the topics of medical informatics, data privacy, and data access. In addition to a search of the electronic databases, a manual search was conducted of high impact journals in the field (e.g., Journal of the American Medical Informatics Association, Journal of Medical Internet Research, and International Journal of Medical Informatics). The keywords used for the search included coded, as well as free-text terms. The former were drawn from MeSH terms (e.g., "privacy", "confidentiality", and "data sharing") while the latter were based on the experience of the section editors (e.g., "de-identification" and "re-identification"). These terms were applied in the search in an iterative fashion to refine the search methodology, so that papers retrieved focused on medical informatics as opposed to more general investigations.

The initial search yielded 465 papers. A manual search yielded an additional 74 papers. The two section editors then performed an initial screening of the titles and abstracts to determine which papers were of potential merit. This screening process led to papers being classified as 1) definitely consider, 2) definitely do not consider, and 3) maybe consider. They then read the "definitely consider" and "maybe consider" papers to reach consensus on a list of 14 candidate papers. Papers were considered with respect to their:

1) Topic's importance to medical and health informatics
2) Scientific and/or practical impact of the paper to the topic
3) Quality of scientific and/or technical content
4) Originality and innovativeness
5) Coverage of related literature
6) Organization and clarity of presentation

In accordance with the International Medical Informatics Association (IMIA) Yearbook selection process, the candidate best papers were assessed by the two section editors and a collection of no fewer than four additional external reviewers. The papers were then discussed at the Yearbook editorial board meeting, and five papers were selected as best papers.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2018 in the special section 'Between Access and Privacy: Challenges in Sharing Health Data'. The articles are listed in alphabetical order of the first author's surname.

Special Section Between Access and Privacy: Challenges in Sharing Health Data
▪ Gilbert M, Bonnell A, Farrell J, Haag D, Bondyra M, Unger D, Elliot E. Click yes to consent: incorporating informed consent into an internet-based testing program for sexually transmitted and blood-borne infections. Int J Med Inform 2017;105:38-48.
▪ Humbert M, Ayday E, Hubaux JP, Telenti A. Quantifying interdependent risks in genomic privacy. ACM Transactions on Privacy and Security 2017;20(1):3.
▪ Ohno-Machado L, Sansone SA, Alter G, Fore I, Grethe J, Xu H, Gonzalez-Beltran A, Rocca-Serra P, Gururaj AE, Bell E, Soysal E, Zong N, Kim HE. Finding useful data across multiple biomedical data repositories using DataMed. Nat Genet 2017;49(6):816-9.
▪ Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Malin B. Expanding Access to large-scale genomic data while promoting privacy: a game theoretic approach. Am J Hum Genet 2017;100(2):316-22.
▪ Yuan J, Malin B, Modave F, Guo Y, Hogan WR, Shenkman E, Bian J. Towards a privacy preserving cohort discovery framework for clinical research networks. J Biomed Inform 2017;66:42-51.

Conclusions and Outlook

The papers from 2017 cover the lifecycle of data collected and shared for biomedical research and healthcare while addressing privacy and security issues. The papers highlight emerging technologies (e.g., secure multiparty computation and blockchain), but also tried-and-true strategies, such as electronic consent forms.

Gilbert et al. [3] investigated the first step of the lifecycle, which corresponds to informed consent (IC) and the initial collection of data from patients. Their study focused on the transition from consent in traditional clinical settings to the at-home setting where consent needs to be solicited online. They specifically investigated Internet-based services associated with testing for sexually-transmitted and blood-borne infections (STBBI). In this environment, they relied on interviews with approximately 15 individuals to evaluate the acceptability of various designs of a mandatory consent page for GetCheckedOnline, an Internet-based STBBI service based in Canada. In general, most participants understood the IC page requirements, while it was found that those with more experience with testing tended to exhibit more comprehensive understanding. One notable aspect of this study was that interviews with gay men indicated this subgroup had a substantive understanding of IC, which suggests it is not merely testing experience but cultural history that influences the acceptability of online IC processes.

IC allows for individuals to enter the system with confidence and for data to be collected about them, after which the data is stored and processed in a relatively isolated fashion. This is an artifact of clinical activities and biomedical research being conducted in a distributed manner. As a consequence, it is challenging to determine what data is available and how to get access to it. Seeking to mitigate this problem, Ohno-Machado et al. [4] introduced DataMed, a data index and search engine that is based on the metadata extracted from a collection of repositories. DataMed is based on the biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE), a Data Discovery Index Consortium initially funded by the Big Data to Knowledge (BD2K) program of the US

National Institutes of Health, but which includes 86 members from 56 institutions in the United States and European Union. The first prototype of DataMed was designed to establish a shallow generic index that, as of May 2017, covered 66 repositories, with more than 1.3 million data sets and 15 types of data. The data is described using the DAtaset Tag Suite (DATS) model, which is based on the principles of PubMed's Journal Article Tag Suite. Based on this index, users can build natural language and structured queries to search for datasets that will be of use in their investigations. In support of this system, the consortium has established use cases, pipelines for data ingress, an interface for query and result retrieval, and pilot studies to illustrate the potential of the system. They also performed a comparison of the tool with more generic web search systems (Google and Bing) to show how DataMed facilitates more precise search results.

Making data available can support more effective clinical care and make biomedical research more efficient. At the same time, there are concerns that potentially sensitive information could be disclosed without the consent of the individuals to whom the data corresponds. In this respect, Humbert et al. [5] considered privacy risks associated with genomic data, which, due to radically decreasing costs in high-throughput technologies, is increasingly collected and utilized in a variety of environments both in and outside of the traditional clinical domain. In this work, the team specifically considered how heritability influences what information can be predicted about an individual given genomic data about their family members (e.g., predicting the genomic status of a child based on her/his parents genomic data). It was shown that a belief propagation algorithm can be applied to perform such inferences to achieve reconstruction attacks (so-named because it is assumed that the targeted individual might be hiding some of her or his genomic information) and that the disclosure rates are non-trivial. They then went on to show how to quantify the extent to which phenotypic information (e.g., disease status) can be predicted based on genomic information. By elucidating these attacks with computational formalisms and publicly accessible data from websites like OpenSNP.org and Facebook, the team was able to illustrate the potential for inferential concern.

One of the reasons why privacy violations, such as inference attacks, are plausible is that individual-level records are observable. Yet, in the context of learning health systems and ever-larger biomedical investigations, the individual record may not be as critical as obtaining statistically-relevant evidence about associations between various factors of interest (e.g., patients' demographics and the efficacy of a drug's ability to mitigate the effects of a certain malady). Given that the data needed to generate such statistics is, as observed earlier, often fragmented across various institutions, Yuan et al. [6] developed a privacy-preserving cohort discovery (PPCD) system for clinical research networks (CRNs). This system is based on elliptic curve cryptography to allow for the comparison of records against queries without disclosing what the values are in the records. To make the process efficient, the protocol invokes the use of a hub site (for centralized management) and blocking that selectively reveals information that is provably limited in the amount of information it discloses. The feasibility of this protocol was illustrated by translating the definition of three cohorts: 1) elderly cervical cancer patients who underwent radical hysterectomy, 2) oropharyngeal and tongue cancer patients who underwent robotic transoral surgery, and 3) female breast cancer patients who underwent mastectomy. These definitions were then tested on an encrypted database of 7.1 million records collected from the Nationwide Inpatient Sample of the US Healthcare Cost and Utilization Project. Using commodity servers, it was shown that queries could be completed within 2 to 4 minutes, but that the problem is parallelizable and can be readily reduced in time by orders of magnitude.

As an alternative to sharing patient-level records in the clear or querying them in an encrypted manner, it can be useful to share aggregate statistics about the data that resides in a repository. In effect, this is a summary of the information, such that it can hide specific records. However, it has been shown that when a user of such a resource is in possession of a named record, then the user can still perpetrate a presence detection attack. In this attack, the user compares the record to the published summary statistics from some resource against those available in some reference population. If the targeted record is sufficiently similar to the resource of interest, then the user has evidence to claim that he or she contributed to (and thus is a member of) the resource. This was first shown by Homer et al. [7], where they assumed users would know the identity of a genomic sequence, which they could then compare against summary statistics of the cases in a genome-phenome study and a reference population of a program like 1000 Genomes Project. While this attack was found to be possible, Wan et al. [8] showed that the extent to which presence detection attacks are likely will be dependent upon a number of factors, including the prior probability that the targeted individual could be in the resource and the incentive one gets from being successful in the attack. Based on this observation, they mapped the presence detection attack into a game theoretic framework and showed how protection could be achieved by 1) changing the cost of using the resource (e.g., penalties for misuse) and 2) suppressing certain information (e.g., regions of the genome). They then illustrated how this approach could be applied to the Sequence and Phenotype Integration Exchange (SPHINX) resource from the Electronic Medical Records and Genomics Network (eMERGE) and how it led to greater amounts of data sharing than would be possible under the more traditional view of attacks as simply possible.

The remaining papers focuse on themes of collection, sharing, and access across the biomedical data lifecycle.

One of the main themes observed was at the beginning of that lifecycle, namely consent and the environment surrounding the consent process. St. John et al. [9] showed that hand-written consent forms for surgical procedures are laden with various problems, including missing and inaccurate information, poor legibility, and high variation. They showed that generating forms through an electronic mechanism addressed these problems, providing evidence for the support of an e-consent process.

Another theme focused on the security of systems that manage patient data while supporting primary care services. Hassidim et al. [10] ran a Facebook survey with medical and para-medical personnel to investigate the

extent to which the users of electronic health record systems kept their credentials, namely user IDs and passwords, confidential. Of 299 responses, over 73% indicated they had obtained the password of another medical staff member. Notably it was found that all of the resident physicians (45 of 45) had done so, while only 57% (38 of 66) of nurses had engaged in this behavior. The notion of managing access control in a distributed setting was addressed by Brandizi et al. [11]. They demonstrated a pilot system that integrated various open-source components, based on digital identity federation, that enables open and restricted access to data associated with biobanking and biosample research.

Other papers focused on how to speed up the processing and consumption of patient data and analytics, an endeavor that is necessary in both primary and in secondary use settings. In this respect, Tafti et al. [12] investigated the accuracy, performance, and efficiency of BigML and Algorithmica machine learning as a service environment with four datasets from the Surveillance, Epidemiology, and End Results (SEER) repository and two datasets from the Knowledge Extraction based on Evolutionary Learning (KEEL) repository. It was shown that these systems have various capabilities and costs, but that there remain concerns over the security of the system and the privacy of data during the computation.

A large number of investigations worked to translate biomedical data processing practices into an encrypted setting. Zhu et al. [13] developed a technique based on bilinear pairing to support classification through a support vector machine with clinical decision support applications. Cetin et al. [14] designed a string matching protocol for genomic data queries based on homomorphic cryptography.

Encrypted biomedical data processing is feasible, but in many instances can be slower than what is needed in a real-time distributed clinical system. In this respect, Brown et al. [15] investigated how to make patient data into use multibit trees to facilitate comparisons and similarity searches. This was specifically designed to support the mapping of patient identifiers, such as name, phone number, and address to enable privacy-preserving record linkage.

Other papers focused on how to thwart attacks on patient data shared in the clear. Raisaro et al. [16], for instance, investigated how to make amendments to basic summary data reporting through noise addition techniques (e.g., differential privacy). They specifically investigated how to thwart a presence detection attack that was designed for the Beacon platform from the Global Alliance for Genomics and Health. Prasser et al. [17] developed open-source software to support the game theoretic perspective of how to amend biomedical data (or influence the overall cost of processing and benefiting from data) to simultaneously support data privacy and utility.

These and other publications show that the biomedical data management lifecycle remains complex and fragmented, but that new technologies are emerging and evolving to support collection, sharing, and use. Still, to bring these technologies into practice will require further evaluation, larger pilot studies, and integration into larger clinical and biomedical research enterprises.

Acknowledgement

We would like to thank John H. Holmes for all of his assistance in moving this work forward. We also wish to thank the external reviewers for the time and energy they devoted during the selection process.

References

1. Louis PCA. Essay on Clinical Instruction. Martin P, trans. London: S. Highley; 1834.
2. Osler W. The influence of Louis on American medicine. In: McGovern JP, Roland CG, editors. The Collected Essays of Sir William Osler, vol. III. Birmingham, Ala.: Classics of Medicine Library; 1985. p. 113-34. Originally published in the Johns Hopkins Hospital Bulletin 1897;77-78(August-September):189-210.
3. Gilbert M, Bonnell A, Farrell J, Haag D, Bondyra M, Unger D, et al. Click yes to consent: incorporating informed Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping arrays. PLoS Genetics 2008;4(8):e1000167.consent into an internet-based testing program for sexually transmitted and blood-borne infections. Int J Med Inform 2017;105:38-48.
4. Ohno-Machado L, Sansone SA, Alter G, Fore I, Grethe J, Xu H et al. Finding useful data across multiple biomedical data repositories using DataMed. Nat Genet 2017;49(6):816-9.
5. Humbert M, Ayday E, Hubaux JP, Telenti A. Quantifying interdependent risks in genomic privacy. ACM Transactions on Privacy and Security 2017;20(1):3.
6. Yuan J, Malin B, Modave F, Guo Y, Hogan WR, Shenkman E, et al. Towards a privacy preserving cohort discovery framework for clinical research networks. J Biomed Inform 2017;66:42-51.
7. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping arrays. PLoS Genet 2008;4(8):e1000167.
8. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Malin B. Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach. Am J Hum Genet 2017;100(2):316-22.
9. St John ER, Scott AJ, Irvine TE, Pakzad F, Leff DR, Layer GT. Completion of hand-written surgical consent forms is frequently suboptimal and could be improved by using electronically generated, procedure-specific forms. Surgeon 2017;15(4):190-5.
10. Hassidim A, Korach T, Shreberk-Hassidim R, Thomaidou E, Uzefovsky F, Ayal S, et al. Prevalence of sharing access credentials in electronic medical records. Healthc Inform Res 2017;23(3):176-82.
11. Brandizi M, Melnichuk O, Bild R, Kohlmayer F, Rodriguez-Castro B, Spengler H, et al. Orchestrating differential data access for translational research: a pilot implementation. BMC Med Inform Decis Mak 2017;17(1):30.
12. Tafti AP, LaRose E, Badger JC, Kleiman R, Peissig P. Machine learning-as-a-service and its application to medical informatics. Proc International Conference on Machine Learning and Data Mining in Pattern Recognition 2017: 206-19.
13. Zhu H, Liu X, Lu R, Li H. Efficient and privacy-preserving online medical prediagnosis framework using nonlinear SVM. IEEE J Biomed Health Inform 2017;21(3):838-50.
14. Cetin GS, Chen H, Laine K, Lauter K, Rindal P, Xia Y. Private queries on encrypted genomic data. BMC Med Genomics 2017;10(Suppl 2):45.
15. Brown AP, Borgs C, Randall SM, Schnell R. Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. BMC Med Inform Dec Mak 2017;17:83.
16. Raisaro JL, Tramer F, Ji Z, Bu D, Zhao Y, Carey K, Lloyd D, et al. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. J Am Med Inform Assoc 2017;24(4):799-805.
17. Prasser F, Gaupp J, Wan Z, Xia W, Vorobeychik Y, Kantarcioglu M, et al. An open source tool for game theoretic health data de-identification. AMIA Annu Symp Proc 2017:1430-9.

Correspondence to:
Bradley Malin, PhD
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN, USA
E-mail: b.malin@vanderbilt.edu

Summary of Best Papers Selected for the 2018 Edition of the IMA Yearbook, Special Section

Humbert M, Ayday E, Hubaux JP, Telenti A

Quantifying Interdependent Risks in Genomic Privacy

ACM Transactions on Privacy and Security 2017;20(1):3

Genomic data and information are often especially sensitive, and research using genomic data poses risks of disclosure without consent of the people from whom the data were derived. This study examines how heritability influences what information can be predicted about a person given genomic data about family members. Investigators show how a belief propagation algorithm can be used to design successful attacks based on the assumption that a targeted person might be hiding some genomic information.

Yuan J, Malin B, Modave F, Guo Y, Hogan WR, Shenkman E, Bian J

Towards a privacy preserving cohort discovery framework for clinical research networks

J Biomed Inform 2017;66:42-51

This article reports on the development of a privacy-preserving cohort discovery system for clinical research networks. Elliptic curve cryptography is used to compare individual patient records against queries without disclosing statistically relevant evidence of associations between such things as demographic data and drug response. The protocol uses a blocking mechanism that selectively reveals provably limited information that is disclosed. Three oncology cohorts are defined and the definitions are tested on an encrypted database of 7.1 million records.

Ohno-Machado L, Sansone SA, Alter G, Fore I, Grethe J, Xu H, Gonzalez-Beltran A, Rocca-Serra P, Gururaj AE, Bell E, Soysal E, Zong N, Kim HE

Finding useful data across multiple biomedical data repositories using DataMed

Nat Genet 2017;49(6):816-9

Patient data is often stored and processed in isolation and across several platforms. This makes it difficult to find out what data is available and how to access it. This innovative study includes the creation of a data index and search engine in which metadata is extracted from a collection of repositories. With these tools, users can build natural language and structured queries to search for datasets to inform their investigations.

Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Malin B

Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach

Am J Hum Genet 2017;100(2):316-22

In "presence detection attacks" one compares a record to published summary statistics of some reference population. If the targeted record is similar enough to a resource of interest, the attacker has evidence to claim that he or she contributed to and therefore is a member of the resource. This report describes a project in which a presence detection attack is mapped into a game theoretic framework, and demonstrates ways protection can be achieved. Applications lead to improved data sharing.

Gilbert M, Bonnell A, Farrell J, Haag D, Bondyra M, Unger D, Elliot E

Click yes to consent: incorporating informed consent into an internet-based testing program for sexually transmitted and blood-borne infections

Int J Med Inform 2017;105:38-48

The valid consent process is traditionally undertaken in clinical settings, but this is increasingly inadequate as more people interact online with health professionals and as more research takes place online. This interesting report examines Internet-based services associated with testing for sexually-transmitted and blood-borne infections, specifically the acceptability of various designs for a mandatory consent page. Findings include that individuals with greater testing experience had better understanding of the consent page and that cultural history influences acceptability of the online process.