Biomedical data privacy: problems, perspectives, and recent advances

Bradley A Malin,1,2 Khaled El Emam,3,4 Christine M O’Keefe5

INTRODUCTION

The notion of privacy in the healthcare domain is at least as old as the ancient Greeks. Several decades ago, as electronic medical record (EMR) systems began to take hold, the necessity of patient privacy was recognized as a core principle, or even a right, that must be upheld.1,2 This belief was re-enforced as computers and EMRs became more common in clinical environments.3–5 However, the arrival of ultra-cheap data collection and processing technologies is fundamentally changing the face of healthcare. The traditional boundaries of primary and tertiary care environments are breaking down and health information is increasingly collected through mobile devices,6 in personal domains (eg, in one’s home7), and from sensors attached on or in the human body (eg, body area networks8–10). At the same time, the detail and diversity of information collected in the context of healthcare and biomedical research is increasing at an unprecedented rate, with clinical and administrative health data being complemented with a range of “omics data, where genomics11 and proteomics12 are currently leading the charge, with other types of molecular data on the horizon.13 Healthcare organizations (HCOs) are adopting and adapting information technologies to support an expanding array of activities designed to derive value from these growing data archives, in terms of enhanced health outcomes.14

The ready availability of such large volumes of detailed data has also been accompanied by privacy invasions. Recent breach notification laws at the US federal and state levels have brought to the public’s attention the scope and frequency of these invasions. For example, there are cases of healthcare provider snooping on the medical records of famous people, family, and friends, use of personal information for identity fraud, and millions of records disclosed through lost and stolen unencrypted mobile devices.15 The danger is that such publicized incidents will erode patient trust over time, and lead to privacy protective behaviors. For example, between 15% and 17% of US adults have changed their behavior to protect the privacy of their health information, doing things such as: going to another doctor, paying out-of-pocket when insured to avoid disclosure, not seeking care to avoid disclosure to an employer, giving inaccurate or incomplete information on medical history, self-treating or self-medicating rather than seeing a provider, or asking a doctor not to write down the health problem or record a less serious or embar- rassing condition.16–18 A survey of service members who had been on active duty found that respondents were concerned that if they received treatment for their mental health problems, it would not be kept confidential and would have a negative impact on future job assignments and career advancement.19 Specific vulnerable populations have reported similar privacy protective behaviors, such as adolescents, people with HIV or at high risk for HIV, women undergoing genetic testing, mental health patients, and victims of domestic violence.20–26 A survey of Californian residents found that discussing depression with their primary care physician was a barrier to 15% of the respondents because of privacy concerns.27

On the other hand, some legal scholars are questioning the survival of conventional privacy expectations.28 Privacy has conventionally been defined as an individual’s ability to control the disclosure of personal facts.29 30 However, privacy is also a multi-dimensional concept31 32 and any shifts in privacy expectations are not homogeneous in direction and intensity across all of these dimensions. Furthermore, advances in informatics that may be eroding individuals’ control over their information are being countered by advances in privacy enhancing technologies, as well as regulatory and policy changes that give individuals back control over their information.

This special issue was established to solicit current research in privacy as it is currently understood and is being redefined for emerging biomedical systems. The selected articles consider the different dimensions of privacy, and describe some novel privacy enhancing technologies and their applications, as well as the governance, regulatory, and policy mechanisms that are being used to manage privacy risks.

Privacy is a major patient, provider, regulator, and legislator concern today. There is therefore a need to address these concerns in a practical way that can be deployed in the short term. Deployment must be preceded by a convincing evidence base demonstrating the rationale, costs, and benefits of an intervention. At the same time, new theoretical models and novel approaches that still need to be evaluated and tested in the field, are also necessary to ensure that the field keeps evolving. In putting together this special issue we attempted to balance these two perspectives, with articles presenting results of immediate relevance and applicability, and material covering theoretical work that remains to be proven in practical settings.

There were 53 papers submitted for consideration in this issue, of which 15 were accepted for publication, for an acceptance rate of 25%. All papers were subject to a rigorous review by at least two referees and oversight by one of the guest editors. The review process for papers authored by guest editors, as well as the editor-in-chief, was handled by an unaffiliated associate editor of the journal. In addition to peer-reviewed manuscripts, two invited papers were solicited for the special issue to address the topics of privacy policy and technical data protection mechanisms.
PRIVACY, ZONES, AND SOCIO-TECHNICAL TRACKS

Privacy is an overloaded and complex term. The concept of privacy often subsumes various constructs, such as anonymity (ie, the ability to hide one’s identity), confidentiality (ie, the ability to share information with a second party without the information being publicly revealed), and solitude (ie, the right to be left alone). Even when the particular construct is unambiguous, it remains difficult to have discussions around the topic of privacy because it is highly contextual, such that the expectations of privacy are often specialized to the situation. For instance, a patient’s expectation of privacy changes when disclosing information to a care provider versus a random person on the street. The expectation is further modified by the perceived sensitivity of the health information in question. And, the extent to which health information (eg, a positive assertion of an HIV diagnosis) is deemed to be sensitive varies from patient to patient. This special issue is organized to trace the lifecycle of biomedical information, which we coarsely partition into three zones that follow the general data lifecycle as follows.

Collection zone

The first zone corresponds to the point at which health information is collected from patients. The collection may occur while an individual is physically located at a healthcare provider or beyond (eg, such as through a website on the internet or an application running on a mobile device). In this zone, privacy tends to be concerned with who can collect health information, how much information should be collected, at what time, and for what purposes. The specific notions of privacy addressed in this zone tend to be associated with anonymity (eg, Is the recipient of the data permitted to know the identity of the information from which it is being collected?), limiting content (eg, What is the minimal amount of information that will satisfy the purpose?), and consent (eg, Did the patient agree to the terms of the data collection?).

Primary use zone

The second zone corresponds to the context in which the data have left the control of the patient and are housed in a system controlled, or accessed by, those who provide a primary service (eg, provision of care, study of biomedical data explicitly solicited for a specific research project). In this zone, privacy tends to be realized through confidentiality (eg, Who is permitted to access or use the data and for what purposes?) and security (eg, How can we ensure that the data are protected from misuse or abuse while at rest in a database or in transit between authorized entities?).

Secondary use zone

The third zone corresponds to the scenario in which biomedical data are utilized for purposes which are different from their primary use. The data may be used by the organization that initially collected the data (eg, repurposing of clinical data for research) or disseminated to external entities (eg, publication of public use datasets) for the performance of certain tasks (eg, evaluation of health policies). In this zone, the privacy issues that tend to arise are anonymity and consent for individuals and groups (eg, Can data collected from a particular ethnic group be reused to study a specific phenotype?). Each of these zones can be partitioned into two interacting, although conceptually distinct, tracks. In the first track, biomedical privacy is defined and regulated via socio-legal mechanisms. This is the arena where the public, ethicists, and policy and law makers come together to define what privacy rights and responsibilities exist. In the second track, technical controls are specified and realized in working information technologies to maintain societal expectations of privacy or requirements specified in policy and law. It is critical to integrate these tracks to ensure that privacy expectations are appropriately represented in technical controls and that policies are designed to realistically account for state-of-the-art technical capabilities. It is further important that societal expectations of privacy and privacy enhancing technologies are current and cognizant of shifts in expectation of technical sophistication.

A CHARACTERIZATION OF THE PAPERS IN THE PRIVACY LIFECYCLE

Papers in the socio-legal track

The socio-legal track of this special issue commences by delving into the desires and expectations society harbors for privacy. As mentioned earlier, privacy is a societal phenomenon, such that the extent to which it is realized is depend-ent on how society chooses to codify the concept in policy and law. This process often begins with field studies and sessions that engage stakeholders in their preferences. In this vein, Caine and Hanania report on a study that asks patients who should control access to health information and what granularity of control is desirable. Often, the expectations of privacy are dependent on the domain in which information is collected. As the traditional boundaries of the healthcare domain expand, it is important to determine how individuals’ perspectives on privacy relate to new technologies. To begin to address this issue, van der Velden and El Emam focus on the use of social media by teenage patients, and how they perceive their health information privacy when interacting online. Insights gained here should inform the more general health data context.

The next set of papers in the socio-legal track move beyond primary uses for biomedical data and into secondary settings. In this environment, it is assumed that policies and laws have been codified. However, policy and law is dependent on the locale, such that it is critical to understand how it guides data management practices. The first paper in this group, by Pencarrick Hertzman, Meagher, and McGrail, presents a case study about how the ‘Privacy by Design’ framework was applied in British Columbia (BC), Canada, to facilitate access to health information in Population Data BC. To date, Population Data BC has facilitated over 350 research studies. This work is followed by the first invited paper by McGraw, which takes a look at the de-identification strategy of the US Health Insurance Portability and Accountability Act (HIPAA). This strategy enables HCOs to disclose information about patients in a manner that is no longer subject to oversight by the regulatory authorities because the risk that they would be individually identifiable is deemed very small. Clarifications to what de-identification means and how it can be achieved in accordance with HIPAA were recently published by the U. S. federal government. The paper by McGraw reports on a workshop on various stakeholders’ support for the current HIPAA de-identification strategy held by the Center for Democracy and Technology and discusses policy proposals to address concerns and improve trust in the process. Peterson and colleagues then recount a recent case before the Supreme Court, Sorell vs IMS Health, and illustrate the challenges associated with selling prescription records for various purposes, such as post-market effectiveness. They highlight some of
the concerns associated with the dissemination of identifiable prescriber and de-identified patient information. While the previous papers focus on traditional health information, the last paper in this track, by Kosseim and colleagues, addresses privacy issues associated with the management of *omics data in particular, with a specific focus on genomics. This paper describes the legal and ethical principles and practices adopted in the Canadian provinces of Newfoundland and Labrador to enable research with genomic, phenomic, and genealogical data.

**Papers in the technical track**

While law and policy codify the rights and requirements for managing data privacy in the biomedical domain, information technology is necessary to uphold and ensure their realization in practice. In this regard, certain aspects of privacy can be achieved through information security. The Security Rule of HIPAA specifies various administrative, physical, and technical safeguards that covered entities must have in place (ie, required controls) or document why such protections are not prudent (ie, addressable controls). For instance, it is required that all covered entities ensure that appropriate authorization is provided before employees of an HCO access a patient’s EMR. By contrast, the encryption of health information at rest within the HCO is addressable, but is not required. Along these lines, the technical track of this special issue begins with a paper by Kwon and Johnson that assesses the extent to which 250 HCOs in the USA have (or have not) adopted various security practices. Their analysis demonstrates patterns of leaders, followers, and laggards in their adoption, and provides recommendations for improving regulatory compliance. Although this work provides a high-level assessment of the adoption of security practices, it does not provide specific guidance on data management strategies. Thus, the next paper, by Fabbrini and Lefevre, discusses a privacy threat encountered on a daily basis in primary care settings, specifically the insider threat. This threat is particularly important to study in the healthcare domain because traditional information security controls (eg, role-based access control) are difficult to realize in care settings due to the highly dynamic nature of healthcare teams. An increasing number of publications have proposed auditing strategies for EMRs, however, this line of work is unique in that it suggests EMR users can be ‘explained’ by the diagnoses that are assigned to the patient records they access. This work suggests that data-driven auditing strategies may help winnow the set of accesses to patient records to a manageable size for review by administrative officials (eg, privacy officers) of HCOs.

The next set of papers in the technical section focus on various strategies that can be invoked to protect patients’ privacy when data are shared for secondary use. However, before presenting specific protection methodologies, this section begins with an illustration of types of research studies that can be enabled through de-identified data. The paper by White and Horvitz integrates web search data from Bing and geocoded data from mobile devices to show how search for health information online correlates with an individual’s physical presence at a healthcare providing facility. This research is performed on data that are stripped of user identifiers and location information prior to the analysis. However, there may be times when it is beneficial to link a patient’s record across multiple healthcare institutions, or within a single institution. To support such efforts without revealing a patient’s identity, there has been a flurry of research in private record linkage (or entity resolution). Such linkage is increasingly based on hashed versions of patient identifiers (eg, personal names) or quasi-identifiers (eg, demographics). The paper from Cassa, Miller, and Mandl suggests a protocol to derive a secure fingerprint from genomic data. They indicate how this approach may be applied to track a patient’s record across the research enterprise in place of explicitly identifying information.

The final set of papers focus on strategies for de-identifying various types of health information. Biomedical data can take a wide array of forms, ranging from free text (eg, natural language clinical notes) to structured information (eg, such as discharge databases) to high-dimensional information (eg, genome-wide scans of single nucleotide polymorphisms). The majority of health information is in free text form and so a significant amount of research over the past several years has investigated how to detect and redact a prespecified set of potential identifiers (such as the list of 18 features in the HIPAA Safe Harbor de-identification standard). The paper by Ferrandez and colleagues provides an example of how rule-based (eg, dictionaries, regular expressions, and rules) and machine learning-based methods (eg, conditional random fields, support vector machines, and naive Bayes classifiers) can be combined to construct a free text scrubber for over 100 different types of Veterans Health Administration clinical notes. The paper by Deleger and colleagues illustrates how machine learning-based text de-identification methodology has negligible impact on clinical concept extraction, in the form of medications, from over 22 note types from the Cincinnati Children’s Hospital Medical Center.

Although the previous papers, and others in the literature, illustrate how identifiers in clinical text can be detected, residual information in the text may still leak inferences or indicators of the corresponding patient (or their relatives). As such, informaticians have worked to develop de-identification strategies that are more formal in their guarantees. These strategies are often applied to more simple data structures, such as field-structured database tuples. The paper by Atreya and colleagues illustrates how a patient’s panel of laboratory test results can be unique and potentially used as a key to track a patient back to their identity. To mitigate this attack, they propose a clinically informed perturbation strategy, which adds noise to the test values. An empirical analysis 61 000 Vanderbilt patients’ records illustrates that such perturbation makes it highly unlikely that a patient’s record could be matched to a group of less than 10 individuals while having minimal influence on clinical interpretation. Although offering some probabilistic protection, this style of noise addition does not guarantee defense against an adversary. In this regard, a significant number of publications have suggested aggregation (ie, generalization) strategies could be applied to ensure that every record corresponds to at least k patients (ie, the k-anonymity principle).

More recently, it has been suggested that de-identification strategies based on the redaction of a prespecified list of features or more formal aggregation strategies may not be an appropriate model of privacy protection because they can leak inferences about the patients from whom the data were collected. While this general notion has been challenged, alternative models have been proposed. In the second invited paper, Dwork and Rettenger describe the notion of differential privacy from a theoretical perspective.

In this model of protection,
researchers are permitted to ask queries of a database, which subsequently responds with a perturbed aggregate response (e.g., a count of 5 may be reported as 6). This response is perturbed such that it is guaranteed that the researcher cannot determine whether a specific individual contributed to the database within a certain probability and that the perturbed answer is within a certain bound of the non-perturbed answer. This model has a number of important strengths, but also faces a number of empirical and practical barriers to its deployment in healthcare settings. The final paper of the special issue, by Gardner and colleagues, provides a more practical perspective on how differential privacy could be applied to databases of health information. They demonstrate how this privacy protection model could be applied to a breast cancer dataset from Emory University, but note there are challenges to applying this approach to high-dimensional data.

NEXT STEPS AND THE FUTURE

As this special issue illustrates, the space of data privacy in the biomedical domain is broad and multi-disciplinary. It crosses ethical, legal, and technical boundaries and is specialized to the type of data and process being supported. Consequently, it is not possible to review the entire field in this issue. As we draw this editorial to a close, we stress that numerous topics (e.g., access control, consent management, statistical disclosure control, and policy specification to manage health information flow) were not addressed, but are no less important than those reported on in this issue. At the same time, we note that new computing infrastructures and high-throughput technologies are creating new challenges to privacy that the biomedical community will need to handle in the not too distant future. One technology that we wish to highlight is cloud computing. As cloud computing costs decline and the amount of data generated by healthcare providers grows, it is increasingly the case that the health information is being stored in systems beyond the direct control and oversight of HCOs, and possibly in foreign jurisdictions with different privacy laws and regulations. We believe this issue demonstrates that appropriate socio-technical protections can be defined for emerging systems and that there is a quite diverse community working on developing them, and are confident that research in this area will lead to new appropriate solutions that balance privacy and data utility and system usability.

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

Received 19 November 2012
Revised 19 November 2012
Accepted 19 November 2012

J Am Med Inform Assoc 2013;20:2–6. doi:10.1136/amiajnl-2012-001509

REFERENCES


