Maintaining Privacy when Searching for Patients Using Electronic Medical Records

Summary
The launch of new drugs is contingent on running safe and thorough clinical trials, but the clinical trials process takes an inordinate amount of time, particularly in patient recruitment. The use of big data analytics methodologies which leverage real databases of electronic medical records is accelerating the clinical trials process and making patient recruitment more efficient. However, use of EHR-databases brings with it the risk of loss of patient privacy. We examine the issues and outline best practice guidelines.

Introduction

Recent developments[1] have shown the real benefits arising from the use of big data analytics techniques to interrogate EHR-databases at hospitals directly. With the right software infrastructure, eligible patients can be discovered based on complex combinations of query criteria, and in real time. Trial managers can unearth more candidates, quicker.

But this development raises questions for patient privacy: using these electronic methods, are an individual's personal details at risk of arriving in unauthorised hands?

In this paper, we will examine the technological and legislative environments, analyse the ensuing risks, and propose some best practice methods for mitigating these risks.

Real-time, EHR-based Patient Search for Clinical Trials

How has patient search been performed until now? Traditionally, the call would go out from the pharmaceutical company to their local affiliates, who would contact local hospitals, where medical researchers would comb paper-based patient records for patients who fit the long list of inclusion and exclusion criteria – a laborious, time-consuming, and costly process.

Systems which run real-time, EHR-based patient recruitment for clinical trials obviate these issues. They allow such criteria to be directly and near-instantaneously queried against medical records data of connected hospitals to obtain a complete and completely up-to-date snapshot of the distribution of eligible patient populations available for a trial given the specified criteria.

Yet it is this very accessing of electronic hospital patient data that creates risks of unauthorised access to private, personal information.

Patient Privacy

As a principle, patients should control how data that is derived from them is utilised. The use and sharing of such data by health plans, healthcare clearing houses, and healthcare providers ("covered entities") in the pursuit of payment, treatment, and operations (PTO) is clearly legislated. Usage beyond PTO is also explicitly regulated and subject to the consent of the patient (except for special circumstances, such as public health and law enforcement exemptions). However, obtaining individual patient consent is not always easy, or possible, in practice; for example, when the data has been collected for one purpose (e.g., PTO), but reused for a different purpose[2].

For our primary usage area of interest, clinical research, it is important to note that multiple studies have illustrated that the individual's choice to consent itself creates a skewed population from the perspective of demographic and socio-economic characteristics[3, 4, 5, 6].

Given all of this, how can patient privacy be balanced with the needs of clinical studies?

Regulations and Directives

In the European Union, the Data Protection Directive 95/46/EC provides a foundational set of guidelines by which all person-specific data is collected, used, and shared.

Regardless of the locale, data protection regulations permit the sharing of de-identified data. For instance, the Data Protection Directive, which strictly prohibits secondary uses of person-specific data without individual consent, provides an exception to the ruling in Recital 26, which states that the: "principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable."

This means that any data, including health information in electronic medical records, can be reused for research purposes once it is anonymised. However, what does it mean for data to be "identifiable"? How do we know when it is no longer identifiable? The Data Protection Directive, and similar directives around the world, do not provide explicit guidelines regarding how data should be protected.

For the purposes of this paper, we rely heavily upon the conceptual principles of identifiability as set forth in the Privacy Rule of the US Health Insurance Portability and Accountability Act of 1996 (HIPAA), while our approaches to personal data protection are guided by the "European Medicines Agency policy on publication of clinical data for medicinal products for human use" (EMA Policy/0070).

HIPAA

In the United States, under the Health Insurance Portability and Accountability Act (HIPAA), the Privacy

Rule designates "protected health information" (PHI) as all health information held, or transmitted by a covered entity, that relates to an individual's health. PHI also covers information that identifies an individual, or can be used to identify the individual. As such, it includes many common identifiers (e.g., name, address, birth date, social security number), but also includes potential "quasi-identifiers" that may permit a recipient of the data to determine the identity of the corresponding subject.

When health information does not identify an individual, and there is no reasonable basis to believe that it can be used to identify an individual, it is said to be "de-identified" and is not protected by the Privacy Rule, which is akin to the notion of being anonymised under the EU Data Protection Directive.

45 C.F.R., section 164.514(a) of the Privacy Rule provides the standard for de-identification of individually identifiable health information: "Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information." Meanwhile, Section 164.514(b) outlines pathways to de-identification of health data.

The first route is the "expert determination" method. Health information can be determined as "not individually identifiable" if an expert "with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable … determines that the risk is very small that the information could be used … to identify an individual," and can document their reasoning and conclusions.

The second is the "safe harbour" method, wherein a list of personal identifiers for the individual, as well as for relatives and employers, are removed from the health record.

The HIPAA Privacy Rule also provides direction with respect to "re-identification": "A covered entity may assign a code or other means of record identification to allow information de-identified under this section to be re-identified, provided that: 1) the code or other means of record identification is not derived from or related to information about the individual and is not otherwise capable of being translated so as to identify the individual; and 2) the covered entity does not use or disclose the code or other means of record identification for any other purpose, and does not disclose the mechanism for re-identification." The use of such a code is the basis for the process known as pseudonymisation.

European Medicines Agency Policy

The European Medicines Agency (EMA) has promulgated new policy on the publication of clinical data generated and/or used in the context of a clinical trial[7] to protect and foster public health, while ensuring there is transparency in clinical trials. As part of this policy, the EMA recommends the protection of personal data. Specifically, it states that: "The secondary analysis of personal data will have to be fully compatible with the individual privacy of clinical trial participants and data protection." To accomplish this goal, data use agreements that prohibit unconsented re-identification are recommended: "[recipients of the data will] not seek to re-identify the trial subjects or other individuals from the Clinical Reports in breach of applicable privacy laws."

The Risks of Re-identification

Many studies have shown that there is a real danger of health information, albeit with explicit identifiers removed, being re-identified again, using complex combinations of statistical methods [4, 8, 9, 10, 11, 12, 13, 14, 15, 16]. The question is, however, one of probability rather than possibility: how likely would such a malicious attack be to occur?[17]. And how likely to succeed? Most data protection directives around the world (including HIPAA's Privacy Rule) do not specify outright that data must not be re-identifiable, but rather that it must not be re-identifiable against reasonable means.

Baseline Risk: Safe Harbour

To determine what is an acceptable level of re-identification risk, we investigated how the residual features which would be permissible to share via a HIPAA safe harbour policy might still permit re-identification.

Taking the limited adversarial model used by privacy risk experts in practice, where the adversary is assumed to be able to have knowledge of between five and seven attributes about the patient[18], analysis of public data from the US Decennial Census[19] based on attacks described in literature[8, 20] produced an estimate for the number of uniques reportable in the population. The estimation process was based on the strategy proposed in[21] and gave the result that 0.48 % of the US population was expected to be unique, or a risk of 0.0048.

Types of Risk

There are several ways by which re-identification risk can be defined, which depend on the knowledge and goals of the attacker. In the literature, these risk metrics are often referred to as 1) prosecutor, 2) journalist, and 3) marketer risks[22].

The first two types of risks correspond to scenarios for the most easily attackable record in the data set. Specifically, the prosecutor and journalist risks correspond to the most re-identifiable person in the published data set (i.e., the sample), and in the broader population (e.g., all individuals in the metropolitan Istanbul, Turkey region), respectively. This attack assumes that the most risky individual is the one of interest for targeting by the adversary. An example of the prosecutor attack is shown in Figure 1.

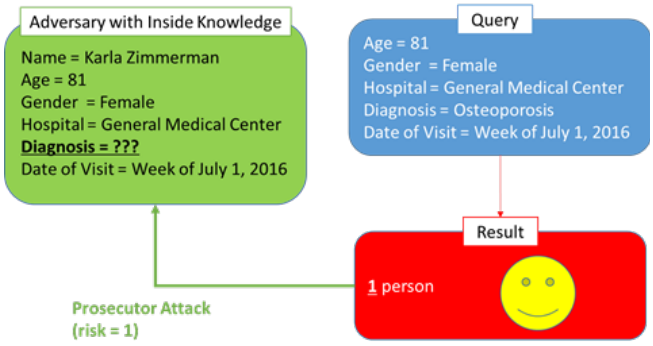

Figure 1. An illustration of a successful prosecutor attack on a query-response data warehouse.

By contrast, the marketer risk is an amortisation of the risk over all individuals in the data set. In this scenario, the adversary aims to re-identify as many individuals as possible, but may not worry about committing a specific targeted identification. An example of the marketer attack is shown in Figure 2. In such a scenario, the marketer risk is defined as the proportion of records that can be correctly re-identified in a data set sampled from a population.

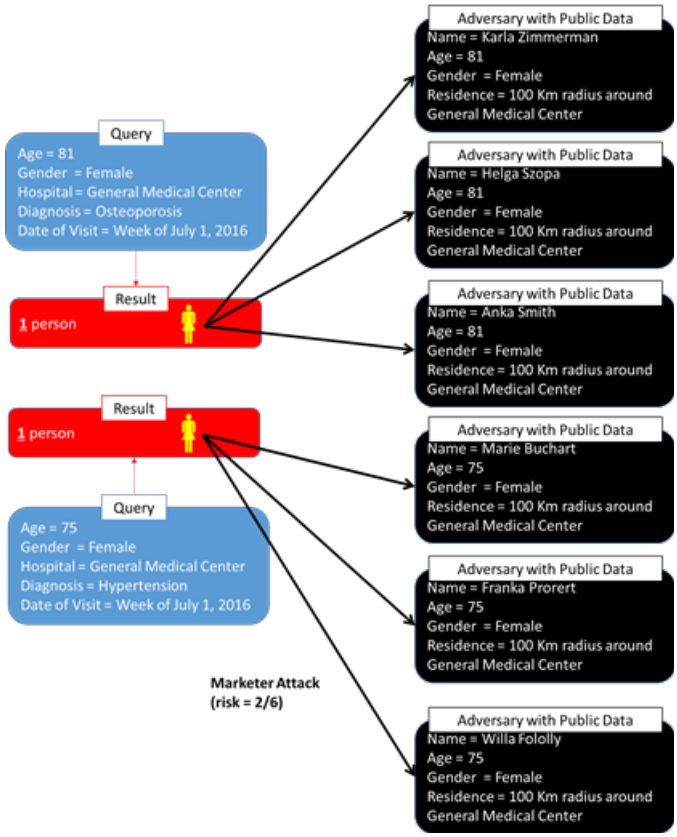

Figure 2. An illustration of a marketer attack on a query-response data warehouse.

Maintaining Privacy: Patient Pseudonymisation

As a first principle, therefore, in maintaining patient privacy, it is important that the data should be de-identified, anonymised, and/or pseudonymised. As we have seen, pseudonymisation is where an individual patient's record is substituted with a value that is unique and consistent for this individual, so that their record may be indexed and managed in the data warehouse, and continue to be updated over time.

There is precedent for incorporating a pseudonym in a record, provided that the recipient was not supplied with the decryption process for the pseudonym. Under the EU Data Protection Directive, the Article 29 Working Group[23] recently stated that: "… using a pseudonym means that it is possible to backtrack to the individual, so that the individual's identity can be discovered, but then only under predefined circumstances. In that case, although data protection rules apply, the risks at stake for the individuals with regard to the processing of such indirectly identifiable information will most often be low, so that the application of these rules will justifiably be more flexible than if information on directly identifiable individuals were processed."

In this model, the only individuals with the ability to translate the pseudonym back to an individual's identity are those who hold the keys to the pseudonymisation process.

As is stated in the Handbook on European Data Protection Law[24], from the European Council (in April 2014) (§2.1.3): "… pseudonymisation of data is one of the most important means of achieving data protection on a large scale, where it is not possible to entirely refrain from using personal data … This is particularly useful where data controllers need to ensure that they are dealing with the same data subjects but do not require, or ought not to have, the data subjects' real identities. This is the case, for example, where a researcher studies the course of a disease with patients, whose identity is known only to the hospital where they are treated and from which the researcher obtains the pseudonymised case histories.

Maintaining Privacy: Mitigation of Risks

Risk itself is a composite of the probability that an attack will be mounted and the probability of the attack's success. The probability of attack can be mitigated by legal, technical, and economic controls, which serve as disincentives to malfeasance in the system and can mitigate the risk of unsanctioned re-identification:

- Economic: recent research has shown that costs can serve as deterrents to potential adversaries and mitigate the chance of re-identification attacks[25, 26]. As a guideline, users of data derived via EHR-based search system must pay a non-trivial monetary sum to provide deterrence in this setting.
- Legal: all users of the system must be made cognisant of acceptable use and confidentiality requirements. While such a policy does not guarantee a user will refrain from violating the terms of service, it does provide the supplier of an EHR-based system the ability to hold users of the system accountable

for their actions in a relevant court of law. Such protections are critical because anonymised (or de-identified) data falls outside the scope of regulation and without such a contract, there would be no accountability for misuse of the data.

- Technical: from a statistical perspective, the data is attenuated to ensure that it does not include sufficient information to facilitate re-identification beyond a certain degree.

While these controls cannot be guaranteed, evidence in information security indicates that they certainly lower the rate at which attacks against such systems are realised[25, 26].

Mitigating Marketer Risk

Running an analysis on US Census estimates, it was found that as the size of a population shrinks, the possibility for re-identification grows. Extrapolating to other countries, the risk of re-identification becomes unacceptably large when the population of the country from which the data is derived drops below three million individuals. As such, this population size can be set as the threshold at which query results from the system must be returned to the end user in an aggregated form only, without disclosing which particular hospital the patients were treated at.

Mitigating Prosecutor Risk

To prevent the suppliers of data to a data warehouse from probing for known individuals upon the integration of such data from other suppliers, the supplier must only return records to suppliers that are aggregated to obscure small values. Each result (or report) returned to suppliers must have a minimum number of patients associated with it during the time period of interest. If this threshold is not satisfied, the report will not be provided.

However, this begs the question of what an appropriate level of aggregation would be.

If we look toward how various agencies apply the "minimum threshold rule" in practice, and relate this to the prosecutor risk, we can determine their threshold of acceptable probability of identification and use it in our own situation. Based on the evidence, we would recommend the use of a threshold of 5. This threshold would be applied in two ways. First, a count of <5 will be displayed for a specific hospital when the user issuing the query works for the institution associated with the count. Second, a count of <5 will be displayed when the total across all hospitals is below the threshold. It is important to recognise that this is a minimum value that is likely to transpire on rare occasions only and that, practically, the risk to an average patient in the system is often significantly smaller.

Conclusion

In this paper, we have attempted to outline the risks to patient privacy stemming from leveraging electronic medical records for the purpose of patient search for clinical trials. We have also proposed best practice methods of mitigating these risks. Protection methods invoked included 1) pseudonymisation, as validated by expert determination, 2) suppression of baseline identity details under safe harbour principles, 3) mitigation of risks via legal, attenuation, and economic controls, 4) aggregated counts for countries with populations below the three million threshold, 5) aggregated counts for results with populations below five. Use of these techniques will mitigate risks that an individual's personal details will arrive in unauthorised hands.

References

1. Beresniak, A., Schmidt, A., Proeve, J., Bolanos, E., Patel, N., Ammour, N., Sundgren, M., Ericson, M., Karakoyun, T., Coorevits, P., Kalra, D., De Moor, G., Dupont, D. "Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the Electronic Health Records for Clinical Research (EHR4CR) European Project," Contemporary Clinical Trials 46 (2016) 85–91.
2. Canadian Institutes of Health Research. CIHR best practices for protecting patient privacy in health research. September 2005.
3. Harris, A., Levy, A., Teschke, K. Personal Privacy and Public Health: Potential impacts of privacy legislation on health research in Canada. Canadian Journal of Public Health. 2008: 293-296.
4. El Emam, K., Kosseim, P. Privacy interests in prescription data, part 2: Research Subject privacy. IEEE Security and Privacy Magazine. 2009; 7(1): 75-78.
5. Kho, M., Duffett, M., Willinson, D., Cook, D., Brouwers, M. Written informed consent and selection bias in observational studies using medical records: systematic review. British Medical Journal (BMJ). 2009; 338: b866.
6. Hill, E.M., Turner, E.L., Martin, R.M., Donovan, J.L. Let's get the best quality research we can: public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study. BMC Medical Research Methodology. 2013; 13: 72.
7. European Medicines Agency. European Medicines Agency policy on publication of clinical data for medicinal products for human use. Policy/0070. October 2, 2014. Available online: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.
8. Sweeney, L. Weaving technology and policy together to maintain confidentiality. Journal of Law, Medicine, and Ethics. 1997; 25(2-3): 98-110.
9. El Emam, K., Jabbouri, S., Sams, S., Drouet, Y., Power, M. Evaluating common de-identification heuristics for personal health information. Journal of Medical

Internet Research. 2006; 8(4): e28.
10. Loukides, G., Denny, J., Malin, B. The disclosure of diagnosis codes can breach research participants' privacy. Journal of the American Medical Informatics Association. 2010; 17(3): 322-327.
11. Sweeney L. Testimony before the National Committee on Vital and Health Statistics Working Group for Secondary Uses of Health Information. August 23, 2007.
12. Brown, I., Brown, L., Korff, D. Limits of anonymisation in NHS data systems. British Medical Journal. 2011; 342: d973.
13. Solomon, A., Hill, R., Janssen, E., Sanders, S., Heiman, J. Uniqueness and how it impacts privacy in health-related social science datasets. Proceedings of the 2nd ACM International Health Informatics Symposium. 2012: 523-532.
14. Cimino, J. The false security of blind dates: chrononymization's lack of impact on data privacy of laboratory data. Applied Clinical Informatics. 2012; 3(4): 392-403.
15. Atreya, R., Smith, J., McCoy, A., Malin, B., Miller, R. Reducing Research Subject re-identification risk for laboratory results within research datasets. Journal of the American Medical Informatics Association. 2013; 20(1): 95-101.
16. Sweeney, L. Matching known patients to health records in Washington state data. Data Privacy Laboratory White Paper 1089-1, Harvard University. June 2013. Available online: http://dataprivacylab.org/projects/wa/1089-1.pdf.
17. Malin, B., Karp, D., Scheuermann, R. Technical and policy approaches to balancing Research Subject privacy and data sharing in clinical and translational research. Journal of Investigative Medicine. 2010; 58(1): 11-18.
18. El Emam, K., Malin, B. Concepts and methods for de-identifying clinical trials data. White Paper for the Institute of Medicine Book, "Sharing clinical trial data: maximizing benefits, minimizing risk". 2015. Available online: http://www.nap.edu/openbook.php?record_id=18998.
19. American Fact Finder. Available online: http://www.factfinder.census.gov.
20. Sweeney, L. K-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems. 2002; 10(5): 557-570.
21. Golle, P. Revisiting uniqueness of simple demographics in the US population. Proceedings of the 5th ACM Workshop on Privacy in the Electronic Society. 2006: 77-80.
22. Dankar, F., El Emam, K. A method for evaluating marketer re-identification risk. Proceedings of the EDBT/ICDT Workshops. 2010: 28.
23. Schaar, P., et al. Article 29 Data Protection Working Party. Opinion 4/2007 on the concept of personal data. 01248/07/EN WP 136. Available online: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf.
24. European Union Agency for Fundamental Rights. Handbook on European data protection law. Available online: http://fra.europa.eu/en/publication/2014/handbook-european-data-protection-law.
25. Khokar, R., Chen, R., Fung, B.C., Lui, S.M. Quantifying the costs and benefits of privacy-preserving health data publishing. Journal of Biomedical Informatics. 2014; 50: 107-121.
26. Wan, Z. Vorobeychik, Y., Xia, W., Clayton, E.W., Kantarcioglu, M., Ganta, R., Heatherly, R., Malin, B. A game theoretic framework for analyzing re-identification risk. PLoS One. 2015; 10(3): e0120592.

Bernhard Bodenmann, PhD, is Lead Analyst at Clinerion and has been working in clinical development and clinical research since 2006, with extensive experience in requirements engineering and system architecture of eClinical solutions.
Email:bernhard.bodenmann@clinerion.com

Le Vin Chin is Head of Marketing and Communications at Clinerion and has been working in communications and marketing for 20 years, in a wide variety of industries, including software and services.
Email: levin.chin@clinerion.com

Bradley Malin, PhD, is an Associate Professor of Biomedical Informatics and Computer Science at Vanderbilt University and has worked in the field of data privacy for over 15 years, with a particular focus on the analysis of electronic medical records.
Email: b.malin@vanderbilt.edu