2015 IEEE 17th International Conference on e-Health Networking, Applications and Services (Healthcom): 2nd International Workshop on Secure and Privacy-Aware Information Management in e-Health

Information Accountability and Health Big Data Analytics: A Consent-Based Model

Daniel Grunwell and Tony Sahama
Science and Engineering Faculty
Queensland University of Technology
Brisbane, Australia
Email: {d.grunwell,t.sahama}@qut.edu.au

***Abstract*—With the ever increasing amount of eHealth data available from various eHealth systems and sources, Health Big Data Analytics promises enticing benefits such as enabling the discovery of new treatment options and improved decision making. However, concerns over the privacy of information have hindered the aggregation of this information. To address these concerns, we propose the use of Information Accountability protocols to provide patients with the ability to decide how and when their data can be shared and aggregated for use in big data research. In this paper, we discuss the issues surrounding Health Big Data Analytics and propose a consent-based model to address privacy concerns to aid in achieving the promised benefits of Big Data in eHealth.**

***Keywords*—Big Data Analytics, Access Control, electronic health records, eHR, eHealth, privacy, security.**

I. Introduction

Health Big Data Analytics has the potential to solve many current and future healthcare problems, from discovering new treatments to improved decision making. Analysis of such large eHealth datasets could enable the discovery of new treatment options [1], improved population health and better policy making [2]. In the current environment, health data is distributed in data silos, and we need to bring the data together to reap the full benefits that big data analytics can provide. However, numerous concerns over the privacy and security of the data hinder such approaches.

In the current eHealth environment, there are conflicting requirements between patients and HCPs, with patients desiring greater control over who can access their information and how it is used, and HCPs wanting easy access to as much medical information as possible to make well-informed decisions. This conflict was highlighted in the recent review of Australia's national Personally Controlled Electronic Health Record (PCEHR) system [3]. In the patient controlled model, HCPs may be discouraged from using such systems because they are unable to rely on eHealth record (EHR) as a complete source of information on a patient they are treating [4, 5]. An appropriate balance of these competing concerns must be achieved so that the full benefits of systems like the PCEHR can be achieved [6].

In order to balance these competing concerns, we devised an Information Accountability Framework (IAF) that applies Information Accountability (IA) protocols to eHealth systems. By applying accountability and transparency for data use, the IAF and so-called Accountable-eHealth systems ensure health information is available to the right person at the right time without rigid barriers while empowering the consumers with control over their information.

The initial model of the IAF was designed for use primarily in shared eHealth Record (SEHR) and local EHR systems, but did not address how the protocols could be applied to Big Data analytics. The possibility of supporting approved research studies while respecting patient privacy preferences through a consent model and ensuring accountability for the information users needs to be investigated and the IAF model needs to be expanded to include these stakeholders.

In this paper, we explore the privacy issues surrounding Health Big Data analytics, and the possibility of using an Information Accountability model and framework to provide a patient consent approach to health data aggregation and study participation. We begin in Section II with a discussion of the background of our work. In Section III, the IAF and IA protocols are explained. In Section IV, we propose a possible consent model for Big Data analytics using the IA protocols. Section V describes some of the major challenges involved in implementing the proposed model. Section VI concludes the paper with a discussion of future work.

II. Background

A. Information Accountability

Information contained in eHealth systems is often very sensitive in nature, and as such, it is vital that access to that information is appropriately managed. When implementing an EHR system, the security of the stored data, access control and access monitoring must all be considered [7]. Traditional preventive access control measures that rigidly deny access to users without appropriate permissions are insufficient on their own in the domains like eHealth with its complex access requirements, and as a result a number of researchers have begun working on augmenting these preventive measures with accountability [8–10].

Information Accountability involves the use of policies and mechanisms to enforce appropriate use through after-the-fact accountability for intentional misuse. Misuse is defined as the unauthorised access, use, modification, or disclosure of information, or other use of information that is not for the purpose for which the information was provided [11, 12]. The

978-1-4673-8325-7/15/$31.00 ©2015 IEEE

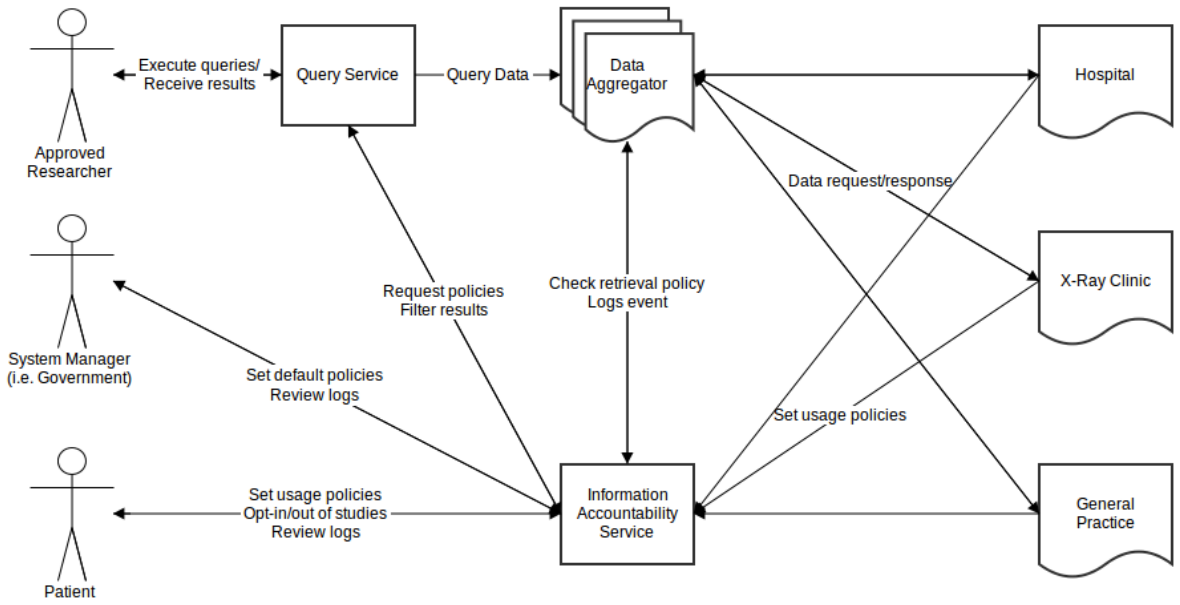

Fig. 1. Information accountability model for health big data analytics

presence of IA mechanisms is intended to act as a deterrent for such misuse [8].

With information dissemination being one of the primary causes of concern among consumers, it is important that it is clear to patients how their information is used, and who it will be disclosed to both now and in the future [13]. With regards to Cloud-based hosting of EHR information, Rodrigues et al. [7] state that appropriate security mechanisms must be put in place while making it transparent to patients how their data is managed. Such transparency is one of the fundamental aspects of Information Accountability [9].

There have been a number of proposed approaches to implementing IA mechanisms, such as Jagadeesan et. al. [14] who attempted to develop a formal foundation for the design of IA systems using privacy policies to define appropriate use of information, focusing on using audit logs that can detect potential policy violations and information misuse. Their approach focused on using audit logs which can detect potential policy violations and information misuse. Weitzner et. al. [9] proposed a transparent audit process that would track all transaction, and make use of policies combined with policy-aware transaction logs and a policy reasoning capability to enable systems to hold users of information accountable. These studies generally focused on IA and accountable systems from a general point of view without consideration for the specific requirements of eHealth systems.

B. The need for accountability in Health Big Data Analytics

In this paper we define "Big Data" in health as a collection of medical data that is so large, so complex, so distributed, and growing so fast that it becomes difficult or impossible to maintain and analyse using traditional software and hardware [15]. In 2012, it was estimated that worldwide digital healthcare data was equal to 500 petabytes and is expected to reach 25,000 petabytes in 2020 [16]. It is believed that analysing Health Big Data can lead to improvements in the quality of healthcare and support better clinical decision making [1, 2].

With the large growth in this health information from the variety of medical systems and sources, there comes significant issues such as interoperability and the creation of data silos [17]. To protect patient privacy, big health data is often scattered and intentionally isolated among institutions [18]. To increase the potential of Health Big Data analytics, we must find ways to address patient privacy concerns while also encouraging the various producers of this health information to share the data. Weber et. al. [18] states that there is a need for a consent mechanism that can enable patients to "decide how and when their data can be shared with or "mashed up" against other databases." This is the gap we believe the use of Information Accountability can fill.

III. Information Accountability Framework

In the Information Accountability Framework devised for use in eHealth systems, four types of users were initially modelled into the framework: data owners (i.e patients), data users (i.e healthcare professionals) using health information for legitimate purposes, data users who misuse health information, and a central health authority (HA) (i.e. a government agency). Data owners have explicit control over which of their preferred HCPs can access their information and are able to set usage policies to grant or limit access further. The HA is in place to ensure that HCPs always have access to the information they need to provide appropriate care through default policies, without unnecessarily hindering the patient's privacy [19].

The IA mechanisms implemented in the IAF enable users who misuse data information to be to be held accountable and

deters those with ill intent through a fear of being caught, with clear messaging of the consequences of actions conveyed to users. Incentives are given to the users to follow the procedures and enforce appropriate use.

A key component of accountable systems are policy-aware transaction logs [9] which provide provenance of the data in the system. Using such logs, the provenance of the data can be compared to usage policies to determine if an action complied with those policies [20]. In the IAF, all information access and other events in the system are logged along with policy used to determine whether the action should be permitted. As a result, the information they contain can also be considered sensitive and must be protected [21]. These logs are made available to the data owner in a user-friendly format which they can review at any time. The IAF actively monitors all actions taken in the system for potential breaches of policy and provides notifications as needed. For example, when a HCP makes an invalid access request, the system notifies the patient of the potential misuse of their eHealth information with a log that can be reviewed and referred to when submitting an inquiry asking the HCP to justify their actions [22].

When the system detects possible misuse of a patient's health data, the patient is able to submit an inquiry asking for a justification of the actions taken by the relevant HCP. The HCP must then provide an explanation to justify their need to access the relevant information. Once this is done, the system uses a semantic reasoner and rules defined by a HA along with the context of the information access, usage policies, and the HCP's justification to determine whether misuse occurred and further investigation is required.

This model has been validated and surveys conducted into user acceptance, but it hasn't been fully implemented and it must be expanded to provide for more diverse users and use cases, including its applicability for use in managing privacy in a Big Data Analytics use case.

IV. Applying the IAF to a HBDA process

In order to reap the benefits of shared eHealth information systems for Big Data Analytics and encourage the sharing of information through providing transparency and accountability to information usage, we have devised a model the sharing of eHealth information that makes use of the principals of the initial Information Accountability Framework. The initial model focused on patient control, but for the purposes of information sharing and enabling big data analytics and research on eHealth information, we must also consider the view point of the producers and providers of eHealth information as stakeholders in the collection and use of this data.

In our IAF model, patients would be able to explicitly consent as to whether or not their data could be aggregated with other patient data in order to conduct analytics. Through accountability mechanisms, they would always be informed how and why their information as being queried and used, and the results shared for the purposes of improved healthcare. In the devised information accountability model for sharing eHealth data, healthcare professionals and other producers of eHealth information are also able to specify policies over how the information they produce is aggregated and used. These policies are then combined with patient policies and policies set by a governing Health Authority to determine which information is aggregated for a patient from that data source. The process for this model is demonstrated in Figure 1.

We define four different types of users to demonstrate this model:

- **Data Owners:** Data owners refers to the individuals to whom the data refers to, i.e. patients.
- **Data Providers:** Data providers refers to the groups and individuals who produce and/or store the information that will be aggregated. Data providers could be various types of healthcare providers such as hospitals, general practitioners, an X-ray clinic, etc.
- **System Manager:** A system manager refers to the organisation responsible for maintaining the shared eHealth information system, and setting appropriate policies and investigate potential misuse. This could be a government department.
- **Data Users:** Data users refers to those who would make use of the aggregated data, such as healthcare professionals, approved researchers, and government studies.

A. Setting policies

1) Data providers: Data providers (i.e. hospitals, specialists, etc.) are able to opt-in to sharing their data and set usage and aggregation policies on the information they produce. For example, a general practice may be willing to share condition and medication summaries about patients, but not detailed notes made by the patients' doctor. A policy depicting this example is represented in Open Digital Rights Language (ODRL)—an open standard rights language capable of expressing a wide range of policy-based information—in Listing 1.

2) Data owners: Data owners (i.e. patients) are able to opt-in to participate in studies through usage policies on their information. Patients maintain control over who has access to their information and in which contexts. When the data is aggregated for the purpose of a study, a filtering stage applying patient usage policies to the information is conducted.

3) System manager: System managers who oversee the shared eHealth data system, such as a government's health department, set default policies and restrictions on data collection and use.

B. Data aggregation

In the model, a data aggregator collects information from the data providers. While doing so, it queries the IA service to retrieve an aggregation policy set made up of data owner and data provider preferences in order to ensure it only aggregates permissible data and avoids patients who have not opted-in to their data being collected.

C. Querying and accessing data

When a data user executes a query in the system, the query service retrieves a policy for the data user. This can include rules regarding which data they can access, how they can use data, and required de-identification of the results.

If they are permitted to perform the query, the retrieved rules are then applied to filter the result set, removing restricted information. The information access request is logged, and the policy versions used to determine the access request is stored with the context-aware log entry.

D. Access to logs

The logs produced in an accountable system can contain sensitive information themselves and must be appropriately protected, including restricting who can view these logs and for what purpose [21].

1) Data providers: Data providers can view log summaries of when and what information was aggregated. The logs maintained by the accountability can also be used for risk management, as if information originating from a data provider is found to have been misused or leaked, they can verify who accessed their information aggregated in the system.

2) Data owners: Data owners can view the log entries for their information. They can review these logs at any time, and submit inquiries for events identified as potential misuse.

3) Data users: Data users will be able to access specific log entries regarding their own access to patient information. They will be able to review the entries when they receive an inquiry requesting that they justify why they needed to access the relevant information in the given situation.

4) System manager: The system manager will be able to view all logs and provenance information for the aggregated data for the purposes for investigating potential misuse detected by the system. They will also need to be able to verify the integrity of the log entries and usage policies.

V. Implementation challenges

Many challenges remain to be investigated in order implement the proposed IAF model when accounting for Big Data Analytics use cases. We believe this provides fertile for future research into this proposed Big Data accountability model to verify its practicability.

A. Scalability and performance

At the scale of Big Data, When performing data analysis at the scale of big data, the complexity of the queries can result in superexponential growth in computing time as the data set increases [15]. With that in mind, it is still important that additional access and privacy controls applied when querying data can scale. In producing a prototype of this model, the efficiency of applying these controls to filter and present results must be considered, as well as techniques for minimising their effects.

```
<o:policy xmlns:o="http://odrlextension.
    org/ns/odrlx/2x" xmlns:eh="urn:ehealth
    .gov" type="http://odrlextension.org/
    ns/odrlx/2x/privacy" uid="policy-use-
    ehr">
  <o:permission>
    <o:asset uid="urn:ehealthSystemData
        :11986" relation="o:target"/>
    <o:asset uid="urn:ehealthSystemData
        :11986" relation="x:collection"/>
    <o:party uid="urn:healthProfessional
        :10946" role="o:assigner"/>
    <o:party uid="urn:ehealthSystem:1458"
        role="o:assignee"/>
    <o:action name="o:aggregate"/>
    <o:constraint name="o:dataType"
       operator="o:isAnyOf"
       rightOperand="eh:prescription
         eh:conditionSummary"/>
  </o:permission>
</o:policy>
```

Listing 1. Example aggregation policy for a general practice represented in ODRL

B. Log storage and presentation

For accountability to work, appropriate provenance information must be stored and verifiable [21]. In a Big Data query, the results must generate policy-aware provenance information that can be used to verify how, why, and when a piece of information was accessed. This creates a challenge of how to efficiently store such data while maintaining privacy and security of the information they contain, as the logs themselves can contain sensitive information [21]. Likewise, a principal of accountability is the transparency of the information use to data owners, so the presentation of this information to patients so they know who accessed their information and under what conditions, provides additional scalability and usability challenges.

When HCPs access health information on individual patients, this can easily be handled; however, when a researcher accesses parts of the data of millions of patients from many data providers, this creates a challenge of how best to store and present the provenance information.

C. Data heterogeneity

Due to the diverse systems that produce health data which come in various formats, the heterogeneity of the data is a major challenge for Big Data Analytics [15]. For the accountability mechanisms to work, the framework must be able to match up data types to in the information to those used to define policies, presenting a challenge of how to normalise the aggregated data.

VI. Conclusion and future work

Health Big Data Analytics has the potential to solve many current and future healthcare problems, from discovering new treatments to improved decision making. However, concerns over patient privacy have hindered the aggregating data from the various sources of eHealth information for such use cases. In this paper, we have proposed an information accountability approach to addressing the privacy concerns of combining data for health data through a patient consent based approach. In the IAF model applied to the Big Data use case, patients are able to opt-in to health trials and decide how and when their data can be shared with or combined with other databases as part of a study, while maintaining accountability and transparency. Additionally, the model aims to manage risk and encourage the sharing of data by healthcare providers by ensuring they have control over how the information they produce is aggregated and used.

Future work will involve prototype implementations of an Information Accountability service for use in this model, verification of the models using health data, further investigations into the challenges of implementing this approach at scale and user testing to verify the usefulness and acceptability of the model.

References

[1] N. H. Shah and J. D. Tenenbaum, "The coming age of data-driven medicine: translational bioinformatics' next frontier," *Journal of the American Medical Informatics Association*, vol. 19, no. e1, pp. e2–e4, 2012.

[2] L. P. Garrison Jr, "Universal health coverage–big thinking versus big data." *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, vol. 16, no. 1 Suppl, p. S1, 2013.

[3] Department of Health, "Personally Controlled Electronic Health Record Review Report," 2014. [Online]. Available: http://www.health.gov.au/internet/main/publishing.nsf/Content/eHealth

[4] S.-T. Liaw and T. Hannan, "Can we trust the pcehr not to leak?" *Aust Fam Physician*, vol. 39, pp. 809–810, 2010.

[5] K. Garrety and I. van Teeseling, "E-Health: are we ready for this brave new world?" ABC – The Drum, 2012. [Online]. Available: http://www.abc.net.au/unleashed/4081982.html

[6] W. M. Tierney, S. A. Alpert, A. Byrket, K. Caine, J. C. Leventhal, E. M. Meslin, and P. H. Schwartz, "Provider responses to patients controlling access to their electronic health records: A prospective cohort study in primary care," *Journal of General Internal Medicine*, vol. 30, no. 1, pp. 31–37, 2015.

[7] J. J. Rodrigues, I. de la Torre, G. Fernández, and M. López-Coronado, "Analysis of the security and privacy requirements of cloud-based electronic health records systems," *Journal of medical Internet research*, vol. 15, no. 8, 2013.

[8] J. Feigenbaum, A. D. Jaggard, and R. N. Wright, "Towards a formal model of accountability," in *Proceedings of the 2011 workshop on New security paradigms workshop*. ACM, 2011, pp. 45–56.

[9] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman, "Information accountability," *Communications of the ACM*, vol. 51, no. 6, pp. 82–87, 2008.

[10] R. H. Sloan and R. Warner, "Developing foundations for accountability systems: Informational norms and context-sensitive judgments," in *Proceedings of the 2010 Workshop on Governance of Technology, Information and Policies*, ser. GTIP '10, ACM. New York, NY, USA: ACM, 2010, pp. 21–26.

[11] Privacy Act 1988, Clth. [Online]. Available: http://www.comlaw.gov.au/Details/C2013C00231

[12] Health Identifiers Act 2010, Clth. [Online]. Available: http://www.comlaw.gov.au/Details/C2010C00440

[13] F. A. Rahim, Z. Ismail, and G. N. Samy, "Information privacy concerns in electronic healthcare records: A systematic literature review," in *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*. IEEE, 2013, pp. 504–509.

[14] R. Jagadeesan, A. Jeffrey, C. Pitcher, and J. Riely, "Towards a theory of accountability and audit," in *Computer Security ESORICS 2009*, ser. Lecture Notes in Computer Science, M. Backes and P. Ning, Eds. Springer Berlin Heidelberg, 2009, vol. 5789, pp. 152–167. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04444-1_10

[15] M.-H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, "Health big data analytics: Current perspectives, challenges and potential solutions," *International Journal of Big Data Intelligence (IJBDI)*, vol. 1, no. 1/2, pp. 114–126, 2014. [Online]. Available: http://www.inderscience.com/info/inarticle.php?artid=63835

[16] J. Sun and C. K. Reddy, "Big data analytics for healthcare," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13, ACM. New York, NY, USA: ACM, 2013, pp. 1525–1525. [Online]. Available: http://doi.acm.org/10.1145/2487575.2506178

[17] R. L. Richesson and C. G. Chute, "Health information technology data standards get down to business: maturation within domains and the emergence of interoperability," *Journal of the American Medical Informatics Association*, p. ocv039, 2015. [Online]. Available: http://jamia.oxfordjournals.org/content/early/2015/05/16/jamia.ocv039

[18] G. M. Weber, K. D. Mandl, and I. S. Kohane, "Finding the missing link for big biomedical data," *JAMA*, vol. 311, no. 24, pp. 2479–2480, 2014.

[19] D. Grunwell, R. Gajanayake, and T. Sahama, "Improving usefulness of ehealth systems through information accountability," *e-Health Technical Committee Newsletter*, vol. 2, no. 6, pp. 3–5, December 2013.

[20] R. Aldeco-Pérez and L. Moreau, "Provenance-based auditing of private data use," in *International Academic Research Conference, Visions of Computer Science*. BCS, September 2008. [Online]. Available: http://eprints.soton.ac.uk/266580/

[21] D. Grunwell, R. Gajanayake, and T. Sahama, "The security and privacy of usage policies and provenance logs in an information accountability framework," in *Eighth Australasian Workshop on Health Informatics and Knowledge Management*, A. Maeder and J. Warren, Eds. Sydney, Australia: Australian Computer Society, 2015, pp. 33–40.

[22] D. Grunwell, R. Gajanayake, and T. Sahama, "Demonstrating accountable-ehealth systems," in *Proceedings of IEEE International Conference on Communications 2014*, IEEE. IEEE, June 2014, pp. 4258–4263.