Operationalizing Privacy Compliance for Cloud-hosted Sharing of Healthcare Data*

A Case Study

Benjamin Eze
School of Electrical Engineering and Computer Science, University of Ottawa, ON, K1N 6N5, Canada
beze080@uottawa.ca

Craig Kuziemsky
Telfer School of Management, University of Ottawa, Ottawa, ON, K1N 6N5, Canada
kuziemsky@telfer.uottawa.ca

Liam Peyton
School of Electrical Engineering and Computer Science, University of Ottawa, ON, K1N 6N5, Canada
lpeyton@uottawa.ca

ABSTRACT

Complex patient health needs and care delivery models such as patient participatory medicine require the ability to share data across multiple touch points. Achieving systematic performance management of care processes require an infrastructure that addresses interoperability and data standardization while supporting data governance and privacy compliance. In this paper, we present a framework for operationalizing privacy compliance for correlated cloud-hosted data using Data Sharing Agreements (DSAs) in support of performance management of community healthcare. Our focus is to show how DSAs can be used to operationalize privacy compliance for a cloud-hosted surveillance and performance management infrastructure by leveraging selective anonymization based on both organizational and patient consents. This allows a cloud-computing infrastructure to configure processes and services, including anonymization to ensure privacy compliance and a systematic approach to data governance.

CCS CONCEPTS

• **Information systems** → **Decision support systems** → Data analytics • **Security and privacy** → **Security services** → Pseudonymity; anonymity and untraceability • **Architectures** → **Distributed architectures** → Cloud computing

KEYWORDS

Data Sharing Agreement, Privacy Compliance, Anonymization, Performance Management, Cloud Computing, Healthcare Data Sharing.

ACM Reference format:

B. Eze, C. Kuziemsky, L. Peyton. 2018. SIG Proceedings Paper in Word Format. In *Proceedings of IEEE/ACM International Workshop on Software Engineering in Healthcare Systems (SEHS), Gothenburg, SWEDEN, May 2018 (GOTHENBURG'18)*, 8 pages.

1 INTRODUCTION

Complex patient health needs and care delivery models such as patient participatory medicine require the ability to share data across multiple touch points [1]. Systematic performance management for complex and comorbid patient care is challenging to the healthcare system. Limited data sharing, usually a consequence of heterogeneous healthcare data silos and privacy regulations, makes correlating healthcare data on complex patients from various healthcare providers difficult and sometimes impossible [2].

A Data Sharing Agreement (DSA) is a fundamental component of cloud-based solutions for supporting connected healthcare delivery [3]. DSAs are very important for describing policies needed for maintaining privacy and confidentiality, especially with a third party data custodian [4–6]. DSAs are an important consideration in a circle of care [7,8] and a necessary requirement for a cloud-based implementation to be HIPAA compliant [9,10].

A DSA is a legal agreement among collaborating data providers, regulating the conditions for data sharing [4,11]. A DSA can also be interpreted as a specification of the set of policies that determine what datasets collaborating organizations are allowed or denied access to with respect to data ownership and use [11]. A DSA specifies the purpose of use, participating organizations, prohibitions on secondary use, data elements to be extracted, formats, meta-data, data classification and organization, quality assurance, storage, security, data recipient responsibilities, intellectual property rights and legal requirements [12,13]. A DSA also provides an agreed-upon mechanism for ensuring privacy compliance with electronic data exchanges [5,14].

Achieving systematic performance management of care processes require an infrastructure that addresses interoperability and data standardization while supporting data governance and privacy compliance [2]. Cloud computing is one potential

SEHS'18, May 27, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5734-0/18/05…$15.00
https://doi.org/10.1145/3194696.3194701

infrastructure for addressing performance management challenges and supporting interoperable healthcare solutions [15–17], across multiple providers if data model standardization and appropriate support for privacy compliance are put in place to protect patient data [18,19].

In this paper, we present a framework for operationalizing privacy compliance for correlated cloud-hosted data using DSAs in support of performance management of community healthcare. Our focus is to show how DSAs can be used to operationalize privacy compliance for a cloud-hosted surveillance and performance management infrastructure by leveraging selective anonymization based on both organizational and patient consents. This allows a cloud-computing infrastructure to configure processes and services, including anonymization to ensure privacy compliance and a systematic approach to data governance.

2 BACKGROUND

2.1 Privacy Compliance in Healthcare

Privacy laws like HIPAA[20] and PIPEDA[21] aim at regulating the use, transmission, and storage of personal health records. Data Sharing Agreements which ensure compliance with such laws are usually signed legal documents, making them difficult to operationalize [12]. Work exists that shows how specification of data sharing agreements in a formal policy language can operationalize compliance in information systems [6]. However, the common approach to privacy compliance in health care is an all or nothing type of approach in which patient data that is excluded from access, is also excluded from aggregation. As a result, exclusion of patient data compromises the accuracy of performance management reports because the aggregate data is not precise [26] even though there are techniques to allow aggregation while ensuring privacy [29].

According to Weber et al. [22], health systems should support data across the patient's life. This would require wide-scale surveillance and data sharing across various healthcare applications. Data surveillance in healthcare requires complex systems that can support various data models, ill-defined workflows and information structures [22]. However, what is prevalent today is that most healthcare systems either avoid wide-scale surveillance to ensure privacy protection and confidentiality or require extensive approval processes to share data. Even when such approvals exist, all patient identifiers may be stripped off, therefore limiting the analytics utility of such datasets [23].

2.2 Privacy Risks with Data Sharing

Protected Health Information (PHI) in a dataset are categorized as 1) Direct Identifiers (DIs) – attributes like names and government identifiers like Canadian Health Card Number (HCN) or Social Insurance Number (SIN) that can identify an individual on their own in the data set. 2) Quasi-identifiers (QIs) – attributes like date of birth, postal code, gender that only identify an individual when used with other quasi-identifiers [24], 3) Sensitive attributes (SAs) – attributes like procedure and drug codes, disease conditions that are not usually public data but sensitive if associated with an individual [18]. Knowledge of QI values could also reveal these sensitive attributes [25], therefore helping an adversary gain more background knowledge on their victims.

Regarding data breaches, there are three types of disclosures: identity, attribute, and membership. Identity disclosure occurs when the record of an individual or an entity in the dataset is re-identifiable [18]. Attribute disclosure occurs when new information can be gained on the sensitive attributes by an attacker or adversary [26]. It is important to note that attribute disclosure is often a consequence of identity disclosure. Membership disclosure is a probabilistic measure of the presence or absence of an individual in a dataset [26]. This knowledge changes the behavior of an adversary towards de-anonymized data. Both identity and attribute disclosures are important in our approach to protecting patient privacy.

2.3 Cloud Computing, Healthcare Interoperability, and Privacy Compliance

Cloud computing is becoming an interesting platform for national and regional healthcare systems interoperability. Torre-Diez et al. [9] explain that cloud computing is well suited for the statewide health infrastructure and platform because of the many advantages: scalability, reach, extensibility, low total cost of ownership and availability [9].

However, privacy and confidentiality are a major concern with cloud computing. While encryption and access control policies help secure data exchanges and storage, they do not protect data from authenticated users that have access to the decrypted data. Encryption also sacrifices the analytical utility of cloud data [26]. Privacy measures, on the other hand, ensure confidentiality of patients and other members of the circle of care of a patient while allowing the data to be available for analytical processing [7,8].
There is the need to address security and privacy-related issues with cloud implementations [9], and this requires cloud providers to be HIPAA compliant [10]. Integration solutions must be HIPAA certified through proper authentication and authorization access control practices, data integrity, accountability and audit trail measures [27].

3 OPERATIONALIZING PRIVACY COMPLIANCE WITH DATA SHARING AGREEMENTS

3.1 Case Study

We present our approach to operationalizing data sharing agreements for performance management of community care in the context of a pilot project for performance management for community care at a Regional Health Authority (RHA) in Canada. The focus of the project was performance management of all community care services provided to patients in the region at their homes, a long-term care facility, or a hospital. Many of the patients have chronic and complex health conditions requiring them to receive care from multiple healthcare providers simultaneously, while some are in palliative care. The RHA works with some 54

Community Support Services (CSS) agencies to provide various needed community services to patients in the region. Services provided by these agencies greatly improve the quality of life of patients especially those with complex and chronic health conditions. Improving the quality of life of these patients also results in major cost savings to the health system by reducing incidents of visits to the emergency room, hospitalization, and the ability to provide care to patients in the comfort of their own homes. The RHA pays the CSS agencies for services within its mandates.

3.2 Privacy Compliance Framework

For this case study, a cloud-based infrastructure for performance management of community care services was implemented to support data hosting, data collection, reporting, analytics, and subscription services.

Fig. 1 shows the cloud-based infrastructure, which hosts all the CSS Agencies local databases. Each stakeholder's operational systems and databases were migrated to a cloud-hosted infrastructure with a data hosting service provided and maintained by the RHA. This was done within a systematic framework of data sharing agreements between each service and the regional health authority in compliance with all relevant privacy regulations. The system is designed to allow each stakeholder their autonomy regarding data management while also providing the infrastructure for privacy compliant data sharing for performance management.

The private cloud infrastructure correlates the data from these internal CSS agency databases into a Shared Services Database that is then used by a Shared Services Reporting Portal to support performance management. A Systematic Data Collection Service pulls data from the CSS agencies local databases into a Staging Database. The very first privacy compliance processing step is to apply organizational consent to incoming data streams within the staging databases. Afterward, the datasets are run through the patient identity matching service that correlates patient profiles across the various data sources using a Match Definitions Document to populate the Aggregate Patient Profile Database. Each patient profile is then assigned a globally unique identifier linking all patient profiles across all CSS agencies that belong to a single patient. This ensures that all the services provided to a patient by each of the stakeholders are grouped under this common patient identifier.

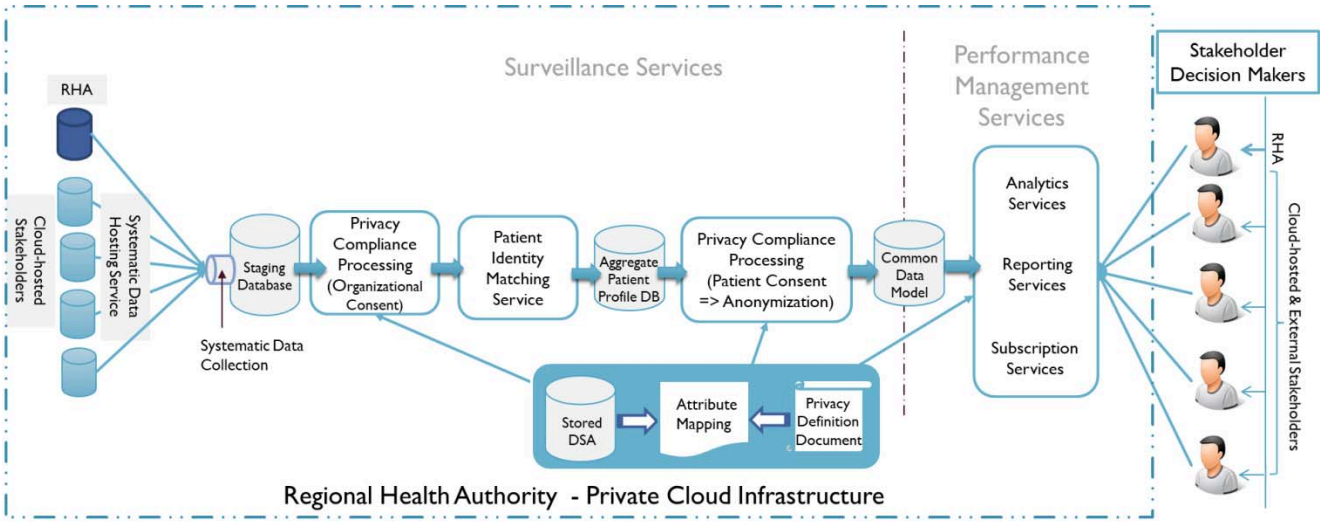

Figure 1: Cloud-based Infrastructure for operationalizing privacy compliance using DSAs.

At this point, the last Privacy Compliance Processing step is then applied on the aggregate patient data based on the DSA for each CSS agency, and the Privacy Compliance Definition Document that describes the anonymization setting for the infrastructure using both organizational and patient consent definitions in the DSAs.

Finally, all data is converted to a logically and semantically equivalent common model. This is needed to ensure a consistent view of information across all data sources [28] and provides the minimum representative set of attributes for performance management of community care services delivery.

4 PRIVACY COMPLIANCE MODEL

4.1 Privacy Compliance Flow

The two major sections of the DSA used for each CSS agency are the Patient and Organizational consent definitions. Organizational consents have local significance for each incoming data stream. Patient consent can have both global (across all data streams), local (specific to each organization), or partial (apply to select data entities and attributes) significance. Enforcing these consents could

result in complete removal of patient data from the common data model. In some cases, there could be full or partial anonymization (data masking, generalization, suppression) of patient data to meet set risk thresholds for the infrastructure.

As shown in Fig. 2, the DSAs organizational consent section allows each organization to include/exclude specific fields and attributes from the data that goes into the performance management infrastructure like those with sensitive financial records and audit data. Each organization is also associated with a defined risk profile for the data they receive from the cloud infrastructure.

Patient consent can be explicitly listed or pulled from each agency local database. In accordance with the design of the CSS Patient Information System, if patient consent definition is not explicitly defined for a patient, an automatic ***Consent Granted*** flag is applied to the patient profile. It is important to understand that this design does not apply to all types of data releases. Rather, it applies only to the collaborating CSS agencies that are part of a patient circle-of-care. Those outside the circle-of-care are only allowed access to fully anonymized patient data irrespective of the patient consent – implied or explicit.

Patient consent applied at this point has a local significance to each CSS agency data. This is because some patients may choose to deny access to performance management processes on some of the community care services they receive and grant consent in others. For example, deny consent to sharing their mental health data but allowing their nursing and other therapy data to be used for performance management. The framework can be configured to exclude non-consenting patients at this point, or allow their data through but have them anonymized before it hits the Common Data Model (CDM) driven Shared Services Database.

The Privacy Compliance Definition apply to the transformed data. It has two major sections – Anonymization settings and CDM Entities definitions. The anonymization settings provide details on patient profiles to target, risk settings, approaches to anonymizing the direct identifiers, quasi-identifiers, and sensitive attributes. The CDM Entities definitions classify each attribute that requires anonymization based on their privacy data type and other details like levels of generalization for quasi-identifiers. This process is described in more details in section 4.4. In line with this model, our cloud framework ensures that various levels of anonymization can be applied to data as needed to protect patient privacy and confidentiality as well as to ensure that participants only have access to reports from the common data model on a need to know principles.

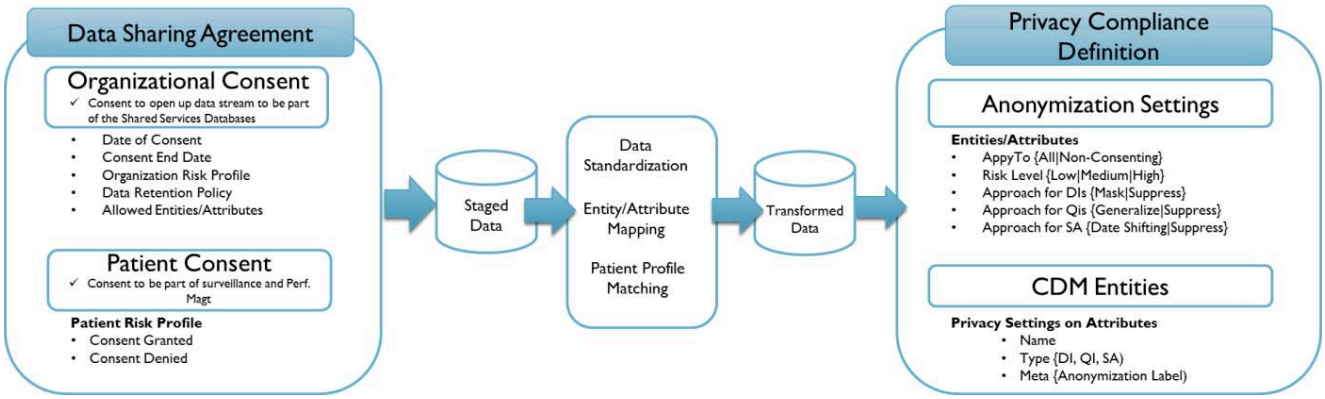

Figure 2: Privacy Compliance Data Flow

4.2 DSA - Organizational Consent

Organizational consent represents each agency's consent and permissions for accessing and using data within its custody. In the context of this case study, a DSA applies to 1) Incoming data streams, 2) Data belonging to each organization within the common data model, and 3) analytics and reporting services outputs. Organizational consent can be in the form of an explicit opt-in, opt-out or anonymize settings for each CSS agency that apply to 1) an entire data feed, 2) specific entities in the data feed, and 3) specific attributes in the data feed.

While none of the analytics and reporting results would be made available to the public without thorough privacy considerations, our framework's DSA allows each participating organization to describe what is allowed or not allowed expressively in conformance with law and their level of comfort regarding utility vs. privacy and confidentiality.

An important consequence of this level of consent is that it is reciprocal. A CSS agency is unable to access data from other participating agencies that it is uncomfortable sharing within its own data stream. For example, failure to share patient demographic details means that reports and analytics or data subscriptions that return data on patient demographic information from other participating organization is explicitly denied for the organization. The justifications for this reciprocal consent is to encourage CSS agency participation in data sharing as the quality and richness of the analytics reports increases with data volume. However, organizations outside the patient circle-of-care are only given access to aggregate analytics from the shared services database with all patient identifying details fully anonymized.

4.3 DSA - Patient Consent

Privacy of patients is an important consideration for large-scale data surveillance and performance management, especially in the healthcare industry. In a cloud computing framework, concerns over the nature and pattern of data sharing and confidentiality of patient sensitive health data must be addressed [29,30]. For this case study, Patient Consent is read from each of the patient profiles associated with each CSS agency internal database.

Patient consent simply indicates whether consent is granted or denied for surveillance and performance management. But, it can mean different things based on the settings within the Privacy Compliance Definition Document (described below) as to whether the patient data is excluded or anonymized. It could result in the patient identifying records stripped from incoming data streams or trigger anonymization of the patient complete or partial profile data.

Anonymization is set up for the entire infrastructure. By centralizing the anonymization settings, we ensure that patient consent is applied consistently in the context of their profiles across all services that they participate in across the region, therefore ensuring the global protection of patient data across the entire infrastructure.

4.4 Privacy Compliance Definition

The Privacy Compliance Definition Document is the anonymization configuration for the common data model. This definition applies to the data in the aggregate patient profile database, the outcome of the patient identity matching service. This XML definition includes privacy definitions for all data sets contained in the Shared Service Database as well as anonymization settings for the database and data attributes. Anonymization is a continuous operation, applied to each incoming data stream. Each applicable data set must be described in the privacy definition document so the attributes are properly anonymized while preserving the analytical utility of the data set.

Fig. 3 shows a simplified sample of this document – showing only the patient and patient address entities. The ***<RunSettings>…</RunSettings>*** show the job management settings. It sets the job admin email, the SMTP server, and notification settings. The ***<AnonymizationSettings>…</AnonymizationSettings>*** definition controls the behavior of the anonymization service within the cloud infrastructure. This then depends on the ***<Entities>…</Entities>*** definitions to determine the tables/attributes defined in the CDM that should be considered for various anonymization processes.

Anonymization processing depends on a complex set of definitions that must be based on the expected risk associated with the data recipient. In this case study, there is no public release of data, so the release of the entire Shared Service Database is very unlikely. Nevertheless, anonymization is necessary to adhere to patient consent while maintaining high analytical utility for the analytics generated from the CDM.

The anonymization settings are described in Table 1 below. The default behavior is to apply anonymization to all incoming data from patients that refused to consent to data sharing. However, it could also be applied to the entire data set when needed, say to release data to an external partner that is not one of the participating stakeholders. One important contribution of this framework is that anonymization is not always applied to the entire data set but to select patient profiles as stipulated by patient consent definitions. But the risk measurement that determines the level of anonymization to apply to these select patient profiles is done using all records in the aggregate database.

```
<PrivacyCompliance>
  <RunSettings
      job_admin="etljobadmins@rha.ca"
      smtp_server="198.22.0.3"
      send_notification="onerror"
      set_defaults="false"
      >
  </RunSettings>
  <AnonymizationSettings apply_to="non-consenting-patients">
    <Risk level="average" />
    <DirectIdentifiers approach="Mask" />
    <QuasiIdentifiers approach="Generalize,Suppress" />
    <SensitiveAttributes approach="none" />
  </AnonymizationSettings>
  <Entities>
    <Entity name="Patients" >
       <PrivacySettings>
        <Attribute name="HCN" type="direct-identifier"
            meta="canadianhealthcardnumber" />
        <Attribute name="Surname" type="direct-identifier"
            meta="lastname" />
        <Attribute name="Firstname" type="direct-identifier"
            meta="firstname" />
        <Attribute name="DOB" type="quasi-identifier"
            meta="dateofbirth" />
        <Attribute name="DeathDate" type="quasi-identifier"
            meta="dateofdeath" />
        <Attribute name="Gender" type="quasi-identifier"
            meta="gender" />
      </PrivacySettings>
    </Entity>
    <Entity name="PatientAddress">
      <PrivacySettings>
        <Attribute name="Address" type="direct-identifier"
            meta="canadianaddress" />
        <Attribute name="Postal_Code" type="quasi-identifier"
            meta="postalcode" />
        <Attribute name="Start_Date" type="quasi-identifier"
            meta="eventdate" />
      </PrivacySettings>
    </Entity>
  </Entities>
</PrivacyCompliance>
```

Figure 3: Privacy Compliance Definition

Irrespective of the risk level, all direct identifiers are always masked on anonymization of a patient profile. For example, the Patients HCN field is masked as a Canadian Health Card Number. Similarly, quasi-identifier risk mitigation is carried out using a combination of generalization/suppression for a *k-anonymity* value determined through risk measurement of aggregate CSS database. For this case study, if the risk level associated with the infrastructure for the data recipient is considered average, a *k-anonymity* value of 5 is applied. If it is high, then a *k-anonymity* value of 10 is applied and when it is low, a *k-anonymity* value of 2 is applied. These risk levels are determined through best practices for data releases based on the intended data recipient. Quasi-identifiers generalization options are set based on the attribute type as identified by the meta attribute definition. Where applicable, the

anonymization engine could apply algorithms like OLA[29] to choose the right generalization for each quasi-identifier. Date Shifting is used to anonymize patient event dates like nursing visits, activity visits, assessment dates while preserving the inter-event time intervals [31].

Table 1: Anonymization Setting

Setting	**Operation**	**Details**
Apply_to	all	Applies to the entire dataset in the CDM
	non-consenting-patients	Applies to only the patients that refused to their consent to data sharing
Risk Level	high	Assume high risk of re-identification (*k=10*)
	average	Assume average risk of re-identification (*k=5*)
	low	Assume a low risk of re-identification (*k=2*)
Approaches	Mask	Applies to DIs only. Masking is done based on the meta attached to the attribute
	Generalize	Applies to QIs and SAs only. Generalization is based on the type of attribute.
	Suppress	Applies to QIs and SAs only. Suppression is applied to all Qis after generalization where applicable.
	Date Shift	Applies to QI event dates. This process ensures that date QI values associated are shifted to a period that ensures their anonymity.

5 RESULTS AND DISCUSSION

To illustrate our framework processes for managing patient consent using anonymization. We will use a small data set of seven patients with three of these patients not-granting consent to data sharing (Fig. 4). Fig. 4a shows the records for those patients with the Consent Granted column indicating if they granted or denied consent to data sharing. In Fig. 4b, an all-or-nothing approach to processing patient consent is applied. In this scenario, all the 3 records belonging to those non-consenting patients are removed. In Fig. 4c, the records belonging to the non-consenting patients are anonymized instead of being removed. Because of space constraints, these sample records included only direct and quasi-identifiers from a single table, with no sensitive attributes.

Anonymization applied to these records include data masking of the direct identifiers like Patient Surname and First name; generalization for the quasi-identifiers like birthdate and postal code attributes. Because of the size of the sample data set, we set the *k-anonymity* threshold value to 2. Based on this risk setting, the birthdate data is generalized to YOB and the Postal codes are generalized further as FSA with the last 3 digits of each postal codes suppressed. Suppression is then applied to the equivalence classes of the quasi-identifiers at a *k-anonymity* value of 2. The anonymization component uses the entire data set to build equivalence classes but applies anonymization only to the non-consenting patient records.

We see that the record belonging to *Terry McCarthy* ended up with a new name of *Patrick Pit*. The equivalence class formed using the Gender, YOB, and FSA in the database was too unique for a *k-Anonymity* value of 2. So the Gender and Postal code values were both suppressed. For *James Elliot* and *Fatima Jones*, the quasi-identifiers were generalized but kept in-place because the equivalence classes have adequate support in the data set.

a) Original Dataset

Surname	First name	Gender	Birth Date	Postal Code	Consent Granted
Smith	John	Male	1965-08-25	K2A 5N6	Yes
Blake	Trevor	Male	1980-10-25	K1B 6N4	Yes
McGregor	Hilary	Female	1965-10-08	K2A 4B6	No
Elliot	James	Male	1980-10-01	K1B 5N4	No
Jones	Fatima	Female	1984-03-10	K2E 2P6	Yes
Wright	Martha	Female	1945-08-23	K1N 5N6	Yes
McCarthy	Terry	Male	1945-06-20	K2A 4N5	No

b) All-or-nothing approach: Non-consenting patients removed.

Surname	First name	Gender	BirthDate	Postal Code	Consent Granted
Smith	John	Male	1965-08-25	K2A 5N6	Yes
Blake	Trevor	Male	1980-10-25	K1B 6N4	Yes
Jones	Fatima	Female	1984-03-10	K2E 2P6	Yes
Wright	Martha	Female	1945-08-23	K1N 5N6	Yes

c) Selective anonymization: Non-consenting patient profiles anonymized

Surname	First name	Gender	BirthDate	Postal Code	Consent Granted
Smith	John	Male	1965-08-25	K2A 5N6	Yes
Blake	Trevor	Male	1980-10-25	K1B 6N4	Yes
Doe⁺	Jane⁺	Female	1965*	K2A*	No
St-Pierre⁺	Peter⁺	Male	1980*	K1B*	No
Jones	Fatima	Female	1984-10-03	K2E 2P6	Yes
Wright	Martha	Female	1945-08-23	K1N 5N6	Yes
Pit⁺	Patrick⁺	***	1945*	***	No

Figure 4: Comparing the all-or-nothing and selective anonymization approaches to process patient consent in DSAs.

The impact of all-or-nothing compliance approach to anonymization is significant to the final aggregate data set. Eliminating all records associated with the non-consenting patients

means that the quality of the analytics and performance management data set is directly dependent on the proportion of patients that consented to data sharing.

However, with selective anonymization, attribute missingness is kept at a reasonable level – therefore keeping the level of missing values in the performance management data set to the minimum required to keep those sensitive records belonging to the non-consenting patients completely anonymized while preserving the analytical utility of the final data sets.

Consequently, this process is then applied to the entire patient demographic data set of 240k patients while simulating the increasing impact of no consents to attribute value missingness. The result is summarized in Fig. 5. Our findings show that if the number of non-consenting patients is low, both approaches yield about the same level of attribute value missingness.

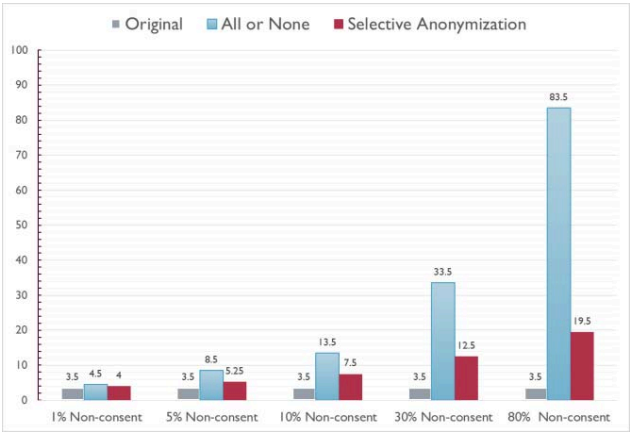

Figure 5: Comparing Data Missingness with the percentage of patients consenting to data sharing

However, if one assumes a maximum 5% suppression, we see that anonymization consistently reduced the overall missingness of the data set even in scenarios with over 80% non-consenting patients. It shows that while the "all-or-nothing" is a good safeguard, it has a substantial negative impact on the analytical utility of the resulting data set to analytics. Anonymization reduces this impact significantly.

6 CONCLUSIONS

Operationalizing privacy compliance, especially with healthcare data sharing, is very important for cloud-enabled data sharing and analytics. The focus of this paper is not on the methods and tools for anonymization, but rather to emphasize that anonymization should be an integral component of privacy compliance implementation, especially for a cloud-hosted surveillance and performance management infrastructure. We have also shown that a DSA can be an important tool for addressing organization and patient consent needs with cloud-hosted data sharing.

One of our conclusions is that healthcare data should not be released publicly except for certain privacy-secure statistics or if the entire data set is anonymized appropriately. Therefore, irrespective of the patient consent, the data custodian has the responsibility to anonymize all publicly released data sets to safeguard patient privacy and confidentiality and to conform to existing privacy compliance regulations.

We have also shown that while most healthcare systems opt for the all-or-nothing approach to handling patient data, eliminating records of non-consenting patients affects downstream analytics processes – making it impossible to see the complete picture of the healthcare system and measurable performance goals like the impact of disease outbreaks, improvements to quality of life for the patient and general well-being of the entire population.

Our work addresses the bigger challenging with operationalizing privacy compliance which is that of setting up the tools for automating continuous anonymization such that data releases are safe from privacy breaches and re-identification.

ACKNOWLEDGMENTS

This work was partially supported by funding from the Canadian Natural Sciences and Engineering Research Council (NSERC) and Ontario Graduate Scholarship (OGS).

REFERENCES

[1] J. Mcgregor, S.W. Mercer, and F.M. Harris. 2016. Health benefits of primary care social work for adults with complex health and social needs: A systematic review. Heal. Soc. Care Community (2016). DOI:https://doi.org/10.1111/hsc.12337

[2] Benjamin Eze, Craig Kuziemsky, and Liam Peyton. 2017. A Patient Identity Matching Service for Cloud-based Performance Management of Community Healthcare. In Procedia Computer Science, 287–294. DOI:https://doi.org/10.1016/j.procs.2017.08.321

[3] Jose Fran Ruiz, Marinella Petrocchi, Ilaria Matteucci, Gianpiero Costantino, Carmela Gambardella, Mirko Manea, and Anil Ozdeniz. 2016. A lifecycle for data sharing agreements: How it works out. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3–20. DOI:https://doi.org/10.1007/978-3-319-44760-5_1

[4] Ilaria Matteucci, Marinella Petrocchi, Marco Luca Sbodio, and Luca Wiegand. 2012. A design phase for data sharing agreements. In Data Privacy Management and Autonomous Spontaneus Security. Springer, 25–41.

[5] Robert Navarro. 2008. An ethical framework for sharing patient data without consent. Inform. Prim. Care 16, 4 (2008), 257–262.

[6] Vipin Swarup, Len Seligman, and Arnon Rosenthal. 2006. A Data Sharing Agreement Framework. Inf. Syst. Secur. (2006), 22–36. Retrieved from http://www.springerlink.com/index/x5100184p5x6u871.pdf

[7] P S Mathew and A S Pillai. 2015. Big Data solutions in Healthcare: Problems and perspectives. Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on, 1–6. DOI:https://doi.org/10.1109/ICIIECS.2015.7193211

[8] Donna Spruijt-Metz, Eric Hekler, Niilo Saranummi, Stephen Intille, Ilkka Korhonen, Wendy Nilsen, Daniel E Rivera, Bonnie Spring, Susan Michie, David A Asch, Alberto Sanna, Vicente Traver Salcedo, Rita Kukakfa, and Misha Pavel. 2015. Building new computational models to support health behavior change and maintenance: new opportunities in behavioral research. Transl. Behav. Med. 5, 3 (September 2015), 335–46. DOI:https://doi.org/10.1007/s13142-015-0324-1

[9] Isabel de la Torre-Díez, Sandra González, and Miguel López-Coronado. 2013. EHR Systems in the Spanish Public Health National System: The Lack of Interoperability between Primary and Specialty Care. J. Med. Syst. 37, 1 (January 2013), 9914. DOI:https://doi.org/10.1007/s10916-012-9914-3

[10] David S. Mendelson, Bradley J. Erickson, and Garry Choy. 2014. Image Sharing: Evolving Solutions in the Age of Interoperability. J. Am. Coll. Radiol. 11, 12 (December 2014), 1260–1269. DOI:https://doi.org/10.1016/j.jacr.2014.09.013

[11] Benjamin Aziz, Alvaro Arenas, and Michael Wilson. 2011. SecPAL4DSA: A policy language for specifying data sharing agreements. In Communications in Computer and Information Science, 29–36. DOI:https://doi.org/10.1007/978-3-642-22339-6_4

[12] Claudio Caimi, Carmela Gambardella, Mirko Manea, Marinella Petrocchi, and Debora Stella. 2016. Legal and technical perspectives in data sharing agreements

definition. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 178–192. DOI:https://doi.org/10.1007/978-3-319-31456-3_10

[13] University of Waterloo. 2017. Elements of a data sharing agreement: An example. Retrieved Feb 1, 2018 from https://uwaterloo.ca/research/office-research-ethics/research-human-participants/pre-submission-and-training/human-research-guidelines-and-policies-alphabetical-list/data-sharing-or-transfer-agreements-what-are-they-and-when/elements-data-sharing-agreemen

[14] Ilaria Matteucci, Marinella Petrocchi, and Marco Luca Sbodio. 2010. CNL4DSA: a controlled natural language for data sharing agreements. In Proceedings of the 2010 ACM Symposium on Applied Computing, 616–620.

[15] François Andry, Richard Ridolfo, and John Huffman. 2015. Migrating healthcare applications to the cloud through containerization and service brokering. In HEALTHINF 2015 - 8th International Conference on Health Informatics, Proceedings; Part of 8th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2015, 164–171. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84938849051&partnerID=tZOtx3y1

[16] Sabishaw Bhaskaran, Girish Suryanarayana, Amarnath Basu, and Roshan Joseph. 2013. Cloud-Enabled Search for Disparate Healthcare Data: A Case Study. In 2013 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 1–8. DOI:https://doi.org/10.1109/CCEM.2013.6684431

[17] Yang Li and Yike Guo. 2015. Wiki-Health: From Quantified Self to Self-Understanding. Futur. Gener. Comput. Syst. (August 2015). DOI:https://doi.org/10.1016/j.future.2015.08.008

[18] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing. ACM Comput. Surv. 42, 4 (2010), 1–53. DOI:https://doi.org/10.1145/1749603.1749605

[19] Konstantinos Perakis, Thanasis Bouras, Dimitris Ntalaperas, Panagiotis Hasapis, Christos Georgousopoulos, Ratnesh Sahay, Oya Deniz Beyan, Cristi Potlog, and Daniela Usurelu. 2013. Advancing Patient Record Safety and EHR Semantic Interoperability. In 2013 IEEE International Conference on Systems, Man, and Cybernetics, 3251–3257. DOI:https://doi.org/10.1109/SMC.2013.554

[20] 2015. Summary of the HIPAA Privacy Rule. Retrieved Feb 1, 2018 from http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/index.html

[21] 2016. The Personal Information Protection and Electronic Documents Act (PIPEDA). Office of the Privacy Commissioner of Canada. Retrieved Feb 1, 2018 from https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/

[22] Jens H. Weber-Jahnke, Morgan Price, and James Williams. 2013. Software engineering in health care: Is it really different? and how to gain impact. In 2013 5th International Workshop on Software Engineering in Health Care, SEHC 2013 - Proceedings, 1–4. DOI:https://doi.org/10.1109/SEHC.2013.6602469

[23] Ann Cavoukian and Khaled El Emam. 2011. Dispelling the Myths Surrounding Anonymization Remains a Strong Tool for Protecting Privacy. Inf. Priv. Comm. Ontario, Canada June (2011). Retrieved from http://www.ipc.on.ca/images/Resources/anonymization.pdf

[24] P Samarati and L Sweeney. 1998. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion. Proc IEEE Symp. Res. Secur. Priv. (1998), 384–393.

[25] Li Ninghui, Li Tiancheng, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. Proc. - Int. Conf. Data Eng. 2 (2007), 106–115. DOI:https://doi.org/10.1109/ICDE.2007.367856

[26] Benjamin Eze and Liam Peyton. 2015. Systematic Literature Review on the Anonymization of High Dimensional Streaming Datasets for Health Data Sharing. Procedia Comput. Sci. 63, (2015), 348–355. DOI:https://doi.org/10.1016/j.procs.2015.08.353

[27] Benjamin Eze, Craig Kuziemsky, Rubina Lakhani, and Liam Peyton. 2016. Leveraging Cloud Computing for Systematic Performance Management of Quality of Care. Procedia Comput. Sci. 98, (2016), 316–323.

[28] Nazanin Sabooniha, Danny Toohey, and Kevin Lee. 2012. An evaluation of hospital information systems integration approaches. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics - ICACCI '12, 498. DOI:https://doi.org/10.1145/2345396.2345479

[29] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, Tyson Roffey, and Jim Bottomley. 2009. A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. J. Am. Med. Informatics Assoc. 16, 5 (2009), 670–682. DOI:https://doi.org/10.1197/jamia.M3144

[30] Jianping Ma, Cong Peng, and Qiang Chen. 2014. Health Information Exchange for Home-Based Chronic Disease Self-Management -- A Hybrid Cloud Approach. In 2014 5th International Conference on Digital Home, 246–251. DOI:https://doi.org/10.1109/ICDH.2014.54

[31] Jianhua Liu, Selnur Erdal, Scott A Silvey, Jing Ding, John D Riedel, Clay B Marsh, and Jyoti Kamal. 2009. Toward a fully de-identified biomedical information warehouse. AMIA Annu. Symp. Proc. 2009, (2009), 370–4.