Patient Controlled, Privacy Preserving IoT Healthcare Data Sharing Framework

Mohammad Jabed Morshed Chowdhury, A. S. M. Kayes, Paul Watters, Patrick Scolyer-Gray, Alex Ng, and Tharam Dillon
La Trobe University, Melbourne, Australia
{m.chowdhury, a.kayes, p.watters, p.scolyer-gray, alex.ng, t.dillon}@latrobe.edu.au

Abstract

Healthcare data personally collected by individuals with wearable devices have become important sources of information for healthcare professionals and medical research worldwide. User-Generated Data (UGD) offers unique and sometimes fine-grained insight into the lived experiences and medical conditions of patients. The sensitive subject-matter of medical data can facilitate the exploitation and/or control of victims. Data collection in medical research therefore restricts access control over participant-data to the researchers. Therefore, cultivating trust with prospective participants concerned about the security of their medical data presents formidable challenges. Anonymization can allay such concerns, but at the cost of information loss. Moreover, such techniques cannot necessarily be applied on real-time streaming health data. In this paper, we aim to analyze the technical requirements to enable individuals to share their real-time wearable healthcare data with researchers without compromising privacy. An extension for delay-free anonymization techniques for real-time streaming health data is also proposed.

1. Introduction

Healthcare data-related acts and regulations worldwide have given rights to patients to request their health records at any time, and often mandate that health care providers must provide data in a format that can be shared with others. This permits patients to become the best aggregators of their medical data users can compile data from each medical centre they attend and the details of every consultation, all of which is stored in electronic health records (EHRs). These data are collected in addition to conventionally recorded information such as fitness data stored on smart devices. Unlike traditional health data, wearable data (both from medical service providers and consumer electronic products, such as Fitbit or Apple Watch) can provide improved measures of everyday behaviour and lifestyle, filling the gaps in more traditional clinical data collection and presenting a more complete picture of health.

This level of access to private health data offers tremendous potential. If researchers can harness the data hubs of individual patients, then there will be no shortage of medical information that can be analyzed, which could ultimately lead to a more efficient care ecosystem. Some studies of longitudinal health and well-being collect and manage data over human lifespans, representing a rich bounty of research data, but also numerous potential risks in the absence of a proper governance framework around confidentiality and data access [26]. The rise of IoT devices in healthcare will provide even more opportunities to collect health data in real-time.

Many patients wish to share their data for medical research. According to various surveys, nine-out-of-ten patients with access to their health data are willing to share that data to support research [1]. However, given the strict qualification criteria imposed by the researchers, only about 5% of candidates eventually constitute the group participating in clinical trials [5]. Long recruitment phases prolong the execution of trials, thus increasing the time it takes for innovative new medicines to be studied and approved, leaving patients to wait years for new treatment options. According to a survey [2], 85% of respondents perceive privacy concerns as a major barrier to sharing health information. It is clear that collected data may be used to extract or infer sensitive information about users private lives, habits, activities and relations, which all refer to individuals privacy [6]. About half of the respondents were either concerned or very concerned about the re-identication of their anonymized health and medical information. If data were irreversibly anonymized, 71% of respondents were willing to share data with researchers.

Therefore, in this work, we have proposed a high level framework for privacy preserving framework for

URI: https://hdl.handle.net/10125/64195
978-0-9981331-3-3

patient controlled data sharing for medical research. There are multiple aspects and components of the overall framework, namely *i) providing consent in terms of access control ii) privacy preserving participants selection, iii) real time data anonymization, and iv) data monetization*. During the recruiting phase, patients matching the needs for the designed trial and willing to participate are recruited. Enrolled patients give their consent to the trusted party so that it can do the anonymization before releasing the data to the researchers. The researchers get access to the anonymized data from the trusted parties and can nally analyze it. This process leverages the ability of modern technologies to communicate over the Internet in order to (a) reach nearly an unlimited number of potential participants and (b) collect relevant data at home without requiring participants to regularly visit the study centers. The downside of this process is the requirement for a trusted third-party.

Structure of the paper. In Section 2, we introduce the requirements that will drive the design of our framework. In section 3, we outline our data sharing framework. We introduce the real-time data anonymization algorithm for streaming health data in section 4. Section 5 discusses related work, and we conclude in section 6.

The main **research contributions** of this work are as follows.

- A privacy preserving healthcare data sharing framework for wearable devices.
- A real-time data anonymization algorithm.

2. Requirement Analysis

In this section, we illustrate the key requirements to develop a patient controlled health record sharing framework.

2.1. Digital Real-time Consent

Health data is privacy critical, and requires strict privacy protection. Moreover, different privacy related legislation (e.g., HIPPA, Health Records Act 2001) empowers the individuals to have privacy protection from the health service provider depending on their jurisdiction. Traditional ”consent” is being taken in ”pen-and-paper” and individual patients do not have much control after they have given the consent. Secondly, it is often time consuming to revoke the consent. The introduction of information technology has made some of the controls available to the patient. However, they are not fine-grained and patient-centric, and often very difficult to use. A good example is the Australian government’s MyHealthRecord system [35], which provides very sophisticated access control levels, but in a way which was scared numerous consumers to opt-out of the system.

Thirdly, with the introduction of IoT and cyber physical systems, we see the rise of real-time, health streaming data. Table 1 shows the list of popular wearable devices and types of information they collect about ourselves.

Table 1. Wearable Devices and Associated Personal Information

Types of Wearable Devices	Types of Contextual Information
Activity Tracker (e.g., Fitbit, Garmin)	• Physical activities (walking, biking, racing) • Body weight and Fat • Heart rate • Sleep
Sports and GPS Watches (e.g., Garmin, Wahoo TICKR)	• Fitness exercise data
Smart Watches (e.g., Apple, Samsung)	• Physical activity • Sleep • Heart rate
Smart Cloths and Shoes (e.g., Nike, Digitsole)	• Body temperature • Physical activities (walking, running)

Traditional access control system often fall short of keeping up the requirements of this type of real-time streaming data. With streaming data, individuals want *context-aware* control and continuous authorization [3, 4]. Finally, we can summarize the requirements as such,

1. Individuals should have full control over the sharing of the data.

2. Individual should have contextual control (e.g., sharing based on time or location) over the sharing of their data.

3. There should a real-time access revocation mechanism, which actually works.

2.2. Identification of Trusted and Untrusted Parties

Our research is informed by privacy concerns about the health information of individuals. As part of the Health Insurance Portability and Accountability Act (HIPAA), introduced in 1996. Additionally, it brought to the forefront some privacy concerns [11]. This study indicate that the lack of trust in ICTs and digital health care affects very seriously any effort to migrate from conventional healthcare procedures to electronic systems.The risk is even greater when data are real-time and streaming in nature. Therefore, a proper trust model is vital while designing a framework to deal with real-time health data.

In general, the term trust implies that the agreement depends on a third party based only on the belief of its integrity and/or benevolence [12]. Trustworthiness has been the fundamental pre-requisite for the progress of commerce and prosperity in human societies and determines to which extent an individual wants to depend on others. The central role of trust as a major type of social capital in online activities is well established. According to the above, any successful data sharing framework should target at increasing a patients trust [21].

In a medical data sharing framework, there are multiple parties involved, such as patients, researchers, wearable device providers, data brokers. Among these parties, patients usually trust hospitals and health service providers. On the other hand, they usually do not trust the researchers with their data (especially corporate researchers aka ”big pharma”). There could be any other government entities which patients usually trust could play the role of data broker between the patient and the researchers, but again, the MyHealthRecord case shows that there is a fair degree of skepticism around government control of personal data.

This identification process is important, because all the anonymization processes will be run on the trusted platform before being released to the health researchers. During this phase, the responsibilities of all involved parities will need to be declared and negotiated. Finally, we can summarize the requirements as such,

1. The system should have clear guidelines about the role of each participant in the system.

2. The trust model should be made clear and based on the trust assumptions the private health data will be exposed to that party.

3. If data is exposed/shared with untrusted parties, the data should be anonymized before being shared with them.

2.3. Data Quality Parameters

Because we are considering the wearable data, a series of characteristics of the data needed to be examined before they could be used by the health research team. The implementation and deployment of effective solutions needs to properly address these parameters.

Accuracy: Heart rate, blood glucose level, the number of steps made per hour, and the daily caloric intake are examples of information that can be gathered from IoHT devices. Like any other information, they have a certain accuracy that characterizes the data. For example, it was studied that heart rate monitoring made by common activity trackers and smartwatches have accuracies that range from 99.9 to 92.8%; thus, in certain scenarios they can be treated as accurate [12]. In [15], the authors measured the performances of a very common activity band with respect to professional calorimeters.

Authenticity: There exist multiple entities that could generate data, so establishing an authentication method to verify the source of data and avoid poor quality data or tampered data is required. Users can be interested in faking data for multiple reasons. An example could be the assumption of opiate drugs that cause dependence. A patient may be willing to fake data in order to receive stronger medications or additional doses of the same drug but for a longer period of time.

Confidentiality: Data confidentiality is mostly achieved through encryption, using algorithms such as AES, DES, or RSA [16]. These algorithms are highly optimized and represent a mature technology, but often they require a conspicuous amount of processing power (it depends also on the parameters for encryption and the strength it is willing to achieve).

Freshness: Some clinical researches require delicate patient monitoring, the delay is a critical requirement. For example, for heart diseases such as arrhythmia, identifying and generating early warnings require very short response times [12].

Availability: In IoT applications, it is common to nd locally centralized systems that send data to the cloud periodically, where the storage solutions are mostly decentralized. This hybrid approach presents strengths such as ease of installation, low maintenance costs,

and simple connection since a single device that acts as a gateway has to be congured and connected to the internet.

Integrity: The concept of integrity is strongly connected to the protection of information from malicious third parties, cybercriminals, or any external interference from the initial transmission to the nal reception of data. The systems must be aware of a threat whenever it tries to tamper with the data. Malicious third parties could be interested in making revenue for their false outsourced data. A solution for such a problem is investigated in [17], where the authors provide an analysis of data integrity verication based on an authenticator suitable for both the cloud and the IoT.

Therefore, the requirements related to data qualities are:

1. data needs to be collected by certified and trusted devices.
2. data needs to be collected in real-time.
3. proper cryptographic mechanism should be in place between the device and the data host, and between the data host and the researchers to ensure the confidentiality and the integrity of the data.

2.4. Privacy Preserving Participants Selection

The researchers usually select the participants for their clinical trials based on some criteria set by themselves for effective study. Participant must have to reveal the information related to those criteria. It may sometimes violate the privacy of the patients. Therefore, the framework need to come up with a mechanism to provide the researchers the ability to search the desired candidate and at the same time, the privacy of the participants also need to be maintained. Therefore, the requirements related to participant section are:

1. researchers should have the ability to search candidate based on their criteria
2. patients' data need to be protected from privacy leakage by ensuring privacy preserving mechanism.
3. If patients agree then only their data will be released to the researchers.

Using meta-data about the health information could be an effective mechanism to protect the privacy of the patients. Some researchers also suggest to use privacy-preserving clustering to handle large datasets.

2.5. Real Time Data Anonymization

Anonymization techniques safeguard the privacy of the individuals when their data is published or shared with others. There are several anonymization approaches available, among them k-anonymization is considered the most popular one. However, it suffers from different limitations and researchers have proposed different extensions such as *l-diversity* and *t-closeness* for k-anonymity [33]. Recently, differential privacy [34] has received much attention from both industry and academia. However, none of these algorithms really address the requirements of real-time streaming data. Real-time streaming data has the following unique characteristics compared to traditional static data. These are:

1. missing data in the stream: sometimes due to network or hardware fault data could not be researched to the anonymization algorithm in real-time.
2. noise in the data stream: there could be noise in the data.
3. delay in the data stream: due to the network delay data could not be reached in due time.

Therefore, we have to come up with an anonymization algorithm which can handle these specific characteristics of the streaming data.

2.6. Data Monetization

Several studies indicates that patients or individuals are not motivated to share their data as they are not aware of the benefit of their contribution or sometime they do not gain from the profit of the pharmaceutical companies [12]. Therefore, there should be a monetization mechanism to allow the individual patients to get financial benefit by sharing their data, subject to appropriate ethical safeguards and controls.

3. Proposed Data Sharing Framework

To realize those requirements outlined in the previous section, we propose a data sharing framework. Figure 1 shows the high level structure of our framework. We have private spaces where patients have access to their devices and data. In the middle, we have the trusted space, where patients will write their data sharing policy to share their healthcare data generated by the devices. These parties are called *"data brokers"* and are usually hospitals or government agencies. We can see them as the *"matchmaker"* between the patients

and the researchers. They also run the anonymization process before releasing patient's data to the researchers. The third column is the researchers who will get access to the data via trusted parties based on the policies defined by the patient. Finally, they can do data analysis on the shared data for their research study. Table 2 shows different actors and their activities in our framework.

3.1. Data Sharing Policy Model

The healthcare data from the wearable will be shared with the researchers based on the conditions defined by the patients. These policies should have the ability to specify the specific researcher (with whom the patient is sharing the data), the contextual conditions, and the privacy requirements. Therefore, the policy model should have three components; namely identity, context, and privacy.

Policy is defined in terms of attributes of the patients (P), researchers(R) and the environment(E). Here environment means attributes which are independent of patient and researchers, such as time. The policy is attached to the data(D) to control its access.

Definition 1 *(Data Sharing Model). The core data sharing model (M) is a tuple.*

$$M = D \times Policy$$
$$Policy = R \times CP \times Decision$$
$$CP = ContextCondition \times Privacy$$

Policy is expressed as the identity of researcher (R) and the CP, where C stands for context condition and P stands for privacy requirements.

Attributes: The data sharing model is designed using the attributes. The attribute is a triples (name, value, type) where the name is a unique identifier and value is an unordered set of atomic values of a given type. Type restricts the data type of the atomic values (e.g. string, integer, boolean, etc.) to a system defined data type. Attributes represent some descriptive characteristic of the entity to which they are assigned.

For example, a researcher's identity may be descried using the email address and the context of the patient may be described as her location. The set of all attributes is divided into four subsets based on their origin and to which entity they may be applied:

Patient Attributes (PA): the set of attribute name, type pairs that may be applied to patient such that $\forall\ a \in PA : a = (name, type)$ and each element of PA has a unique name (i.e. there cannot be two elements/attributes with the same name).

Researcher Attributes (RA): the set of attribute name, type pairs that may be applied to researcher such that

Table 2. Role and Activities in Data Sharing Framework

Role	Activities
Patient	1. Register with the Data Broker. 2. Provides meta-data about types of data they want to share (e.g., blood pressure) and their personal information (e.g, name, age, sex) to the data broker. 3. Integrate their cyber physical device/s with the data broker. 4. Specify whether they want to donate (Sharing data without any financial gain) or sell their data. 5. Define policy specifying their conditions of sharing. 6. Accept the request (from researchers) to share. 7. Get payment for sharing their data.
Data Broker	1. Do match making between patients and researchers. 2. Allow the patients to integrate their cyber physical devices through APIs with the broker system. 3. Allow the patients to define data sharing policies. 4. Do anonymization before releasing the data to the researchers. 5. Provide payment mechanism for both the patients and researcher.
Researcher	1. Register with the data broker 2. Search for their appropriate candidates based on their criteria in the data broker. 3. Integrate their system with the data broker (may be by APIs) to access real-time streaming data. 4. Send request to the patient to share or make an offer (if commercial researchers) to buy the data. 5. Recruit the participants for their study. 6. Access the shared data and do analytic on the data. 7. Pay the participants (if commercial sharing)

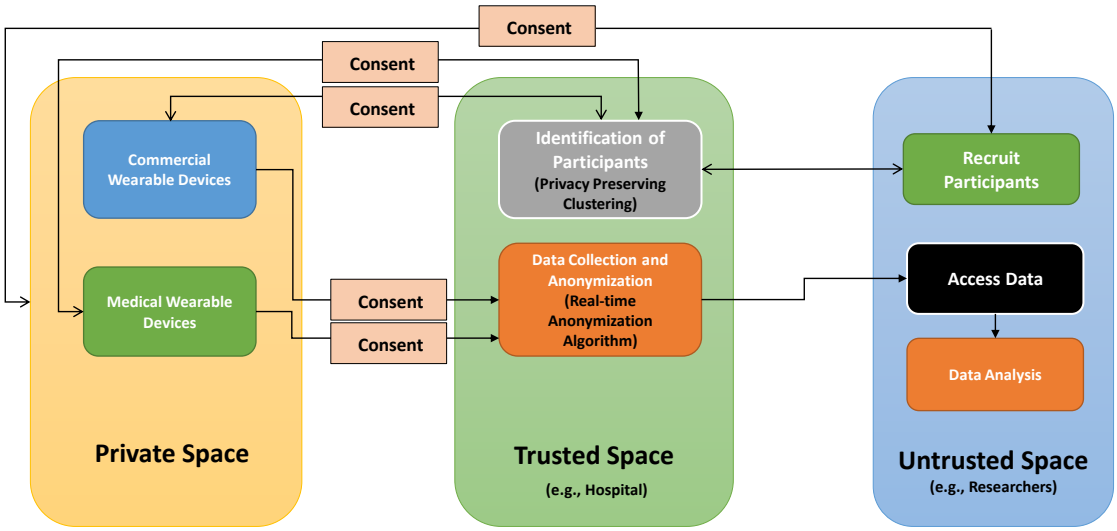

Figure 1. High Level Process for Patient Oriented Clinical Trials

$\forall\ a \in RA : a = (name, type)$ and each element of RA has a unique name (i.e. there cannot be two elements/attributes with the same name).

Environment Attributes (EA): the set of attribute name, type pairs that are independent of patient and researcher such that $\forall\ a \in DSA : a = (name, type)$ and each element of EA has a unique name (i.e. there cannot be two elements/attributes with the same name).

Definition 2 *(Identity of Researcher). Let I denote the set of identities of all researchers. The set of identities includes email, username, user-id.*

$$\text{I} \in \{\ \textit{email, username, userid}\ \}$$

The identity of the *researchers* is expressed as the *email, username, or userid* and their corresponding value.

Definition 3 *(Context Condition). Context is represented by the dynamic attributes of the entities. Context depends on the dynamic condition of the patient, researcher and environment. It is expressed as the attributes of the one or more attribute-value pairs of the data sharing entities (e.g., patient, researcher and environment).*

$$\textit{Context} \subseteq \text{PA x EA}$$

In the above equation, $\forall\ att\_name \in \{\text{P.PA} \vee \text{E.EA}\}$ and $\forall$ att_value $\in \{\text{PAA} \vee \text{EAA}\ \}$.

Definition 4 *(Continuous Authorization). The continuous authorization is defined by using a special attribute in context condition, namely interval. This attribute is used to by the patient to define the time interval between two authorization check.*

For instance, if the *interval* is set at 5 second by the patient, then the context condition will be checked at every 5 minutes. The default value of the *interval* is set to *zero*. That means the context condition will only be checked at once, at the beginning of the access (when access request is made). Secondly, one *interval* will only be associate with one contextual attributes. This is very vital to protect the privacy of the patient as the wearable data are real-time and streaming in nature.

$$< e.cs, rel.op, v, interval >$$

In the above tuple, $e \in Environment$, $c_s \in C_s$, $rel.op \in \{<, \leq, >, \geq, =, \neq\}$, and $interval \in Time$.

Definition 5 *(Privacy). Privacy is expressed as the specific privacy related attributes. There are 2 attributes related to privacy (e.g., purpose and anonymization). purpose express the motive to use to share their data. The value of this attribute could be "donate" or sell. The value for the attribute anonymization could be either true or false. If true is selected then data will be released as anonymized form.*

$$\textit{Privacy} \subseteq \{\ \text{purpose} \vee \text{anonymization}\ \}$$

Figure 2 shows the meta-model of the data sharing policy for sharing real-time streaming data from the wearable devices with the researchers.

3.2. Privacy Preserving Participants Selection

Selecting the right kind of participants for the clinical trial is vital. The section process is usually done by lookup into the entries in the database in the data broker. Data broker provides the searching interface to the researchers to search participants based on their criteria. However, database query based searching has its limitations as it only find the participants who are

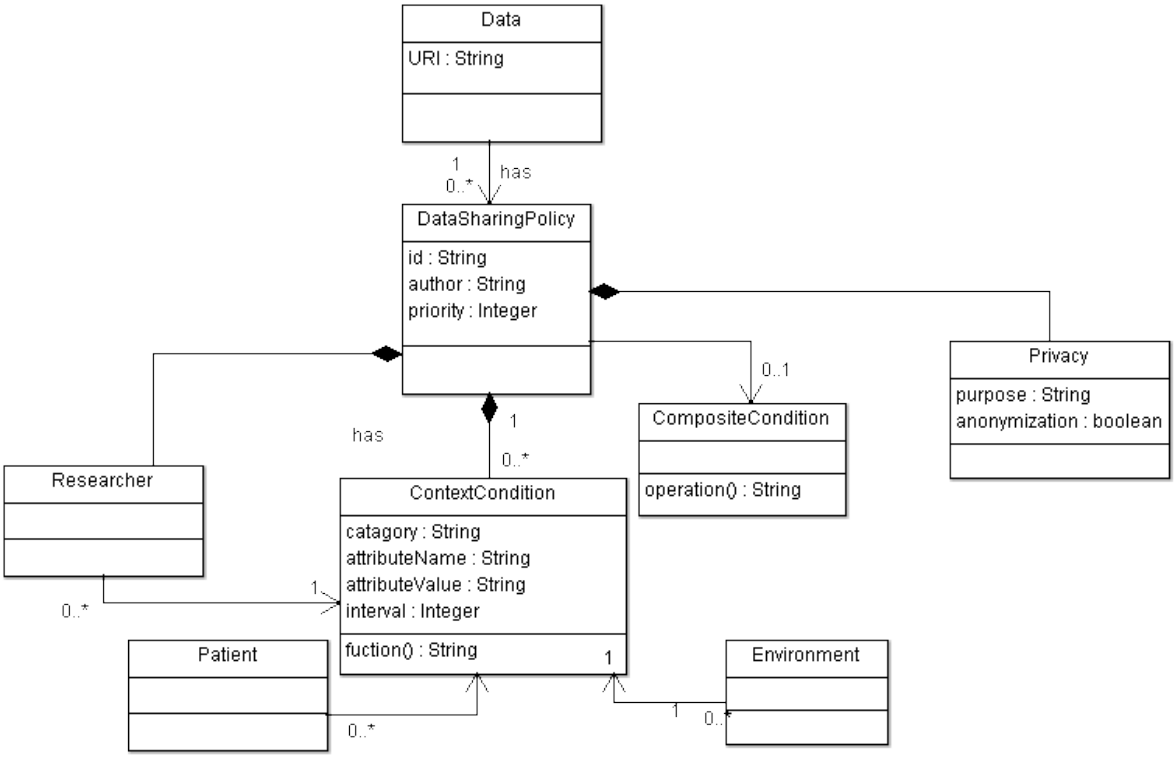

Figure 2. Meta-model of Data Sharing Policy

matching the exact criteria. In recent, time there has a rise in clustering algorithms for big dataset. Clustering algorithms can group the items in the database based on their similarity. Therefore, this can help the researchers find the participants which are not exact match but very similar to their criteria.

Among the clustering algorithm, k-means algorithm is extensively used in industry and academia for its efficiency. The k-means algorithm is used to partition a given set of observations into a predefined amount of k clusters. The algorithm as described by [14] starts with a random set of k center-points (μ). During each update step, all observations x are assigned to their nearest center-point (see equation 1).

$$S_i^{(t)} = \left\{x_p : \left\|x_p - \mu_i^{(t)}\right\|^2 \leq \left\|x_p - \mu_j^{(t)}\right\|^2 \forall j, 1 \leq j \leq k\right\} \quad (1)$$

After several iterations, it can group the similar items in the dataset, which means similar type of participants in terms of our use-case. However, K-mean algorithm may reveal some of the private information of the patients. Therefore, the data broker should use privacy-preserving k-mean algorithm [13] to ensure that no private data is revealed to the researchers without the consent of the patients.

4. Anonymization Algorithm for Real-time Streaming Data

Most of the existing privacy-preserving techniques, such as k-anonymity methods, are designed for static data sets. As such, they cannot be applied to streaming data which are continuous, transient, and usually unbounded. Moreover, in streaming applications, there is a need to offer strong guarantees on the maximum allowed delay between incoming data and the corresponding anonymized output. Recently, researchers have proposed anonymization algorithms which address the *"delay"* factor in streaming data [22]. However, in best of our knowledge, there is no streaming data anonymization techniques available which can address the issues of *i) noise* and *ii) missing data*. We propose an anonymization techniques that can address these two issues on top of delay factor in streaming data.

For health data, Personal Health Information (PHI) attributes are categorized under one of three categories: i) Direct Identifier attributes– that can directly identify the person, such as name ii) Quasi Identifier attributes, on their own, cannot identify an individual, when quasi-identifiers (QI) are combined, they behave like direct identifiers. Most re-identification efforts link

QI values to publicly available data repositories to re-identify individuals in an anonymized dataset; iii) Sensitive Identifier Attributes (SI) are not usually public data but are sensitive if associated with an individual, such as heart condition is a sensitive attributes.

We assume real-time health data streams containing personal information. Among the various attributes in a data stream, we focus on only three main attributes in order to simplify the problem: (tupleID, QI, and SI). According to the attributes, a data stream can be described as *(tupleID, QIs, SI)*.

TupleID indicates the unique number of a tuple. It is usually removed before releasing the data stream, because it can be an identifier of the tuple. The *QI* is an attribute of an individual in the tuple, such as age, sex, nationality, and zipcode. The *SI* is the private attribute, which should not be directly related to the individual, such as blood pressure and medical condition. In the following definition of the data stream, we omit the tupleID, since it is consequentially removed in the anonymization process.

Definition 6 *(Data stream). Let QI be the quasi-identifier attributes of a tuple, where* $QI=\{qi_1,qi_2,...,qi_n\}$*, and let SI be the sensitive information of a tuple. We define a data stream S as a set of tuples (QI,si).*

Definition 7 *(Noisy Data in the Stream). Given a tuple t(QI,si) from a data stream and a domain of the sensitive information* D_{SI}*. If any si is above or less than the thresholds set by the algorithm then the si is marked as noise. An machine learning mechanism, called Oracle* (σ) *is used to find the appropriate value from the historical value and replace the incoming si with the calculated si. If the Oracle* (σ) *cannot find the a si with certain confidence* (δ)*, then that particular record is discarded.*

Definition 8 *(Missing Data in the Stream). Like the previous step, the si is checked, if data is missing in the stream, the Oracle* (σ) *will generate si from the historical data with certain confidence*(δ) *set by data broker or researchers.*

We can use the following algorithm (Algorithm 1) with existing delay-free data streaming anonymization algorithm [23,24] to cover all three aspects of streaming data discussed in section 2.5.

```
Result: Anonymized data
initialization QI - ϕ; SI - ϕ;
while t ∈ T ≠ endtime do
    instructions;
    if lowerBoundary ¡ si ¡ upperBoundary then
        anonymizeTouple (qi, si);
    else
        if si == null then
            si =oracle(t);
            anonymizeTouple (qi, si);
        else
            si =oracle(t);
            anonymizeTouple (qi, si);
        end
    end
end
```

Algorithm 1: Anonymization technique for steaming data with noise and missing data

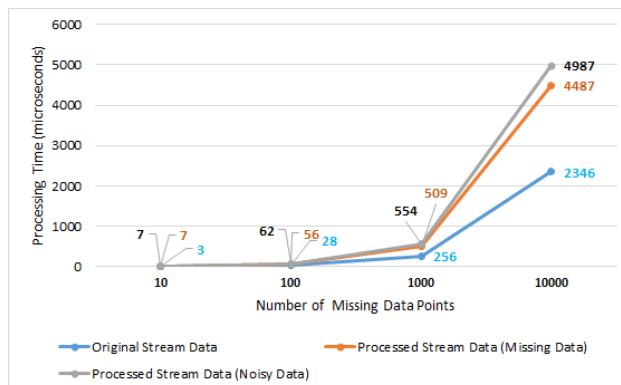

Figure 3. Processing time for missing and noisy data

4.1. Experiment

We have done an experiment to measure the delay in case of noisy or missing data in the data stream. We have used Fitbit heart rate data as the streaming data source. Fitbit heart rate API provides the heart rate of the individual as a time series data. We have only measured the time to calculate new sensitive data (si) in case of missing or noisy sensitive data (si). The implementation of the full anonymization algorithm is our future work. Figure 3 shows the time taken to calculate the new sensitive value from the historical data. In our experiment, we have found that generating a new data from the historical data is quite efficient and can be done without adding much delay in the anonymization process. We understand that the performance and accuracy of the algorithm hence the anonymized dataset will vary based on the wearable input datasets. However, we suggest that this can provide guidelines for other researchers to consider missing and noisy data while anonymizing real-time streaming data.

5. Related Work

5.1. Medical Data Access and Sharing Approaches

Sharing medical data for the research activity can improve overall health care sector. Researchers have proposed different mechanism to share their data based on the sentiment of the patients captured by different surveys [2, 5].

Different researchers have taken different approaches to enable individual patients to share their data with the researchers. Malin et al. have proposed a policy based approach to share the medical data with the researcher [6]. Yang et al. and Doel et al. have proposed medical data sharing framework using cloud infrastructure [7, 8].

In addition to cloud, we have seen several research efforts aiming to leverage the blockchain technology to tackle the aforementioned issues. Examples include MedRec from MIT [9] and others [10, 18, 19] for sharing medical data, and for improving the transparency in clinical trials [20].

A successful history of developing a family of Context-Aware Access Control (CAAC) approaches [27–32] has been proposed in the last few years, to access data and information resources from different data sources (e.g., providing patients' health records access to different hospital users). Towards this end, the authors have proposed - a fuzzy context information system to deal with imprecise contexts [28], a global CAAC model to apply a single set of policies for accessing data from multiple sources [32] [27], a new CAAC model for managing critical situations dealing with IoT-based data resources through dynamic contextual roles [29]. However, these CAAC approaches and frameworks are not adequate for patient-controlled medical data sharing through dealing with anonymization for real-time streaming data.

5.2. Data Anonymization Approaches

Several researchers have proposed anonymization techniques to overcome the delay factor in the data streaming anonymization. Cao et. al. has proposed an scheme called CASTLE which generates clusters in which the tuples of data streams are accumulated [23]. Then, it releases the clusters when the size of the accumulation reaches the delay-constraint. Basically, CASTLE guarantees k-anonymity, and also provides an algorithm for l-diversity. The method presented in [25] uses a probability function to determine the release of data streams; it also accumulates data from data streams and decides on the release time according to this function. The function promotes releases when the accumulated data experience less information loss but a longer delay. The authors of SABRE [24] presented an algorithm for anonymizing microdata to satisfy t-closeness, and also extended the algorithm to anonymize data streams. The SABRE framework maintains sliding windows that function as buffers of the input tuples. When tuples in a certain window are expired due to newly inserted tuples, old tuples should be released. These tuples are anonymized by SABRE and released to an output stream to overcome the delay in the data. All the above mentioned mechanisms have been proposed to overcome delay and does not cover missing or noisy data in the data stream.

6. Conclusion and Future Research Directions

Data availability for health researchers is critical for sustaining the momentum of successful innovations in healthcare technology. Wearable devices allow individuals to generate and share their health related data. However, individuals are concerned about the privacy of their data. Although research has addressed privacy issues in traditional health care data, preserving the privacy of health data (especially for cyber-physical based streaming data) represents a blind-spot in scholarly knowledge. In this work, we have outlined the requirements for a privacy-preserving framework. The requirements for real-time streaming data anonymization were discussed, and an algorithm was presented, demonstrated and found suitable for the task. This framework and algorithm will help individual patients share their data with prospective researchers without the risk of compromising their privacy.

We propose that future research should continue experimenting with the proposed algorithm, ideally with larger samples and heterogeneous selections of health-related IoT devices, to explore the full potential and additional applications of the findings presented here. Further opportunities are presented by research into the robustness of our claim to risk reduction. Finally, we strongly encourage the pursuit of research that tests and measures the efficacy of improving research participation rates and levels of participant confidence in data security that might stem from the inclusion of the data anonymization technology discussed in this paper.

References

[1] PatientsLikeMe, "PatientsLikeMe Survey Shows Vast Majority of People With Health Conditions Are Willing To Share Their Health

Data,"https://tinyurl.com/y32nhf94' (accessed on 5th of June, 2019)

[2] Weitzman, Elissa R and Kaci, Liljana and Mandl, Kenneth D, "Sharing medical data for health research: the early personal health record experience", Journal of medical Internet research, Vol-12, Issue-2, Year-2010.

[3] Chowdhury, Mohammad Jabed Morshed and Colman, Alan and Han, Jun and Kabir, Muhammad Ashad, "A Policy Framework for Subject-Driven Data Sharing", In Proceedings of the 51st Hawaii International Conference on System Sciences, HICCS, 2018

[4] Chowdhury, Mohammad Jabed Morshed and Colman, Alan and Han, Jun and Kabir, Muhammad Ashad, "A system architecture for subject-centric data sharing", In Proceedings of the Australasian Computer Science Week Multiconference, ACM, 2018, p-4

[5] Weitzman, Elissa R and Kelemen, Skyler and Kaci, Liljana and Mandl, Kenneth D, "Willingness to share personal health record data for care improvement and public health: a survey of experienced personal health record users", vol-12, issue-2, BMC medical informatics and decision making, 2012, pp-39

[6] Malin, Bradley and Karp, David and Scheuermann, Richard H, "Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research", Journal of Investigative Medicine, vol-58, issue-1, 2010, pp-11-18

[7] Yang, Ji-Jiang and Li, Jian-Qiang and Niu, Yu, "A hybrid solution for privacy preserving medical data sharing in the cloud environment", Future Generation Computer Systems, vol-43, pp-74-86, 2015.

[8] Doel, Tom and Shakir, Dzhoshkun I and Pratt, Rosalind and Aertsen, Michael and Moggridge, James and Bellon, Erwin and David, Anna L and Deprest, Jan and Vercauteren, Tom and Ourselin, "GIFT-Cloud: A data sharing and collaboration platform for medical imaging research", computer methods and programs in biomedicine, vol - 139, 2017, pp-181-190

[9] Azaria, Asaph and Ekblaw, Ariel and Vieira, Thiago and Lippman, Andrew, "Medrec: Using blockchain for medical data access and permission management", International Conference on Open and Big Data (OBD), 2016, pp-25-30.

[10] Liu, Paul Tak Shing,"Medical record system using blockchain, big data and tokenization" , International Conference on Information and Communications Security, 2016, pp-254-261.

[11] Annas, George J and others, "HIPAA regulations-a new era of medical-record privacy?", New England Journal of Medicine, vol-348, issue-15 2003, pp-1486–1490.

[12] Angeletti, Fabio and Chatzigiannakis, Ioannis and Vitaletti, Andrea, "Towards an Architecture to Guarantee Both Data Privacy and Utility in the First Phases of Digital Clinical Trials", Journal Sensors, vol-18, issue-12, 2018, pp-4175.

[13] Meskine, Fatima and Bahloul, Safia Nait, "Privacy preserving k-means clustering: a survey research.", Int. Arab J. Inf. Technol. vol-9, issue-2, 2012, p: 194-2000.

[14] Huang, Zhexue and Ng, Michael K,"A note on k-modes clustering", Journal of Classification, vol-20, issue-2,2003, p:257-261.

[15] Adam Noah, J.; Spierer, D.K.; Gu, J.; Bronner, S. Comparison of steps and energy expenditure assessment in adults of Fitbit Tracker and Ultra to the Actical and indirect calorimetry. J. Med. Eng. Technol. 2013, 37, 456462.

[16] Suo, H.; Wan, J.; Zou, C.; Liu, J. Security in the internet of things: A review. In Proceedings of the 2012 International Conference on Computer Science and Electronics Engineering (ICCSEE), Hangzhou, China, 2325 March 2012; Volume 3, pp. 648651.

[17] Liu, C.; Yang, C.; Zhang, X.; Chen, J. External integrity verification for outsourced big data in cloud and IoT: A big picture. Future Gener. Comput. Syst. 2015, 49, 5867

[18] Xia, Qi and Sifah, Emmanuel Boateng and Smahi, Abla and Amofa, Sandro and Zhang, Xiaosong,"BBDS: Blockchain-based data sharing for electronic medical records in cloud environments", Multidisciplinary Digital Publishing Institute, 2016.

[19] Dubovitskaya, Alevtina and Xu, Zhigang and Ryu, Samuel and Schumacher, Michael and Wang, Fusheng, "Secure and trustable electronic medical records sharing using blockchain", American Medical Informatics Association, 2017, p-650.

[20] Nugent, Timothy and Upton, David and Cimpoesu, Mihai, "Improving data transparency in clinical trials using blockchain smart contracts", F1000Research, vol-5, 2016.

[21] H. Tran, P. Watters, M. Hitchens and V. Varadharajan, "Trust and authorization in the grid: a recommendation model". In ICPS'05. Proceedings. International Conference on Pervasive Services, 2005. (pp. 433-436), 2005.

[22] Kim, Soohyung and Sung, Min Kyoung and Chung, Yon Dohn, "A framework to preserve the privacy of electronic health data streams", Journal of biomedical informatics, vol-50, 2014, p:95-106.

[23] J. Cao, B. Carminati, E. Ferrari, K.-L. Tan Castle: continuously anonymizing data streams IEEE Trans Dependable Secure Comput, 8 (3) (2011), pp. 337-352

[24] J. Cao, P. Karras, P. Kalnis, K.-L. Tan Sabre: a sensitive attribute bucketization and redistribution framework for t-closeness, VLDB J, 20 (1) (2011), pp. 59-81

[25] B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao, Y. Jia Continuous privacy preserving publishing of data streams Proceedings of the 12th international conference on extending database technology: advances in database technology, ACM (2009), pp. 648-659

[26] P. Watters, D. Kuh, S. Latham, I. Shah, and K. Garwood, "Enabling access to british birth cohort studies: A Secure Web Interface for the NSHD (SWIFT)", Proceedings of the 11th IEEE International Conference on e-Health Networking, Applications and Services, Healthcom, 2009.

[27] A. Kayes, J. Han, W. Rahayu, T. Dillon, M. Islam, and J. Han, "A Policy Model and Framework for Context-Aware Access Control to Information Resources", The Computer Journal, volume 62, issue 5, pp. 670-705, Oxford University Press, 2018.

[28] A. Kayes, W. Rahayu, T. Dillon, E. Chang, and J. Han, "Context-aware access control with imprecise context characterization for cloud-based data resources", Future Generation Computer Systems, volume 93, pp. 237-255, Elsevier, 2019.

[29] A. Kayes, W. Rahayu, and T. Dillon, "Critical situation management utilizing IoT-based data resources through

dynamic contextual role modeling and activation" Computing, volume 101, issue 7, pp. 743-772, Springer, 2019.

[30] A. Kayes, W. Rahayu, T. Dillon, E. Chang, and J. Han, "Context-Aware Access Control with Imprecise Context Characterization Through a Combined Fuzzy Logic and Ontology-Based Approach", Proceedings of the 25th International Conference on Cooperative Information Systems (CoopIS 2017), pp. 132-153, Springer, 2017.

[31] A. Kayes, J. Han, A. Colman, and M. Islam "Relboss: A relationship-aware access control framework for software services", Proceedings of the 22nd International Conference on Cooperative Information Systems (CoopIS 2014), pp. 258-276, Springer, 2014.

[32] A. Kayes, W. Rahayu, T. Dillon, and E. Chang, "Accessing Data from Multiple Sources Through Context-Aware Access Control", Proceedings of the 17th IEEE International Conference On Trust, Security And Privacy in Computing And Communications (TrustCom 2018), pp. 551-559, Springer, 2018.

[33] Li, Ninghui and Li, Tiancheng and Venkatasubramanian, Suresh, "t-closeness: Privacy beyond k-anonymity and l-diversity", IEEE 23rd International Conference on Data Engineering, 2007, pp:106-115

[34] Dwork, Cynthia, "Differential privacy", Journal of Encyclopedia of Cryptography and Security, 2011, p:338-340

[35] MyHealthRecord, Url: https://www.myhealthrecord.gov.au/ (accessed on 14th of March, 2019)