

Mobile Anonymization and Pseudonymization of Structured Health Data for Research

Stella Dimopoulou², Chrysostomos Symvoulidis^{1, 2}, Konstantinos Koutsoukos^{1, 2},
Athanasios Kiourtis¹, Argyro Mavrogiorgou¹, Dimosthenis Kyriazis¹

Affiliation 1: Dept. of Digital Systems, University of Piraeus

Piraeus, Greece, {simvoul, konkoutsoukos, kiourtis, margy, dimos}@unipi.gr

Affiliation 2: BYTE Computer S.A.

Athens, Greece, {sdimopoulou, csymvoulidis}@byte.gr

Abstract—Healthcare Organizations need to share the health data of the patients with Research Centers in order to fulfill research purposes and improve the healthcare services provided to the patients. However, the information being processed by the Research Centers includes personal and/or sensitive data, which puts the privacy of the individuals at stake. To mitigate the risk of identity disclosure and privacy violation, a variety of privacy mechanisms, such as anonymization and pseudonymization, can be applied to the personal data of the data subjects. In this paper a mobile library is presented in order to either anonymize or pseudonymize the individuals' personal information which follows the Fast Healthcare Interoperability Resources protocol. To evaluate the implementation and the functionalities of the library two case studies are described – one for each privacy mechanism.

Index Terms—anonymization, pseudonymization, health data, privacy, mobile

I. Introduction

Nowadays the Healthcare Organizations collect the patients' health data through the Electronic Health Records (EHRs) in order to improve the patients' aid, assist them more effectively and provide them with better treatment [1], [2]. Along these lines, the quality of the services is enhanced while the diagnoses are more precise and accurate [1], [2]. However, the patients' data should be confidential and private and only the authorized parties, who are involved with this data, should be able to extract knowledge from the patients' dataset [2], [3].

The ability of identifying the sickness of the patient or the likelihood to disclose the individuals' identity according to their dataset are two privacy concerns which play a key role when processing sensitive data [2]. At this point it is worth noting that the personal data is a broader category of data which also includes the sensitive data. Examples of sensitive data are the health-related data, religious beliefs, political opinions of the individuals etc [4]. When it comes to data sharing with other parties, such as researchers, the

General Data Protection Regulation (GDPR) stipulates some principles that needs to be followed. The explicit consent of the data subject and the legal basis of the purpose of processing are among these principles [5]. Additional techniques, such as data anonymization and data pseudonymization privacy mechanisms, can be applied to the personal data of the data subject – before its processing [4] – in order to protect and maintain the privacy of the individuals.

As already mentioned the privacy of the individuals plays a key role when processing their personal data, and even more when processing their health data. This happens due to the fact that the exposure of this kind of data may lead to violation of human rights and the freedom of the individual [5]. To preserve the privacy of the individuals a mobile library has been implemented in order to anonymize or pseudonymize their data so as not to share it with other parties, such as researchers, in its original form. In brief, the logic of the library is as follows.

First of all, a health-related dataset, which is represented in JavaScript Object Notation (JSON) format, is requested from the citizen in order to be sent to the Research Center. Thereupon either of the aforementioned privacy mechanisms should be applied to the individuals' personal data, according to the needs of each research study. The anonymization and pseudonymization processes are implemented locally on their mobile phone, without exposing the original data anywhere, and the anonymized or pseudonymized data are finally sent to the Research Center in order for the researchers to process them accordingly. It is important that the data subjects should be well-informed and provide their consent in order for their data to get processed (anonymized or pseudonymized). In addition, in case of data pseudonymization the consent should also include how the stakeholders may use this data since according to GDPR pseudonymized data constitutes personal data even though it is pseudonymized

[5].

To sum up, when it comes to anonymized data – after the anonymization process has been implemented – the data can be used without the consent of the individuals. On the contrary, in data pseudonymization, both the processing of pseudonymization and the pseudonymized data per se require the explicit consent of the data subjects since the pseudonymized data are essentially personal data [4].

This paper analyzes the privacy issues and concerns which privacy experts are confronted with when Healthcare Organizations and citizens share their personal data with third parties, such as researchers, in order to satisfy specific purposes of processing. It also presents a mobile library which can be used as an approach in order to mitigate the risk of identity disclosure and safeguard the privacy of the data subjects. The remainder paper has been divided into four sections as follows: Section two describes the definition of the data anonymization and data pseudonymization processes as well as their main differences, their methods of applicability in health-related state-of-the-art research as well as the publicly known implementations on the mobile phone. Section three presents the scenario description and the overall architecture of the library. Section four discusses the functionalities and the results of the mobile library. Section five sets out the conclusions and the future steps.

II. Theoretical Background

A. Data Anonymization and Data Pseudonymization

Data anonymization is a one-way process in which all personal data of the data subject is deleted or modified in order not to be retained the linkage between the natural person and its dataset. The anonymized data should be anonymized with probability of one and by no means should the stakeholders be able to identify the person to whom the data belongs [4], [5], [6].

On the contrary, data pseudonymization is a process in which all personal data of the data subject is deleted or modified so as to maintain the association between the data subjects and their dataset. To achieve this association, there is some additional information within the pseudonymized dataset – the pseudonyms – which can identify a natural person indirectly if needed. This indirect identification is fulfilled through the mapping table which retains a) the pseudonym, which is placed in the pseudonymized dataset, and b) the personal data of the data subject, which is required for the re-identification. However, the identification of the natural person should not be feasible without the utilization of the mapping table with probability of one. In addition, the mapping table should be maintained independently

from the pseudonymized data and accessed only by the data controller who is the authorized party for the re-identification of the natural person [4].

The key difference between data anonymization and data pseudonymization is that the former renders the re-identification impracticable whereas the latter maintains information leading to identity disclosure of the data subject [4], [5], [6].

B. Privacy in Health-related Data

In this day and age Electronic Health Records are utilized for a diverse range of purposes [1], [7]. Several parties, such as hospitals, medical and pharmaceutical researchers and patients, can be benefited from the exchange and the interoperability of the EHRs [1], [7]. Patients are able to share their data with other stakeholders, which can process the patients' personal data and extract knowledge from this information. More specifically, medical experts and pharmaceutical researchers analyze the individuals' health-related data in order to be led in informed decision making and provide the patients indirectly with better services in the future [1]. Since such data may include sensitive information, it should be accessible to as few stakeholders as possible and not compromise the privacy of the individuals [1], [7]. Concerning the information security, limited access of the data can be achieved through the access control mechanisms. On the other hand, the privacy of the individuals can be preserved through the data anonymization and data pseudonymization privacy mechanisms, which are widely used so as to prevent the identity disclosure of the data subject [1], [4].

With regard to data anonymization, the most widely used techniques applied to the personal data are k -anonymity and ℓ -diversity [8].

1) *k-anonymity*: In k -anonymity, there is a number of at least k individuals who have the same attribute values – the same personal information – within a dataset [9]. The procedure of k -anonymity consists of the following steps.

First of all, all direct identifiers – all information which can identify independently a natural person – should be removed from the dataset. As a next step, all attributes which can identify an individual indirectly – all indirect identifiers or quasi-identifiers – should get modified and generalized in broader categories so as to have the same values for at least k individuals [8], [9]. The remaining attributes which constitute sensitive data will remain the same within the dataset. As a result, the personal data of an individual cannot be distinguished from at least $k - 1$ data of the data subjects which belong in the same group [9]. Due to the fact that the probability of re-identifying an individual

within the dataset is associated with the k -value, the k -value plays a key role when implementing k -anonymity to a health-related dataset.

2) ℓ -diversity: An extended approach of k -anonymity is ℓ -diversity. ℓ -diversity is essentially the same as k -anonymity, but with an additional feature. ℓ -diversity defines as q -block a set of records which have the same quasi-identifier values – *this was virtually the logic behind each group in k -anonymity in the previous sub-section*. A q -block of a dataset is considered as ℓ -diverse if it contains at least ℓ different/distinct and well-distributed sensitive data values [8]. Therefore, the entire dataset is ℓ -diverse if all q -blocks are ℓ -diverse.

C. Privacy Mechanisms on the Mobile Phone

In modern times the number of wearable devices has been increased gradually. These devices tend to collect the individuals' private information in order to provide them with the ability to monitor their fitness levels and track their location with GPS. However, it is vital the fact that wearable devices gather and process the personal data of the data subject, since processing may lead to privacy violation [10]. To counterbalance the utilization of wearable devices and the possible privacy threats which may occur, there is a variety of privacy measures which can be implemented.

Due to the fact that many devices collect personal data through sensors, the need for implementing techniques such as data anonymization or data pseudonymization is increased rapidly [10]. However, even though these techniques are state of the art for the privacy of the individuals, only a few publicly known implementations have been published with regard to the data anonymization on the mobile phone.

Data anonymization can be applied, for instance, to location-based services. Location-based services, which collect information through the mobile phone, can be utilized in order to provide GPS directions and/or answers to queries such as "which is the closest hospital to me". The result of this query will be a visual map with all hospitals located around the target point. What is worth noting is that the location of the individuals comprises personal information and should be protected with terms of privacy when being shared with the service provider [11]. Even though the individuals do not provide personal information, which may lead directly to identification, to the service, an unauthorized entity may obtain additional information through location tracking or space and time correlation inference [12]. As a result, the complete movements of the data subject may be exposed and in this case the privacy of the individual will be violated [11].

To overcome the issue of the information leakage, the location-based information of the data subject will be anonymized through k -anonymity technique as presented in the previous section [11], [12]. In k -anonymity a data subject is deemed k -anonymous if its location cannot be distinguished from the location of at least $k - 1$ data subjects located at the same region [12]. Thus, the service provider, and an unauthorized entity in general, will not know the precise location of the individual, but the approximate region instead [11].

III. Proposed Architecture

A. Scenario Description

The aim of the application is to collect [13] the individuals' data – from the various sources [14] – in order for them to participate in health-related research. However, to achieve the data retrieval there are some intermediate steps which should be followed.

First of all, the application user should get informed in general about the research studies that the application supports, and provide his explicit consent, in order to participate in future studies – in order to get notified about future research studies. For each study, if the data subject meets the criteria to participate in this specific survey, its explicit consent is requested so that it can be enrolled at the study. After the user provides his consent to participate in the study, his data is requested as well in order to be sent to the Research Center (RC). What is worth noting is that only the minimum necessary dataset is retrieved (data minimization principle [5]), and this data gets anonymized or pseudonymized, and afterwards encrypted locally on the phone before being transmitted to the RC.

Data anonymization and data pseudonymization privacy mechanisms are used in order to protect the privacy of the individuals and prevent the identification of the data subject. Only in case of pseudonymization an authorized entity could be led in identity disclosure if necessary. Along with the opt-in consents, at the beginning of the scenario, there is an opportunity to withdraw the consents as well. The data subject can cancel its participation in a specific survey or exclude itself from all future studies. If this is the case, the data stop being collected while at the same time all previously collected data is deleted permanently.

B. Library Functionalities

The aim of the library is to anonymize or pseudonymize the individuals' health data before being sent to the researchers conducting health-related research. The prerequisites which should be satisfied before the library processes the personal data are the following: According to GDPR [5], a) the purpose of

processing should be lawful and b) the data subject should be i) well informed regarding its data which will get processed, ii) aware of the purpose of processing, and iii) should provide its explicit consent before the data processing. If and after these requirements have been satisfied, the library can collect the patients' personal data and process them accordingly.

The library supports two privacy mechanisms which safeguard the individuals' privacy. These mechanisms are the data pseudonymization and data anonymization privacy mechanisms. The former maintains a linkage between the natural person and its personal data, whereas the latter maintains no association at all. Pursuant to the needs of the research study either of the above-mentioned mechanisms will be applied to the personal data. However, the procedure in both cases is similar.

First of all, the users' data is collected through the application and the library process them locally on the phone based on the terms of the study. These terms are stated within the Research Definition Document (RDD). Along with other information/rules about the survey, the RDD records which privacy mechanism will be implemented to the data as well as whether a pseudo-identity or a pseudonym will be utilized in case of pseudonymization.

In the event of data anonymization all personal information – all direct and indirect identifiers – which may lead to the identity of the natural person is deleted from the dataset. In addition, the “need-to-know” principle is followed, and only the minimum and necessary information which is related to the current research is maintained. On the other hand, while in data pseudonymization all unnecessary information is deleted as well, an association between the data subject and the corresponding personal data is also preserved through the mapping table. The mapping table contains essentially all personal data which describe an individual, and a unique identifier which is different for each individual and for each research study. This identifier can be either a pseudo-identity or a pseudonym, and either of them is sent along with the minimum dataset to the RC.

The dataset format which describes the personal data of the data subject follows a standard named Fast Healthcare Interoperability Resources (FHIR). These FHIR resources are basically JSON files, and the library supports fourteen different resources. There is a list of attributes defined within the library, all of which must be deleted from the resources in order to achieve either data anonymization or data pseudonymization. More specifically, each time the library receives one of the profiles as an input, it can anonymize or pseudonymize

the dataset by deleting the aforementioned set of attributes.

Considering that an association between the natural person and its health data should be maintained, the privacy mechanism which should be implemented is data pseudonymization. In such a circumstance, there are two alternatives of replacing all personal information with a unique identifier. The first alternative is the pseudo-identities whereas the second one is the pseudonyms. On the other hand, when the identity disclosure of the individuals should be prevented, the data anonymization privacy mechanism should be applied to the data.

IV. Experimental Evaluation

As already mentioned in the previous section, the library supports two privacy mechanisms, data anonymization and data pseudonymization. Depending on the content of the Reference Definition Document, which states the terms of each research study, either of these mechanisms is going to be applied to the individuals' health data in order to safeguard their privacy.

First of all, the Research Center, which is responsible for the formation of the RDD, selects the privacy mechanism implemented on the personal information of the data subjects. On the assumption that the data subjects should not be identifiable throughout the research study, the mechanism which will be implemented is data anonymization. On the contrary, if the re-identification is legitimate according to the purpose of processing, the privacy mechanism which will be selected is data pseudonymization.

Since the suitable privacy mechanism has been documented in the RDD, and the explicit consent of the data subject has been collected through the application, the personal data of the data subject can be retrieved and processed locally on its mobile phone, in order to be sent anonymized or pseudonymized at the Research Center.

As depicted in the snippet below (Snippet 1), the FHIR resource which represents the individuals' personal data – and more specifically the patient's personal information – is as follows.

```
{
  "resourceType": "Patient",
  "id": "3466",
  "language": "it-IT",
  "text": {
    "status": "generated",
    "div": "<div xmlns='http://www.w3.org/1999/xhtml' xml:lang='it-IT' lang='it-IT'><p><b>Generated Narrative</b></p><p><b>identifier</b>: id: Patient/MS01</p><p><b>name</b>: Markus Smith </p><p><b>gender</b>: male</p><p><b>birthDate</b>: 2013-12-05</p><p><b>address</b>: Rome IT (HOME)</p><p><b>generalPractitioner</b>: <a href='\"Organization-34432.html\"'>Generated Summary: language: it-IT; id: Organization/FTGM01; name: Fondazione Gabriele Monasterio</a></p></div>"
  },
  "identifier": [
    {
      "system": "http://interopEHRate.eu/fhir-resource/",
      "value": "Patient/MS01"
    }
  ]
}
```

```

},
"name": [
  {
    "family": "Smith",
    "given": [
      "Markus"
    ]
  }
],
"gender": "male",
"birthDate": "2013-12-05",
"address": [
  {
    "use": "home",
    "type": "physical",
    "city": "Rome",
    "country": "IT"
  }
],
"generalPractitioner": [
  {
    "reference": "Organization/34432"
  }
]
}

```

Snippet 1. Sample of FHIR resource in its original form

A. Data Anonymization

The patient profile, represented in Snippet 1, consists of a variety of key-value pairs. To achieve the anonymization operation of FHIR resources, all key values which represent personal information should be deleted from the patient's dataset. An anonymization extension declaring the privacy mechanism which is applied to the profile is required as well.

According to the example, the key values id, text, identifier, name, birthDate, address, and generalPractitioner should be removed since they are either direct or indirect identifiers for the current patient. However, in order for the profile to be valid, some attributes cannot be entirely deleted. In such cases, these attributes should get modified so as not to contain personal information. The attributes in this example are the id, the identifier, the name, and the birthDate. What is worth noting is that the id and the identifier should have also the same – random – value.

Pursuant to the above sample of the FHIR resource the anonymized version of Snippet 1 is illustrated in Snippet 2 as follows.

```

{
  "resourceType": "Patient",
  "id": "1467746362506598617",
  "language": "it-IT",
  "extension": [
    {
      "url": "http://interopehrate.eu/fhir/StructureDefinition/AnonymizationExtension-IEHR",
      "valueCoding": {
        "system": "http://interopehrate.eu/fhir/CodeSystem/AnonymizationType-IEHR",
        "code": "anonymization",
        "display": "Anonymization"
      }
    }
  ],
  "identifier": [{
    "value": "1467746362506598617"
  }],
  "name": [{
    "family": "Anonymous"
  }],
  "gender": "male",
  "birthDate": "2013"
}

```

Snippet 2. Anonymized version of FHIR resource

B. Data Pseudonymization

Similarly as in data anonymization, all direct and indirect identifiers of the patient should be deleted or modified, as well as the corresponding pseudonymization extension should be appended. In addition, to retain the association between the data and the data subject either a pseudo-identity or a pseudonym should replace the value of the keys named id and identifier.

In consideration of the fact that the documented privacy mechanism of the RDD is data pseudonymization with pseudo-identities, and the pseudo-identity is 96845138c8261dfc466861ba1890e9ae0d7b92f246bec37e67bd38547014a76b202b3e, the key value of id and identifier should be replaced with this pseudo-identity. The attributes text, address, and generalPractitioner should be removed whereas the attributes name and birthDate should be modified, as in case of data anonymization.

Based on the sample of the FHIR resource in Snippet 1, the pseudonymized version of this profile is depicted in Snippet 3 as follows.

```

{
  "resourceType": "Patient",
  "id": "96845138c8261dfc466861ba1890e9ae0d7b92f246bec37e67bd38547014a76b202b3e",
  "language": "it-IT",
  "extension": [
    {
      "url": "http://interopehrate.eu/fhir/StructureDefinition/AnonymizationExtension-IEHR",
      "valueCoding": {
        "system": "http://interopehrate.eu/fhir/CodeSystem/AnonymizationType-IEHR",
        "code": "pseudonymization",
        "display": "Pseudonymization"
      }
    }
  ],
  "identifier": [
    {
      "value": "96845138c8261dfc466861ba1890e9ae0d7b92f246bec37e67bd38547014a76b202b3e"
    }
  ],
  "name": [
    {
      "family": "Anonymous"
    }
  ],
  "gender": "male",
  "birthDate": "2013"
}

```

Snippet 3. Pseudonymized version of FHIR resource

C. Evaluation of Results

To evaluate the efficiency of data anonymization and data pseudonymization on the mobile phone, five out of fourteen FHIR resources have been anonymized and pseudonymized. The time needed for implementing these two privacy mechanisms is depicted in Table 1. More specifically, Table 1 indicates the name and the size of each profile as well as the time needed for the anonymization and pseudonymization processes. From the table it is clear that as the size of the profile is increased, the process of data anonymization requires more time than that of data pseudonymization. On the contrary, as the size of the profile is decreased, the process of data anonymization requires less time than that of data pseudonymization. However, the duration

of the anonymization and pseudonymization privacy mechanisms highly depends on the size of the profile.

TABLE I
EVALUATION OF PRIVACY MECHANISMS

Table FHIR Profile <small>(size)</small>	Privacy Mechanisms <small>(in sec)</small>	
	Anonymization	Pseudonymization
Composition Confidentiality <small>(132 KB)</small>	0.887420	0.764434
Diagnostic Report Imaging <small>(131 KB)</small>	0.470300	0.446620
Document Reference <small>(126 KB)</small>	0.342727	0.353317
Patient <small>(1,07 KB)</small>	0.005492	0.024307
Practitioner <small>(504 byte)</small>	0.001545	0.010481

V. Conclusions

The scope of the paper was to present both data anonymization and data pseudonymization privacy mechanisms, along with their utility and usability when processing health-related data on the mobile phone. The structured health data, which was used in the mobile library, follows the FHIR protocol in JSON format. According to the privacy preference, which was stated within the RDD, either of the aforementioned privacy mechanisms was applied to the personal data locally on the citizen's phone. As a result, the library generated an anonymized or pseudonymized dataset, and sent it to the RC. What is worth noting is that data anonymization is an irreversible process, and the anonymized data is not associated with an individual anymore. On the contrary, data pseudonymization is a reversible process, and if an authorized entity needs to re-identify the data subject for a specific and legitimate purpose, he/she can reverse the process in order to be led to the identity of the natural person.

As future work, additional functionalities will be implemented in order for the FHIR resources to be anonymized or pseudonymized based on a set of attributes which will be stated within the RDD. More specifically, along with the privacy mechanisms and the preference of pseudo-identities/pseudonyms, the author of the RDD can also define the list of attributes which will be needed for conducting a specific research study – *the remaining attributes must be deleted*. In addition, more profiles will be supported, and a list of attributes per profile will be generated in order for the process to be more precise and accurate.

Acknowledgment

The research leading to this result has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826106 (InteropEHRate project).

References

- [1] A. Aminifar, Y. Lamo, K. I. Pun, and F. Rabbi, "A practical methodology for anonymization of structured health data," 2019.
- [2] M. Jayabalan and M. E. Rana, "Anonymizing healthcare records: a study of privacy preserving data publishing techniques," *Advanced Science Letters*, vol. 24, no. 3, pp. 1694–1697, 2018.
- [3] M. Vardalachakis, H. Kondylakis, L. Koumakis, A. Kouroubali, and D. G. Katehakis, "Shinyanonymizer: A tool for anonymizing health data." in *ICT4AWE*, 2019, pp. 325–332.
- [4] "Information technology — security techniques — privacy framework," International Organization for Standardization, Geneva, CH, Standard, 2011.
- [5] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. European Commission. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [6] Pseudonymisation techniques and best practices. European Union Agency for Cybersecurity. [Online]. Available: <https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices>
- [7] S. Neto, F. S. Ferraz, and C. A. G. Ferraz, "Towards identity management in healthcare systems," in *Proceedings on the International Conference on Internet Computing (ICOMP)*. The Steering Committee of The World Congress in Computer Science, Computer ..., 2016, p. 157.
- [8] I. E. Olatunji, J. Rauch, M. Katzensteiner, and M. Khosla, "A review of anonymization for healthcare data," *CoRR*, vol. abs/2104.06523, 2021. [Online]. Available: <https://arxiv.org/abs/2104.06523>
- [9] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. A. Kuhn, "Flash: efficient, stable and optimal k-anonymity," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*. IEEE, 2012, pp. 708–717.
- [10] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in *Proceedings of the International Conference on Internet of Things Design and Implementation*, ser. IoTDI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 49–58. [Online]. Available: <https://doi.org/10.1145/3302505.3310068>
- [11] M. Mano and Y. Ishikawa, "Anonymizing user location and profile information for privacy-aware mobile services," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, ser. LBSN '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 68–75. [Online]. Available: <https://doi.org/10.1145/1867699.1867712>
- [12] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," *25th IEEE International Conference on Distributed Computing Systems (ICDCS'05)*, pp. 620–629, 2005.
- [13] C. Symvoulidis, A. Mavrogiorgou, A. Kiourtis, G. Marinos, and D. Kyriazis, "Facilitating health information exchange in medical emergencies," in *2021 International Conference on e-Health and Bioengineering (EHB)*, 2021, pp. 1–4.
- [14] C. Symvoulidis, A. Kiourtis, A. Mavrogiorgou, and D. Kyriazis, "Healthcare provision in the cloud: An ehr object store-based cloud used for emergency," in *HEALTHINF*, 01 2021, pp. 435–442.