Anonymizing clinical and genetic data of patients with minimum information loss

Dimitrios Dimitrios[1] and Athanasios Mpouras[1]

[1] UBITECH Ltd

dntalaperas@ubitech.eu

Abstract. While collaboration in research requires the publication of data as well as exchange of them between institutions, these data often contain personal information that according to ethical requirements and existing legislature, are not allowed to be disclosed. Anonymization of these data is therefore mandatory. In previous work we have presented a framework for anonymizing patient data using Data Cubes. Though efficient for anonymization, the Data Cubes approach often lacks in flexibility. In the current work, we present an alternative approach which is based on disclosing a row based anonymized version of the original data set. The methodology is more versatile, while it also preserves the statistical characteristics of the original data set. We demonstrate this by considering an SVM predictor that tries to estimate the value of Breslow's depth, based on the values of another clinical variable, namely Clark's level, and the expression count of a skin cancer related gene (CDKN2A). The predictions are shown to have the same characteristics for both the original and the anonymized data sets.

Keywords: Data Anonymization, Privacy Models, ARX framework

1 Introduction

While publication and exchange of clinical and genetic data is crucial for research, it is often the case that data contain sensitive information; information that, if disclosed, may cause harm to the patients. It is therefore mandatory, both in ethical and in legal terms, to respect and protect the individual's right to anonymity for the patients whose data are used for research. With the advent of GDPR[1] in particular, which is binding for all member states of the EU, the set of rules that govern patient personal data is clearly defined both for Data Controllers and for Data Processors. GDPR requires that, except for some specific cases, sensitive and personal data to be anonymized when they are disclosed to the public. Anonymized, in this sense, means that an attacker cannot reverse engineer the anonymized data set in order to retrieve the original data containing the sensitive information.

In previous work[2], we have demonstrated a methodology that allowed the anonymization of data, by using the aggregate structure known as a Data Cube. This methodology produced completely anonymized data sets, which maintained the statistical properties of the original data set, thus being effective for the purposes of analysis. It

relied upon identifying attribute combinations of interest, and then aggregating over the patients that shared the same values for these combinations. If, for example, the attributes of *diabetes* and *smoking* were to be measured, a two-dimensional data cube would have been created. If the data cube is denoted as *DC*, then *DC[0][0]* would contain the number of patients that had no diabetes and were not smoking, *DC[1][0]* the number of non-smokers who had diabetes, and so on. The above methodology, while producing sound results, suffers from two main drawbacks

The first one was that of inflexibility. In order to produce a Data Cube, a set of attributes and a categorization schema for each one of these attributes was chosen. If, however, other attributes were to be added or removed, the Data Cube needed to be reconstructed from start.

The second one was the one of performance. When combining clinical and genetic data, which are typically stored in different files, the amount of join operations led to a growth of the required amount of time needed to prepare data for processing. This growth, while polynomial, still poses a limiting factor when producing large Data Cubes. Although it has been demonstrated that by using High Performance Computing techniques, this overhead can be reduced by a cubic factor[3], the order of the polynomial overhead depends on the amount of attributes needed to be included. Thus, after a point the data pre-processing will become slow even by using HPC techniques.

In this work we introduce a new methodology that performs anonymization directly in the patient data. The transformation may be applied to the whole data set, which is then distributed in row format. Since the transformation is applied at the row level, it is linear in time and attributes of interest can be removed/added in a time efficient manner. The methodology will be applied in a use case that involves data of patients treated for melanoma.

This work was funded in the context of the SAGE-CARE[4] and ChildRescue[5] Project, both funded under the Horizon 2020 Programme.

2 Theory

In the most direct sense, a set is to be considered anonymized if no personal information of the data subjects is exposed to potential attacker. In the case of row-based format of data, this means that an individual cannot be linked to her corresponding entry in the data set. In a typical case however, especially those involving patient data, anonymization must be forced in the stricter sense, so that not only personal but also sensitive information is protected[6]. To make this clearer, we consider two more disclosure cases, against which an anonymization algorithm must protect the data subject.

Membership disclosure, involves the knowledge of whether data regarding an individual are contained in the dataset or not. *Attribute* disclosure, involves the knowledge of whether a specific attribute, has a specific value for an individual. Membership and attribute disclosure allow an attacker to discern sensitive information regarding individual(s), even without re-identifying the data set. For example, if an attacker knows that a subject is part of a double-blind placebo control clinical trial for a new drug

developed for melanoma, then she knows that the subject has melanoma with a probability ~50%.

The different disclosure cases force the categorization of subject attributes into four types:

- Identifying attributes, that contain subject's personal identifiable information (PII) and which are, in any case, removed from the set.
- Quasi-identifying (QIA) attributes which though they cannot be used on their own to identify a subject, they can be used, in combination with other variables to identify subjects.
- Sensitive attributes (SA), which contain information about subjects with which the subject should not be associated with. The presence of melanoma referred to earlier is a typical example of a sensitive attribute. Genetic information, which can now be linked to existing or potential health issues, such as predisposition to cancer, is also considered sensitive information.
- Non-sensitive attributes, which are irrelevant for the purposes of re-identification

QIA are transformed, while sensitive attributes are transmitted as is, provided that certain constraints are held. These constraints are enforced to prevent re-identification and are defined by the so-called privacy models

2.1 Privacy Models

Privacy models concerning anonymization can be very diverse; for the purposes of the present work we consider only the ones that are relevant to the use case of melanoma patients to be presented.

- *k-anonymity*[7]: This the most basic model; all other models are typically an extension of k-anonymity. Let a group of entries that shares the same values of QIA be called an equivalence class. k-anonymity enforces each equivalence class to have a size of at least k.
- *l-diversity*[8]: This model demands that each SI has at least l distinct values in each equivalence class.
- *t-closeness*[9]: this model demands that the distribution of the SI in each equivalence class is similar to the one of the whole dataset.

Figure 1 depicts a simple example demonstrating the meaning of *k* and *l* values. Anonymization with the desired characteristics is achieved with performing transformation on the QIA; the most typical ones are generalization, where values are grouped under categories, and suppression, where part of the information is masked.

Race	Birth	Gender	Postal Code	Diabetes
White	1980	male	20*	Yes
White	1980	male	20*	No
White	1980	male	20*	Yes
White	1982	female	18*	Yes
White	1982	female	18*	Yes
Black	1982	male	18*	Yes
Black	1982	male	18*	No

Figure 1: Data set with a k-value equal to 2. Postal codes and birth dates are suppressed to achieve the required k-value. Since the 2nd equivalence class has only one distinct value for the sensitive information of diabetes, the l-value is equal to one.

3 Application

In the case of patient data, anonymization needs to be such that no sensitive information may be attributed to a single patient. The transformed data set however, needs to exhibit the same statistical characteristics as the original one; otherwise the dataset will lose its validity and will be of no scientific value to the entities that share the information. For the purposes of the current work, we considered datasets that contain data of patients that were treated for melanoma. These data were retrieved in the context of the SAGE-CARE project and correspond to real patient data that are stripped from any PII. They contained both phenotypical information, concerning the clinical picture of the patient as well as genetic information corresponding to expression counts of the patient's genes. Satellite data, contained subject's information such as height, ethnicity etc. are also included in the dataset. These satellite data will be considered QIA in our model.

For anonymizing the dataset, we transform the set by requiring a k-value of 5, l-value of 2 and t-closeness of 0.2. The transformation is carried out by using the ARX framework[6]; for each of the QIA a hierarchy was defined that was used for implementing the generalization strategy. Figure 2 depicts a sample listing of the code used to anonymize the data set.

```
ARXConfiguration config = ARXConfiguration.create();
config.addPrivacyModel(new KAnonymity(5));
config.setSuppressionLimit(0.02d);

data.getDefinition().setAttributeType("breslow-depth", AttributeType.SENSITIVE_ATTRIBUTE);

config.addPrivacyModel(new EqualDistanceTCloseness("breslow-depth", 0.2d));

// l-diversity
config.addPrivacyModel(new DistinctLDiversity("breslow-depth ", 2));
```

Figure 2: Sample code listing depicting the data set anonymization. A k-value of 5 is defined with a suppression limit equal to 0.02. The sensitive attributes are then set (breslow depth is depicted in the example). The required t closeness and l values is then applied to each sensitive attribute (again, this is depicted for the Breslow depth attribute in the sample code).

3.1 Demonstration

To demonstrate that the transformation allows the data set to keep its main predictive features, we will show how a typical *Support Vector Machine* [11] (SVM) classification performs under the two data sets. We consider the original data set and take as an example the case of three sensitive information, namely the Breslow's depth, the Clark's level and the expression count of the *Cyclin-dependent kinase Inhibitor 2A* (CDKN2A) gene. The Clark's level is a measure of the depth that the melanoma has grown into the skin and of the levels of the skin that are affected. It is used as a prognosis factor in melanoma. Breslow's depth is another prognostic factor that measures the depth of the tumor using ocular micrometer. Breslow's depth is a more accurate prognostic factor that Clark's level; for bigger values of Breslow's depth specifically, the Breslow depth has a significantly better value as a predictor. The CDKN2A gene and its mutations on the other hand, has been shown to be linked with the appearance of skin cancers [12]. When the weaker predictor (Clark's level) is combined with the gene count, we expect the results to have a better predictive value. Figure 3 depicts the results of an SVM classification for the Breslow's depth of a patient's melanoma based on the values of the Clark's level and the CDKN2A expression count. The results suggest that for low expression counts of CDKN2A, the Clark's value is suggestive of the Breslow value. When the expression increases however, the Clark's level is no longer a good predictor for the Breslow's depth; indeed, a high expression of CDKN2A seems to be correlated with low Breslow's depth.

Figure 4 depicts the same classification when applied in the anonymized data set. Taking into account the difference in the scaling produced between the two cases, the classification has the same characteristics. It is to be noted, that the aim of the above experiment is not to extract any significant medical conclusions; indeed there are far more elaborate statistical models to demonstrate correlations between various phenotypical and genetic attributes. The main goal of the experiment is to demonstrate that

the anonymized data set has the same statistical characteristics and that any conclusions drawn from the original data set, persist to the anonymized one.

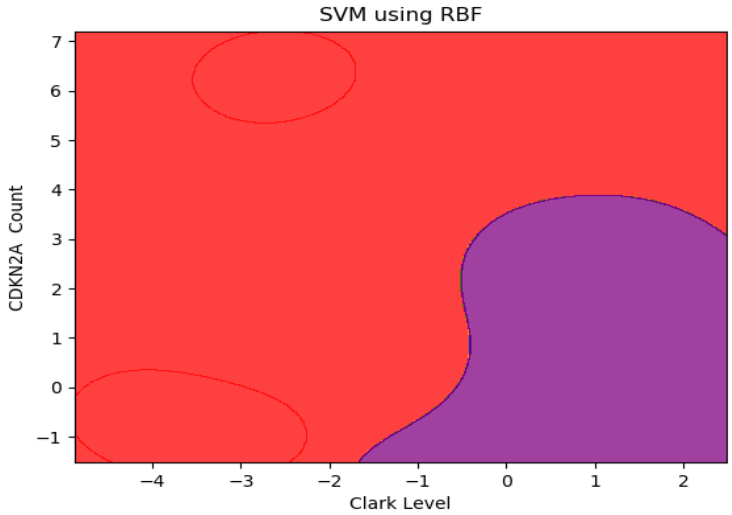

Figure 3: SVM classification for the Breslow-level for the original data set. Red(purple) areas depict low(high) values. The values are depicted scaled. For low Clark level (left part of the diagram), the Breslow's depth is also low. When the Clark's level is high, the Breslow's depth is also high only for low expressions counts of the CDKN2A gene (lower right part of the picture).

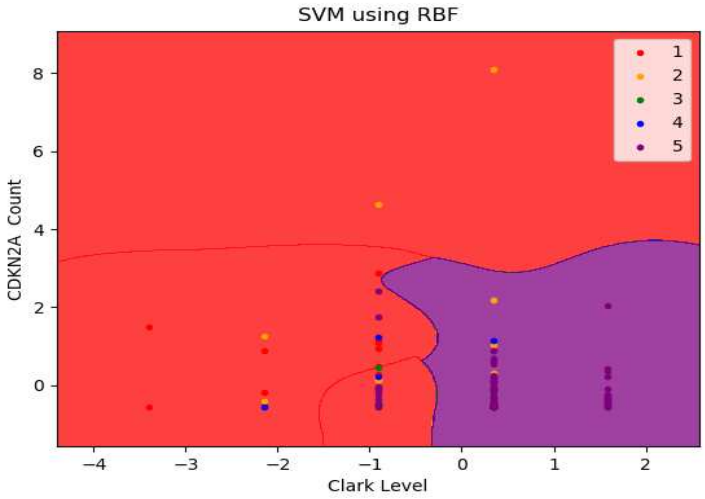

Figure 4: SVM classification for the Breslow-level for the anonymized data set. Red(purple) areas depict low(high) values. The values are depicted scaled. The test data are also depicted, with 5 ranges for Breslow's depth values as shown in the legend. When the Clark's level is high, the Breslow's depth is also high only for low expressions counts of the CDKN2A gene (lower right part of the picture).

4 Conclusions and further work

In the present paper we demonstrated that datasets containing phenotypical and genetic data can be effectively anonymized, with any information loss not leading to a substantial change to the data's statistical characteristics. The methodology can be applied to patient data and these can be shared between parties in a manner that does not compromise patients' personal and sensitive information and that is also compliant with the requirements of GDPR. The transformed data moreover, exhibit the same statistical

characteristics as the original data; their scientific value is therefore not compromised by the anonymization transformations.

Further work involves demonstrating the same result by using a model that will be proven to be applicable to a generic case, involving arbitrary number of attributes; the model will also provide metrics that will indicate the amount of information loss imposed by the transformation, thereby quantizing the efficiency of the anonymization.

References

1. EUR-Lex, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679, last accessed 2018/22/12
2. Ntalaperas D., A. Mpouras A.: An approach for anonymization of sensitive clinical and genetic data based on Data Cube Structures, in: Collaborative European Research Conference, 2016
3. Ntalaperas D., Mpouras A. Utilizing High Performance Computing Techniques for efficiently anonymizing sensitive patient data, in: Collaborative European Research Conference, 2017
4. SAGE-CARE Project Homepage, https://www.sage-care.eu, last accessed 2018/22/12
5. ChildRescue Project Homepage, https://www.childrescue.eu, last accessed 2018/22/12
6. GDPR Recital 51, https://gdpr-info.eu/recitals/no-51, last accessed 2018/22/12
7. Samarati P, Sweeney L:. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression" (PDF). Harvard Data Privacy Lab. 1998
8. Aggarwal C., Yu P.: Privacy-Preserving Data Mining – Models and Algorithms. Springer. pp. 11–52.
9. Ninghui L., Tiancheng L., Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering, 2007.
10. Prasser F, Kohlmayer F., Lautenschlaeger, Kuhn K. A.: ARX – A Comprehensive Tool for Anonymizing Biomedical Data. Proceedings of the AMIA 2014 Annual Symposium, 2014
11. Cortes C., Vapnik V. N.: Support-vector networks, Machine Learning. 20(3), 273–297 (1995).
12. Helgadottir H., Höiom V., Tuominen R., Jönsson G., Månsson-Brahme E., Olsson H., Hansson J.: CDKN2a mutation-negative melanoma families have increased risk exclusively for skin cancers but not for other malignancies, Int J Cancer. 137(9) (2015).