Test-Driven Anonymization in Health Data: A Case of Study on Assistive Reproduction

Cristian Augusto†
Department of Computing
University of Oviedo
Gijón, Spain
augustocristian@uniovi.es

Miguel Angel Olivero
ISTI-CNR- Pisa, Italy
IWT2 - University of Seville -
Sevilla, Spain
miguelangel.olivero@isti.cnr.it

Jesús Morán
Department of Computing
University of Oviedo
Gijón, Spain
moranjesus@uniovi.es

Leticia Morales
Web Engineering and Early Testing Research Group
University of Seville
Sevilla, Spain
leticia.morales@iwt2.org

Claudio de la Riva
Department of Computing
University of Oviedo
Gijón, Spain
claudio@uniovi.es

Javier Aroba
Department of Information Technologies
University of Huelva
Huelva, Spain
aroba@dti.uhu.es

Javier Tuya
Department of Computing
University of Oviedo
Gijón, Spain
tuya@uniovi.es

***Abstract*—Artificial intelligence (AI) is a broad field whose prevalence in the health sector has increased during recent years. Clinical data are the basic staple that feeds intelligent healthcare applications, but due to its sensitive character, its sharing and usage by third parties require compliance with both confidentiality agreements and security measures. Data Anonymization emerges as a solution to both increasing the data privacy and reducing the risk against unintentional disclosure of sensitive information through data modifications. Despite the anonymization improves privacy, the diverse modifications also harm the data functional suitability. These data modifications can affect to the applications that employ the anonymized data, especially those that are data-centric as the AI tools. To obtain a trade-off between both qualities (privacy and functional suitability), we use the Test-Driven Anonymization (TDA) approach, which anonymizes incrementally the data to train the AI tools and validate with the real data until maximize its quality. The approach is evaluated in a real-world dataset from the Spanish Institute for the Study of the Biology of Human Reproduction (INEBIR). The anonymized datasets are used to train AI tools and select the dataset that gets the best trade-off between privacy and functional quality requirements. The results show that TDA can be successfully applied to anonymize the clinical data of the INEBIR, allowing third parties to transfer without transgressing the user privacy and develop useful AI Tools with the anonymized data.**

Keywords*—*Anonymization, Software Testing, Artificial intelligence, Health-Care Data, k-Anonymity

I. Introduction

Nowadays, the data transferring/publishing from the data owners to third parties is a common collaborative practice in the software development process. These data may have sensitive information that transgresses the privacy laws of the body to which they belong. There are several techniques that allow the private transfer/publication of sensitive data [1], among which arise the anonymization. The anonymization techniques suppress/replace the data with the objective of increasing the privacy reducing the data heterogeneity. To measure how much anonymized a dataset is, there are a number of techniques like K-Anonymity [2] ensuring that each individual is un-distinguished from other K-1 individuals. The alterations introduced into the data by the anonymization harm the data quality, largely affecting the data-centric developments like artificial intelligence tools. Some authors have studied the effects of the anonymization into the data, concluding that anonymized data may be used in AI developments [3]. Other lines of research have focused on optimizing the anonymized data quality but, they only concern about utility or privacy, not for both [4], [5].

In this paper, we evaluate our previous approach Test-Driven Anonymization (TDA) [6] into a real-world health dataset from the Spanish Institute for the Study of the Biology of Human Reproduction (INEBIR). TDA aims to anonymize the data incrementally by testing each of these anonymization efforts to obtain the dataset that achieves a tradeoff between the functional (functional suitability) and non-functional (privacy) quality. The contributions of this paper are: (1) An evaluation of TDA in a real-life scenario related to health data, (2) An anonymized health dataset that achieves a tradeoff between both privacy requirements and functional suitability that can be used to train AI tools. The remainder of this paper is organized as follows: the approach is evaluated with the real-world dataset into Section II and the conclusions are presented in Section III.

II. Evaluation

The INEBIR dataset used to evaluate the TDA has information about patient treatments, age, identification number, levels of hormones as the Estradiol, and the pregnancy result, among others. This dataset is used in an ERP application, but its potential is reduced, due to the sensible data that don't allow transfer them without any privacy protection to third parties. TDA emerges as a feasible solution to incrementally obtain an anonymized dataset that reaches a trade-off between the data utility and privacy, checking against the original if the data are still useful for production.

In this section, we evaluate TDA to select the anonymized dataset that ensured that the AI models predict as close as the original if the patient is going to need medication: *Meriestra* in terms of their previous treatments and hormones levels. In the first place, an analysis of the data identifies sensible fields that may harm patient privacy, for then apply our TDA approach. We suppress the fields that may identify a patient: name, contact information, surnames and address. Secondly, we identify the fields that each one cannot identify a patient, but an aggregation of them can (pseudo-identifiers): *Estradiol level* and *Age*. TDA applies the anonymization method called generalization over the Age and Estradiol pseudo-identifiers, producing several anonymized datasets with different K-anonymity levels (privacy). TDA analyzes the anonymized datasets obtaining functional and non-functional qualities. The non-functional quality (Q_{NF}) is measured through the K-Anonymity levels. The functional quality (Q_F) measures the functional suitability of the AI models developed (in this case based on a *Bagging* algorithm), with the Precision, Accuracy, R-Squared, among others. With each AI model developed with the anonymized data, we want to classify the patients into "require" or "not require" the medication. The functional quality (Q_F) of each anonymization is obtained validating the AI model's performance with the Precision metric using the original data. The Quality (Q) of the dataset is obtained with the sum of the Q_F and Q_{NF} weighted with the user requirements (α).

The trend of Quality is depicted in Fig. 1.a (green color) as the weighted sum of Q_F (blue color) and Q_{NF} (red color). The maximum quality is obtained by the fifth anonymization effort (16-anonymity). The *Bagging* models trained with the first fifth anonymization efforts ({4, 10, 16}-anonymity) have similar behavior as Fig. 1 b, c, d, and e depicts. However, the AI model starts to not predict well from the sixth anonymization effort onwards, similar as depicted in Fig. 1.f (28-anonymity). Despite both Q_F and Q_{NF} are important, in our case study about health the Q_{NF} is a little bit more important than the Q_F. The best trade-off is obtained in the 5 effort, that has a privacy level of 16-Anonymity. It implies that against a data disclosure, the privacy of the users remains protected because the individual's data are indistinguishable from others.

III. Conclusions

The results show that the INEBIR center may release the anonymized dataset obtained with the TDA, so that third parties could develop AI tools with it, since (1) the data are roughly as useful as the original and (2) they complies with a high level of privacy (each patient cannot be distinguished from other 15 individuals). We also have validated the TDA approach with other Artificial intelligence algorithms, achieving similar results as shown with the *Bagging* algorithm. As conclusion, TDA may enabled the safe data transference between the INEBIR and the thirds parties, while the developments incurred with this data remain useful.

As future work, we pursue to apply TDA into more real-world health-care datasets or open the scope to more fields.

Acknowledgments

This work was supported in part by the Spanish Ministry of Economy and Competitiveness under TestEAMoS (TIN2016-76956-C3-1-R) and POLOLAS (TIN 2016-76956-C3-2-R) projects, and ERDF funds. We want to thank the collaboration of INEBIR (inebir.com) and G7 Innovation (g7innovation.com) in the development of this scientific project

References

[1] C. C. Aggarwal and P. S. Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," in *Privacy-preserving data mining*, Springer, 2008, pp. 11–52.

[2] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, 2001.

[3] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in *Proceedings - Int. Conf. on Data Eng.*, 2009, pp. 429–440.

[4] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 279–288.

[5] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W. C. Fu, "Utility-based anonymization using local recoding," in *Proceedings of the ACM SIGKDD Int. Conf. on KDD Mining*, 2006, vol. 2006, pp. 785–790.

[6] C. Augusto, J. Morán, C. De La Riva, and J. Tuya, "Test-driven anonymization for artificial intelligence," in *Proceedings - 2019 IEEE Int. Conf. on Artificial Intelligence Testing, AITest 2019*, 2019, pp. 103–110.

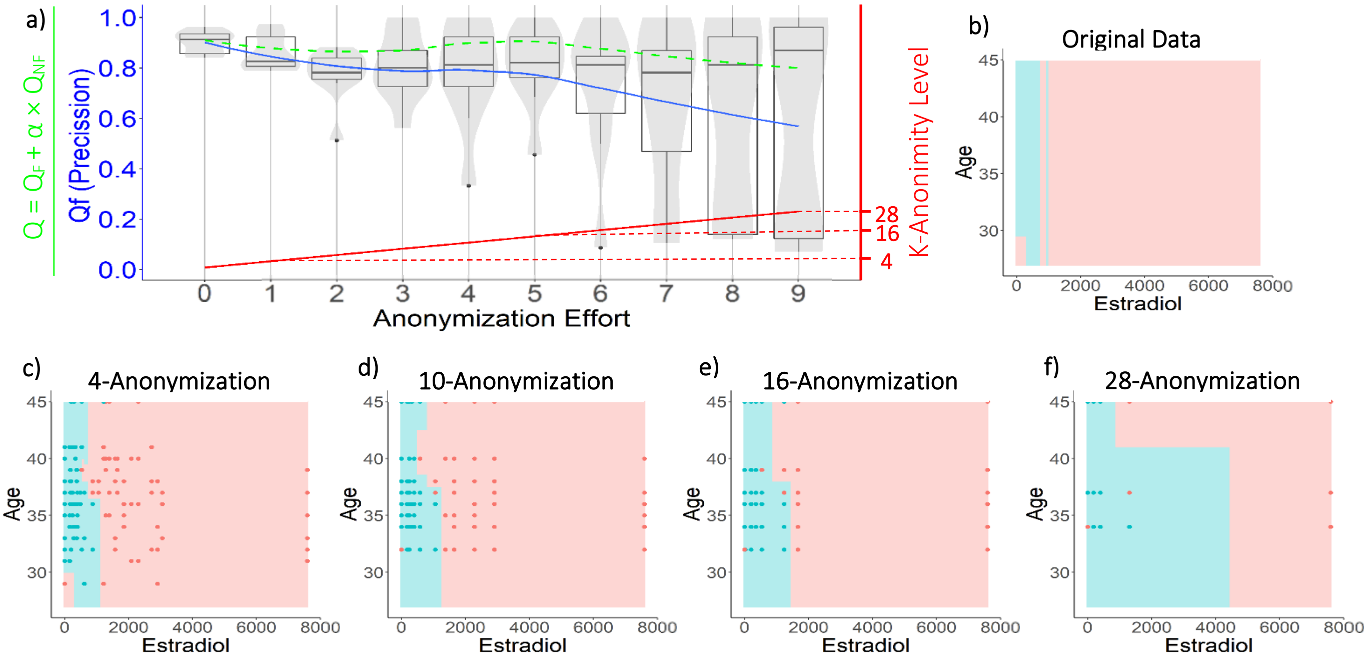

Fig. 1 Experimentation results