Electronic health record from a lab information system

EMMA SLOOT, University of Twente, The Netherlands

Electronic health records in a lab information system contain data that is privacy sensitive. Currently, it is not clear how this data is shared throughout the event of a lab test. This research focuses on reviewing the process of a blood test from start to finish. This is done through synthetic data generation and process mining. To obtain this goal, there is literature review on both synthetic data generation and privacy in healthcare systems. Secondly, a diagram to show the relations is made. Lastly, the results from the process mining are interpreted. The literature research is then compared to the diagram and results from the process mining.

Additional Key Words and Phrases: Lab information system, privacy evaluation, synthetic data generation, process mining

1 INTRODUCTION

Data science in general is important for the field of healthcare. The book of Consoli et al discusses the importance of data science within healthcare and the risks that the current influx of available data can bring. Healthcare workers might get overwhelmed and spend more time with the data than patients. Consoli et al states that the solution to this issue can be found in artificial intelligence and data science[7]. An example of how data science can help giving insights in healthcare and usefulness of treatments can be found in Sloot et al.[15].

Within data science, a point of focus is privacy. Privacy of individuals in Europe is protected through the GDPR[1]. The GDPR does offer options for data usage for scientific purposes, as can be found in article 89. However, as van der Aarst et al.[16] mentions in their paper, data science should be used responsibly. Where possible, the privacy of individuals should be maintained. This goes beyond direct data protection through the creation of synthetic data. The processes themselves need privacy protection. Healthcare processes often can be traced back to the original person through metadata. There are steps taken to anonymise the data of these processes, but there are still challenges in this field, as outlined in Pika et al[14].

In this paper, the privacy of a electronic health record from a lab information system will be evaluated through a synthetic data set. Lab information systems are an essential part of healthcare that handle sensitive data. The scope of this research will be limited to the process surrounding a blood test. It will follow the path of a general practitioner (GP) requesting a blood test for their patient, the blood being taken, and the analysis of the test tube.

2 PROBLEM STATEMENT

There has been previous research regarding synthetic data production in healthcare[6, 8, 17]. However, this data has mostly been focused on improving procedures within healthcare. One of the other issues within a healthcare system is data retention. What part of the information gets sent to different parties? Who knows which part of the data? Evaluating the privacy of data within healthcare, specifically a lab information system, is the goal of this paper. This goal can be summarized in the main research question, namely:

- **RQ** In what way can the privacy of an electronic health record from a lab information system be evaluated?

To achieve the goal, the main research question is supported by four secondary research questions that should guide the process

2.1 Secondary research questions

- **SQ1** What current methodologies and procedures exist for data generation, especially in the healthcare domain?
- **SQ2** What are the relations within lab research regarding blood work?
- **SQ3** How do the aforementioned relations within lab research translate to a data set?
- **SQ4** What is the contribution of the generated mimic data set in relation to privacy evaluation of a lab information system?

The secondary research questions have different contributions to the research. SQ1 focuses on the theoretical background and previous research as to create a basis for SQ2, which results in an Entity Relations Diagram. Based on this ERD, the answer to SQ3 is formed. It will follow the data structure of the ERD from SQ2. Lastly, SQ4 reflects on the process and what its impact is. SQ4 and SQ1 together should form the basis for the answer to the main research question.

3 METHODS OF RESEARCH

3.1 Literature review

To answer SQ1, literature review will be the main source of information. It will be less relevant for the other questions, but still used to an extent to improve the generated results. For the literature review, scholar.google.com will be used to create a basis of knowledge for further research through Scopus. The research will be done based on multiple keywords. For SQ1, these keywords are: *"Synthetic data generation", "Data production" , "Electronic health record"*.. These keywords will be combined with *"Healthcare", "Lab information systems"* to obtain results within the scope of this project. For the other secondary questions the literature review is mostly to obtain a clear view of how the processes regarding a lab information system work. Therefore, the keywords are: *"Lab information systems", "Blood work", "Lab research", "privacy evaluation"*. These keywords are used in varying combinations to obtain a variety of results.

The results of this literature research will only be used if they are peer reviewed. Next to this, the results of research regarding details of the lab information system itself will mostly be used to create a better understanding of the system itself. Its relevance for the research apart from this is very limited.

TScIT 37, July 8, 2022, Enschede, The Netherlands

3.2 Data generation

The data generation required for this research will be divided in two big elements. Firstly, an ERD will be created to review the structure that the data should be following. After this, a synthetic data set will be generated with Python.

3.2.1 ERD. To create the ERD, Visual Paradigm will be used. Next to literature research regarding the workings of a lab information system, an unpublished diagram by S.A. Sohail will be the basis of the ERD. Due to the importance of the ERD in the general process of the research, it will be discussed and reviewed with others multiple times.

3.2.2 Synthetic data generation. To create a synthetic data set, Python will be used. Each entity in the ERD will be created as a separate class. Then, these entities will be added to a list. These lists will differ from length depending on the entity. There are some relations that will always be the same. These are the relationships that in the ERD are either "One to One" or "One to many". To preserve these relationships, the necessary data will first be first matched in new lists, and then different items in these lists will be randomly picked for a final result. The goal is to create approximately 1500 iterations of the ERD. This will be done by reiterating the previously described process 1500 times.

3.3 Process mining

To get insights in what the added value of the generated data is, process mining through ProM will be used. Through process mining, the data set generated for SQ3 will be evaluated. The process mining will result in multiple models showing the complexity of the data, or lack thereof. Next to this, they will review the data retention through the process.

3.4 Comparison

To review the impact of the generated data set, it will be compared to the literature review done for SQ1. These two items should give insights that can answer SQ4. Next to this, the evaluation of the data from SQ2 and SQ3 should give insights that can help with the final conclusion for this research

4 RELATED WORK

The related works have two main sub-divisions. Firstly, research regarding data generation in healthcare in general. This research is used for a better insight in how to generate data for healthcare purposes. The second part of this theoretical background is aimed towards privacy evaluation in healthcare. This part of the related works will be used to evaluate the practical research.

4.1 Data generation in healthcare

In H.-M et al.[9], the multiple-channel latent Dirichlet allocation is used to model healthcare data regarding diagnosis, medication, and contextual information. This research originates from 2016 and shows success rates for predictive medicine around 50-60 percent, which were outstanding for that time, but shows room to improve. The literature research in Pant et al.[12] shows multiple articles regarding modelling in healthcare data. These articles refer to models

Fig. 1. A visualisation of the dataset through ProM

with a higher accuracy level than the previously mentioned Dirichlet allocation, for example Mashete et al.[10] shows a 99% accuracy rate for predictions of heart attacks. Although the accuracy is higher, the scope of these projects is very limited.
There are multiple systems within healthcare to generate synthetic data. Synsys [8], for example, is a machine learning-based synthetic data generation method. It generates realistic sequences of data with timestamps that fit reality. Synsys strives to create large amounts of data for easier anomalies detection in a health environment. Next to Synsys, Synthea[17] is a system that creates fake health dossiers of patients. The system runs, but shows inaccuracies in the prevalence of different diseases. There has been research in using AI to create deep generative models. The article by Chen et al.[6] also refers to the privacy issues that deep fakes have on healthcare, as "algorithms for the generation of deep-fakes can also be used to potentially impersonate patients and to exploit PHI, to falsely bill health insurers relying on imaging data for the approval of insurance claims".
The paper of Bhanot et al.[4] focuses on the fairness of synthetic data generation and highlights issues with regards to the representation of minorities. It shows multiple models that can be used to highlight the different possibilities for data generation.

4.2 Privacy evaluation in healthcare

Privacy in healthcare systems has been researched extensively before and multiple frameworks to ensure privacy for patients have been set up, with varying degrees of success and strictness[3]. Next to this, big data has evolved to a level where data flows in from multiple sources, giving opportunities to link information together that breaches the privacy of a patient[13]. Multiple options for improvements in privacy are being practiced, including encryption, data masking, de-identification, and access control[2].

5 DATA REVIEW

5.1 Entity Relationship Diagram

The Entity Relationship Diagram can be found in Appendix A. This diagram shows the flow of the process regarding a test tube. The diagram starts at a contact between the patient and GP. After this contact, the GP will request a test at the biolab, which will be taken by an employee (tester) of the biolab. As soon as the test is taken, the test tube is created. The content of the tube will be analysed by another employee of the biolab (analyzer). This analyzer will generate a test result. The test colour is a separate entity, as to simplify the code that needs to be written. Next to this, the biolab has multiple machines to use on the test tubes, but these are not used in the process as the machines have no impact on the privacy evaluation. They are referenced in the ERD, but not used further in the data set.

Table 1. Impact of different ProM plugins

ProM plugin	Fitness	Precision
None	1	1
Anonymise event log	0.6439	1
Add noise (20%)	1	0.7923
Reverse log	1	0.6877
Filter event log	1	0.8333

5.2 Process mining

Process mining tool ProM was used to evaluate the data generated. One of the ways the data was evaluated was checking the health of data after adding noise to preserve privacy[11]. The general flow of the process according to ProM can be found in figure 1. The different plugins that were used and their influence on the fitness and/or precision of the data can be found in table 1. To check how to interpret the fitness of data, the paper of Buijs et al. proved incredibly useful[5].

As table 1 shows, the data set without any plugin has both a fitness and precision of 1. Therefore, the event log has a good ability to be replayed, and is not underfitting. What is shown is that anonymising the event log gives less fitness, but the precision stays the same. For all other changes, the fitness remains 1, whereas the precision lowers significantly. The anonymous log creates a difficulty for the replay of the event log. The addition of noise, reversing the log, and filtering the events lessens the precision. As these plugins interfere with the data set as a whole, rather than the log itself, it is logical that the impact is mostly in the non-visible parts of the event log. The ProM plugins generally see the data as healthy. The data quality check gives an average of 9.4, with only consistency being below 10.

6 DISCUSSION

6.1 Limitations

Due to the scope of the research, the data set and ERD were limited in its complexity. For example, the first step may not always be that a GP requests a blood test, but it could be that the GP gets a request from another healthcare party like a psychiatrist for a blood test for the patient. There would already be more data sharing amongst these parties than between the current parties. Anothe rexample is that the GP currently sends the request, but never gets the feedback of the result. This is to avoid loops in the diagram, but it is not realistic. The research focused more on a simplification of reality due to time constraints, but this harmed the outcome severely. Within ProM, most process mining attempts either did not compute, or gave an overview confirming the simplicity of the data. These overviews had little added value to the research, which is why only one of these models is added to the paper, as figure 1. Lastly, a limitation was found in the knowledge regarding healthcare. Although there were multiple attempts at researching a detailed description of the process that this paper researched, most descriptions were either lacking, conflicting with others, or both.

6.2 Future recommendations

For future research on this subject, expanding the steps and details within the process is recommended. By adding extra steps and sacrificing simplicity for realism, more accurate research can be done, which would greatly benefit the general research.

Another recommendation would be to research the loops within the system more detailed. The current diagram explicitly excludes any loops, but in doing so, it strays away from the reality.

Lastly, it would be interesting to have a healthcare professional included in future research. They could provide insights in which knowledge is actually accessible. Next to this, they could prove to be a help in finding proper research regarding the topic.

7 CONCLUSIONS

The lab information system of which a privacy evaluation was done shows a decent data retention when privacy enhancing process mining is applied. However, due to the influence that anonymising the log has on the fitness of the data, it is clear that the process is reliant on knowing personal data to be able to execute the steps. When looking at the process through data analysis, the same conclusion is reached. If the results of the blood work cannot be matched back to a patient, the blood work is not relevant in this system anymore. There should be at least one actor that knows which test tube belongs to which person. In this case, the actor is the GP. The privacy evaluation of an electronic health record from a lab information system can be done through synthetic data generation and process mining. However, the manual analysis of the process gives insights that do not show through process mining.

ACKNOWLEDGMENTS

I would like to thank both S.A. Sohail and F.A. Bukhsh for their guidance during this research project. They went above and beyond to answer any questions I had and support me in resolving issues. Next to this, I would like to thank Valeria Pinus for her insights.

REFERENCES

[1] 2019. General Data Protection Regulation (GDPR). https://eur-lex.europa.eu/legal-content/NL/TXT/?uri=CELEX%3A32016R0679
[2] Karim Abouelmehdi, Abderrahim Beni-Hessane, and Hayat Khaloufi. 2018. Big healthcare data: preserving security and privacy. *Journal of big data* 5, 1 (2018), 1–18.
[3] Sasikanth Avancha, Amit Baxi, and David Kotz. 2012. Privacy in mobile technology for personal healthcare. *ACM Computing Surveys (CSUR)* 45, 1 (2012), 1–54.
[4] Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. 2021. The problem of fairness in synthetic healthcare data. *Entropy* 23, 9 (2021), 1165.
[5] Joos C. A. M. Buijs, Boudewijn F. van Dongen, and Wil M. P. van der Aalst. 2012. On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery. In *On the Move to Meaningful Internet Systems: OTM 2012*, Robert Meersman, Hervé Panetto, Tharam Dillon, Stefanie Rinderle-Ma, Peter Dadam, Xiaofang Zhou, Siani Pearson, Alois Ferscha, Sonia Bergamaschi, and Isabel F. Cruz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 305–322.
[6] Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F. Williamson, and Faisal Mahmood. 2021. Synthetic data in Machine Learning for medicine and Healthcare. *Nature Biomedical Engineering* 5, 6 (2021), 493–497. https://doi.org/10.1038/s41551-021-00751-8
[7] Sergio Consoli, D Reforgiato Recupero, and Milan Petkovic. 2019. *Data science for healthcare*. Springer.
[8] Jessamyn Dahmen and Diane Cook. 2019. SynSys: A synthetic data generation system for healthcare applications. *Sensors* 19, 5 (2019), 1181.
[9] Hsin-Min Lu, Chih-Ping Wei, and Fei-Yuan Hsiao. 2016. Modeling Healthcare data using multiple-channel latent Dirichlet allocation. https://www.sciencedirect.

com/science/article/pii/S1532046416000253
[10] Hlaudi Daniel Masethe and Mosima Anna Masethe. 2014. Prediction of heart disease using classification algorithms. In *Proceedings of the world Congress on Engineering and computer Science*, Vol. 2. 25–29.
[11] Kato Mivule. 2013. Utilizing noise addition for Data Privacy, an overview. https://arxiv.org/abs/1309.3958v1
[12] Diva Pant, Vishal Kumar, Jaydeep Kishore, and Ritu Pal. 2017. Healthcare data modeling in R. In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*. 230–233. https://doi.org/10.1109/ICISIM.2017.8122178
[13] Harsh Kupwade Patil and Ravi Seshadri. 2014. Big data security and privacy issues in healthcare. In *2014 IEEE international congress on big data*. IEEE, 762–765.
[14] Anastasiia Pika, Moe T Wynn, Stephanus Budiono, Arthur HM Ter Hofstede, Wil MP van der Aalst, and Hajo A Reijers. 2020. Privacy-preserving process mining in healthcare. *International journal of environmental research and public health* 17, 5 (2020), 1612.
[15] Sarah Sloot, Yian A. Chen, Xiuhua Zhao, Jamie L. Weber, Jacob J. Benedict, James J. Mulé, Keiran S. Smalley, Jeffrey S. Weber, Jonathan S. Zager, Peter A. Forsyth, and et al. 2017. Improved survival of patients with melanoma brain metastases in the era of targeted Braf and immune checkpoint therapies. *Cancer* 124, 2 (2017), 297–305. https://doi.org/10.1002/cncr.30946
[16] Wil MP van der Aalst. 2016. Responsible data science: using event data in a "people friendly" manner. In *International Conference on Enterprise Information Systems*. Springer, 3–28.
[17] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, Scott McLachlan, and et al. 2017. Synthea: An approach, method, and software mechanism for generating synthetic patients and the Synthetic Electronic Health Care Record. *Journal of the American Medical Informatics Association* 25, 3 (2017), 230–238. https://doi.org/10.1093/jamia/ocx079

A ENTITY RELATION DIAGRAM

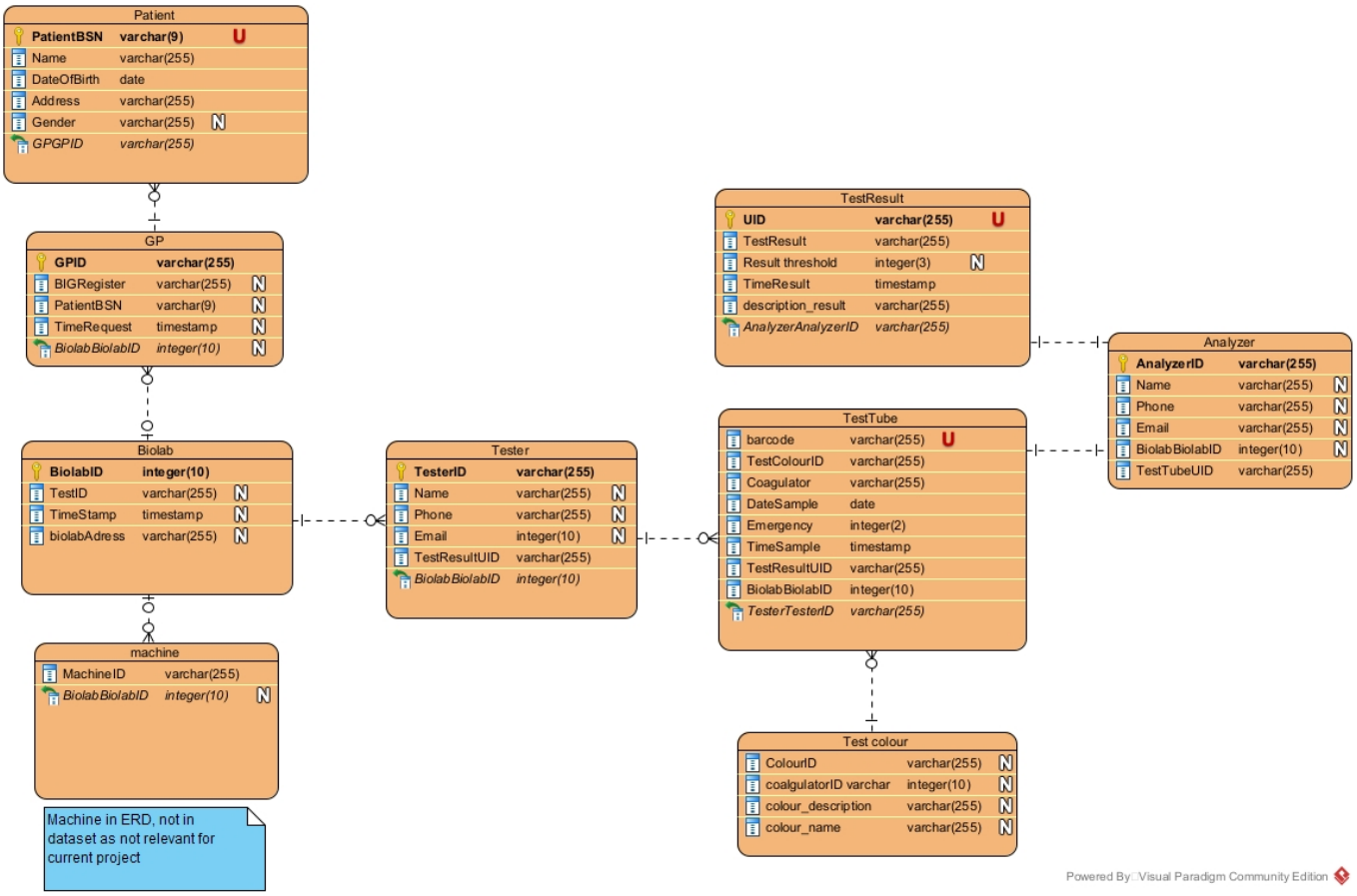

Fig. 2. Entity relationship diagram of a lab information system