German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210051

Generating Enriched Synthetic German Hospital Claims Data – A Use Case Driven Approach

Sven HELFER[a,1] Michéle KÜMMEL[a], Franziska BATHELT[a], and Martin SEDLMAYR[a]
[a]*Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany*

Abstract. Clinical data and above all individual patient data are highly sensitive. All the more it is important to protect these critical information while analyzing and exploring their specifics for further research. However, in order to enable students and other researchers to develop decision support systems and to use modern data analysis methods such as intelligent pattern recognition, the provision of clinical data is essential. In order to allow this while completely protecting the privacy of a patient, we present a mixed approach to generate semantically and clinically realistic data: (1) We use available synthetic data, extract information on patient visits and diagnoses and adapt them to the encoding systems of German claims data; (2) based on a statistical analysis of real German hospital data, we identify distributions of procedures, laboratory data and other measurements and transfer them to the synthetic patient's visits and diagnoses in a semi-automated way. This enabled us to provide students a data set that is as semantically and clinically realistic as possible to apply patient-level prediction algorithms within the development of clinical decision support systems without putting patient data at any risk.

Keywords. Synthetic Data Generation, German Claims Data, Education, Data science, Privacy, Medical Informatics Applications

1. Introduction

Concepts like artificial intelligence, big data analyses, and decision support systems show high potential to improve health care and thereby patients' health, but also come with major risks regarding privacy and security of patient data. In this paper, we present an approach for generating enriched synthetic German claims data that can be used in educational and research settings while completely protecting real patient data.

1.1. Background

The last decades of digitalization have brought a large amount of data to the health care sector. With the risks of this accumulation of mostly highly sensitive data becoming

[1] Corresponding Author, Sven Helfer, Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany E-mail: sven.helfer@tu-dresden.de.

clearer, privacy law in many regions aggravated barriers and responsibilities for data processing [1,2]. Unfortunately, these barriers may also affect innovation by restricting access to data that is highly needed for education and the development of new technologies. Given the current situation, there is a need for anonymous health data. The generation of synthetic health data is a promising way to deliver realistic data with no re-identification risk [3]. Furthermore, synthetic data could be adopted to any use case necessary with minimal effort. Thereby it could support the innovation of new health care products by delivering a privacy compatible way to develop data-driven technology in the health sector and by enabling teachers to let their students learn with realistic data.

1.2. Requirements

The primary focus of our project was to create data to share with our students without any risk of disclosing private patient data. We are conducting a course on practical medical informatics at the Technical University Dresden focusing on secondary health data use to develop patient-level prediction algorithms. To reach that goal it is necessary to create data that is semantically and clinically realistic. In this context, we understand semantically realistic data as data sets that (1) mimic the clinical parts of German claims data (as defined by §301 SGB V and § 21 KHEntGG [4]), (2) have data formats that one would find in real data sets, and (3) have missing values and other minor issues as one would expect in real-world data. On the other hand, clinically realistic data include data that (1) mimic real patient stays, (2) portray realistic frequencies of diagnoses and procedures, (3) show similar correlations between diagnoses and procedures as real data, and (4) have realistic distributions of measurements and laboratory values. Finally, our solution should be released under an open-source license.

2. State of the art

Several projects provide a solution to data generation [5–14]. All solutions use different approaches to the generated data or the generation mechanism. Some of the tools use simple pseudo-random sampling mechanisms [6], other use machine learning algorithms [8] and some are based on complex state-machine-like models [5,10]. An overview of tools, methods, and projects is available in [5] and [15]. Most projects use a single input table to generate a single output table and often these tables have to be strictly formatted. This reduces the variety of the generated data. It also increases the amount of pre- and post-processing necessary to generate data sets from more complex sources and to achieve results in the needed format. Finally, these approaches are often limited when generating more-dimensional data like data sets including multiple hospital stays with information on multiple measurements, procedures or diagnoses. In addition, many of these projects are in development stage and documentation is limited.
To our knowledge, none of the existing projects provides an accessible way to produce data as complex and versatile as claims data, let alone in a format resembling German claims data.

3. Concept

We developed a process that combines freely available US American patient data with distributions taken from real German patient data and enrich them with randomly defined laboratory results. An overview of this process is shown in Figure 1.

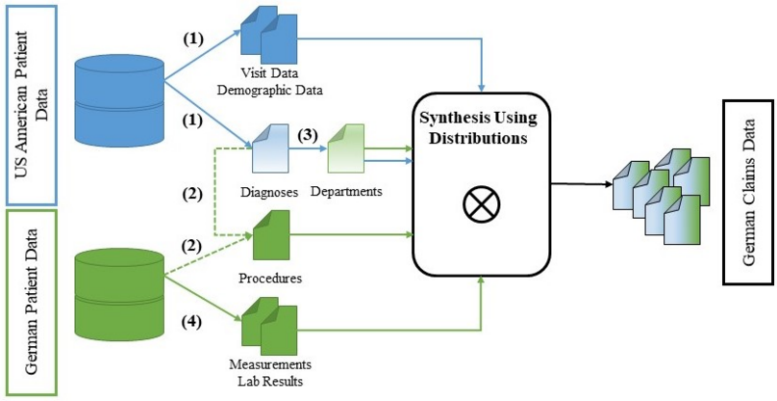

Figure 1: Overview of the Data Generation Process – (1) Extraction of OMOP SynPUF 1k Data; (2) Generating Procedure Data; (3) Assignment of corresponding specialty departments; (4) Creation of Clinical Measurements and Laboratory Results

3.1. Extraction of OMOP SynPUF Data ("FALL", "ICD")

To generate data from patients with multiple stays and realistic diagnosis data, we extracted the information given in the OHDSI OMOP Common Data Model (OMOP CDM) [16] "SynPUF 1k" data set which is based on anonymized American claims data [17]. We used an automated "extraction, transformation, load" (ETL) process to create two tables containing information on hospital visits ("FALL") and diagnoses ("ICD"). We chose this process because we had already used an ETL process to transform German claims data to OMOP CDM. The main reason not to generate this data from scratch was the complexity that different hospital stays with multiple differences e.g. in the length of stay would have posed to a generation process. We did not find this longitudinal aspect of health care data covered very well in any of the available tools. We added any data missing in the "FALL" and "ICD" tables after extraction. A medical expert (SH) determined the distributions of these variables.

3.2. Generating Procedure Data ("OPS")

To generate data on medical procedures, we used the primary diagnosis as a determinant for possible procedures. We therefore used the distributions of procedures for each primary diagnosis based on the claims dataset of our university hospital. We then generated the procedure data according to these distributions and developed a transformation process to create data conformant to §21 procedure data and linked it to the corresponding demographic and encounter data already created. This process seems the most approachable way to create data with dependencies. It does not require special machine learning techniques and can be implemented using standard tools in data

science. Additionally, we were able to remove combinations occurring less than a given threshold to further reduce any remaining re-identification risk.

3.3. Assignment of corresponding specialty departments ("FAB")

To allow a realistic assignment of patient cases to treating departments, a reference table was defined semi manually by a medical expert (SH) that correlates departments and primary diagnoses. Using this reference table each encounter was enriched with the corresponding department identifier. We then used an automated ETL process to extract the corresponding data table ("FAB"). Although possible, we decided not to use the same approach to create department data than procedure data. For our use case the information about the treating departments was less relevant so we decided to define mappings of main diagnoses to departments to decrease implementation time and increase generation speed.

3.4. Creation of Clinical Measurements and Laboratory Results

To further enrich the patient data created, we added laboratory results as well as body height and weight measurements. Latter were generated using German body mass index (BMI) distributions. For each laboratory parameter, a medical expert defined the properties necessary to randomly create result values. Patients were randomly sampled to have these parameters determined and then random "result" values were sampled from normal distributions defined by the properties given to each laboratory parameter. Laboratory results and measurements are in general not part of German hospital claims data but were necessary for our use case. The distribution of BMI values in Germany is freely available [19] and was sufficient for our use case. For laboratory data, similar data could have been gathered in different sources but without any additional value. We decided not to have distributions of all laboratory results for all main diagnoses created in our data integration center because the immense effort did not seem worth the benefits.

4. Implementation

4.1. Extraction of OMOP SynPUF Data ("FALL", "ICD")

We were able to reverse the ETL process, which we had already designed and implemented to load German claims data into the OMOP CDM [19] using semantic mappings of German patient data to standardized OMOP Concepts. The ETL process was designed with the help of Pentaho Data Integration [20]. We generated values in columns not extracted by the ETL process using a python script. For our use case, most variables could be set to constants or sampled from simple distributions. The only parameter that we needed to be distributed in a more complex way was duration of ventilation. Here we used age as a factor to determine if a ventilation was administered and length of stay as a factor to determine the duration of ventilation with some random noise. Table 1 shows key characteristics of the two source data sets and the generated value set.

Table 1: Number of entries in the US DE-SynPUF, German Claims (§21) and Synthetic Data Set.

Number of entries	SynDat	Claims Data	Synthetic Data
Patients	842	~50,000	842
Cases	47,457	~70,000	47,457
Diagnoses	99,622	~300,000	99,622
Procedures	0	~400,000	5,149
Laboratory	0	0	169,337

4.2. Generating Procedure Data ("OPS")

We received relative frequencies of procedures occurring with all primary diagnoses from our data integration center (limited to combinations with 5 or more occurrences for privacy reasons). We used a script written in python to pivot the table with each row containing all relative frequencies of all procedure codes for one diagnosis code. Figure 2 shows relative frequencies of procedures for a given primary diagnosis exemplary for the 100 most common combinations. We merged this table with the "ICD" table and then used the relative frequencies as probability to sample if the given procedure occurred. After unpivoting the resulting table, we added timestamps within the duration of stay and added the missing columns using data from "FALL". We set some columns to constants that were not necessary for our use case.

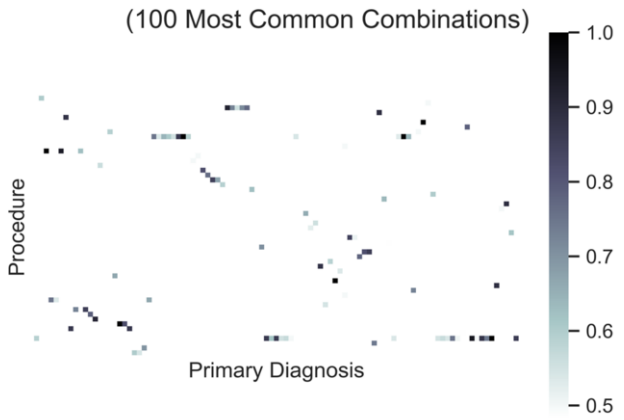

Figure 2: Heatmap of procedure codes by primary diagnosis. Each pixel represents the relative frequency of a procedure given a primary diagnosis.

4.3. Assignment of corresponding specialty departments ("FAB")

The department file "FAB" closely resembles the case file "FALL" so most columns were copied from there, merging the main diagnosis information from "ICD". Afterwards the manually created mapping table was used to replace diagnosis codes with corresponding department ("FAB-") codes.

4.4. Creation of Clinical Measurements and Laboratory Results

To create body height and weight, BMI was sampled from the publicly available data. We generated body height randomly sampling a normal distribution centered at 170 cm and calculated the body weight that would result in the body mass index generated before. (Figure 3, left) We generated laboratory values in a two-step process. First, we randomly sampled binary values for each case corresponding to blood count, liver and kidney parameters and CRP being determined in the respective case. These values were estimated by our medical expert (SH). In the second step, the corresponding laboratory results were randomly sampled from normal distributions. (Figure 3, right)

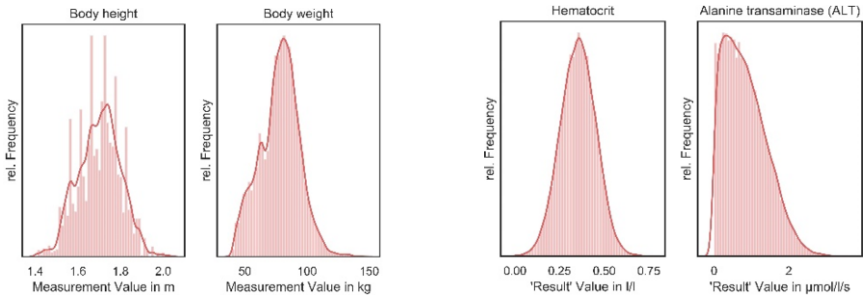

Figure 3: left: Exemplary synthetic measurements distribution; right: Laboratory result distribution

5. Lessons learned (Discussion)

During the implementation of our data generation process we had several insights. First and foremost, we realized that the generation of synthetic health data is not a trivial task and most of the tools available do not fulfill the requirements that we anticipate in our use cases. To create data as realistic as possible, we planned to use real patient data to determine all distributions of major interest to our use case. We were able to enrich American anonymized data with German procedure data. At this point, we had met almost all requirements stated above (i.e. requirements necessary for our use case) but we needed more resources than planned. Therefore, during implementation, we decided to omit the consideration of laboratory and measurement distributions. Moreover, as these parameters are highly dependent on very complex patient characteristics. Thus, a discrete concept is necessary to identify important parameters and synthesize fully realistic laboratory data and measurements while keeping calculation time reasonable. This came at the cost of our students not being able to use machine learning algorithms that would deliver realistic results (which was not a primary goal of our course) but only results that corresponded to the realism of our data. To use our approach in a research context focused on German characteristics, we aim to substitute the use of US American patient data by realistic data that is synthesized using distributions of German encounter data. Additionally, we aim to reduce manual effort and the need for medical expertise while becoming more realistic regarding e.g. laboratory data.

One major drawback of our approach is the probable loss of information included in multi-dimensional patterns. In our approach we used the combination of time-dimensional data from the SynPUF data set and combined it with distributions from

German claims data. Most other data was derived randomly or using more simple distributions.

6. Conclusion

With our use case driven approach based on US American and German patient data we were able to synthesize sufficiently realistic hospital claims data within a relatively short period. This data set is useful in an educational context without putting patient privacy at any risk. Despite the given limitations and need for future work, we are very confident that we will be able to provide unrestricted access to realistic German patient data in the near future that can then be used in education and development e.g. of applications of artificial intelligence in medicine. We believe that our use case driven approach enables us to bridge the gap between solutions based on predefined models, calculated distributions and random sampling.

Conflict of Interest

The authors state that they have no conflict of interests.

References

[1] 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 (General Data Protection Regulation) (Text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed March 21, 2020).

[2] Office for Civil Rights (OCR), Summary of the HIPAA Security Rule, *HHS.Gov*. (2009). https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html (accessed March 21, 2020).

[3] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K.P. Bennett, Privacy Preserving Synthetic Health Data, *ESANN 2019 - European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. (2019). https://hal.inria.fr/hal-02160496 (accessed October 16, 2019).

[4] InEK GmbH, Data Set Description, (n.d.). https://www.g-drg.de/Datenlieferung_gem._21_KHEntgG/Dokumente_zur_Datenlieferung/Datensatzbeschreibung (accessed March 21, 2020).

[5] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan, 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 (2018) 230–238. doi:10.1093/jamia/ocx079.

[6] U. Kartoun, A Methodology to Generate Virtual Patient Repositories, (2016). http://arxiv.org/abs/1608.00570 (accessed March 21, 2020).

[7] T. Inbar, and E.J. Dann, Preoperative Anemia and Blood Transfusion Requirement during Hip Surgery: Synthetic and Real Patient Cohort Data, *Blood*. 134 (2019) 3693–3693. doi:10.1182/blood-2019-125252.

[8] E. Choi, S. Biswal, B. Malin, J. Duke, W.F. Stewart, and J. Sun, Generating Multi-label Discrete Patient Records using Generative Adversarial Networks, (2018). http://arxiv.org/abs/1703.06490 (accessed November 4, 2019).

[9] M. Mannino, and A. Abouzied, Is This Real?: Generating Synthetic Data That Looks Real, in: Proceedings of the 32Nd Annual ACM Symposium on User Interface Software and Technology, ACM, New York, NY, USA, 2019: pp. 549–561. doi:10.1145/3332165.3347866.

[10] H. Ping, J. Stoyanovich, and B. Howe, DataSynthesizer: Privacy-Preserving Synthetic Datasets, in: Proceedings of the 29th International Conference on Scientific and Statistical Database Management - SSDBM '17, ACM Press, Chicago, IL, USA, 2017: pp. 1–5. doi:10.1145/3085504.3091117.

[11] N. Patki, R. Wedge, and K. Veeramachaneni, The Synthetic Data Vault, in: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016: pp. 399–410. doi:10.1109/DSAA.2016.49.

[12] MDClone, *MDClone*. (n.d.). https://www.mdclone.com (accessed March 21, 2020).

[13] Mostly GENERATE, (n.d.). https://mostly.ai/mostly-generate.html (accessed March 21, 2020).

[14] FINRAOS/DataGenerator. https://github.com/FINRAOS/DataGenerator (accessed March 21, 2020).

[15] H. Surendra, and H. Mohan, A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing, (2019). https://www.ijstr.org/final-print/mar2017/A-Review-Of-Synthetic-Data-Generation-Methods-For-Privacy-Preserving-Data-Publishing.pdf (accessed November 4, 2019).

[16] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, and P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers, *Studies in Health Technology and Informatics*. 216 (2015) 574–578.

[17] Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF), (2013) 5.

[18] Statistisches Bundesamt Deutschland - GENESIS-Online, (2020). https://www-genesis.destatis.de/genesis/online (accessed March 21, 2020).

[19] M. Kümmel, I. Reinecke, M. Gruhl, F. Bathelt, and M. Sedlmayr, Transition Databse for a harmonized mapping of German patient data to the OMOP CDM, in: 2020 OHDSI European Symposium, 2020. (accepted March, 2020)

[20] M. Carina Roldán, Pentaho Data Integration Beginner's Guide, Packt Publishing, 2013. http://www.myilibrary.com?id=537754 (accessed March 13, 2020).