Improving Access to German Health Claims Data through Synthetic Data Generation: A methodological comparison of different approaches

RWD196

Tobias Heidler [1], Michael Schultze [2], George Kafatos [3], Bagmeet Behera [4], Caroline Lienau [5], Alexander Unger [5], Valentina Balko [5], Julius Brandenburg [5], Zhenchen Wang [6], Philipp Großer [7], Adam Hilbert [8], Nils Kossack [1], Marc Pignot [2]

1 WIG2 GmbH, Leipzig, Germany
2 ZEG Berlin GmbH, Berlin, Germany
3 Amgen Limited, Uxbridge, UK
4 Amgen Research (Munich) GmbH, Munich, Germany
5 AstraZeneca GmbH, Hamburg, Germany
6 Medicines and Healthcare products Regulatory Agency (MHRA), London, UK
7 Limebit GmbH, Berlin, Germany
8 ai4medicine UG, Berlin, Germany

Introduction

- Despite ongoing demand for high-quality, representative data for scientific research, privacy concerns limit access to health data, including claims data [1,2].
- Synthetic data is data created artificially and mirrors the statistical characteristics of real data. Typically, a model is trained on actual data to produce a new dataset that captures the key features of the original data.
- Synthetic data generation presents a promising solution to improve access. Unlike anonymized or de-identified data, which still contain inherent risks of re-identification, synthetic data offers an enhanced level of privacy-protection [2].
- For this study, a systemic lupus erythematosus (SLE) — a rare and complex autoimmune disease-— patient cohort was used for synthetic data generation training.

Objective

The aim of the study was to establish a holistic framework for the evaluation of generated synthetic claims data in four key aspects: **Privacy**, protecting sensitive information from exposure; **fidelity**, the statistical resemblance of the original data; the **robustness and scalability** to larger, heterogeneous populations and the **utility**, the practical application in different evidence generation scenarios. This framework is to be applied on generated synthetic data of SLE patients.

Methods

Data Source

- WIG2 Benchmark database: A large, longitudinal medical claims database that is a representative sample of insured patients in Germany with approximately 4 million insured individuals available from 2014 until end of 2021 [3].

Study Design & Population

- The study cohort consisted of 6 743 retrospectively identified patients with any diagnosis of SLE (ICD-10 GM code *M32*.-), including confirmed, suspected and exclusion diagnoses.
- Patient claims data, consisting of characteristics, outpatient visits and inpatient admissions, including prescriptions, surgeries and fees, was utilized as a reference to train various models to reproduce this dataset.

The following samples were used as part of the study framework:

- SLE-Sample: A general population of patients with any SLE-diagnosis used for synthetic data training and evaluation.
- RWE-cohort: Continuously insured patients with confirmed SLE-diagnosis used specifically for evaluation of RWE replication in a sample that uses common inclusion and exclusion criteria.

An overview of the study design is shown in **Figure 1**.

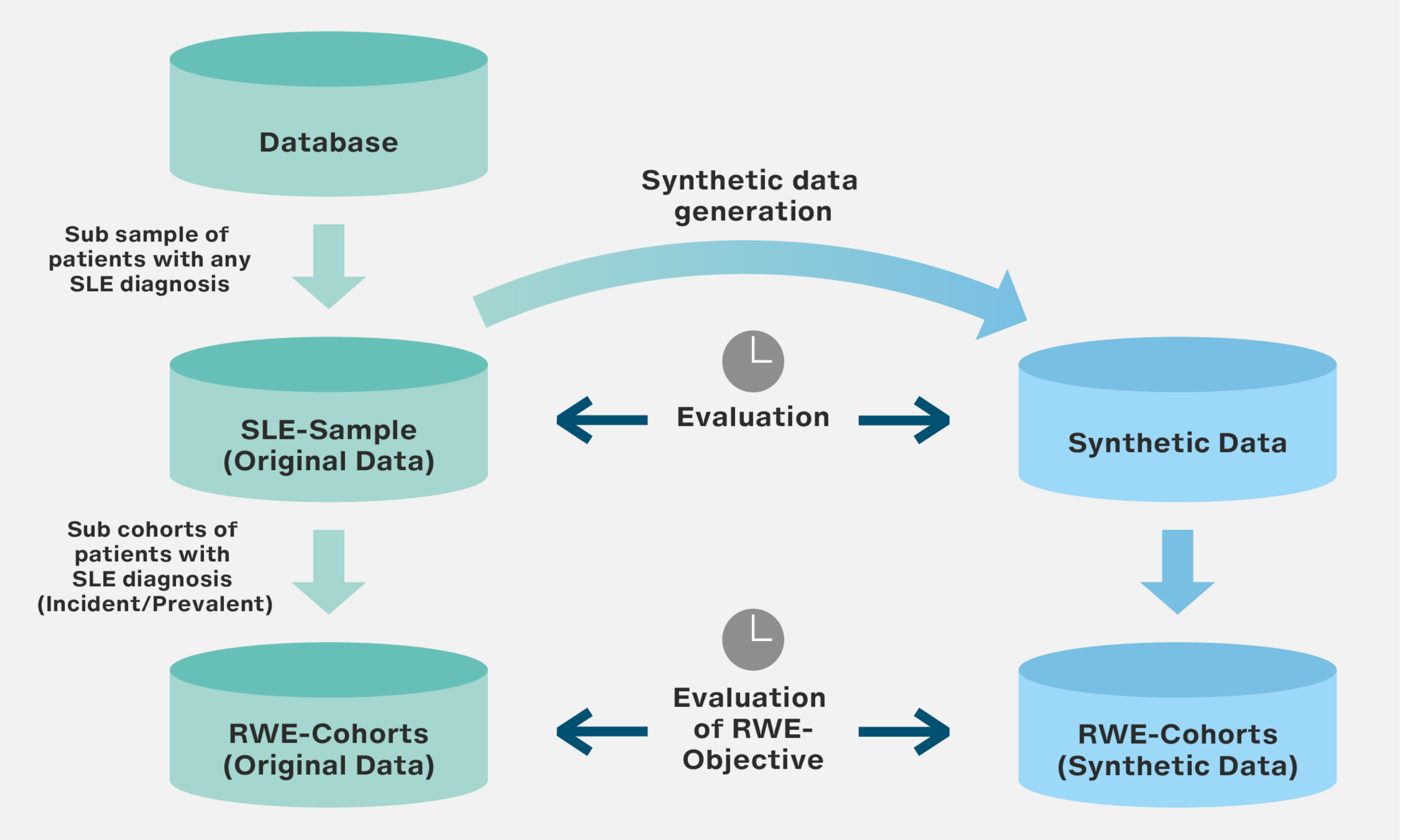
Database
Sub sample of patients with any SLE diagnosis
Synthetic data generation
SLE-Sample (Original Data)
Evaluation
Synthetic Data
Sub cohorts of patients with SLE diagnosis (Incident/Prevalent)
RWE-Cohorts (Original Data)
Evaluation of RWE-Objective
RWE-Cohorts (Synthetic Data)

Figure 1: Data Sampling and Utilization Flowchart for SLE patient data synthesis and evaluation

Synthetic data was generated using the following methods as described in **Table 1**.

Table 1: Methods used for synthetic data generation

Method	Description
Adversarial Random Forests (ARF)	A machine learning framework that combines adversarial methods with random forests refining the data generation utilizing iterative feedback loops.
Bayesian Networks (BN)	A probabilistic graphical model that represents variables and their conditional dependencies using directed acyclic graphs for generating synthetic data.
Generative Adversarial Networks (GAN)	A deep learning framework where two neural networks, a generator and a discriminator, compete to generate realistic synthetic data.

Evaluation Framework

Privacy

- A distance-based approach will be used that will not require manual selection of sensitive attributes to assess **duplicate records** and the **robustness to privacy attacks** in contrast to traditional privacy metrics. A discussion on **shareability** will be performed using these metrics as well as insights gained from fidelity and utility evaluation and synthetization method inherent privacy guarantees.
- Ideally, shareability of the synthetic data seems plausible as there are no or very little privacy concerns. A full anonymization can be assumed.

Robustness & Scalability

- A common technical setup across all methods was used for training and data generation. An overall assessment of **computational efficiency** included reports of CPU and RAM-utilization during training and data generation to extrapolate expenses to bigger, more diverse datasets. We documented and evaluated manual interventions needed for training and data synthesis to assess the **generalization** capabilities of every approach.
- Ideally, the expansion to multiple diseases or a complete health claims data set is feasible without manual intervention.

Fidelity

- This assessment will cover univariate, bivariate, and high-dimensional distributions, focusing on **distributional closeness** and **high-dimensional dependencies**. Additionally, **temporal consistency** within the synthetic data will be evaluated using metrics such as the Kolmogorov-Smirnov Test (KS-Test) and advanced techniques like Uniform Manifold Approximation and Projection for dimensionality reduction (UMAP).
- Ideally, the synthetic data exhibits a medium to high fidelity to the original data set, capturing both univariate and multivariate statistical properties as well as temporal trends.

Utility

- The focus is on determining the data's utility in enhancing technical capabilities in data analysis and scripting within the context of healthcare analytics in terms of **analysis and script development** by adhering to the technical rules of the original data and performance evaluation within several common and multi-facetted RWE-scenarios (**RWE replication**) from baseline characteristics to complex health economics and outcomes research.
- Ideally, the data facilitates the creation and enhancement of complex analytical methods and scripts, and a multitude of analyses can be performed with a reasonable closeness to the original data.

Limitations

- Evaluation of synthetic data is a complex multi-dimensional task. The balance between privacy and utility must be assessed taking the problem at hand into account.
- The synthetic data is only tested on an excerpt of possible RWE-scenarios, also limited by the applicability of the underlying population, thus the correct mimicking of other analytical tasks is not guaranteed.
- The opaque nature of models like GANs ("Blackbox models") makes it difficult to comprehend their inner mechanisms fully. This lack of transparency can be an issue a concern, especially in situations demanding clear decision-making insights and robust, high-level of privacy.

Conclusion

- There is increasing interest in the use of synthetic data to provide data insights and facilitate access.
- This study is unique as it aims to generate synthetic data using different approaches.
- The evaluation of the four key aspects of synthetic real-world data generation is pivotal in understanding strengths and weaknesses of various synthetic data generation methods.
- We established a comprehensive evaluation framework for synthetic data, a critical step towards facilitating access to German health claims data.
- **Next Steps**: Ongoing efforts aim to balance privacy with utility, ensuring scalable and practical applications for broader health research. Future work will involve presenting and discussing the synthetic data generated using this evaluation framework.

References

[1] Gill, J., Avouac, B., Duncombe, R., Hutton, J., Jahnz-Rozyk, K., Schramm, W., Spandonaro, F., Thomas, M., & Kanavos, P. (2016). The use of Real World Evidence in the European context: An analysis of key expert opinion. DOI: <https://doi.org/10.21953/LSE.68442>.

[2] Alloza, C., Knox, B., Raad, H., Aguilà, M., Coakley, C., Mohrova, Z., Boin, É., Bénard, M., Davies, J., Jacquot, E., Lecomte, C., Fabre, A., & Batech, M. (2023). A Case for Synthetic Data in Regulatory Decision-Making in Europe. Clinical pharmacology and therapeutics, 114(4), pp. 795–801. DOI: <https://doi.org/10.1002/cpt.3001>.

[3] Ständer, S., Ketz, M., Kossack, N., Akumo, D., Pignot, M., Gabriel, S., & Chavda, R. (2020). Epidemiology of Prurigo Nodularis compared with Psoriasis in Germany: A Claims Database Analysis. Acta dermato-venereologica, 100(18), pp. adv00309. DOI: <https://doi.org/10.2340/00015555-3655>.

Disclosure and Funding Statements

This study and publication was funded in equal shares by Amgen Ltd. and AstraZeneca GmbH.

T. Heidler, N. Kossack declare no conflicts of interest with respect to the research, authorship, and/or publication of this article; they are employees of WIG2 GmbH. M. Schultze and M. Pignot are employees of ZEG Berlin GmbH, which received funding from Amgen and AZ to conduct this study. G. Kafatos and B. Behera are employees of Amgen Limited, Uxbridge, UK and Amgen Research (Munich) GmbH respectively. C. Lienau, A. Unger, V. Balko and J. Brandenburg are employees of AstraZeneca GmbH. Synthetic data generation was performed by Z. Wang (MHRA), P. Großer (Limebit GmbH), A. Hilbert (ai4medicine UG). Corresponding author: T. Heidler (WIG2 GmbH), tobias.heidler@wig2.de