Improving Access to German Health Claims Data through Synthetic Data Generation: Findings and Achievements of a Holistic Evaluation

RWD100

Tobias Heidler [1], Michael Schultze [2], George Kafatos [3], Bagmeet Behera [4], Caroline Lienau [5], Alexander Unger [5], Valentina Balko [5], Julius Brandenburg [5], Lea Grotenrath [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

- Access to real-world health claims data is often restricted by stringent privacy regulations, which limits research and innovation potential in healthcare analytics [1,2].
- Synthetic data provides a promising solution by replicating the statistical properties of original datasets while protecting insurants privacy [1,2].
- Previous work has established a holistic evaluation framework to assess privacy, fidelity, scalability, and utility of synthetic datasets in health research contexts [3].
- Longitudinal relational claims data present additional challenges for synthesis due to multi-table structures, temporal dependencies, and diverse clinical coding systems.
- Systemic lupus erythematosus (SLE) is a rare, clinically heterogeneous, and resource-intensive condition, making it a suitable case study for evaluating synthetic data generation methods.

Objectives

- Apply and compare multiple synthetic data generation methods on a German longitudinal health claims dataset consisting of insurants with SLE diagnoses.
- Assess performance across the dimensions of **privacy**, **scalability & robustness**, **fidelity**, and **utility** dimensions using the previously developed evaluation framework [3].
- Investigate the feasibility of using synthetic data for both basic analytical script development and complex real-world evidence (RWE) studies, including disease prevalence, treatment pathways, and healthcare utilization analyses.
- Provide guidance on method selection and dataset tailoring to balance privacy protection with analytical fidelity in German health claims research.

Methods

Data source: WIG2 benchmark database (*DS-WIG2*) - longitudinal German health claims.
Cohort: 6,743 insurants with SLE diagnosis between 2014 - 2021 spanning 11 different tables.
Synthetic data generation methods: Generative Adversarial Networks (*DS-GAN*), Adversarial Random Forests (*DS-ARF*), and two Bayesian Network (*DS-BNN-Kaur*, *DS-BNN-WangTucker*).
Hardware: Off-the-shelf computing resources.
Evaluation metrics:

- **Privacy**: Absence of duplicates between original and synthetic insurants and resistance to re-identification attacks using Normalized Compression Distances (NCDs).
- **Scalability & Robustness**: Ability to generate complete datasets utilizing off-the-shelf hardware with minimal input data simplifications. No SLE-specific information should be incorporated manually into the synthetic data model and output data.
- **Fidelity & Utility**: Uni-, bi- and multivariate statistical alignment by, e.g., standardized mean differences (SMDs), discriminator model performance (ROC AUC), and temporal consistency (weighted R^2 of feature onset and discontinuation probabilities). Overall, suitability for basic analysis scripting and Real-World-Evidence (RWE) synthesis.

Results

Privacy

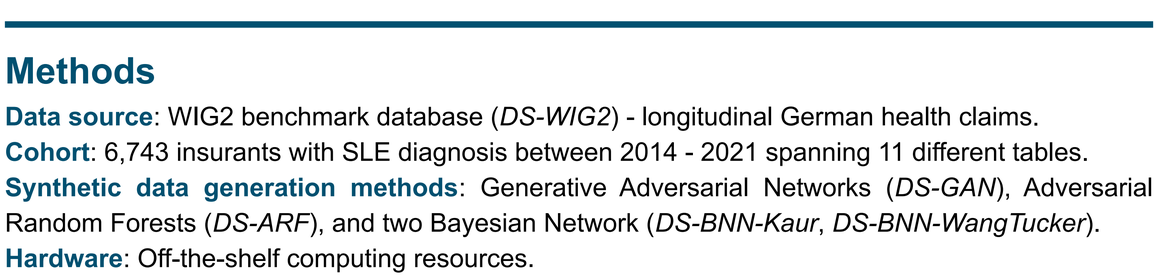

Figure 1: Median, IQR and range of NCD distances between original data insurants

High median total NCD (≥7.84, see Figure 1) were observed across all datasets compared to self-comparison (1.22), indicating low similarity between original and synthetic insurants. The total distances to the closest original insurant did not fall below 7.35 (5.52 in DS-WIG2), except for DS-ARF (3.85).

Privacy attack robustness testing showed minimal leakage risk for DS-BNN-WangTucker (98.9% of insurants without NCD <0.25 across ≥2 tables) and low rates for DS-BNN-Kaur (23.8%) and DS-GAN (22.4%). DS-ARF showed low distances across sensitive tables for 1.65% of DS-WIG2 insurants, but manual review confirmed no sensitive data overlap.

Scalability & Robustness

Model	Scaling	Training data reduction	Limited by	Training time
DS-ARF	Logarithmic	None	-	7 d
DS-BNN-Kaur	Exponential	Severe (12.8%)[1]	RAM	8 h
DS-BNN-WT	Logarithmic	None	RAM	20 h
DS-GAN	Cubic	Minimal (81.3%)[1]	GPU	54 h

Table 1: Scalability of synthetic data generation methods

[1] Percentage of insurants used for training.

DS-ARF successfully processed the full dataset but had the longest runtime. DS-BNN-Kaur exhibited severe scalability limitations due to RAM constraints. DS-BNN-WangTucker also processed the full dataset but was slowed by single-threaded execution. DS-GAN leveraged GPU acceleration and trained on slightly less than the full dataset due to the validation split, but exhibited cubic scaling complexity (see Table 1).

Fidelity & Utility

Insurant counts were well reproduced, but large discrepancies in row and unique case counts occurred, particularly in DS-GAN and DS-BNN-Kaur (2 - 798% of original), whereas DS-ARF most closely matched the training data in volume and composition. DS-BNN-WangTucker suffered major identifier linkage issues, hindering several cross-tabular evaluations. DS-ARF matched missing value rates and avoided missing or hallucinated codes, unlike DS-GAN and DS-BNN-Kaur.

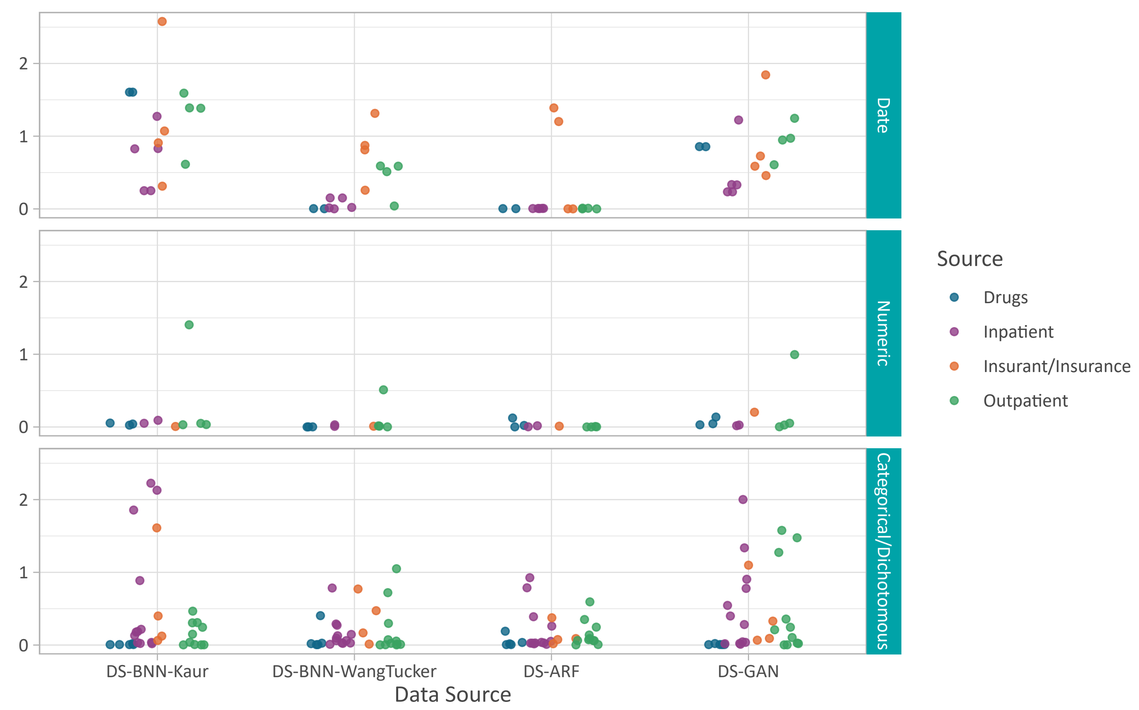

Figure 2: Absolute SMDs of synthetic dataset features compared to original data

Numerical features were generally well replicated (absolute SMDs ≤0.20); however outpatient case years showed high SMD values across three synthetic datasets, as shown in Figure 2. Underperformance in TOP-15 categorical feature replication based on original data frequencies can be seen across all datasets in various features.

Date features were well replicated in DS-BNN-ARF. Nevertheless, continuous insurance period date generation was poorly aligned across all datasets (SMD ≥ 1.38). Claims in within insured periods were comparable to DS-WIG2 in DS-BNN-Kaur, and DS-GAN, but less so in DS-BNN-WangTucker (17 - 55%) and DS-ARF (59 - 70%). Feature onset and discontinuation rates were poorly replicated. (weighted R^2 0.16 - 0.43)

Overall, nearly perfect discriminator model performance (ROC AUC ≥0.99), inability to replicate weaker associations and temporal dependencies hint that while medium univariate fidelity was achieved, multivariate and time-dependent fidelity remains a challenge.

Subsequent RWE generation showed serious flaws in generating insurant populations with confirmed SLE diagnosis and sufficient follow-up times. Baseline characteristics, prevalence and incidence estimates and other RWE analysis results differed substantially from the original data.

Conclusions

- **Synthetic data methods enable mid-fidelity data generation with strong privacy preservation, even with limited resources.**
- They can support hypothesis generation and analytic script development where access to real data is restricted; however, correlations between variables on synthetic data should be examined carefully and all findings should be validated on real data.
- Approaches vary in fidelity, scalability, and privacy, so methods must be matched to the use case and data type.
- A pragmatic data generation strategy involves aligning methods with application, making evaluation results available, applying pre- and post-processing to enhance quality and usability, and iteratively refining models and expectations.
- Future work should improve multivariate and temporal fidelity and extend methods for complex relational data.

References

1. Kafatos G, Levy J, Jose S, Hindocha P, Archangelidi O, Vernon S, et al. Leveraging Synthetic Data to Facilitate Research: A Collaborative Model for Analyzing Sensitive National Cancer Registry Data in England. Therapeutic innovation & regulatory science. 2025;.
2. Pezoulas VC, Zaridis DI, Mylona E, Androutsos C, Apostolidis K, Tachos NS, et al. Synthetic data generation methods in healthcare: A review on open-source tools and methods. Computational and structural biotechnology journal. 2024;23:2892–910.
3. Heidler T, Schultze M, Kafatos G, Behera B, Lienau C, Unger A, et al. RWD196 Improving Access to German Health Claims Data Through Synthetic Data Generation: A Methodological Comparison of Different Approaches. Value in Health. 2024;27:S612.

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, L. Grotenrath 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