Distributional Fidelity of Synthetic German Claims Data Under Real-World Study Designs: Implications for HCRU and Cost Analysis

RWD149

Tobias Heidler [1], Michael Schultze [2], George Kafatos [3], Bagmeet Behera [4], Caroline Lienau [5], Alexander Franz 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, United Kingdom
4 Amgen Research (Munich) GmbH, Munich, Germany
5 AstraZeneca GmbH, Hamburg, Germany
6 Medicines and Healthcare products Regulatory Agency (MHRA), London, United Kingdom
7 Limebit GmbH, Berlin, Germany
8 ai4medicine UG, Berlin, Germany

Introduction

- Synthetic health data are proposed as privacy-preserving alternatives to real-world German statutory health insurance (SHI) claims data, supporting research where data access is restricted [1].
- Prior work established mid-fidelity generation in raw population comparisons, but the impact of a real-world study design-cohort entry, disease criteria, continuous insurance on distributional fidelity across Healthcare Resource Utilization (HCRU) domains remains underevaluated [2].
- Fidelity is usually assessed via first- and second-moment comparisons, e.g. mean, standard deviation and error-metrics. These are inherently blind by construction to shape, skew, and tail behavior.
- A rigorous evaluation instead requires quantile-wise comparison across the full distribution, tail-index and extreme-value estimation, and concentration measures alongside distributional distance metrics capable of detecting divergence in shape, not just location and scale.
- 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

- Quantify distributional similarity between four synthetic SHI datasets and ground truth (*DS-WIG2*) across **three HCRU/cost domains**: drug, outpatient, and inpatient utilisation.
- Compare fidelity under two analytical scenarios: a **raw population** comparison (all insurants, 2015–2021) and an **applied SLE cohort** (systemic lupus erythematosus diagnosis, ≥1 year continuous pre-index insurance, age ≥18).
- Assess whether cost distributions are adequately reproduced across the full range of HCRU outcomes - not just one domain - with particular attention to the extreme-cost tail.

Methods

Ground truth: WIG2 SHI benchmark database (*DS-WIG2*), ~ 6,743 insurants, 2015–2021; 4,267 with SLE diagnosis.
Synthetic generators: Generative Adversarial Networks (*DS-GAN*), Adversarial Random Forests (*DS-ARF*), and two Bayesian Networks (*DS-BNN-Kaur*, *DS-BNN-WangTucker*).
Analysis units: Patient-year; outcomes (cost, number of visits, length of inpatient stay) aggregated per insured person per calendar year.
Metrics: Distributional Similarity = 1 - KS-D (Kolmogorov-Smirnov; 0 = maximally different, 1 = identical), Normalised Mean Difference (NMD), Wasserstein-1 distance, Gini coefficient, coefficient of variation (CV), and tail quantile ratios (synthetic ÷ reference P90/P95/P99), top-5% cost/HCRU share.
Strata: Full and conditional (non-zero utilisers only) distributions were evaluated separately to isolate zero-utilisation effects from conditional cost shape.

Results

Cohort Attrition

Step	DS-WIG2	DS-ARF	DS-BNN-WT	DS-BNN-Kaur	DS-GAN
Initial pool	6,743	7,000	6,743	6,742	6,743
SLE diagnosis	4,267 (63%)	2,878 (41%)	1,764 (26%)	5,324 (79%)	5,179 (77%)
≥1 yr insurance *	4,071 (60%)	**585 (8%)**	**580 (9%)**	5,292 (79%)	**3,977 (59%)**
Age ≥18 & sex	3,990 (59%)	571 (8%)	568 (8%)	5,268 (78%)	3,976 (59%)

Table 1: Attrition table
* **Key step:** *DS-ARF* and *DS-BNN-WangTucker* retain only 8% vs 60% in *DS-WIG2* - a 7× gap that drives severe downstream bias across all HCRU domains. DS-GAN matches reference attrition (59%).

Raw fidelity is achievable

DS-ARF produced raw HCRU/costs that showed **no statistically significant differences from the original data**, reaching similarity scores of 0.968–0.990 (drugs), 0.994–0.995 (outpatient) and 0.963–0.997 (inpatient), as shown in Figure 1. Metrics show high similarity in skew, kurtosis, Gini, CV, TOP-5% cost and HCRU share, zero-rate, well replicating WIG2 distributions (see Figure 2). **Drug cost similarities were high** across all datasets (0.723–0.972), domain and feature specific differences in generational quality should be noted. **HCRU/Cost bias varied by an order of magnitude across generators.** Raw drug cost NMD ranged +12% (*DS-ARF*) to −68% (*DS-GAN*), outpatient cost +3% to −87%, and inpatient cost similarity 0.65–0.96 - i.e. the same dataset can be adequate in one HCRU/cost domain and unsuitable in another.

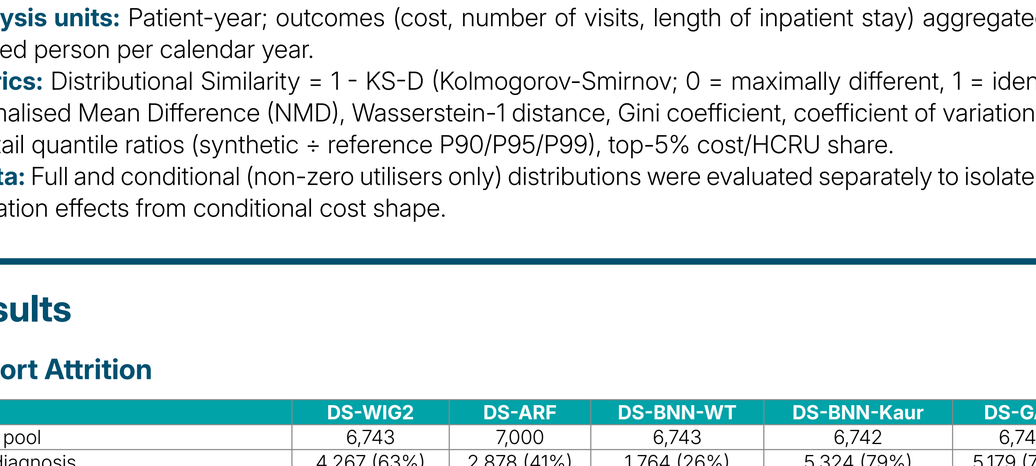


Figure 1: Distributional similarity vs tail accuracy.

Study design substantially alters distributional fidelity

Skewed similarity distributions were seen under the sample definition scenario in *DS-ARF*, retaining only 8% of all patients after attrition (see Table 1), with degradation in specific sectors in other generators. Interestingly, the lower-fidelity dataset *DS-BNN-Kaur* improved similarity substantially in several domains. After applying the study design criteria, *DS-GAN* showed limited improvements as well.

The mechanism is a lost selection gradient: on cohort entry the reference **enriches** for costly patients (median drug cost ×1.50), all generators attenuate this (×1.06–×1.14), and *DS-ARF* reverses it (×0.39). Eligibility criteria are informative in real data and substantially less informative in synthetic data. Cohort metrics rest on as few as 990 patient-years in *DS-ARF* (reference: 20,601).

Several failure modes identified

- **Tail truncation vs. fabrication**. P99 drug costs ranged from 21% (*DS-GAN*) to 113% (*DS-ARF*) of reference, with comparable truncation in outpatient and inpatient domains. *DS-BNN-WangTucker* inflates inpatient HCRU (P99 12.5×, max 7,854 days per patient-year) while **deflating** inpatient cost (P90–P99 ratios 0.37–0.45) - quantile ratios and concentration measures can disagree in sign.
- **Extensive margin and validity**. *DS-BNN-Kaur* and *DS-GAN* misplace the utiliser/non-utiliser split in both directions (outpatient zero-rate 0.60/0.64 vs 0.10; 99% of *DS-BNN-Kaur* patient-years hospitalised, inpatient cases +1,111%). All generators emit values the claims process cannot produce - negative costs in up to 7.9% of patient-years, and stays exceeding the 366-day accumulation window.

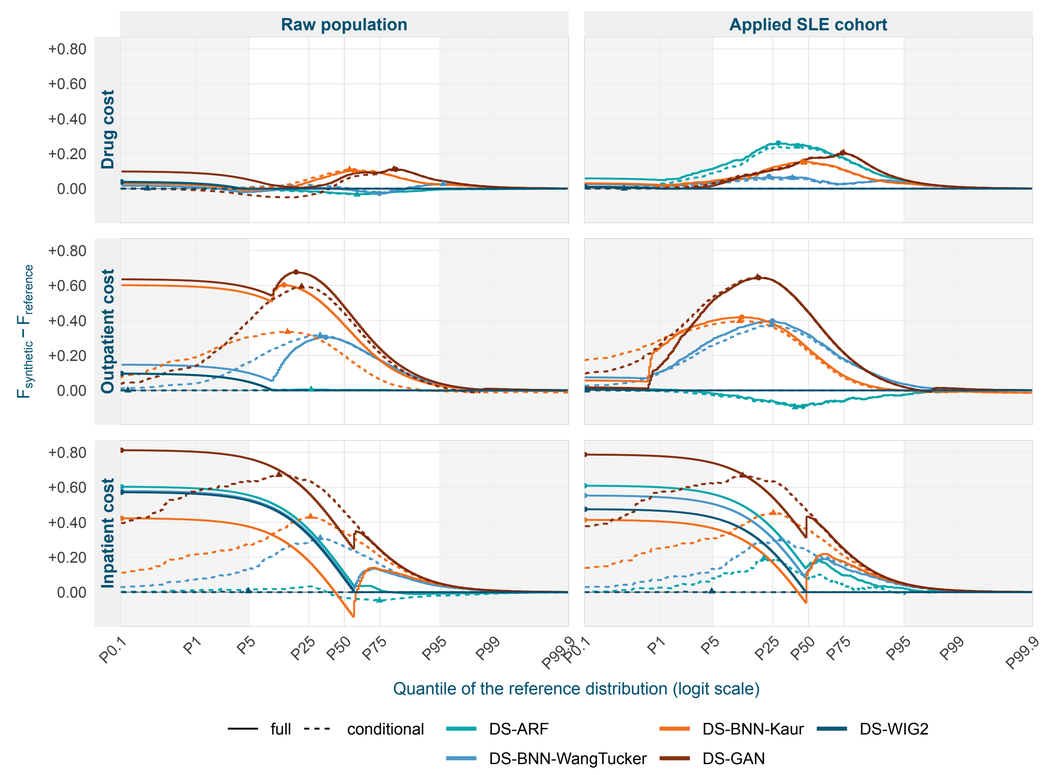


Figure 2: Divergence of synthetic to reference distributions in cost. Vertical axis: Δ between the synthetic and the reference cumulative distribution function at each reference quantile; zero = identical. Positive values indicate excess probability mass at low costs, i.e. cost under-generation. Horizontal axis: quantile of *DS-WIG2* on a logit scale, expanding both tails. Dots mark each dataset's maximum deviation (= KS-D, the basis of the similarity score); shaded bands mark the tails (<P5, >P95). Full vs. conditional (>0) distributions.
Note: all curves return to zero beyond ~P95 because both CDFs converge to 1 by construction - KS-based metrics are structurally blind to the extreme-cost tail.

Why Multiple Metrics Are Required

A single summary statistic can completely mask distributional failure modes, and the two blind spots are complementary. For example:

The mean can be exactly right. *DS-BNN-Kaur*'s raw mean outpatient cost (€1,116) lies within €2 of the *DS-WIG2* reference (€1,114). The distribution behind it is not: a 6-fold inflated zero-utilisation rate (60% vs 10%) is offset by +128% cost inflation among actual utilisers, yielding a median of €0 (vs €719) and a Gini of 0.96 (vs 0.56). NMD indicates similar mean costs in this dataset; but KS-similarity (0.40) and a Wasserstein-1 distance of €1,333 highlight substantial differences.

KS-similarity can be equally right, and equally wrong. For raw drug cost, *DS-BNN-WangTucker* attains the highest similarity of any generator (0.972, above *DS-ARF*'s 0.968) while **missing 53% of the P99** and 88% of the maximum, at a mean bias of −32%.

Consequently, NMD and Wasserstein distance indicate cost-level bias invisible to shape metrics; zero-rate separates the extensive from the intensive margin; and tail quantile ratios directly measure the representation of high-cost patients - the quantity that determines whether a synthetic dataset can support realistic budget-impact or cost-effectiveness analysis.

Conclusions

- **Impact on HCRU/cost estimates** varies by generator and sector. Budget-impact and cost-effectiveness models built on these datasets could under- or overestimate expenditure depending on domain. *DS-ARF* is conditionally suitable for raw-population analyses; cohort-based studies require generators that additionally reproduce attrition dynamics.
- **Fidelity is not a scalar** as generator ranking inverts across cost & HCRU domains and across metrics - no single score supports a claim of fitness for use. Multiple metrics should be used.
- **Evaluation and impact assessment is essential** before synthetic data are rolled out for health economic analyses or decision-making. The magnitude of observed fidelity differences may vary across indications, while the methodological implications extend beyond SLE.

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;. https://doi.org/10.1007/s43441-025-00820-z
2. 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. https://doi.org/10.1016/j.jval.2024.10.3753

Disclosure 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. F. 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