STATISTICAL ANALYSIS PLAN

|  |  |
| --- | --- |
| Title | **SETTING UP A SYNTHETIC DATA ACCESS SOLUTION**  Synthetic Data Generation Using a German Claims Dataset for a Proof-of-Concept Study in Systemic Lupus Erythematosus |
| Protocol version | 1.0, 2023-12-15 |
| Active substance | - |
| Medicinal product | - |
| Marketing authorization holder | - |
| Author | WIG2 GmbH  Markt 8  04109 Leipzig  Germany  *Tobias Heidler* |

Table of Contents

[2 Abbreviations 4](#_Toc157493836)

[3 Introduction 5](#_Toc157493837)

[3.1 Rationale and Background 5](#_Toc157493838)

[3.2 Research Questions and Objectives 5](#_Toc157493839)

[3.3 Protocol Version and Amendments 5](#_Toc157493840)

[4 Study Objectives 6](#_Toc157493841)

[5 Study Design 8](#_Toc157493842)

[**1.** 9](#_Toc157493843)

[**2.** 9](#_Toc157493844)

[5.1 Data Generation Methods 9](#_Toc157493845)

[5.2 Study Population 10](#_Toc157493846)

[5.2.1 Inclusion Criteria 11](#_Toc157493847)

[5.2.2 Exclusion Criteria 11](#_Toc157493848)

[5.3 Participant Follow-Up 11](#_Toc157493849)

[5.4 Study Variables 12](#_Toc157493850)

[6 Power and Sample Size 13](#_Toc157493851)

[7 Statistical Methodology 14](#_Toc157493852)

[7.1 Analysis of Population Characteristics 14](#_Toc157493853)

[7.1.1 Descriptive Statistics 14](#_Toc157493854)

[7.1.2 Stratifying Factors 14](#_Toc157493855)

[7.2 Evaluation across Tables 14](#_Toc157493856)

[7.3 Normalization 16](#_Toc157493857)

[7.3.1 Min-Max Scaling 17](#_Toc157493858)

[7.3.2 Z-Score Standardization 17](#_Toc157493859)

[7.4 Dimensional Reduction 18](#_Toc157493860)

[7.5 Analysis of Primary Endpoints (Objective 1) 19](#_Toc157493861)

[7.5.1 Privacy (Objective 1.1) 19](#_Toc157493862)

[7.5.2 Robustness & Scalability (Objective 1.2) 22](#_Toc157493863)

[7.5.3 Fidelity (Objective 1.3) 22](#_Toc157493864)

[7.6 Analysis of Secondary Endpoints (Objective 2) 24](#_Toc157493865)

[7.6.1 Utility (Objective 2.1) 24](#_Toc157493866)

[7.7 Missing Data 29](#_Toc157493867)

[8 Data Handling 30](#_Toc157493868)

[8.1 Data Coding 30](#_Toc157493869)

[8.2 Variable Definitions 30](#_Toc157493870)

[9 Interim Analyses and Data Monitoring 31](#_Toc157493871)

[9.1 Data Monitoring 31](#_Toc157493872)

[9.2 Interim Analysis 31](#_Toc157493873)

[10 Structure of the Analysis Tables 32](#_Toc157493874)

[11 References 33](#_Toc157493875)

[12 Appendix 34](#_Toc157493876)

[I List of derived variables 34](#_Toc157493877)

[II Mock Tables 34](#_Toc157493878)

# Abbreviations

|  |  |
| --- | --- |
| **Abbreviation or special term** | **Explanation** |
| AI | Artificial Intelligence |
| ATC | Anatomical Therapeutic Chemical classification system |
| BN | Bayesian Network |
| CPRD | Clinical Practice Research Datalink |
| DDD | Daily Defined Dose |
| EBM | German Uniform Assessment Standard (Einheitlicher Bewertungsmaßstab) |
| EU | European Union |
| GAN | Generative Adversarial Networks |
| GEV | Generalized Extreme Value |
| GM | German Modification |
| GPT | Generative Pretrained Transformer |
| HCRU | Healthcare Resource Use |
| ICD-10 | 10th revision of the International Statistical Classification of Diseases and Related Health Problems |
| IQR | Interquartile Range |
| LSTM | Long Short-Term Memory |
| MIAD | Mean Integrated Absolute Difference |
| NLP | Natural Language Processing |
| OPS | The German procedure classification (Operationen- und Prozedurenschlüssel) |
| PCA | Primary Component Analysis |
| PPC | Pearson pairwise correlation |
| PZN | Pharma central number (Pharmazentralnummer) |
| RNN | Recurrent Neural Network |
| RWD | Real-World Data |
| RWE | Real-World-Evidence |
| SAP | Statistical Analysis Plan |
| SHAP | SHapley Additive exPlanations |
| SLE | Systemic Lupus Erythematosus |
| t-SNE | t-Distributed Stochastic Neighbor Embedding |
| TVD | Total Variation Distance |
| UMAP | Uniform Manifold Approximation and Projection |
| WIG2 | Wissenschaftliches Institut für Gesundheitsökonomie und Gesundheitssystemforschung |
| XGBoost | eXtreme Gradient Boosting |
| ZEG | Zentrum für Epidemiologie und Gesundheitsforschung |

# Introduction

## Rationale and Background

The increasing role of Real-World Evidence (RWE) in healthcare decision-making is constrained by data privacy concerns and accessibility challenges. Synthetic data, designed to preserve privacy without compromising data quality, offers a promising avenue. However, the existing literature lacks a comprehensive evaluation of synthetic data generation methods, particularly in the context of German health claims data. This study will focus on Systemic Lupus Erythematosus (SLE) cases, chosen for its clinical complexity and data richness.

The study aims to fill a critical gap in the validation of synthetic data by comparing various data generation methods (Bayesian Network, Generative Adversarial Networks, and Generative Pre-trained Transformer) on German health claims data, both inpatient and outpatient. It will examine trade-offs between privacy, robustness, scalability, fidelity and utility. The evaluation criteria include attack based privacy tests for privacy assessment, generalizability tests for robustness and scalability, and statistical similarity tests for fidelity as well as utility assessment using various RWE-scenarios.

## Research Questions and Objectives

Primary Objectives

Evaluate methods for generating synthetic data on Systemic Lupus Erythematosus patients across three domains:

* Privacy-preservation
* Scalability & robustness
* Fidelity

Privacy preservation: Assess risks of patient re-identification in synthetic data.

Robustness & scalability: Examine computational resource requirements and adaptability to other claims databases.

Fidelity: Explore statistical similarity between synthetic and training data.

Secondary Objectives

Utility: Assess applicability of synthetic data in various real-world evidence (RWE) scenarios.

## Protocol Version and Amendments

Final study protocol version 1.0, 2023-12-15

# Study Objectives

|  |  |
| --- | --- |
| **Objective** | **Endpoint** |
| **Primary Objective** |  |
| 1.1.1 Duplicate records: Verify absence of duplicate records between the training and synthetic data sets through a deterministic or probabilistic matching algorithm. | Number of duplicate records between synthetic and original data |
| 1.1.2 Robustness to privacy attacks: Measure the robustness of the generated synthetic data to common privacy attack scenarios (e.g., single out, linkeability, inference). | k-anonymity, l-diversity, and t-closeness |
| 1.1.3 Shareability: Evaluate the potential for open access data sharing of the generated data sets, emphasizing that privacy-preserving characteristics must be confirmed prior to any sharing or publication. | The dataset is shareable to third parties: Yes/No |
| 1.2.1 Computational efficiency: Measure the computational resources and time required for both training and synthesizing the data, ensuring it meets acceptable standards for scalability. Benchmarks should include CPU utilization, memory usage, and data throughput rates. | Run-Time  RAM & CPU usage  Incorporated domain knowledge (e.g.: manual and/or post-processing restrictions) |
| 1.2.2 Generalization: Evaluate whether the synthetic data model is extensible to multiple diseases or a complete health claims data set without significant alterations or manual inputs. | The data generation method is generalizable to other diseases: Yes/No  The data generation method is generalizable to a complete claims data population: Yes/No |
| 1.3.1 Distributional closeness: Evaluate the statistical similarity between the distributions of the synthetic and real-world data sets across key variables univariate and bivariate. | Univariate Assessment  Descriptive Statistics (e.g., frequency, mean, median, SD, IQR); Statistical Test (e.g., t-test, Wilcoxon, Kolmogorov-Smirnov), Total Variation Distance (TVD), Visualizations  Bivariate Relationship  Spearman's Rank Correlation, Visualizations |
| 1.3.2 High-dimensional dependency: Analyze whether the synthetic data preserves complex relationships among multiple variables present in the original data that could span across multiple tables. | Visual and metrics based evaluation after dimensional reduction, discriminator based assessment (metrics e.g., Cohen’s kappa, accuracy), key discriminators feature assessment (SHapley Additive exPlanations) |
| 1.3.3 Temporal consistency: For longitudinal data, assess the preservation of temporal relationships and trends in the synthetic data compared to the original dataset. | Pearson Pearwise Correlation, time-dependent discriminator based data assessment (metrics e.g., Cohen’s kappa, accuracy) |
| **Secondary Objectives** |  |
| 2.1.1 Analysis and script development: Evaluating the extent to which the synthetic data facilitates technical advancements. This includes, but is not limited to, the creation and refinement of complex analytical methods and the development of scripts for data processing and analysis. The focus is on determining the data's utility in enhancing technical capabilities in data analysis and scripting within the context of healthcare analytics. | Number and percentage, significancy of:   * Missing or inappropriate linkages to tables * Claims in uninsured periods * Continuous insurance periods * Inappropriate code mappings |
| 2.1.2 RWE replication: Comparing the results of analyses conducted with the synthetic data to those obtained from the original data to evaluate their similarity. The following analyses are to be performed | Number and percentage of prevalent and incident cases of SLE  Stratified by year, gender, age  Baseline demographics and clinical characteristics of patients with SLE   * Age at index (= first diagnosis) (10-year age brackets) * Gender (Male/Female) * Index year (2015 - 2021) * First SLE-diagnosis code (ICD-10-GM diagnosis codes: M32.0, M32.1, M32.8 and/or M32.9) * Charlson Comorbidity Index (ICD-10-GM diagnosis codes) * SLE-specific treatments (ATC-codes) * SLE-specific comorbidities (ICD-10-GM diagnosis codes) * TOP-10 comorbidities (ICD-10-GM diagnosis codes) * TOP-10 mediciations (ATC-codes)   Describe cost and HCRU of SLE patients   * Number of outpatient/inpatient visits * Duration of inpatient visits * Inpatient/outpatient/medication cost   Characterize the treatment of newly diagnosed SLE patients  Number and percentage of patients with SLE-treatment per treatment line  Describe real-world clinical outcomes in newly diagnosed SLE patients  Time to death/first treatment/first SLE-hospitalization/systemic glucocorticoid prescription/first flare treatment, Mean Integrated Absolute Difference |

# Study Design

This is an observational retrospective cohort study on patients diagnosed with Systemic Lupus Erythematosus (SLE) from the WIG2 Benchmark database and will include the creation of synthetic data and benchmarking of the different methods involved.

The WIG2 Benchmark database will be used to train different models that generate synthetic data. Furthermore, the WIG2 Benchmark database will be used as a reference throughout the study to evaluate the generated synthetic datasets in terms of privacy, fidelity and utility. Utility will be evaluated by assessing the results of different common RWE scenarios.

As summarized in Figure 1 an overarching SLE-Sample is drawn from the WIG2 Benchmark database with minimal restrictions on inclusion and exclusion criteria mimicking an unbiased draw of SLE-patients from the data base. RWE-Cohorts with stricter definitions, resembling typical applied RWE criteria are used for evaluating utility.

Figure 1: Data Sampling and Utilization Flowchart for SLE Study

Ein Bild, das Text, Screenshot, Schrift, Design enthält.

Automatisch generierte Beschreibung

The methods to generate this synthetic data will be Bayesian, GANs and GPT. The Bayesian approach uses Bayes' theorem to generate synthetic data based on probabilistic relationships. GANs, on the other hand, consist of two models that are trained simultaneously - the Generator, which generates the synthetic data, and the Discriminator, which tries to distinguish between the real and synthetic data. GPT is a transformer based model which uses attention mechanisms to weight input features during training.

Privacy, fidelity, robustness and scalability will be evaluated across all data generation methods given common metrics. Utility will be evaluated on similar RWE-Cohorts using the same restrictions on the synthetic data as on the original data. The Workflow for generating and evaluating the synthetic data is depicted in **Figure 2**.

Figure 2: Synthetic Data Generation and Evaluation Workflow

Ein Bild, das Text, Screenshot, Schrift, Design enthält.

Automatisch generierte Beschreibung



The WIG2 Benchmark database is a large, longitudinal medical claims database that is a representative sample of insured patients in Germany with approximately 4 million insured individuals. Data is currently available from 2014 until end of 2021.

The WIG2 Benchmark data source has the link between the outpatient and inpatient care settings. It combines full data access describing patient treatment pathways and health care resource utilization across the full spectrum of medical care.

The dimensionality of the data used for training and evaluation will be reduced to decrease the time and complexity of training. First, the selection of variables was confined to outpatient and inpatient data that are processed electronically for billing purposes. Secondly, the decision to exclude certain variables was made through a careful and selective process, informed by extensive expert knowledge. This process specifically focused on removing variables that are seldom used in typical health claims data analysis.

Additionally, ATC-Codes and Daily Defined Doses (DDDs) are provided for drug dispenses through the reference database of ABDATA, as these are usually not part of the original underlying data sources but are often used in health claims data analyses.

For a more detailed description of the variables included in the data set see **Section 5.4**, **Table 1**.

## Data Generation Methods

Three vendors will provide synthetic datasets derived from the WIG2 Benchmark database data. This strategy ensures a rich diversity in the synthetic data, as different vendors may apply unique methodologies and optimizations in their models. The use of multiple vendors allows for cross-validation and enhances the robustness and reliability of the synthetic datasets. Additionally, it mitigates risks associated with vendor-specific biases or limitations, ensuring a broader and more comprehensive representation of the health claims data.

**ai4medicine**

Three synthetic datasets will be generated by ai4medicine:

*BN:* Drawing a baseline with out-of-the-box libraries and one of the most used traditional methods, Bayesian Networks.

*GAN*: Using a custom Generative Adversarial Neural Network (GAN) and exploring different GAN variants, including additional functionalities for robust and stable training such as Spectral normalization (Miyato et al. 2018) or Wasserstein loss (Arjovsky et al. 2017).

*GPT*: Exploring the possibility to generate synthetic data using a custom Generative Pre-trained Transformer (GPT) based approach to generate the data.

**CPRD**

One synthetic dataset will be generated by CPRD:

*BN*: Using a Bayesian full joint probability distribution method to represent a set of variables and their conditional dependencies in a probabilistic graphical structure as described in Wang et al. (2021) and Tucker et al. (2020).

**Limebit**

Three synthetic datasets will be generated by Limebit:

*BN*: Using a Bayesian full joint probability distribution method to represent a set of variables and their conditional dependencies in a probabilistic graphical structure as described in Kaur et al. (2021).

*GAN*: Using a GAN as a function approximation method leveraging Gated Recurrent Units (GRUs) to solve the challenge of temporal dynamics in the data as described in Lu et al. (2023).

*GPT*: Exploration of approaches leveraging the architecture and technology of GPT3.

## Study Population

SLE patients will be identified retrospectively using ICD-10 GM codes. The study cohort will consist of patients with a diagnosis of SLE. Based on a feasibility study we expect to include about N = 6,700 patients with SLE in this study.

Two distinct populations will be used in this analysis:

1) An overarching population of all **patients with SLE-diagnosis** (*SLE-Sample*) in the WIG2 Benchmark database from 2014 to 2021 with at least one inpatient or outpatient diagnosis of SLE (ICD-10 GM: M32.-). This population is used to generate the synthetic data and to assess the Primary Objectives of this study. This population represents the whole body of patients with SLE diagnosis in claims and synthetic data. As such, no further inclusion or exclusion criteria are applied.

2) **Patients for RWE-analysis** (*RWE-Cohorts*) – Consisting of a subset of *SLE-Sample* according to the following inclusion and exclusion criteria. This population will be used to conduct analyses in a RWE setting and will further be split into Incident and Prevalent patient cohorts. This population is used to evaluate RWE-analyses conducted in the Secondary Objective, resembling application of typical inclusion and exclusion criteria applied in those type of studies.

The data sampling and utilization of the samples and cohorts is summarised in **Figure 1.**

### Inclusion Criteria

No further inclusion criteria are applied for the *SLE-Sample*.

*RWE-Cohorts* fall into the cohorts of **Prevalent** and/or **Incident** patients, that fulfil the following criteria:

* **Prevalent** patients will consist of all patients with at least one confirmed outpatient diagnosis and/or one inpatient diagnosis of SLE in 2014 to 2020.
* **Incident** patients will consist of all patients with at least one confirmed outpatient diagnosis and/or one inpatient diagnosis of SLE in 2015 to 2020. No such SLE diagnosis must have occurred in 2014.

**Prevalent** and **Incident** cohorts are not mutually exclusive. All **Incident** patients are automatically **Prevalent** patients as well, though not all **Prevalent** patients are Incident.

**Prevalent** and **Incident** patients must be continuously insured - without any gap - at least for one year prior to the index date (= baseline period). As such, **Prevalent** patients that died in 2014 will not be included in this analysis.

### Exclusion Criteria

No exclusion criteria are applied for the*SLE-Sample*.

The*RWE-Cohort*and subsequenty the cohorts of **Prevalent** and **Incident** patients will be excluded when any of the following conditions are fulfilled:

* Age 17 or less at index date
* Gender not “male” (gender column value: 2) or “female” (1), as the number of patients identifying as non-male or non-female (3, “diverse”) in german health claims data is very low.

## Participant Follow-Up

The patients will be included from the start up until the end of data availability, usually the end of the observational period, the death of the patient or the switch to a sickness fund not included in the WIG2 Benchmark database.

The data will furthermore include patients that switched between included sickness funds, with respective gaps in the data availability and claims that have been redeemed after death of patients.

For the patients within the *SLE-Sample*, the entire span of data availability is pertinent, allowing for an unfiltered observation of their healthcare interactions. In contrast, for patients categorized within the *RWE-Cohorts*, specific time frames are delineated to align with the research objective of typical RWE-analyses:

Subjects in the **Prevalent**/**Incident** cohorts will be *followed-up* from index date up until the end of observation, insurance gap of 14 days, or death, whatever comes first, unless stated otherwise.

The *index date* is usually the date of the first confirmed outpatient and/or inpatient SLE diagnosis. For **Prevalent** patients the index date is at least 2015-01-01, even when the first diagnosis should occur earlier, allowing for a 12-months baseline period as defined below.

Baseline characteristics for these cohorts will be reported for the *12-months baseline-period*, prior to the index date. Outpatient characteristics that are available only in quarters (e.g. diagnoses) will be reported using the three quarters prior to the index date quarter, excluding the index quarter itself.

The *complete follow-up period* for each patient is defined from the index date until the end of data availability. Accordingly, the index quarter will be included in the complete follow-up period as well.

## Study Variables

For a description of the WIG2 database see **Section 5.2**. The original data set consist of the following variables:

Table 1: Variables in the Training Dataset

|  |  |
| --- | --- |
| Type | Included Information |
| Patient Characteristics | Year of birth  Gender (female = 1, male = 2, diverse = 3)  Date of death  Continuous insurance periods |
| Drugs | Date of prescription & dispense  Pharma central number  Speciality of prescriber  Physicians & practice code  Quantity prescribed  Amount due  ATC-code  Daily Defined Dose (DDD) |
| Inpatient | Date and cause of admission & discharge  Inpatient or outpatient treatment indicator (yes = 1, no = 0)  Department of admission & discharge  Diagnosis code (ICD-10 GM)  Type of diagnosis (Treatment, Admission, Referral, Follow-Up, Main, Secondary, Additional, Department main diagnosis, Operation, Extension, Co-occuring)  Billing related information (periods, codes, amount, quantity)  Procedure codes (OPS) related information (codes, date of procedure, localisation) |
| Outpatient | Practice code  Treatment period  Year and quarter of outpatient case  Diagnosis related information (ICD-10 GM code, qualification, localisation)  Billing (EBM) related information (codes, physician, speciality, quantity, date)  Procedure codes (OPS) related information (codes, date of procedure, localisation) |

# Power and Sample Size

The study's primary focus is on exploring and qualitatively analyzing various methods for generating synthetic German claims data, as opposed to testing specific hypotheses in a quantitative manner. This exploratory nature necessitates a different approach from the standard power and sample size calculations typically seen in confirmatory studies. Since our aim is to compare and describe the strengths and weaknesses of different synthetic data generation methods, traditional power calculations, which are primarily designed for hypothesis testing, are not applicable in this context.

Regarding the sample size for the generation of synthetic patient data, our strategy is to align it closely with the size of the original dataset. This decision is driven by the need to ensure that the synthetic data is representative of the original data, facilitating an accurate and meaningful comparison. A larger sample size might allow for a more comprehensive analysis of outliers and extreme values. However, the primary goal of our study is not to extend beyond the scope of the original data, but to mirror its characteristics as closely as possible. This approach helps in creating a synthetic dataset that serves as an effective and comparable proxy to the 'ground truth' provided by the original data.

In doing so, we acknowledge the limitations of not expanding the sample size to address potential outliers and extreme values. The focus remains on striking a balance between the fidelity of the synthetic data to the original dataset and the practical considerations such as computational resources and the inherent characteristics of the data itself. By maintaining this balance, the study aims to provide a thorough understanding of the methodologies for synthetic data generation, within the constraints of the available data.

# Statistical Methodology

## Analysis of Population Characteristics

### Descriptive Statistics

Descriptive statistics will be reportedfor several outcomes. Within each descriptive statistic the synthetic data set will be compared to the original dataset. Descriptive statistics include continuous and dichotomous/categorical variables.

For continuous variables, we will report the mean, standard deviation (SD), median, and interquartile range (IQR). These measures were chosen for their ability to provide a comprehensive view of the central tendency and variability within the data. This approach will help in understanding the distribution and spread of continuous variables in both synthetic and original datasets. In the case of dichotomous and categorical variables, we will calculate and report the number and percentage of each category. This method is effective in illustrating the distribution and prevalence of categories within these variables.

A key aspect of our analysis will involve assessing the differences between synthetic and original data. To achieve this, we will use standard mean differences (SMD, Cohen’s d) as our primary tool. Cohen's d is particularly suitable for this study as it provides a standardized measure of effect size, which is crucial for comparing the efficacy of different synthetic data generation methods. Effect sizes will be described using the “rule of thumb” interpretation of Cohen’s d: small = 0.20; medium = 0.50; large = 0.80.

Furthermore, we will employ appropriate statistical tests to evaluate the significance of our findings. For continuous variables, depending on the underlying data distribution and computational considerations, we will choose between unpaired two-samples Wilcoxon or t-tests. Similarly, for dichotomous or categorical variables, we will use either Fisher’s exact test or Chi²-test, selecting the most appropriate test based on the data characteristics.

Variables and cohorts that are crucial for a statistical description will be detailed in their respective objective assessment methodologies.

### Stratifying Factors

Population characteristics will be stratified by the source of the variables, comparing synthetic agains original data sets. Other stratification factors that might be used will be detailed in their respective section of the SAP.

## Evaluation across Tables

In the objectives Privacy and Fidelity associations across several tables are to be evaluated. For this purpose the tabular structured, rowwise fashion of the data is less ideal than a patient centric broad data structure. For instance, a single patient's data combining ten different prescriptions, including one row with metformin, with ten different diagnoses, one is type 2 diabetes, results in one hundred possible combinations. However, among these, only a select few combinations are likely to demonstrate meaningful associations. In this example only one in one hundred rows represents the association of diabetes prescriptions and diagnoses. This methodology significantly narrows down the scope of relevant associations in health claims data analyses, focusing on the most outstanding connections.

A patient-centric dataset is, therefore, employed to facilitate a more effective evaluation of associations across tables. This dataset structure is characterized by a comprehensive representation of patient characteristics across multiple columns, as opposed to a row-based arrangement that could dilute the clarity of associations. This format enables a more efficient identification of clinically meaningful patterns and correlations.

On the other hand, certain intricate characteristics of the original dataset will be lost. For example the resolution of such a table in the temporal dimension must be reduced to gather enough prescriptions and diagnoses to be meaningful. Some associations are also hard to come by, for example a representation of all pharma central numbers (PZN) yields many columns that are particulary sparse in data. This sparsity presents a challenge for traditional data analysis techniques, as most columns will have a vast majority of zeros, indicating the absence of the prescription, and very few ones.

However, it is important to acknowledge certain limitations inherent in this approach. Specifically, the broad table format necessitates a reduction in the temporal resolution of the data to accumulate a sufficient number of prescriptions and diagnoses for meaningful analysis. This could lead to the loss of some detailed characteristics present in the original dataset. Additionally, certain associations, such as those involving pharmaceutical central numbers (PZN), can be challenging to represent due to the resulting sparsity of data in many columns. This sparsity, characterized predominantly by zeros, poses significant challenges for conventional data analysis techniques and may lead to issues like the 'curse of dimensionality'. This phenomenon occurs when the increase in data dimensions results in sparse data distribution, complicating the establishment of meaningful variable relationships and increasing the risk of overfitting in predictive models.

To mitigate these challenges, the dataset columns included in this broad format are selected based on their clinical relevance. Less frequently occurring diagnoses, prescriptions, and procedures will be excluded. Additionally, as detailed in **Section 7.4**, dimensionality reduction techniques may be employed to manage the dataset's high dimensionality, albeit at the expense of interpretability.

The dataset includes patient-centric data compiled for each year. For every year that a patient has recorded claims or is insured for at least one day, a new entry is generated. This entry encompasses patient demographics (age, gender, last region of residence), the number of days insured, diagnoses, prescriptions, and procedures undertaken, along with healthcare resource utilization (HCRU) and associated costs by each sector (prescriptions, inpatient, outpatient services). Detailed specifications of this data structure are provided in **Table 2**.

Table 2: Representation of Variables used in a Patient Centric Dataset

|  |  |
| --- | --- |
| Variable | Representation |
| Year | Year with at least one Claim or within insured period |
| Age | Difference between the year with claims and birth year |
| Gender (Female) | Female = 1, Male and others = 0 |
| Last Region | The last region that was encoded for the patient in the respective year. Fallback: In case a region in the preceeding years is available the region of the last year is chosen. Otherwise the fist succeeding region is chosen. In case no region is available the region is randomly chosen, given the underlying distribution of regions of the data set. |
| Number of Days Insured | Zero for years without any day insured. Otherwise difference between end and start of insurance period within the respective year (At least 1, at most 365/366) |
| Diagnoses | Broad numeric representation of diagnoses (ICD-10 GM codes) in the respective year. For inpatient diagnoses the start date of hospitalization will be used to anchor diagnoses to the year(s) in question. The first three letters of the ICD code (ICD-3) will be used to determine the column to be filled. Only one diagnosis per year per patient will be used. For this, diagnosis types will be sorted accordingly and the highest value in the brackets for each patient and each year will be used to fill the respective column for the patient:  Exclusion (-1) > inpatient (1) > confirmed (1) > condition after (1) > suspicion of (.25). No diagnosis will be filled with zero. E.g., an exclusion diagnosis will always be preferred over an inpatient or confirmed diagnosis. ICD codes that occur less than 100 times in the original data set will be removed. |
| Prescriptions | Broad numeric representation (has prescription = 1, has no prescription = 0) of prescriptions (ATC codes) in the respective year. For prescriptions the date of dispense will be used to anchor the prescription to the respective year in question. The first five letters of the ATC-Code will be used to determine the column to be filled. ATC codes that occur less than 100 times in the original data set will be removed. |
| Procedures | Broad numeric representation (has procedure = 1, has no procedure = 0) of procedures (OPS codes) in the respective year. For procedures the date of procedure will be used to anchor the procedure to the respective year in question. The first four letters (excluding “-“) of the OPS-Code will be used to determine the column to be filled. OPS codes that occur less than 100 times in the original data set will be removed. |
| outpatient HCRU | Number of outpatient cases in the respective year |
| Inpatient HCRU | Number of inpatient cases, the start date of hospitalization will be used to anchor HCRU to the year(s) in question. |
| PRESCRIPTION HCRU | Number of distinct ATC-codes per patient in the respective year |
| OUTPATIENT COST | Sum of all outpatient Fees of all cases within the respective year |
| Inpatient Cost | Sum of all inpatient Fees of all cases within the respective year |
| Prescription Cost | Sum of all prescription Cost of in the respective year |

## Normalization

Normalization of continuous data is a critical step in many types of analyses, including those utilized in the analyses performed in this study. Normalization is essential to ensure comparability and consistency across different scales and units in a dataset. In analyses involving synthetic data, especially where comparisons are made between synthetic and original datasets, disparities in scales can lead to misleading conclusions. For instance, one variable might be measured in thousands while another in fractions, making direct comparisons without normalization problematic. By bringing all variables to a common scale, normalization eliminates these inconsistencies, allowing for a more accurate and meaningful comparison.

Furthermore, normalization is particularly crucial when employing distance-based metrics, such as Dynamic Time Warping (DTW) or dimensional reduction techniques (e.g., PCA/t-SNE/UMAP), in the analysis. Since these metrics are sensitive to the scale of the data, unnormalized data can disproportionately influence the distance calculation, where variables with larger scales dominate the outcome. Normalizing ensures that each variable contributes equally to the distance measure, providing a more balanced and representative analysis. In the context of privacy preservation, normalization plays a pivotal role in maintaining the integrity of the analysis. It helps in assessing the true similarity or dissimilarity between synthetic and original data.

Overall, normalization is not just a procedural step, but a foundational aspect of ensuring accuracy, fairness, and reliability in data analysis, particularly when dealing with complex datasets where privacy and correlation preservation is a key concern.

Normalization in this study is not performed within a data set (e.g., separately for each synthetic data set), but involves the prior concatenation of all data sets (used in Min-Max Scaling) or defining key normalization parameters on the original dataset, that are then applied to the synthetic datasets in question (used in Z-Score Standardization). This ensures that the normalization parameters (like the minimum and maximum values for Min-Max Scaling, or the mean and standard deviation for Z-Score Standardization) are consistent across all datasets. This uniformity is crucial when comparing synthetic datasets with the original dataset, as it guarantees that any differences observed are not due to variations in normalization scales but are inherent in the data itself.

### Min-Max Scaling

The process of Min-Max Scaling involves transforming each feature of the dataset so that its minimum value is transformed to 0 and its maximum value to 1. The transformation is performed using the formula:

One key advantage of Min-Max Scaling is its simplicity and ease of interpretation. The scaled data is bounded within a fixed range, making it straightforward to understand and visualize. However, it's important to note that Min-Max scaling is sensitive to outliers. Extreme values in the data can compress the majority of the data into a narrow range, potentially impacting an analysis. Thus, Min-Max Scaling will only be used for an easy interpretation of the data and will only be used after combining the results of all original and synthetic data sets to bring them to an equal scaling.

### Z-Score Standardization

Z-Score Standardization, also known as standard scaling, is a technique used to normalize the features of a dataset. Unlike Min-Max scaling, which normalizes data into a range between 0 and 1, Z-Score Standardization transforms the data such that it has a mean of 0 and a standard deviation of 1. This method is particularly useful in scenarios where we want to understand the position of a data point relative to the standard distribution of the dataset.

The Z-Score of a data point is calculated using the formula:

In this formula, the mean and standard deviation are calculated from the dataset, and the original value is then transformed into its Z-Score. This score represents the number of standard deviations by which the data point is above or below the mean of the dataset.

One of the primary advantages of Z-Score Standardization is its resilience to outliers. Since it involves standard deviation, this method is less likely to be skewed by extreme values compared to Min-Max scaling. This makes it particularly suitable for datasets with significant outliers or non-uniform distributions. Thus, Z-Score Standardization is the standard method used for standardization/normalization within this study, unless it is stated otherwise.

For our study, the parameters essential for Z-Score Standardization – the mean and standard deviation (SD) – will be derived exclusively from the columns of the original dataset as an ‘intrinsic standard’. This decision is taken to ensure a consistent baseline is used across both the original and synthetic datasets. By employing the mean and SD from the original data, we are establishing a standard against which all synthetic datasets are scaled.

This approach is pivotal in maintaining consistency and comparability in our analysis. It ensures that the scaling of the synthetic data directly corresponds with that of the original data, allowing for a more meaningful and accurate comparison. This methodological choice guarantees that any differences or similarities observed in our analysis are attributable to the characteristics of the synthetic data generation methods themselves, rather than being influenced by variations in the data scaling process.

## Dimensional Reduction

Dimension reduction techniques are vital in data analysis, particularly when dealing with large and complex datasets like those involving multiple synthetic datasets and original data. These techniques simplify the data by reducing the number of variables under consideration, while retaining as much of the significant information as possible. Thus, computational expenses are reduced while highlighting the importance of aspects of the underlying data. Even allowing for (visual) interpretation of high-dimensional spaces.

The method of choice for this analysis is Uniform Manifold Approximation and Projection (UMAP), a dimension reduction technique similar to t-SNE, but with the advantage of higher computational speed and better preservation of the global data structure. It facilitates a high-dimensional graph representation of the data that is reduced to a low-dimensional representation of the underlying structure that is meant to resemble the datas underlying structure as close as possible.

However, despite its strengths, UMAP has certain limitations that must be considered. One of the primary downsides of UMAP is its reliance on hyperparameter choices, which can significantly influence the resulting data representation. Parameters like the number of neighbors, minimum distance, and the choice of metric can alter the outcome, sometimes making it challenging to choose the optimal settings without substantial experimentation. Additionally, UMAP's non-linear nature can complicate the interpretation of the resulting dimensions, as they do not retain the same meaning as in the original high-dimensional space. This can make it difficult to draw direct, interpretable correlations between features in the reduced space.

Another limitation is the potential for overfitting, especially when working with small datasets. UMAP can sometimes capture noise as if it were a significant structure, leading to misleading representations. This is particularly concerning in health claims data, where the accuracy and reliability of interpretations are paramount.

Given these considerations, the use of Principal Component Analysis (PCA) alongside UMAP may still be beneficial in this analysis. PCA, being a linear dimensionality reduction technique, provides a more straightforward interpretation of the data, as the principal components are linear combinations of the original variables. This makes PCA particularly useful for identifying the most significant features contributing to variance in the dataset. Moreover, PCA's simplicity and the linear nature offers a more stable and reliable reduction for smaller datasets or when interpretability is a primary concern.

Therefore, in this analysis, we propose a dual approach where both UMAP and PCA are utilized. Initially, PCA can be applied to reduce dimensionality while retaining interpretability and identifying key features. Subsequently, UMAP can be employed to explore more complex, non-linear relationships in the data. This combined approach allows for a comprehensive analysis, leveraging the strengths of both techniques to provide a deeper understanding of the underlying structures in the (synthetic) health claims data.

All columns that contain patient identifiers (*pid*), case ids (*case\_id*), data sources (e.g., original, CPRD – BN, …) and physicians identifiers will not be involved in dimensional reductions and will not be changed by the following data processing steps described. Prior to dimensional reduction numerical features, including dates, will be normalized using z-Score Standardization according to **Section 7.2.2**. All remaining categories need to be re-encoded into numerical variables. This involves removing aesthetical symbols, cutting the width of the categories text and padding characters with white spaces prior to splitting the column into individual character columns, then one-hot encoding is applied. Zero or low-variance features might be removed to reduce the dimensionality of the data set.

Similar to the approach described in **Section 7.2.2**, UMAP or PCA is initially applied to the original dataset. This initial application is crucial as it helps in establishing an intrinstic and static standard with baseline parameters that are then applied to the synthetic datasets. By using the same dimensional reduction model parameters across the datasets, we ensure that the dimensional reduction is consistent, thereby providing a robust foundation for comparing the effectiveness of the synthetic data generation methods against the original data.

Hyperparameters available to UMAP calculations are tuned on the original data set as well with the goal of producing GPT a low-dimensional representation that accurately captures the underlying structure of the data. Though, initial hyperparameter tuning might be performed on a randomly sampled, reduced data set to speed up the hyperparmeter tuning process. The tuning of hyperparameters is a critical step in the use of UMAP, as it significantly influences the quality and interpretability of the results. By adjusting these parameters on the original dataset, we aim to achieve an optimal balance between preserving the local and global structure of the data in the reduced dimensional space.

The tuning of hyperparameters are guided by the specific characteristics and requirements of our original dataset. Key parameters such as the number of neighbors, minimum distance, and metric choice are carefully adjusted to ensure that the dimensionally reduced representation is as faithful as possible to the original high-dimensional structure. This process involves iterative testing and validation to identify the hyperparameter settings that best reflect the intrinsic patterns and relationships within the data.

## Dynamic Time Warping

DTW is a technique used to measure similarity between two temporal sequences which may vary in speed or length. It aligns these sequences to identify the optimal match between them, even if they are out of phase in the time dimension. The process involves constructing a matrix where each element represents the distance between points in the two sequences and finding the path through this matrix that minimizes the total cumulative distance. This path is the "warping path," and the sum of distances along this path gives the DTW distance. The smaller this distance, the more similar the two sequences are, accounting for possible stretching or compressing in time.

The utilization of DTW allows us to capture complex temporal relationships and patterns that might be missed by more traditional distance metrics, especially when temporal alignment is not guaranteed. This is particularly important in datasets where the timing of events or measurements can be as crucial as the events or measurements themselves.

Since many datapoints need to be processed in this analysis a custom implementation of DTW is used that is implemented in Julia. The full algorithm will be provided, but the key function is shown in **Figure 3**. A window will be used to reduce the number of comparisons per row in the n to 100, thus reducing the complexity of the algorithm to . This does limit the ability of the algorithm to align highly out of phase matrizes with high differences in the number of rows, but greatly reduces the computational expense. <

Figure 3: DTW Algorithm Implementation in Julia

function dtw(s :: Matrix{Float64}, t :: Matrix{Float64}) :: Float64

    n, m = size(s, 1), size(t, 1)

    w = 100

    dtw\_matrix = fill(Inf, (n+1, m+1))

    dtw\_matrix[1, 1] = 0.0

    for i in 2:n+1

        for j in max(2, i-w):min(m+1, i+w+1)

            dtw\_matrix[i, j] = euclidean\_distance(s[i-1, :], t[j-1, :]) +

                min(dtw\_matrix[i-1, j], dtw\_matrix[i, j-1], dtw\_matrix[i-1, j-1])

        end

    end

    return dtw\_matrix[n+1, m+1]

end

## Analysis of Primary Endpoints (Objective 1)

### Privacy (Objective 1.1)

#### Duplicate Records (Objective 1.1.1)

To quantify the similarity between patients in a synthetic dataset and their counterparts in an original dataset we employ a distance-based metric. For each table, every patient in the synthetic data set and every patient in the original data set the distance to one another is calculated.

Distances will be calculated using Dynamic Time Warping (DTW). DTW is a technique used to measure similarity between two temporal sequences which may vary in speed or length, as detailed in **Section 7.5**.

Prior to distance calculation the tables and columns are reduced to the columns most relevant for privacy. This ensures that differences in information non-relevant in terms of privacy will not influence the distance measurement, allowing us to focus on the aspects that are most critical for evaluating the synthetic data in preserving privacy and to reduce the computational expenses. Only distinct rows of synthetic and original data will be compared against each other. Each row will be sorted according to a temporal, alphabetic and/or numerical feature to provide rows for DTW in a reproducible and deterministic way, in case multiple rows occur. That approach mimics a temporal dimension in otherwise sparse temporal data set (e.g., in outpatient setting there is only year and quarter available). Columns that contain patient identifiers (pid), case ids (case\_id) and data sources (e.g., original, CPRD – BN, …) will not be involved in distance calculations and will not be changed by the following data processing steps described.

Numerical values, including dates, will be normalized using z-Score Standardization according to **Section 7.2.2**, taking into account all original as well as synthetic data sets. Columns that contain text will be uppercased and have all characters removed that are aesthetical (e.g., procedure-/OPS-codes are structured like x-yyy.zz, the “-“ and “.” can be savely removed). The following regular expression will be used to keep relevant characters: [A-Z0-9]\*. Text will then be shortened according to the specifications as defined in **Table 2**, left-padded with spaces to these specified number of characters and then split into individual columns per character. One-hot encoding will be used on each of the newly created character columns to create a numeric feature matrix that can be used in DTW. In case a patient has no entries in a particular table (e.g. no inpatient stay) the patient is compared to an equally sized one row zero-vector. Dimensional reduction techniques (PCA/t-SNP) might be employed to reduce the dimensionality of the data and thus the computational expense. Dimension reduction, if used, will be performed across the original data and across all synthetic data sets by concatenation as described in **Section 7.3**.

Table 3: Length of codes used in duplicate record assessment

|  |  |  |
| --- | --- | --- |
| **Code** | **Number of characters** | **Example** |
| ATC-Code | 7 | A10BA02 |
| Diagnosis | 4 | I10.0 |
| Procedures | 4 | 1-100 |
| EBM/GOP | 5 | 10330 |

Assessments and involved columns most relevant for privacy and their respective sorting in ascending order are described in **Table 3**. The mean distances between the assessments per patient pair is calculated. For each patient in the original data set the synthetic patient with the overall minimal mean distance to a patient in the synthetic data set is kept. Thus, the number of reference patients is static. Euclidean metrics will be utilized for calculating the distances in DTW.

After calculating the distances for all synthetic data the individual tables are merged and normalized using min-max normalization. Normalization will be performed according to **Section 7.2.1** with one deviation: The minimal value of each column is not calculated but kept static at zero. Since distances in DTW are strictly positive, this allows a comparison of closeness of each table within values of 0 (= zero distance across all entries in the assessment/column) and 1 (= maximum distance from entries in the assessment/column over all data sets).

Table 4: Privacy critical columns and sorting used in duplicate record assessment

|  |  |  |
| --- | --- | --- |
| **Assessment** | **Columns** | **Ordered by** |
| Insurant | Age, Gender | Age, Gender |
| Insurance Periods | Insurance start & end dates | Insurance start & end |
| Prescriptions | Date of dispense, ATC-Code | Date of dispense, ATC-Code |
| Inpatient Duration | Start & end of hospitalization | Start & end of hospitalization |
| Inpatient Diagnoses | Start of hospitalization, diagnosis code | Start of hospitalization, diagnosis code |
| Inpatient Procedures | Start of hospitalization, procedure code | Start of hospitalization, procedure code |
| Outpatient Diagnosis | Year, Quarter, diagnosis code | Year, Quarter, diagnosis code |
| Outpatient Procedures | Year, Quarter, procedure code | Year, Quarter, procedure code |
| Outpatient Fees | Year, Quarter, EBM-code | Year, Quarter, EBM-code |

The distances between these paired synthetic and original patients in all assessments as well as the overall mean distances will be described using the measures for continuous variables as defined in **Section 7.1.1**. Furthermore, to evaluate and compare the performance of multiple synthetic datasets, we will conduct an Analysis of Variance (ANOVA). ANOVA will allow us to test for significant differences in the mean distances across different synthetic datasets, providing a statistical basis to determine if the variations among them are greater than would be expected by chance.

In cases where the assumptions of ANOVA are not met, particularly regarding the normality of distribution or homogeneity of variances, we will opt for a non-parametric alternative such as the Kruskal-Wallis test. This test is well-suited for comparing multiple groups when the data does not follow a normal distribution, ensuring that our results are robust and reliable.

The outcome of this analysis will provide semi-quantitative insights into the consistency and reliability of the synthetic datasets. By identifying significant differences among them, we can better understand the variability in how these datasets replicate the original data, and thus assess their utility and accuracy in research contexts where privacy preservation is paramount.

In case one or more patients are identified that have overall mean distances of zero or close to zero an individual assessment of the criticality of possible privacy violations is performed. As such, no patient duplication should occur in the synthetic data set. Though, a patient in the original data with low healthcare resource use could be generated by the synthetic method by pure chance, without actual violation of the privacy criterion.

#### Robustness to Privacy Attacks (Objective 1.1.2)

**k-Anonymity** ensures that each record is indistinguishable from at least k-1 other records concerning certain identifying attributes. These identifying attributes are defined in **Table 4**. Specific codes to identify each attribute will be defined in the annex.

Table 5: Attributes and Classes used as Quasi-Identifiers

|  |  |
| --- | --- |
| **Attribute** | **Class** |
| Age | Patients Demographics |
| Gender | Patients Demographics |
| Region | Patients Demographics |
| Psychological Diseases | Critical Illnesses |

After identifying the quasi-identifiers calculation is as follows and is applied to each data set:

1. **Grouping Records:** Group the dataset by the quasi-identifiers.
2. **Counting Records:** For each group, count the number of records.
3. **Ensuring k-Anonymity:** If any group has fewer than k records, apply generalization or suppression to the quasi-identifiers until each group has at least k records.

The overall distribution is described using descriptive statistics, as described in **Section 7.1.1**. The lowest values for k-anonymity will be identified across all datasets and will be reported separately for each dataset.