|  |  |  |
| --- | --- | --- |
|  | Non-Interventional, Secondary Data Study Protocol | |
| Home - ZEG Berlin | Study Code | D2287R00203 |
|  | Version | 1.0 |
|  | Date | 14.02.2024 |
|  |  |  |
|  |  |  |
|  |  |  |
|  | | |
| **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   |  |  | | --- | --- | | Sponsors: | AMGEN  AstraZeneca | | Authors: | Tobias Heidler (WIG2)  Nils Kossack (WIG2)  George Kafatos (AMGEN)  Bagmeet Behera (AMGEN)  Alexander Unger (AstraZeneca)  Caroline Lienau (AstraZeneca)  Michael Schultze (ZEG Berlin)  Marc Pignot (ZEG Berlin) | | | |

|  |  |
| --- | --- |
| TABLE OF CONTENTS | PAGE |

Press F9 to update Table of Contents

TABLE OF CONTENTS 2

List Of Abbreviations And definition of terms 4

Responsible Parties 6

Protocol Synopsis 7

amendment history 12

Milestones 13

1. Background and Rationale 14

1.1 Background 14

1.2 Rationale 14

2. Objectives and Hypotheses 16

2.1 Primary Objective(s) & Hypothesis(es) 16

2.1.1 Privacy preservation 16

2.1.2 Robustness & scalability 16

2.1.3 Fidelity 17

2.2 Secondary Objective(s) & Hypothesis(es) (Optional) 17

2.2.1 Utility 17

2.3 Exploratory Objective(s) & Hypothesis(es) (Optional) 17

3. Methodology 17

3.1 Study Design – General Aspects 17

3.1.1 Data Source(s) 18

3.1.2 Data generation methods 19

3.2 Study Population 20

3.3 Inclusion Criteria 20

3.4 Exclusion Criteria 21

3.5 Participant Follow-up 21

4. Variables and Measurements 22

4.1 Exposures 22

4.2 Outcomes 22

4.2.1 Primary Outcomes 22

4.2.2 Secondary Outcomes 23

4.2.3 Exploratory Outcomes 24

4.3 Other Variables and Covariates 24

5. Statistical Analysis Plan 24

5.1 Statistical Methods – General Aspects 24

5.1.1 Primary Objective(s): Calculation of Epidemiological Measure(s) of Interest (e.g., descriptive statistics, hazard ratios, incidence rates, test/retest reliability) 25

5.1.1.1 Privacy 25

5.1.1.2 Robustness & scalability 25

5.1.1.3 Fidelity 26

5.1.2 Secondary Objective(s): Calculation of Epidemiological Measure(s) of Interest (e.g., hazard ratios, incidence rates, test/retest reliability) 27

5.1.2.1 Utility 27

5.1.3 Exploratory Objective(s): Calculation of Epidemiological Measure(s) of Interest (e.g., hazard ratios, incidence rates, test/retest reliability) 35

5.2 Bias 35

5.2.1 Methods to Minimize Bias 35

5.2.2 Adjustment for Multiple Comparisons 35

5.2.3 Strengths and Limitations 35

5.3 Interim Analyses 37

5.4 Sample Size and Power Calculations 37

6. Study Conduct AND REGULATORY DETAILS 37

6.1 Study Conduct 37

6.1.1 Quality Control 37

6.2 Protection of Human Subjects 37

6.3 Collection and Reporting of Adverse Events/Adverse Drug Reactions 38

6.3.1 Reporting of Adverse Events 38

6.4 PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS 38

7. List of References 39

8. Appendices 41

8.1 Variables in the Training Dataset 41

8.2 Statistical Measures to Assess Privacy 42

8.3 Statistical Measures to Assess Fidelity 43

9. Attachments 44

10. Signatures 45

List Of Abbreviations And definition of terms

| 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 |
| 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 |
| WIG2 | Wissenschaftliches Institut für Gesundheitsökonomie und Gesundheitssystemforschung |
| XGBoost | eXtreme Gradient Boosting |
| ZEG | Zentrum für Epidemiologie und Gesundheitsforschung |
|  |  |
|  |  |

Responsible Parties

| Name | Professional Title | Role in Study | Affiliation | Email Address |
| --- | --- | --- | --- | --- |
| George Kafatos | Director Data & Analytics | Project Lead | AMGEN | gkafatos@amgen.com |
| Bagmeet Behera | Sr Manager - Data Science | Project Lead | AMGEN | bagmeet.behera@amgen.com |
| Alexander Unger | Director Data Insights, Data Science & AI | Project Lead | AstraZeneca | alexander.unger@astrazeneca.com |
| Caroline Lienau | Data Scientist | Project Lead | AstraZeneca | caroline.lienau@astrazeneca.com |
| Marc Pignot | Head of Real-World Evidence & Safety Data Analytics | Project Director | Zentrum für Epidemiologie und Gesundheitsforschung Berlin (ZEG Berlin) | m.pignot@zeg-berlin.de |
| Michael Schultze | Director of Real-World Evidence & Safety Data Analytics | Project Manager | ZEG Berlin | m.schultze@zeg-berlin.de |
| Nils Kossack | Head of Analytics | Lead Analyst | Wissenschaftliches Institut für Gesundheitsökonomie und Gesundheitssystemforschung, (WIG2) | nils.kossack@wig2.de |
| Tobias Heidler | Analyst | Analyst | WIG2 | tobias.heidler@wig2.de |
| Dietmar Frey | CEO | AI / synthetic data Consultant | Ai4medicine | dietmar.frey@ai4medicine.com |
| Zhenchen Wang | Data Science Specialist | AI / synthetic data Consultant | CPRD | zhenchen.wang@mhra.gov.uk |
| Philipp Koch | Data Scientist & Machine Learning Engineer | AI / synthetic data Consultant | Limebit | philipp.koch@limebit.de |
|  |  |  |  |  |

Protocol Synopsis

|  |
| --- |
|  |
| **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** |
|  |

**Background/Rationale:**

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.

**Objectives and Hypotheses:**

**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.

**Methods:**

**Study design:**

This is an observational retrospective cohort study focused on patients with Systemic Lupus Erythematosus (SLE) sourced from the WIG2 Benchmark database. The study involves the generation of synthetic data using Bayesian methods, GANs, and GPT. Metrics for privacy, fidelity, robustness, and scalability will be assessed across all data generation techniques. The WIG2 Benchmark database will serve both as a training ground for models and a reference for evaluating synthetic datasets, particularly in terms of privacy and fidelity. Fidelity will further be gauged through common RWE scenarios.

**Data Source(s):**

The study will utilize the WIG2 Benchmark database, a longitudinal medical claims dataset of approximately 4 million insured individuals in Germany. The dataset covers the period from 2014 to 2021 and integrates both outpatient and inpatient care settings. For efficiency, the dataset’s dimensionality will be reduced by selecting variables processed electronically for billing and guided by expert-driven exclusion criteria. Additional data variables like ATC-Codes and DDDs will be sourced from ABDATA’s reference database.

Synthetic datasets will be generated by three vendors: ai4medicine, CPRD, and Limebit.

* *ai4medicine* will generate two datasets using custom GAN and GPT-based methodologies, incorporating Spectral normalization (Miyato et al. 2018) and Wasserstein loss (Arjovsky et al. 2017).
* *CPRD* will employ a Bayesian full joint probability distribution method based on works by Wang et al. (2021) and Tucker et al. (2020).
* *Limebit* will also produce two datasets, one using a Bayesian approach and the other utilizing a GAN with Gated Recurrent Units to account for temporal dynamics in data, as per Lu et al. (2023).

**Study Population:**

Two distinct populations will be used in this analysis:

1. An overarching population of all **Patients with SLE-diagnosis**in the WIG2 Benchmark database from 2014 to 2021 with at least one inpatient or outpatient diagnosis of SLE (ICD-10 GM: M32.-). No further inclusion or exclusion criteria are applied. This population is used to generate the synthetic data and to assess the Primary and Secondary Objectives of this study.
2. **Patients for RWE-analysis** – Consisting of a subset of *Patients with SLE diagnosis* 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.

*Inclusion Criteria*:

**Prevalent**

* At least one confirmed outpatient and/or one inpatient diagnosis of SLE in 2014 to 2020.
* Continuously observable for at least one year before the index date, defined as no gaps (less than 14 days) between insurance periods (= start and end date). The index date is defined as the first SLE diagnosis or 2015-01-01, whatever comes last. Patients that are diagnosed after the 2015-01-01 have their index date on the date of diagnosis, while patients diagnosed before will have their index date set to this fixed date. Thus, allowing at least a complete 12-months baseline period for every Prevalent patient and removing periods in which no or few Claims should occur in the follow-up after the index. Patients that died in 2014 will subsequently be excluded.

**Incident**

* First confirmed outpatient and/or inpatient diagnosis of SLE in 2015 to 2020.
* Continuously observable for at least one year before the index date (= first diagnosis date).

*Exclusion Criteria*:

* Age less than 18 years at index date.
* Gender is not male or female.

*Participant Follow-Up*: Patients with SLE-diagnosis will be followed for the whole period in which they appear in the data set from 2014 to the end of the observational period, including uninsured periods or post-mortem claims. The data and observational window will encompass both **Prevalent** and **Incident** cohorts cases, 12-months before index (= first diagnosis) to either end of the observational period, patient death, or transfer to a non-participating sickness fund. Data may have gaps due to fund switching or mortality.

**Outcome(s):**

To assess the primary objectives the following outcomes will be evaluated in three primary categories: Privacy, robustness & scalability, and fidelity & utility:

*Primary Outcomes*

*Privacy*:

Duplicate records: Absence of duplicates between training and synthetic data will be verified through deterministic or probabilistic algorithms.

Robustness to Privacy Attacks: The synthetic data will be tested for susceptibility to standard privacy attack scenarios.

Shareability: The data’s eligibility for open access sharing will be assessed, contingent upon privacy-preserving features.

*Robustness & scalability:*

Computational efficiency: Metrics such as CPU usage, memory, and data throughput will be benchmarked for scalability.

Generalization: The model’s adaptability to various diseases or full health claims data sets will be evaluated.

*Fidelity*:

Distributional closeness: Univariate and bivariate distributional similarities with real-world data will be assessed.

High-dimensional dependency: The preservation of complex multivariate relationships will be evaluated.

Temporal consistency: For longitudinal data, temporal trends and relationships will be compared to the original dataset.

*Secondary Outcomes*

*Utility:*

Real-World Evidence (RWE) Analyses: The synthetic data’s applicability for latitudinal (cross-sectional) and longitudinal RWE studies will be measured, considering statistical validity, accuracy, and representativeness.

**Sample Size Estimations:**

An estimated N = 6,700 patients are expected to be included, spanning years 2014 to 2021.

**Statistical Analysis:**

*Privacy*: Multiple distance measures and attack scenarios (singling out, linkability, inference) are used to evaluate the data’s resilience to privacy attacks. Success rates of attacks help quantify privacy risk.

*Robustness & scalability*: No statistical measures will be conducted for robustness & scalability. Metrics like RAM and CPU usage during training and synthetization are considered. Any manual interventions required and implemented constraints are also documented and evaluated.

*Utility*: Real-world evidence analyses are to be conducted to assess the utility of the synthetic data, focusing on various aspects like epidemiology, patient characteristics, and healthcare resource utilization in the context of Systemic Lupus Erythematosus (SLE), comparing the results of the synthetic data sets to the WIG2 Benchmark database results.

*Fidelity*: To demonstrate statistical similarity to the WIG2 Benchmark Database, both descriptive analysis and a range of statistical methods are employed. This includes, but is not limited to, using two-sample Kolmogorov-Smirnov tests to evaluate distributional closeness or the deployment of discriminator models to assess the preservation of high-dimensional dependencies and temporal consistency within the data.

amendment history

| Date | Section of study protocol | Amendment or update | Reason |
| --- | --- | --- | --- |
| << >> |  | << >> |  |
|  |  | N/A |  |

Milestones

| Milestone | Planned date |
| --- | --- |
| Protocol Development  Training/Creation  Analysis  Evaluation  Reporting | October 15, 2023  November 30, 2023  February 15, 2024  March 31, 2024  May 31, 2024 |
|  |  |

# Background and Rationale

This study aims to create and evaluate multiple synthetic datasets using different methods trained on a German proprietary sickness fund database data, to assess the trade-offs between privacy, fidelity, and robustness of these approaches. The rationale is driven by the rising importance of synthetic data in healthcare research, the demand for secure, high-quality data, and the potential to elevate the throughput, quality, and performance of Real-World Data (RWD) through the means of synthetic datasets.

## Background

Real-World Evidence (RWE) are gaining importance in the decision-making process of healthcare policy, regulatory approvals, and clinical practice. The analyses required to generate RWE often require large and representative datasets, which can be hard to come by because of governance rules that restrict access mainly due to privacy issues.

Synthetic data is a rapidly expanding field of research and promises a new paradigm of privacy-preserving research in the healthcare industry. Unlike anonymized or de-identified data, which still contain inherent risks of re-identification, synthetic data offers an enhanced level of protection against this vulnerability. The mitigation of re-identification risk in synthetic data is sufficiently robust to be negligible or non-existent. This significantly reduces privacy concerns and can contribute to greater trust among data contributors and users.

However, the need for preserving privacy and ensuring the required fidelity of the generated synthetic data, as well as the scalability of the method and process chosen depends largely on the problem at hand. For example, if the aim is to extract univariate descriptions and simple sample statistics then a low fidelity synthetic data set and the associated methods will suffice. On the other hand, if the goal involves exploring the development of predictive models, for instance for predicting risk of specific health outcomes in patients, e.g., stroke or myocardial infarction, then a medium or high fidelity synthetic dataset (and methods capable of generating these) may be more appropriate.

While synthetic data can potentially assuage data-privacy concerns and ameliorate challenges of data accessibility, it is imperative to comprehensively evaluate their quality and performance in typical RWE-analysis scenarios.

The goal of this study is to compare and evaluate different approaches for creating synthetic data and describe the trade-off profiles between privacy, fidelity, scalability and robustness as well as utility.

## Rationale

The rationale for this study is grounded in several factors. There is an ongoing demand for high-quality, representative datasets to conduct RWE analyses. However, confidentiality concerns often restrict access to RWD (especially within certain regions, such as the EU). The inability to freely access and share health data hampers research progression. Synthetic data, designed to mimic real-world data, while not containing any personally identifiable information, can provide insights on the patient population and help accelerate data access. However, their usage necessitates validation of their inherent tradeoffs between privacy and fidelity (Bullward et al. 2023).

Past efforts include the Simulacrum dataset (Health Data Insight CiC) that demonstrated the feasibility of creating synthetic data from cancer registries in the UK. Multiple approaches have been made to generate synthetic data from electronic health records, like Synthea (Walonoski et al. 2018) that aims to model lifespans of patients with the ten most frequent primary care encounters or I-safe (Yoon et al. 2022) demonstrating the generation of high-fidelity and privacy-preserving inpatient datasets. Helfer et al. (2021) made significant progress in creating synthetic electronic health (RecordIEHR) data derived from German hospital claims data, though also limited to data originating from inpatient hospital settings. SyH-DR (Katz et al. 2021) is an all-payer, nationally representative synthetic claims database consisting of inpatient, outpatient, and prescription drug claims, including utilization, payment, and enrollment data, for people insured by Medicare, Medicaid, or commercial health insurance in 2016.

This study will be focusing on German health claims data, using both inpatient and outpatient information, for all patients that received a diagnosis of Systemic Lupus Erythematosus (SLE). SLE is a rare multi-systemic autoimmune disorder with a variety of clinical presentations, multi-faceted treatment regimens and a heterogenious population. This results in a rich, high-dimensional, though small dataset.

Different methods and algorithms are available for creating synthetic longitudinal data, like traditional Bayesian approaches (e.g., PrivBayes), deep-learning methods like Generative Adversarial Networks (GAN, I., I-M-GAN, DoppelGANger, medGAN) or more recent natural language processing (NLP) based approaches using transformer models like GPT (Generative Pretrained Transformer).

However, there's a lack of comprehensive comparative studies assessing the quality and performance of algorithms creating synthetic claims data across various RWE-analysis scenarios on health claims data.

This study aims to fill that gap by describing and evaluating different synthetic data generation methods (Bayesian Network, GAN & GPT). As such, the data generation methods chosen for this study represent a broad spectrum of current synthetic data generation techniques, from traditional statistical methods to cutting-edge deep learning, generative approaches. These were chosen to provide a comprehensive view of the current capabilities of synthetic data generation.

Furthermore, by examining the strengths and weaknesses of various synthetic data generation methods regarding data privacy, robustness & scalability, fidelity and utility, we can identify the optimal synthetic data generation technique depending on the requirements. For example, a low-fidelity synthetic dataset may be sufficient if the intended use is to generate programming code whereas a high-fidelity synthetic dataset may be needed if the intention is to build predictive models. It is important to note that higher fidelity in synthetic datasets might compromise privacy, as more detailed and accurate representations could inadvertently reveal sensitive information. This can guide future development in synthetic data generation, leading to more reliable and accurate synthetic datasets that are fit-for-purpose.

Privacy will be evaluated by multiple tests for differential privacy, robustness & scalabilty characterized by evaluating the generalization of the methods to reproduce the results on various diseases or bigger data sources, ideally with modest computational resources and minimal manual interventions in the training and data generation processes. The synthetic datasets resulting from this study will be evaluated for privacy risk, and only when there are no critical concerns (i.e, risks are adjuged to be minimal and thus acceptable), these data sets might be shareable or even become available for open access. This approach will demonstrate a mechanism for generating and releasing synthetic data derived from specific real-world sources, thus making it available for use by the wider research community. To evaluate fidelity, we will evaluate the statistical similarity of synthetic data to the original data. We will also evaluate the utility of the generated data - the usability of synthetic data in various typical RWE-analysis scenarios as well as the technical suitability to generate analysis scripts or complex analyses.

By addressing these points, this study will contribute substantially to the emerging field of synthetic data in healthcare research, potentially lessening the burden of accessing German health claims data for a broader scientific audience and thus accelerating data access, leading to richer data insights.

# Objectives and Hypotheses

## Primary Objective(s) & Hypothesis(es)

The primary objectives are the following:

* To test different methods for generating synthetic data for patients with Systemic Lupus Erythematosus and evaluate the data in terms of privacy-preservation, scalability & robustness and fidelity.

The strengths and limitations of each approach will be described. No hypothesis-testing will be undertaken in this study.

### Privacy preservation

To assess the risk of re-identifying patients due to the generation of synthetic data that may too closely resemble real patients’ characteristics.

### Robustness & scalability

To descriptively evaluate the computational resource requirement during training of the algorithms for the generation of synthetic data and the data generation itself. The overall need to manually fine-tune the models to assess whether they can be easily utilised for other claims databases.

### Fidelity

To explore and access the statistical similarity of the generated synthetic data to the original data.

## Secondary Objective(s) & Hypothesis(es) (Optional)

### Utility

To evaluate the suitability to use the synthetic data in different RWE-scenarios.

## Exploratory Objective(s) & Hypothesis(es) (Optional)

N/A

# Methodology

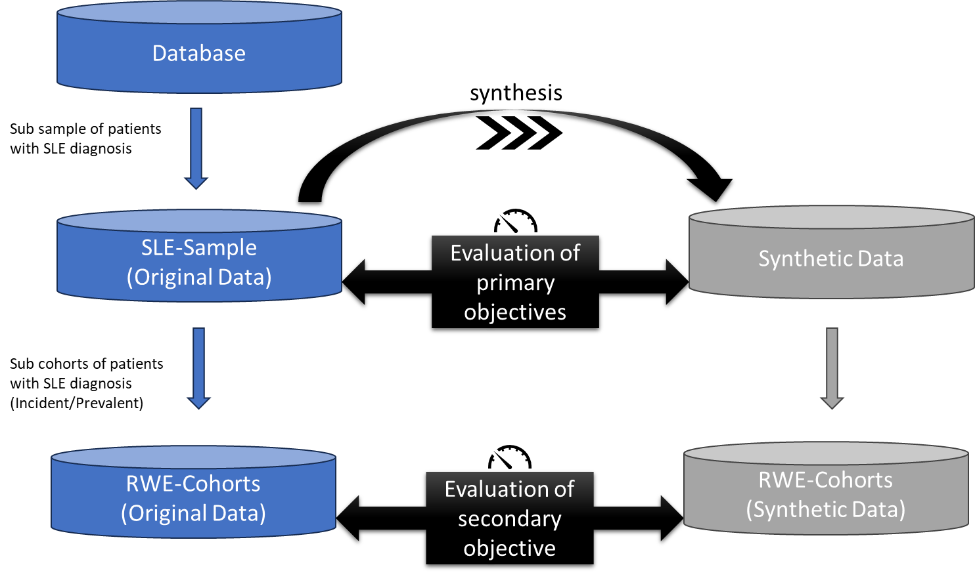
## Study Design – General Aspects

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.

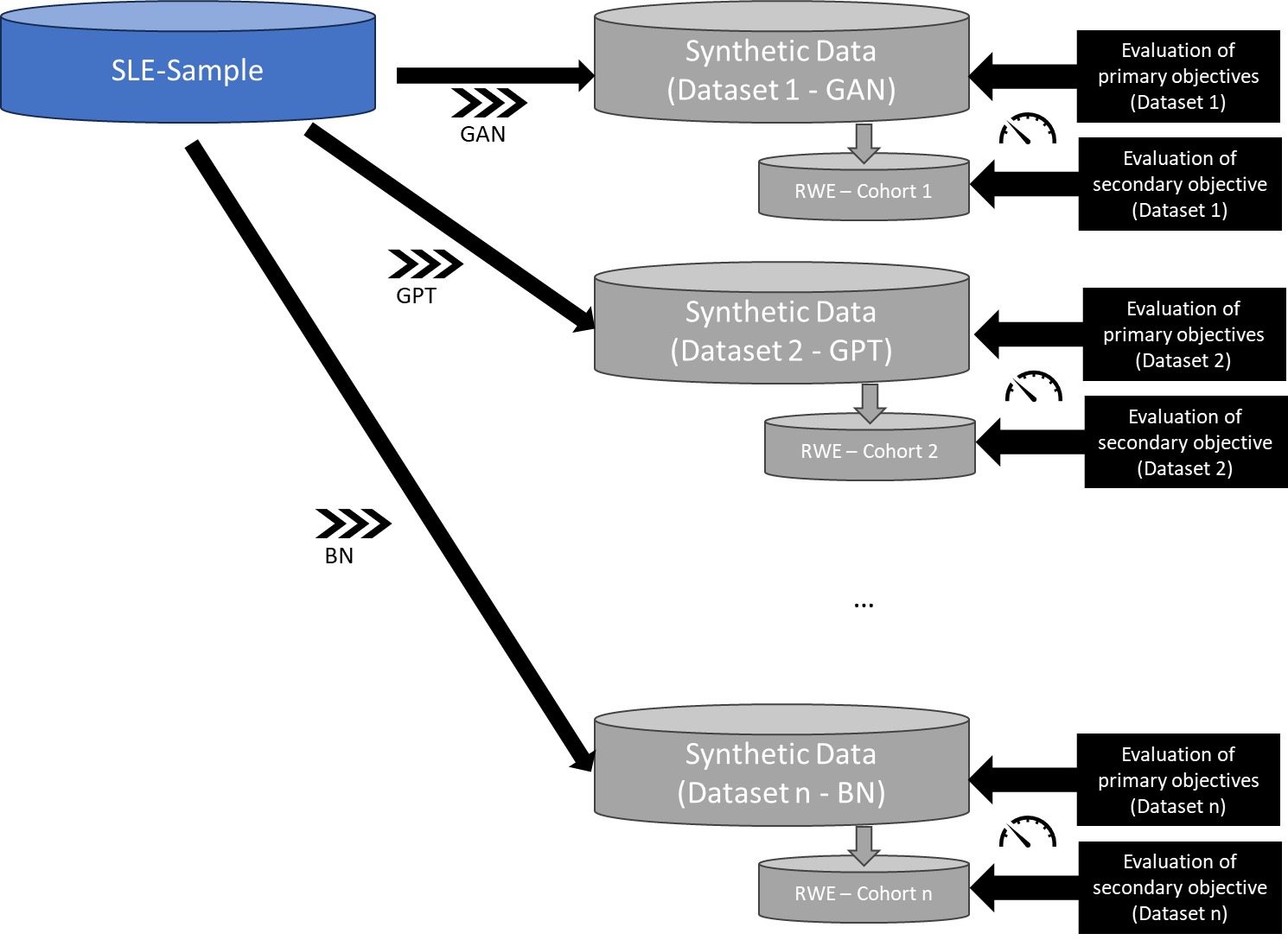
Figure 1: Data Sampling and Utilization Flowchart for SLE Study



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



### Data Source(s)

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. This was already performed.

German claims data about (outpatient) prescription do not contain any information about which agent was prescribed and what dose is given to the patient. But the data contains a column which is capable of linking these information to the claims, the pharma central number. We are unsure if the methods to synthetize the data are capable of reproducing the pharma central numbers. Thus, ATC-Codes and Daily Defined Doses (DDDs) are directly linked to the provided drug dispense data using the reference database of ABDATA. Both information are often used in health claims data analyses of pharmaceuticals and allow for a broader range of possible analysis, even when pharma central numbers cannot reliably be generated by the methods chosen.

For a more detailed description of the variables included in the data set see **Section 8.1, Table 4.**

### 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 theSLE-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. Continuous insurance is defined as a maximum of 14 days between current end date and successive start date of an insurance period as provided by the insurance table. This table is also part of the data that is provided and is to be generated by the synthetic data generation methods. No or very few claims (e.g., doctor’s note) for patients should occur in gaps between these periods.

## 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” or “female”, as the number of patients identifying as non-male or non-female (“divers”) 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. These gaps can be identified by the date differences of insurance start and successive end dates as provided by the WIG2 original dataset. Data will be provided as is and the data generation methods should generate such gaps appropriately. No imputation is performed for these periods.

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. The first date of diagnosis is assessed as defined in the inclusion criteria and could therefore span from 2015-01-01 to 2020-12-31.

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.

# Variables and Measurements

For the robustness & scalability no statistical analysis will be conducted. The data will be presented in a descriptive manner, serving solely for qualitative interpretation.

The number of patients with missing data will be provided; no imputation methods will be used to handle missing data.

## Exposures

Not applicable

## Outcomes

### Primary Outcomes

To assess the primary objectives the synthetic data will be evaluated using the following outcomes:

**Privacy:**

*Duplicate records*: Verify absence of duplicate records between the training and synthetic data sets through a deterministic or probabilistic matching algorithm.

*Robustness to privacy attacks*: Measure the robustness of the generated synthetic data to common privacy attack scenarios (e.g., single out, linkeability, inference).

*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.

Ideally a shareability of the synthetic data seems plausible as there are no or very little privacy concerns. A full anonymization can be assumed.

**Robustness & scalability:**

*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.

*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.

Ideally a representation of multiple diseases or a complete health claims data set is feasible without manual intervention.

**Fidelity:**

*Distributional closeness*: Evaluate the statistical similarity between the distributions of the synthetic and real-world data sets across key variables univariate and bivariate.

*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.

*Temporal consistency*: For longitudinal data, assess the preservation of temporal relationships and trends in the synthetic data compared to the original dataset.

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.

### Secondary Outcomes

**Utility**

Overall, the synthetic data resembles the real data close enough to realize a variety of latitudinal and longitudinal RWE analyses.

*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.

*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:

* To describe prevalent and incident cases of SLE
* To describe the baseline demographics and clinical characteristics of patients with SLE
* To describe cost and healthcare resource use (HCRU) of SLE patients each year
* To characterize the treatments of newly diagnosed SLE patients
* To describe real-world clinical outcomes in newly diagnosed SLE patients

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.

### Exploratory Outcomes

Not applicable

## Other Variables and Covariates

Not applicable

# Statistical Analysis Plan

## Statistical Methods – General Aspects

Descriptive analyses will be conducted in this study. Discrete variables will be summarized using appropriate measures of central tendency and dispersion, such as frequencies and proportions. Continuous variables will be described using suitable summary statistics, which may include but are not limited to means, standard deviation, medians, and interquartile ranges, depending on the data distribution and research objectives.

The synthetic datasets will be evaluated using the *SLE-Sample* drawn from the WIG2 database as a reference (original data). Appropriate statistical measures and tests to compare the different synthetic data sets to the reference will be applied and may include 95 % confidence intervals if appropriate. If tests are applied, a significance level of 5 % will be used. No adjustment for multiple testing will be conducted.

*RWE replication* will be measured by comparing the results of these RWE analyses between the synthetic datasets and *RWE-Cohorts* (original data). The evaluation of how close the analyses results are between the synthetic and original data is a comparative analysis between synthetic and original data.

The initial evaluation of the data will be conducted by WIG2, which will rigorously assess compliance with privacy standards for the synthetic dataset. This assessment will ascertain that all privacy benchmarks are met, ensuring the synthetic data retains no trace of personal identifiers. WIG2 will also verify that data types and data structures are preserved as well as the data’s distribution characteristics to confirm that it is a representative and safe surrogate for the original dataset, thereby permitting its secure dissemination for analytical purposes.

Frequency distributions, measures of central tendency, and variability will be ascertained for key variables. Results will be presented in tabular and graphical formats to facilitate a clear understanding of the findings. Subsequently, a comprehensive discussion will ensue, integrating the descriptive results with existing literature to provide a nuanced interpretation. This discussion will aim to contextualize the findings within the broader scientific and clinical landscape, highlighting their implications and potential for informing future research and practice.

### Primary Objective(s): Calculation of Epidemiological Measure(s) of Interest (e.g., descriptive statistics, hazard ratios, incidence rates, test/retest reliability)

#### Privacy

Privacy assessment will utilize a subset of methods as described in the annex (**Section 8.2**, **Table 5**), employing a range of methods to ensure the privacy of the data.

*Absence of duplicate records* will be demonstrated using pairs of patient records in the real and synthetic data. Each pair of subjects will be compared against one another using appropriate methods, e.g. distance functions like Gowers distance or Hamming distance. The premise being that there exists no or low levels of patient level records in the synthetic data tables that resemble the data present in the original data and as such having a distance of or close to zero.

The assessment of *robustness to privacy attacks* will include considerations of potential Singling Out, Linkability, and Inference attacks, alongside associated measures such as k-anonymity, l-diversity, and t-closeness. Adaptations of these scenarios and measures will be informed by methodologies discussed in current literature, including but not limited to Giomi et al. (2023). Reports will detail the success rates of simulated attacks and outcomes from the related privacy-preserving measures to provide an evaluation of the privacy risk.

*Shareability* will be evaluated descriptively based on the *absence of duplicate records* and the results of *robustness to privacy attacks.* The assessment will also encompass a detailed justification of the privacy guarantees offered by differential privacy mechanisms, where applicable. This discussion might be supplemented with an explanation of the operational parameters of the differential privacy algorithm, including the noise distribution and sensitivity settings, and how these concretely uphold the privacy guarantee as provided by the vendors.

Further details will be provided in the SAP.











# List of References

Albrecht, Katinka; Redeker, Imke; Aringer, Martin; Marschall, Ursula; Strangfeld, Anja; Callhoff, Johanna (2021): Comorbidity and healthcare utilisation in persons with incident systemic lupus erythematosus followed for 3 years after diagnosis: analysis of a claims data cohort. In: *Lupus science & medicine* 8 (1). DOI: 10.1136/lupus-2021-000526.

Arjovsky, Martin; Chintala, Soumith; Bottou, Léon (2017): Wasserstein GAN. Online verfügbar unter <https://arxiv.org/pdf/1701.07875.pdf>.

Bullward, Alistair; Aljebreen, Abdulaziz; Coles, Alexander; McInerney, Ciarán; Johnson, Owen (2023): Research Paper: Process Mining and Synthetic Health Data: Reflections and Lessons Learnt. In: Marco Montali, Arik Senderovich und Matthias Weidlich (Hg.): Process Mining Workshops. ICPM 2022 International Workshops, Bozen-Bolzano, Italy, October 23–28, 2022, Revised Selected Papers, Bd. 468. 1st ed. 2023. Cham: Springer Nature Switzerland; Imprint Springer (Lecture Notes in Business Information Processing, 468), S. 341–353.

Giomi, Matteo; Boenisch, Franziska; Wehmeyer, Christoph; Tasnádi, Borbála (2023): A Unified Framework for Quantifying Privacy Risk in synthetic data. In: *PoPETs* 2023 (2), S. 312–328. DOI: 10.56553/popets-2023-0055.

Health Data Insight CiC: The Simulacrum. A synthetic dataset developed by Health Data Insight CiC and derived from anonymous cancer data provided by the National Disease Registration Service, NHS England. Online available under <https://simulacrum.healthdatainsight.org.uk>, last checked on 2023-10-05.

Helfer, Sven; Kümmel, Michéle; Bathelt, Franziska; Sedlmayr, Martin (2021): Generating Enriched Synthetic German Hospital Claims Data - A Use Case Driven Approach. In: *Studies in health technology and informatics* 278, S. 58–65. DOI: 10.3233/SHTI210051.

Katz, Slava; Karaca, Zeynal; Wong, Weihuang; Hesgrove, Brandon; Wong, Herbert; Jang, Don; Resnick, Dean (2021): Using National synthetic data to Conduct Health Services Research. In: *Health Services Research* 56 (S2), S. 26. DOI: 10.1111/1475-6773.13754.

Kaur, Dhamanpreet; Sobiesk, Matthew; Patil, Shubham; Liu, Jin; Bhagat, Puran; Gupta, Amar; Markuzon, Natasha (2021): Application of Bayesian networks to generate synthetic health data. In: *Journal of the American Medical Informatics Association : JAMIA* 28 (4), S. 801–811. DOI: 10.1093/jamia/ocaa303.

Lu, Chang; Reddy, Chandan K.; Wang, Ping; Nie, Dong; Ning, Yue (2023): Multi-Label Clinical Time-Series Generation via Conditional GAN. In: *IEEE Trans. Knowl. Data Eng.*, S. 1–13. DOI: 10.1109/TKDE.2023.3310909.

Miyato, Takeru; Kataoka, Toshiki; Koyama, Masanori; Yoshida, Yuichi (2018): Spectral Normalization for Generative Adversarial Networks.

Quan, Hude; Sundararajan, Vijaya; Halfon, Patricia; Fong, Andrew; Burnand, Bernard; Luthi, Jean-Christophe et al. (2005): Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. In: *Medical care* 43 (11), S. 1130–1139. DOI: 10.1097/01.mlr.0000182534.19832.83.

Ständer, Sonja; Ketz, Miriam; Kossack, Nils; Akumo, Divine; Pignot, Marc; Gabriel, Sylvie; Chavda, Rajeev (2020): Epidemiology of Prurigo Nodularis compared with Psoriasis in Germany: A Claims Database Analysis. In: *Acta dermato-venereologica 100* (18), adv00309. DOI: 10.2340/00015555-3655.

Tucker, Allan; Wang, Zhenchen; Rotalinti, Ylenia; Myles, Puja (2020): Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. In: *NPJ digital medicine* 3 (1), S. 147. DOI: 10.1038/s41746-020-00353-9.

Walonoski, Jason; Kramer, Mark; Nichols, Joseph; Quina, Andre; Moesel, Chris; Hall, Dylan et al. (2018): Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. In: *Journal of the American Medical Informatics Association : JAMIA* 25 (3), S. 230–238. DOI: 10.1093/jamia/ocx079.

Wang, Zhenchen; Myles, Puja; Tucker, Allan (2021): Generating and evaluating cross‐sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. In: *Computational Intelligence* 37 (2), S. 819–851. DOI: 10.1111/coin.12427.

Yoon, Jinsung; Mizrahi, Michel; Ghalaty, Nahid; Jarvinen, Thomas; Ravi, Ashwin; Brune, Peter et al. (2022): EHR-Safe: Generating High-Fidelity and Privacy-Preserving Synthetic Electronic Health Records.

# Appendices

## Variables in the Training Dataset

Table 4: Variables in the training dataset

|  |  |
| --- | --- |
| Type | Included Information |
| Patient Characteristics | * Year of birth * Gender * 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 * DDD |
| Inpatient | * Date and cause of admission & discharge * Inpatient or outpatient treatment indicator * 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 * 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) |

## Statistical Measures to Assess Privacy

Table 5: Statistical Measures to Assess Privacy

|  |  |  |
| --- | --- | --- |
| **Category** | **Assessment** | **Tests** |
| **Duplicate Records** | Distance Based | * Gowers distance * Hamming distance * Euclidean distance * Mahalanobis distance |
| **Robustness to Privacy Attacks** | Singling Out | * Machine Learning * k-anonymity |
|  | Linkability | * Machine Learning * l-diversity |
|  | Inference | * Machine Learning * t-closeness |
| **Shareability** | Differential Privacy | * Specification of Privacy Budget (ε) * Operational Parameters * Noise distribution * Sensitivity Settings |

## Statistical Measures to Assess Fidelity

Table 6: Statistical Measures to Assess Fidelity

|  |  |  |
| --- | --- | --- |
| **Category** | **Assessment** | **Tests** |
| **Distributional Closeness** | Distributional Centers | * t-test * Wilcoxon rank-sum * Fisher’s-exact * Chi² |
|  | Differences / Distances | * SMD / Cohen’s d * Total Variation Distance * Maximum Mean Discrepancy * Contingency Similarity |
|  | Distributional Resemblance Tests | * Kolmogorov-Smirnov test * Wasserstein Distances * Jensen-Shannon Divergence |
|  | Heavy Tails and Extreme Values | * Hill’s Estimator * Generalized Extreme Value Distribution |
|  | Bivariate Relationships | * Spearman’s Rank Correlation * Bivariate Kolmogorov-Smirnov Test * Bivariate Kernel Density Estimation * Cross-correlation |
| **High Dimensional Dependency** | Dimensional Reduction | * PCA * t-SNA |
|  | Discriminator Models | * Random Forrest * XGBoost * Decision Trees * Neural Networks |
| **Temporal Consistency** | Statistical Methods | * Pearson pairwise correlation |
|  | Discriminator Models | * Neural Networks |

# Attachments

# Signatures

|  |  |  |  |
| --- | --- | --- | --- |
| SPONSOR Signature(s)Please remove this page if not applicable | | | |
|  | | | |
|  | | | |
| <<Study Description>> | | | |
|  | | | |
|  | | | |
| *<<This Non-Interventional, Secondary Data Study Protocol >> <<has/have>> been subjected to an internal Sponsor review>>*  I agree to the terms of this Study protocol. | | | |
| Sponsor representative |  |  |  |
|  | *<<Name, title>>*  *<<Email address and telephone number>>* | | Date (Day Month Year) |
|  |  | | |
| This document contains confidential information, which should not be copied, referred to, released or published without written approval from Sponsor. Investigators are cautioned that the information in this protocol may be subject to change and revision. | | | |