SYNTHO

Synthetic Data in HealthCare Report

Cedars Sinai · Centraal Bureau voor de Statistiek · Erasmus MC · KVK · LUMC Leiden University Medical Center · lifelines · NS · umcg

Table of content

Introduction

Healthcare organizations rely heavily on data to make evidence-based medical decisions, tailor treatments to individual patients, and drive medical research, all of which contribute to better patient outcomes, operational efficiency, and advancements in medical knowledge and technology.

Synthetic data offers a valuable solution for healthcare organizations, ensuring data privacy while still enabling the generation of realistic and non-sensitive datasets. Synthetic data empowers researchers, clinicians, and data scientists to innovate, validate algorithms, and conduct analysis without compromising patient privacy. Given the complex and constantly changing regulatory landscape in healthcare, artificial intelligence has become increasingly appealing to these organizations as a means to overcome these challenges effectively.

$67.4bn

expected AI Healthcare market value by 2027

95%

identity theft cases specifically target health records

85%

of machine learning models don't make it into production

What do the privacy regulators say about synthetic data?

Under the **EU General Data Protection Regulation**, data synthesis would benefit from the same reasoning as the case for **not needing a legal basis for anonymization**.

EU Data Protection Work Party IAPP 1 & 2

Under the **U.S. Health Insurance Portability and Accountability Act**, synthesis would similarly be deemed to be a form of treatment, planning and operations and therefore does **not require additional authorization by a covered entity**.

US Department of Health & Services HHS

EU DPAs are open **to labeling synthetic data anonymous** under data protection standards.

A guide to the EU's unclear anonymization standards (2021)

Synthetic data in Healthcare

Healthcare organizations and the role of data

Healthcare organizations' data usage is important as it enables evidence-based medical decisions, personalized treatments, and medical research, ultimately leading to enhanced patient outcomes, improved operational efficiency, and advancements in medical knowledge and technologies. Synthetic data can significantly benefit healthcare organizations by providing privacy-preserving alternatives. It enables the creation of realistic and non-sensitive datasets, empowering researchers, clinicians, and data scientists to innovate, validate algorithms, and conduct analysis without compromising patient privacy.

77%

Of financial institutions afraid of losing competition without leveraging Big Data

60%

consumers lack sufficient access to patient data

95%

identity theft cases specifically target health records

60%

healthcare IT will use AI for automation & decision-making by 2024

Hospitals

- Improve Patient Care
- Reduce the time required to access data
- Protect Personal Health Information (PHI) from the Electronic Health Record System (EHR , MHR)
- Increase data utilization and predictive analytics capabilities
- Address the lack of realistic data for software development and testing

Pharma & Life Sciences

- Share data and collaborate efficiently with health systems, payers, and related institutions to solve bigger problems faster
- Overcome data silos
- Perform studies and clinical trials to understand the drug product's impact (efficacy) on this new disease
- Complete a full analysis in less than a month, with less effort

Academic Research

- Accelerate the pace of data-driven research by providing the ability to access data faster and easier
- Access to more data for hypothesis evaluation
- Solution for generating and sharing data in support of precision healthcare
- Check project feasibility before submitting for original data access

Why do health organizations consider synthetic data?

Privacy-sensitive data

Health data is the most privacy-sensitive data with even stricter (privacy) regulations.

Urge to innovate with data

Data is a key resource for health innovation, as the health vertical is understaffed, and over-pressured with the potential to save lives.

Data quality

Anonymization techniques destroy data quality, while data accuracy is crucial in health (e.g. for academic research and clinical trials).

Data exchange

The potential of data as a result of collaborative data exchange between health organizations, health systems, drug developers, and researchers is enormous

Reduce costs

Healthcare organizations are under extreme pressure to reduce costs. This could be realized with being smarter with current resources, e.g. via analytics, for which data is needed.

Synthetic data is essential for addressing various challenges in data-driven fields:

Unlock data & valuable insights

Organizations today are collecting vast amounts of data. However, not all of it can be used as it is sensitive and contains personal information. Consequently, this data is “locked” and cannot simply be used. This is challenging as **data-driven tech is only as good as the data it can utilize**. This is where AI-generated synthetic data comes in.

Gain digital trust

Customers want to know that **their personal data is safe and secure** and that the organizations they do business with are transparent and honest. One way that companies can build digital trust is by using AI synthetic data.

Drive industry collaborations

Organizations are constantly looking for ways to collaborate and share data internally or maybe even externally to drive innovation and gain a competitive edge. However, **privacy concerns and data silos can make it difficult** to work with sensitive data across departments, companies, and industries.

Introduction

Synthetic data:

- Artificially generated data that **mimics the statistical characteristics and patterns of real-world data**
- Created using algorithms or models based on existing data, **without containing any actual information from individuals or entities**
- Synthetic data is commonly used in various fields, including machine learning, data analysis, and software testing, to enhance privacy, data security, and **overcome limitations in accessing or sharing real data**

Types of synthetic data generation

1. Fully AI-generated synthetic data

This type of synthetic data is built from scratch using ML algorithms. The machine learning model trains on actual data to **learn about the data's structure, patterns, and relationships**. Generative AI then uses this knowledge **to generate new data that closely resembles the original's statistical properties** (again, while making it unidentifiable).

2. Synthetic mock data

This synthetic data type refers to artificially created data that **imitate the structure and format of real data but doesn't necessarily reflect actual information**. It helps developers ensure their applications can handle various inputs and scenarios without using genuine, private, or sensitive data and, most importantly, without relying on real-world data. This practice is essential for testing functionality and refining software applications in a controlled and secure manner.

3. Rule-based synthetic data

Rule-based synthetic data is useful for **generating customized datasets based on predefined rules, constraints, and logic**.
This method provides flexibility by allowing users to configure data output according to specific business needs, adjusting parameters such as minimum, maximum, and average values.

SYNTHO

About Syntho

Syntho is revolutionizing the HealthCare industry with a Synthetic data platform.

About Syntho

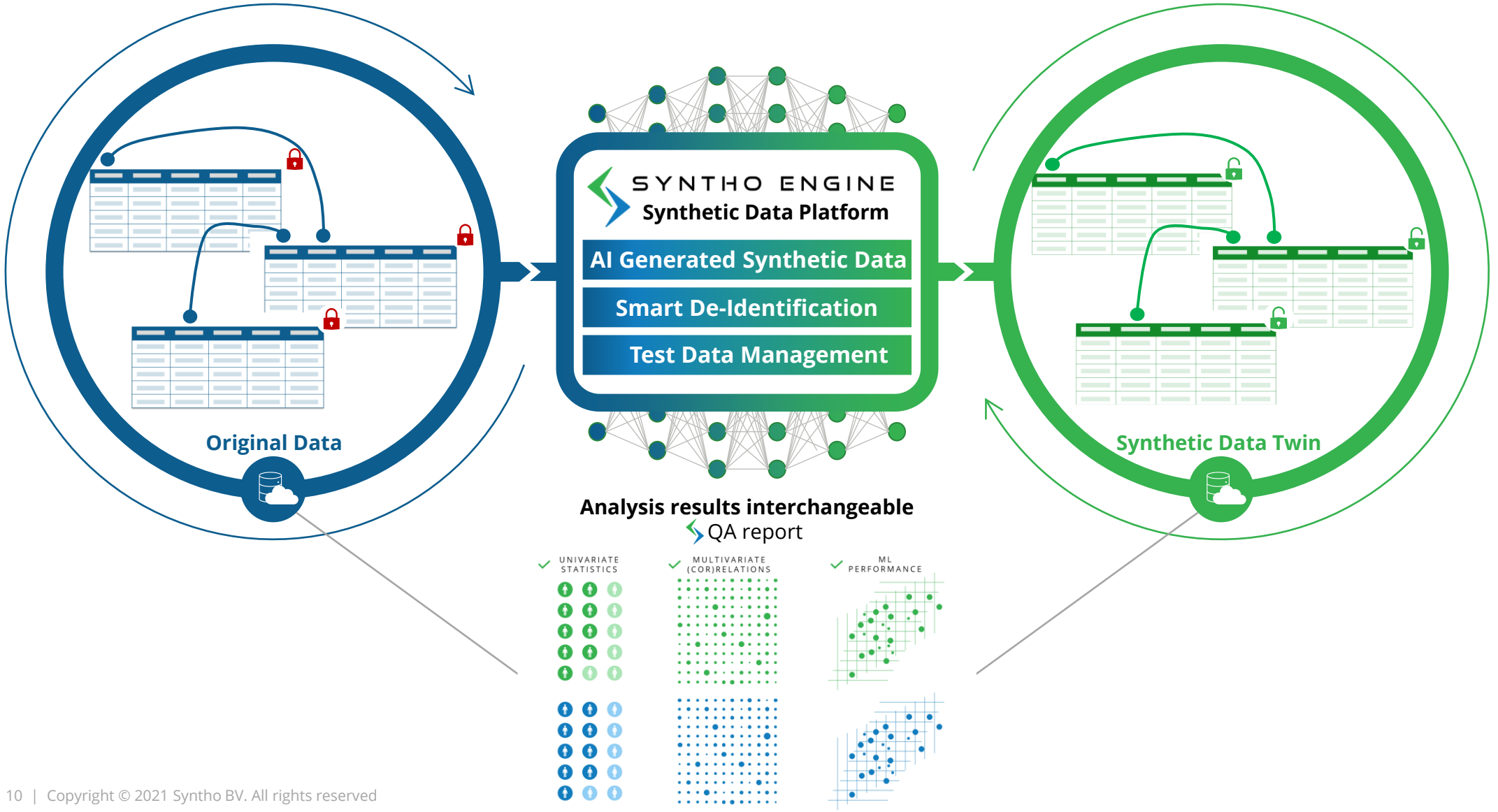
Founded in 2020, **Syntho** is the Amsterdam-based scaleup that is revolutionizing the tech industry with AI-generated synthetic data for analytics and test data management.

As a leading provider of synthetic data software, Syntho's mission is to empower businesses worldwide to generate and leverage high-quality Synthetic Data at scale.

Through our innovative solution, **The Syntho Engine**, we are accelerating the data revolution by unlocking privacy-sensitive data. Dramatically reducing the time to data and providing high-quality test data.

By doing so, we aim to foster an open data economy where information can be freely shared and utilized without compromises on privacy.

All Synthetic Data Generation approaches in one platform



Syntho solutions

AI Generated Synthetic Data

Mimic statistical patterns of original data in synthetic data with the power of artificial intelligence

- ✓ *Quality assurance report:* Assess generated synthetic data on accuracy, privacy, and speed
- ✓ *External evaluation by SAS:* our synthetic data is assessed and approved by the data experts at SAS
- ✓ *Time series synthetic data:* Synthesize time-series data accurately with Syntho

Smart De-Identification

Protect sensitive information by removing or modifying personally identifiable information (PII)

- ✓ *PII Scanner:* Identify PII automatically with our AI-powered PII Scanner
- ✓ *Synthetic mock data:* Substitute sensitive PII, PHI, and other identifiers
- ✓ *Consistent mapping:* Preserve referential integrity in an entire relational data ecosystem

Test Data Management

Create, maintain, and control representative test data for non-production environments

- ✓ *De-identification and synthetization:* generate test data used for comprehensive testing and development in representative scenarios
- ✓ *Rule-based synthetic data:* Generate synthetic data to mimic real-world or targeted scenarios using predefined rules and constraints
- ✓ *Subsetting:* Reduce records to create a smaller, representative subset of a relational database while maintaining referential integrity

Main use cases

Synthetic Data as Test Data

Using personal or original production data as test data is not allowed.

Our solution

By using synthetic test data, we can provide production-like synthetic test data.

Synthetic Data for Analytics

For many organizations, data cannot simply be used and shared.

Our solution

We help you gain easy and fast access to AI-generated synthetic data that is as good as real data.

Synthetic Data as Data Sharing

Data-sharing issues (i.e., legal delays, untapped valuable data, lack of a solid framework) cause project setbacks.

Our solution

Using synthetic data instead of real data reduces data-sharing obstacles.

Synthetic Data for Product demo

Your demo data may be suboptimal, leading to missed opportunities during product demonstrations.

Our solution

Impress your prospects with exceptional product demos featuring AI-generated synthetic data tailored to their needs.

Synthetic Data for Data monetization

Data monetization faces significant challenges, including ensuring data privacy and compliance, maintaining data quality and integrity, and implementing robust data governance practices.

Our solution

Synthetic data allows you to create a data marketplace with a data catalog and share it among internal and external stakeholders.

Synthetic Data for AI modelling

Ensuring equal access to high-quality data and mitigating biases to influence the accuracy and fairness of predictive outcomes.

Our solution

A synthetic data platform allows you to create high-quality synthetic data, which will capture all variables from the original data and allow you to upsample underrepresented data.

Clients and Recognitions

Client references

Cedars Sinai

lifelines

Centraal Bureau voor de Statistiek

Leids Universitair Medisch Centrum

Erasmus Universiteit Rotterdam

umcg

Erasmus MC University Medical Center Rotterdam

KVK

Industries

- HealthCare
- Public Organizations
- Financial Services

Awards

Philips Innovation Award

SAS Global Hackathon in Healthcare & Lifescience

Selected by NVIDIA as leading Generative AI scale-up

UNESCO Gender Bias challenge

SYNTHO

Case studies

This use case section highlights Syntho's current clients and their success stories.

Faster clinical research: Case Study

The use of synthetic data enables researchers to access sensitive data fast and perform quick prototyping. It allows researchers to efficiently organize and access data, explore new ideas, and uncover insights that drive research, ultimately leading to improved patient outcomes and significant healthcare innovation. By maximizing data utility while preserving patient privacy, synthetic data access broadens the availability of information to more people with fewer restrictions. This accelerates the discovery process from years to mere days or even hours.

Case Study 1

Predicting deterioration and mortality as part of cancer research on synthetic data

The goal

To predict deterioration and mortality as part of cancer research for a leading hospital. Synthetic data was used as algorithm training data.

The process

1. Generated Synthetic data out of real data, and ran a QA report to evaluate that all statistical patterns (e.g. correlations, AUC, and variable importance) were preserved

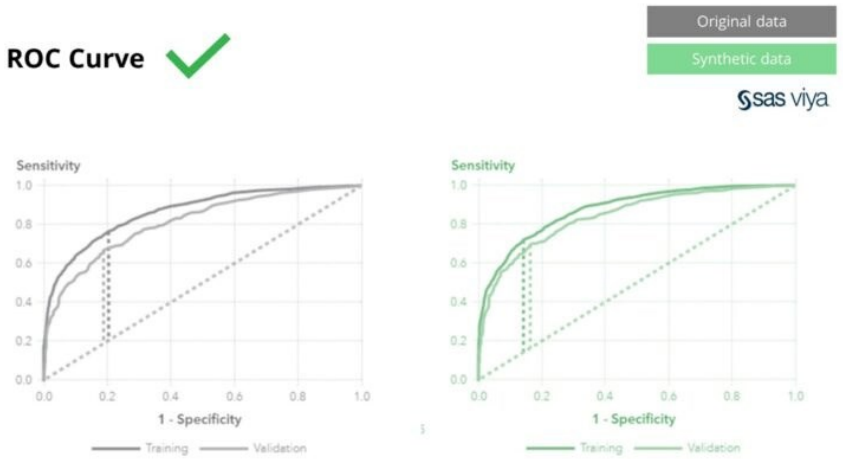
2. Trained the model on synthetic data and validated the results

3. Combined 2 synthetic data sets and evaluated model performance

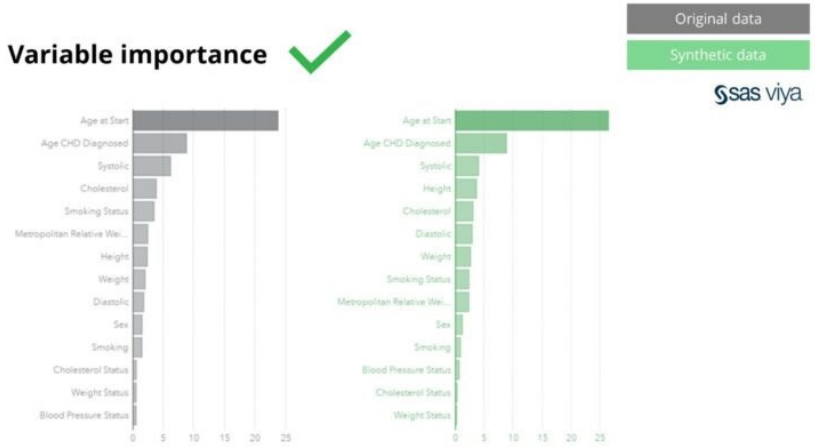
Faster clinical research: the accuracy of synthetic data is evaluated and demonstrated

Correlations ✓

ROC Curve ✓



Variable importance ✓



[Download Syntho's Quality Assurance Report]

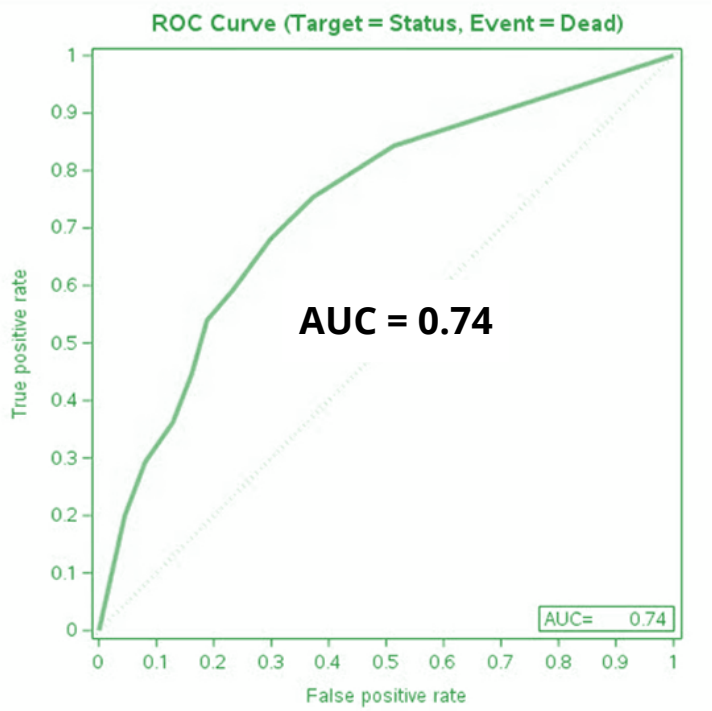
 |

Faster clinical research: Model performance

The application of synthetic data led to the development of a model capable of predicting deterioration and mortality, achieving an Area Under the Curve (AUC) of 0.74. Additionally, the combination of synthetic data from multiple hospitals resulted in a remarkable boost in predictive power from 0.74 to 0.78. These outcomes underscore the transformative potential of synthetic data in generating data-driven insights, the relevance of data availability and advancements within the field of healthcare.

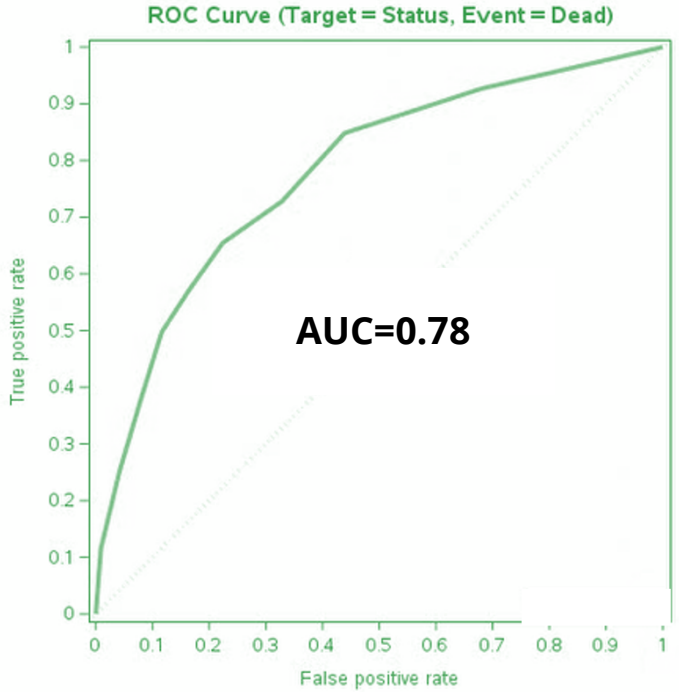
The result for one leading hospital

- An AUC of 0.74 and a model that is able to predict deterioration and mortality



The result for two hospitals:

- An AUC of 0.78, showing that more data results in better predictive power of those models



Faster clinical research: Results

Results, future steps, and implications

1. Syntho, a cutting-edge synthetic data generation tool was successfully used for the generation of synthetic data.
2. The accuracy of the synthetic data was thoroughly validated, as models trained on this data exhibited comparable scores to those trained on the original data.
3. This milestone furthered cancer research by enabling predictions of deterioration and mortality using synthetic data.
4. By combining synthetic data from two hospitals, an increase in the area under the curve (AUC) was realized from 0.74 to 0.78.

The next steps involve expanding collaborations with more hospitals, exploring diverse use cases, and extending the application of synthetic data across various sectors. With techniques that are sector-agnostic, we aim to unlock data and realize data-driven insights in healthcare.

Read more

Data availability for academic research: Case Study

Many organizations strive for data-driven innovation, where data plays a vital role and often requires sharing internally or externally with third parties. Without data, there can be no data-driven innovation or collaboration opportunities. A robust foundation for accessing and sharing relevant data is crucial for realizing data-driven innovation. However, accessing this data, whether by colleagues or third-party providers, presents challenges, and awareness of potential data-sharing solutions is limited. Synthetic data could be used to enhance data availability.

Case Study 2

Synthetic patient EHR data for advanced analytics with Erasmus MC

The goal

Healthcare data is the most privacy-sensitive data and is therefore locked. This privacy-sensitive data: is time-consuming to access, requires extensive paperwork, and cannot simply be used. This is challenging as data-driven insights could improve patient care and efficiency, potentially saving lives. Erasmus MC is an academic hospital and the medical faculty of Erasmus University Rotterdam, students engagement in the medical research project. The Smart Health Tech Center (SHTC) of Erasmus MC aims at the integration, development, testing, and validation of technologies for health, such as advanced AI-based technologies

The solution

The Smart Health Tech Center (SHTC) of Erasmus MC recently organized the official kick-off for synthetic data. At Erasmus MC, it will be possible to request synthetic data through the Research Suite.

Data sharing for academic research: Results

The benefits

1 Analytics with synthetic data

AI is used to model synthetic data in such a way that the statistical patterns, relationships, and characteristics are preserved to such an extent that the generated synthetic data can even be used for analytics. Especially in the model development phase, Erasmus MC will prefer using synthetic data and always challenge data users with the question: “Why use real data when you can use synthetic data?”

2 Enlarge data for testing purposes (upsampling)

By making smart use of generative AI in the creation of synthetic data, it is also possible to enlarge and simulate datasets, especially when there is insufficient data (data scarcity)

3 Get started faster

By using synthetic data as an alternative for real data, Erasmus MC can reduce risk assessments and similar time-consuming processes. Synthetic Data enables Erasmus MC to unlock data. Additionally, Erasmus MC can accelerate data access requests. Accordingly, Erasmus MC builds a strong foundation to accelerate data-driven innovation.

4 Enlarge data for testing purposes

Data augmentation techniques can be used to generate and simulate data that can be used for testing purposes.

Read more

Data distribution with synthetic data

Data distribution is a crucial component in modern research and development, particularly in domains that demand substantial quantities of high-quality data. Traditional data distribution methods often face hurdles, such as concerns over privacy, limited access to comprehensive datasets, and the logistical complexities of sharing sensitive or regulated data. These challenges can impede the flow of information necessary for innovative research and precise model development.

Case Study 3

Realistic synthetic patient data for clinical research with Lifelines

lifelines

The goal

Lifelines, conducts a multigenerational cohort study since 2006 with over 167,000 participants to collect relevant data and biosamples. This data is related to lifestyle, health, personality, BMI, blood pressure, cognitive abilities, and more. Lifelines offers this valuable data, making it an essential resource for national and international researchers, organizations, policymakers, and other stakeholders that typically focus on preventing, predicting, diagnosing, and treating diseases.

The solution

Lifelines partners with Syntho to synthesize data, thereby enhancing its accessibility and preserving the privacy of participants. As an alternative to using real data, everyone has now the possibility to work with synthetic data. Unlike open-source and other commercial solutions, Syntho's platform has consistently delivered better results in terms of accuracy, privacy, and usability, particularly when handling geographical location and longitudinal data, effectively demonstrating that synthetic data is as good as real.

Data distribution with synthetic data: Results

lifelines

The benefits

1 Faster access to data

Synthetic data allows for faster access to data by minimizing compliance paperwork and procedures. This enables data users for quicker analysis, faster hypothesis testing, and earlier results, without delays caused by compliance procedures.

2 Preserve the privacy of participants

By incorporating synthetic data, participant information remains secure, safeguarding their sensitive details effectively. Privacy-enhancing techniques, like synthetic data, improve confidence in participants that their data is protected, encouraging their active participation in research projects. This fosters trust in this biobank as a reliable and trusted resource, further accelerating participant engagement.

3 Increased accessibility to data

Synthetic data opens new possibilities for sharing information with organizations that might not prefer to access real data or might have access to minimal data. This approach allows for increased data accessibility while mitigating risks associated with sharing actual data.

4 Preview data before buying with a data catalog

With data commercialization, potential buyers often prefer to preview the data before making a purchase in something like a sandbox environment. However, using real data for previews becomes problematic due to compliance paperwork requirements and the risk of devaluing the data if exchanged beforehand. One could overcome these challenges by employing a synthetic data catalogue, allowing prospective buyers to preview data conveniently, thereby enhancing the commercialization process.

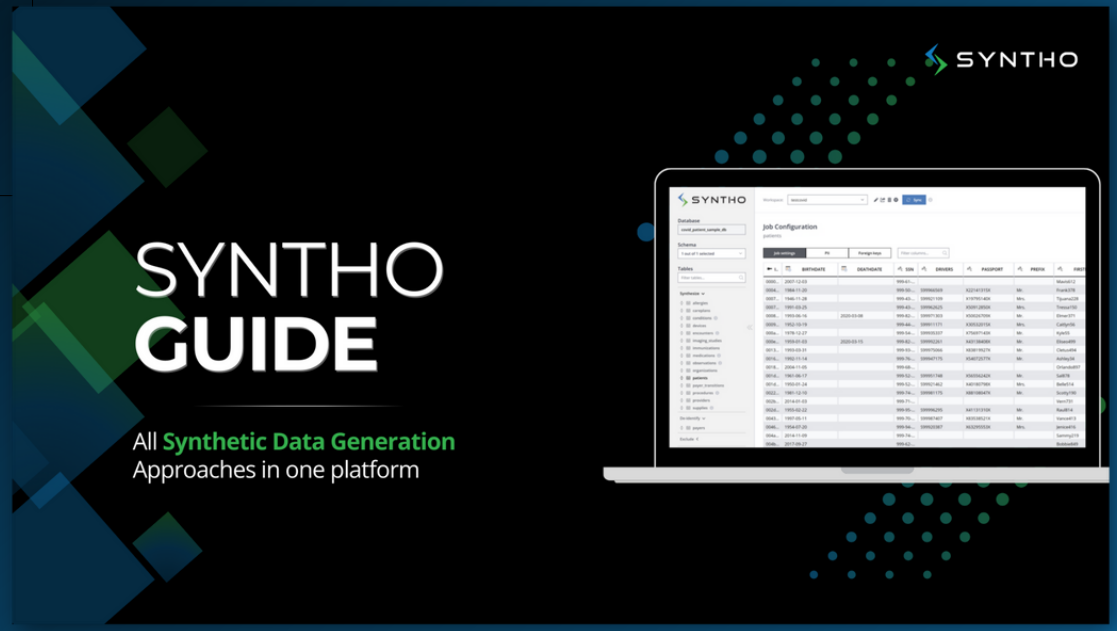
Read more

SYNTHO

Syntho Guide

- What is synthetic data
- How does it work
- Why do organizations use it
- How to get started

Access Syntho Guide



SYNTHO

Syntho's user documentation

- Getting started
- Deployment and connectors
- User interface
- Features
- User roles and support

Access User Documentation

Contact us

Wim Kees Janssen

CEO

Contact: kees@syntho.ai

Schedule meeting: syntho.ai/kees

Uliana Matyashovska

Business Development Executive

Contact: uliana@syntho.ai

Schedule meeting: syntho.ai/uliana