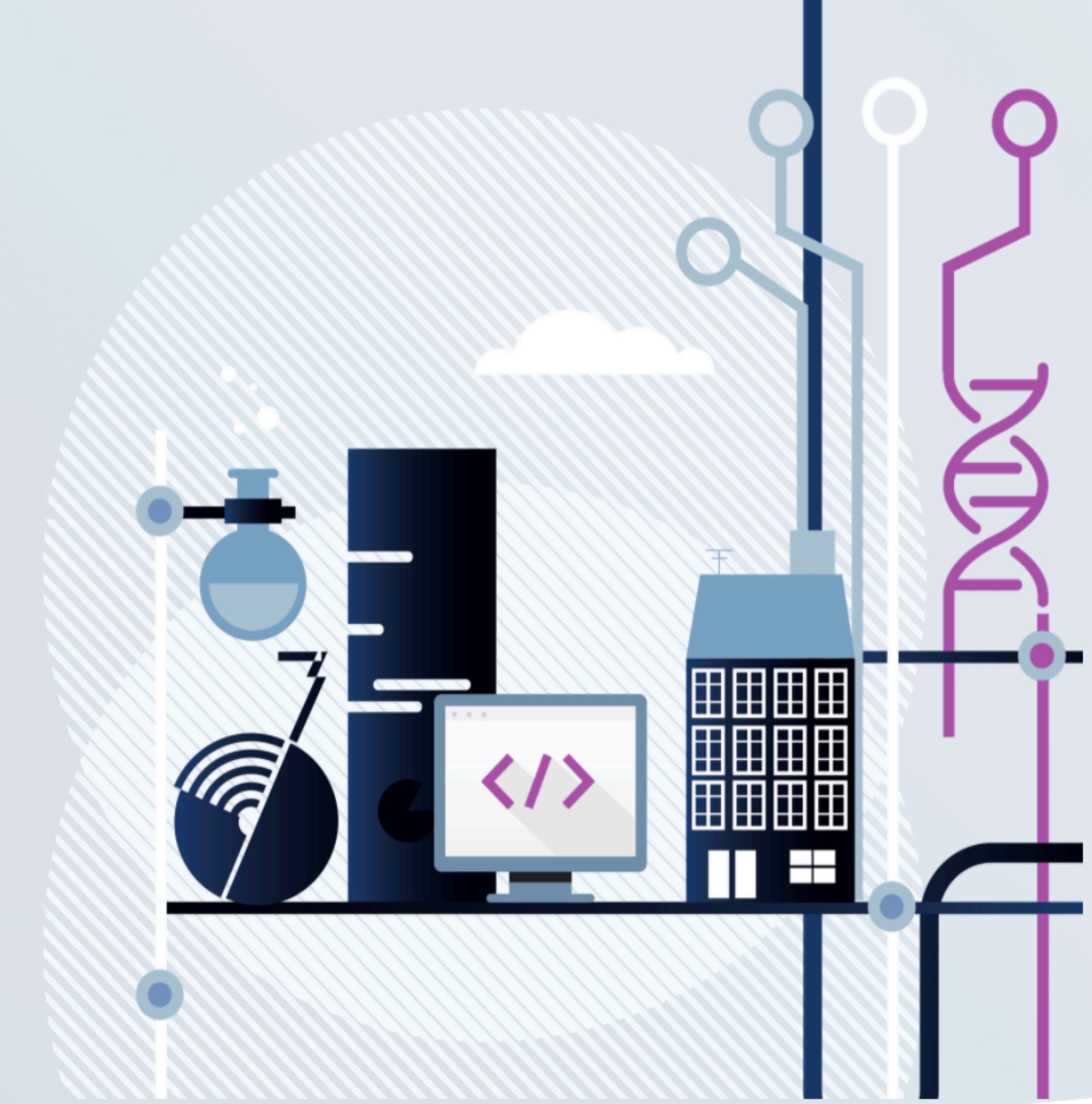
Deloitte. REGION H Rigshospitalet

Digital Hub Denmark

JANUARY 2021

Synthetic Health Data Hackathon

Summary and findings from a virtual hackathon conducted on November 27-29, 2020



Contents and purpose

'I definitely see the value in providing synthetic data for ***privacy*** *preservation, as well as* ***accelerating*** *development and* ***augmenting*** *scarce data sources.'*
– Hackathon participant

'Being able to generate synthetic genomes ***could break a lot of regulatory barriers*** *for genome centres around the world.'*
– Hackathon judge

Exploring the potentials of using synthetic health data to fuel innovation

Synthetic data is emerging as a promising technique that offers a way to accelerate research and innovation by enabling faster access to fictional yet useful data sets. Synthetically generated data does not entail the usual privacy risks. Instead of sharing sensitive data, a synthetic population is generated, which maintains most of the deep statistical properties of the real population.

Synthetic data is thus one of several techniques that aim to bridge the gap between privacy and utility. These techniques are maturing rapidly due to a combination of the general evolution of advanced analytics and the specific interest in finding secure ways to work with sensible data, not least health data.

However, these techniques share a common fate of being difficult to understand for people outside the field and hence of being difficult for authorities to regulate and for organisations to adopt. We need awareness and a better understanding of relevance, potential use cases and limitations.

Therefore, this first synthetic health data hackathon in Denmark was planned with an intention to explore value, use cases and limitations and to help raise awareness of the method and its possible role in the advancement of research and innovation in healthcare. Twenty-two teams encompassing 79 researchers and students from across the globe spent a weekend in November 2020 working on challenges related to diabetes and Alzheimer's based on synthetic data sets. The teams were mentored by a set of international experts within the fields of data science and bioinformatics.

The overarching finding was that synthetically generated data sets were considered a valuable new addition to the toolbox with multiple use cases. The evaluation resulted in a number of learnings and pointed towards next steps. We look forward to seeing the method evolve and demonstrate results in healthcare.

Henning Langberg, *Professor, Department of Public Health, Copenhagen University, Chief Innovation Officer, Rigshospitalet*

Thor Hvidbak, *Healthcare Client Relationship Executive, Deloitte*

Martin Closter Jespersen, *Senior Data Scientist, Deloitte*

Executive summary

Background

Health data is a resource for vital knowledge, and Denmark has some of the world's highest-quality health data. Secure access to work with these data is paramount for research and a key to unlocking innovative digital health solutions that can help improve diagnoses, treatments and ultimately patients' quality of life. However, as these data hold our most sensitive information, it is naturally strictly regulated and protected. Due to regulatory barriers and a complex actor ecosystem of data owners and custodians, health data are rarely shared across public and private organisations in other formats than in aggregated form with low utility.

Synthetic data is an emerging technology aiming to close the gap while preserving privacy, enabling sharing of data with high utility and with minimum risk of privacy loss. To raise awareness and investigate potential use cases, a virtual hackathon was hosted by Rigshospitalet in a partnership with Digital Hub Denmark and the project SHARED (https://shared.landen.co/). A total of 79 people with a wide range of backgrounds and nationalities participated in three different challenges: a diabetes bioinformatical track, a diabetes machine learning track and an Alzheimer's MRI images machine learning track.

Main findings

Participants showed a great interest in working with the synthetic data sets. The hackathon demonstrated the value of using synthetic data for educational purposes and idea generation.

The synthetic data sets proved to have a series of different potential uses, and some use cases are less vulnerable to limitations in data quality than others. Teams with domain knowledge quickly identified limitations in the diabetes synthetic data, as some expected correlations were lost in the synthetisation process. This loss highlights the importance of meticulously evaluating the synthetic data's quality to ensure the data provide the declared and expected value for researchers.

Participants also pointed to the value of using synthetic data in combination with the real data in order to increase data set size and thereby improving the performance of their artificial intelligence (AI) models.

Recommendations and future perspectives

For synthetic data to become integrated as a standard tool for supplement data or anonymizing sensitive health data to share data with researchers, pharma companies and across country borders, more awareness and documentation is needed – both for organizations to adopt and for regulators to regulate the method:

- For researchers, trust in the utility of the synthetic data is paramount. Therefore, sufficient automated documentation of the synthetic data's utility should be made a routine part of the data generation pipeline.
- To provide audit trails and transparency, documentation of the risk of privacy leaks as well as the generation process will also be needed. For this, methods to evaluate privacy risks of synthetic data should be developed together with regulators, hospital data stewards and experts within machine learning and data privacy.
- Explore how synthetic data can play a supporting role in with other privacy and data protection technologies
- Finally, we recommend evaluating on a larger scale how well synthetic data can replicate studies already conducted on the original data in order to fully understand synthetic data's full potential and current limitations.

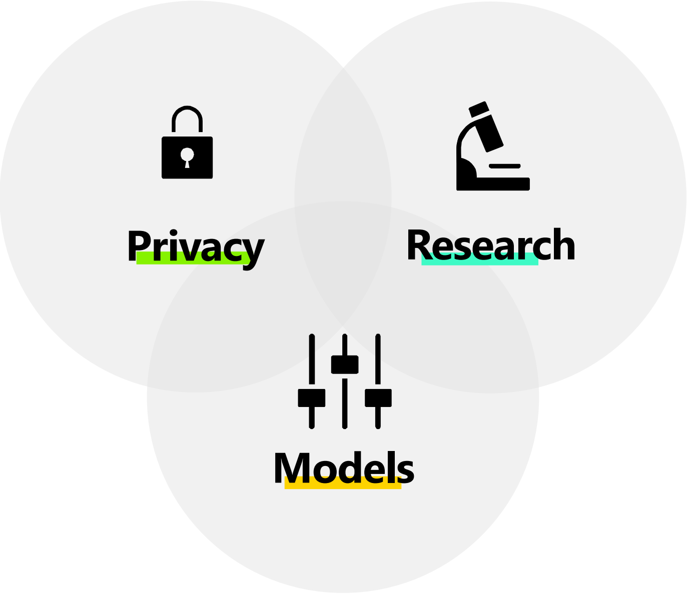
Synthetic data explained

What are synthetic data, how do you generate synthetic data and how do synthetic data provide value?

What if data could be shared freely between researchers?

Health data is a key enabler for improving the quality of care and for developing innovative digital health solutions. But health data is subject to strict privacy regulations – and with good reason. Synthetic data is emerging as a privacy technique that makes it possible to work more freely with data.

The advancement of AI in healthcare is beginning to unlock new potentials and can assist our health systems in responding to the major challenges they face[1]. But data is a key limiting factor, both in terms of data quality and data accessibility. Developing and improving AI models often require access to large, diverse and detailed sets of sensitive data and thus come at high risks in terms of privacy – and, consequently, strict regulations. It is essential for research and innovation to respect these regulations and safeguard privacy. Therefore, researchers are working to develop methods that make it possible to securely train and develop models on data without running the privacy-related risks of conventional methods. However, it is important to stress that there is no such thing as absolute anonymous data with high data utility, making it paramount to find the right balance between the risk of leaking sensitive data and the possibility of obtaining the right data utility. Synthetic data offer a vision where data utility and privacy preservation are better balanced, thus achieving high data utility with significant privacy guarantees.



Privacy

Research

Models

Strict privacy regulations

Rising restrictions on data rights make it very difficult to share sensitive patient data securely to promote research and accelerate innovation. Synthetic data provides very high privacy guarantees without sacrificing too much usefulness (as conventional anonymisation does).

Accelerate research

Getting access to real data can take many months and thereby slow down research and digital innovation. Large and complex data sets required for deep learning take even longer to obtain. There are thus obvious use cases for exploratory purposes, hypothesis building, model generation and practice while waiting for the real data.

More accurate AI models

Synthetic data enable off site analytics where a combination of multiple data sources could result in more accurate models. Combining data sources from different regions or countries allows researchers to build AI models which can take into account variance in biological profiles, measurement equipment and treatment strategies.

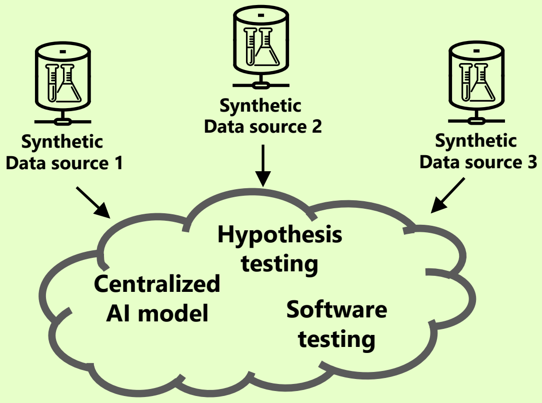
Synthetic data is one of multiple innovative privacy-preserving techniques

There is an emerging and rapidly maturing set of privacy-preserving techniques with different potential uses, all aimed at addressing the gap between privacy and utility.

The increasingly sharp focus on data privacy leads to more rigorous regulations. Getting access to real sensitive data in a secure manner can take a long time, slowing down both research and innovation of digital solutions. Identifying privacy-preserving techniques to share useful data is therefore key to reaccelerate innovation[2]. Three main techniques are presented below. A further layer of privacy can be added to the techniques through differential privacy, which blurs the real data by adding random noise.

Synthetic data

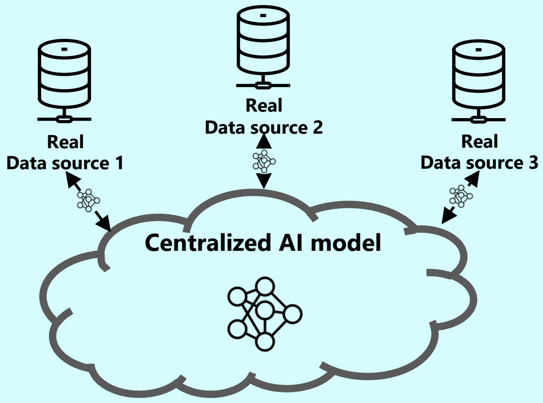
For many data sets, synthetic data can largely retain the statistical properties and thus be used for research and machine learning. Synthetic data has been widely used in machine learning where data were scarce. Synthetic data is often used for explorative analyses and testing but is limited by the technology's generally lower quality than real data.



Centralized platform

Federated learning

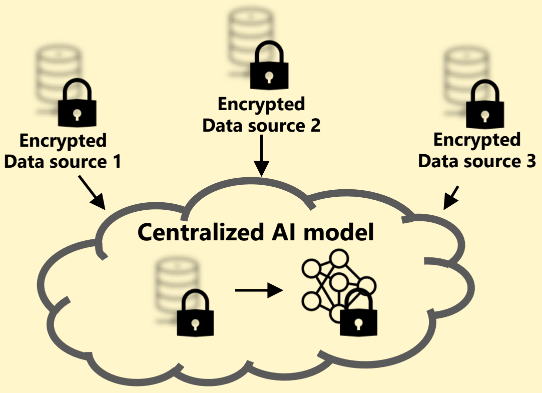
Federated learning is a way to train AI models on multiple data sources without any data leaving any source. It has been developed for autocorrect on phones by Google. NVIDIA has recently demonstrated its use in healthcare for COVID AI[3]. Federated learning is a secure way to work on the real data enabling high-quality AI without direct access to view the data. However, as data is not anonymised, it is still limited by consent or other data permits.



Centralized platform

Encrypted computations

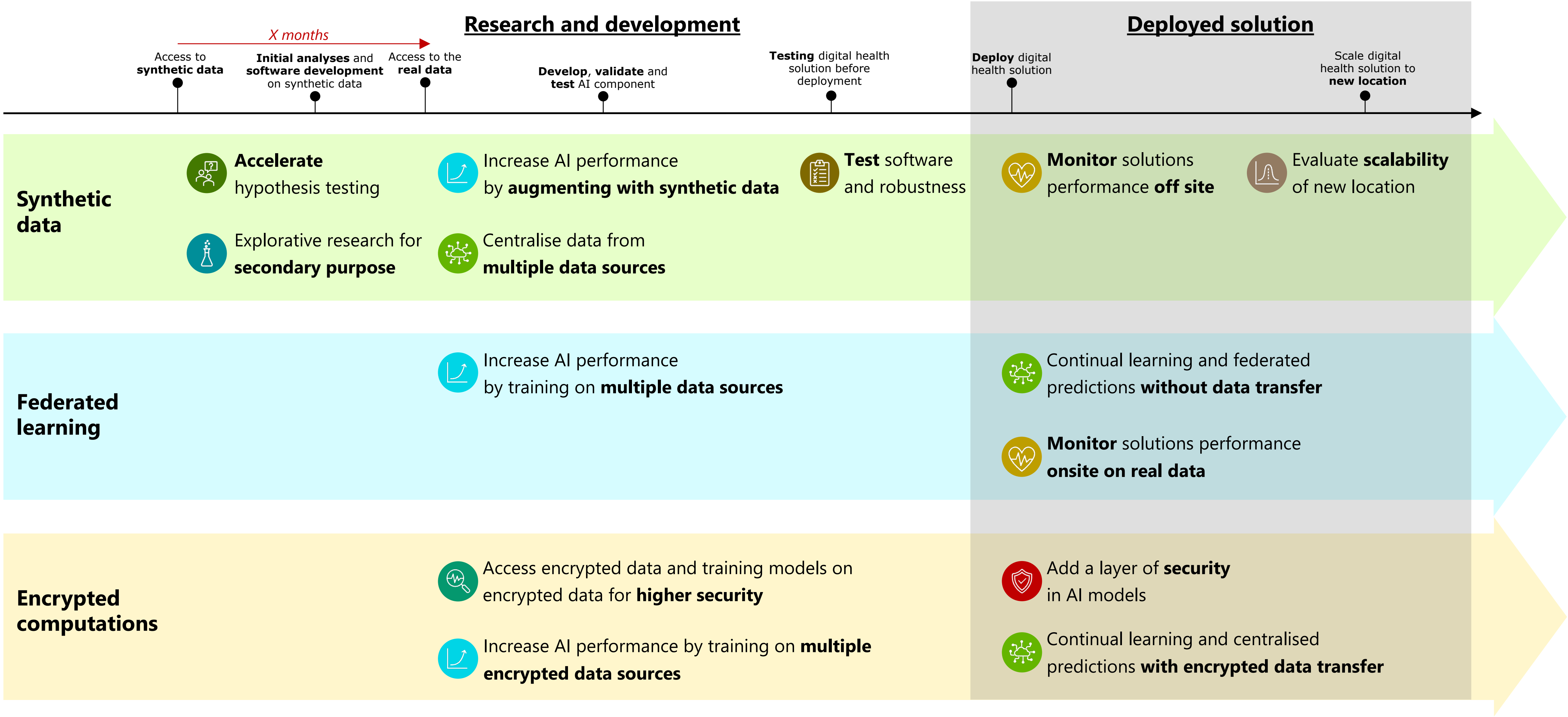
Encrypted computations allow scientists to do mathematical functions and AI on non-readable encrypted data. This technique has been implemented to train AI models to increase the protection of data. IBM is spearheading a lot of the research in this field. Encrypted machine learning enables training of AI models without access to look at the real data. Like federated learning, this technique is limited by the given consent (or other forms of data permits).



Centralized platform

A synergistic relationship between the three technologies

Synthetic data, federated learning and encrypted computations could work in synergy and by supporting different parts of the lifecycle.



Research and development

X months

Access to **synthetic data**

Initial analyses and **software development** on synthetic data

Access to the **real data**

Develop, validate and **test** AI component

Testing digital health solution before deployment

Deployed solution

Deploy digital health solution

Scale digital health solution to **new location**

Synthetic data

Accelerate hypothesis testing

Explorative research for **secondary purpose**

Increase AI performance by **augmenting with synthetic data**

Centralise data from **multiple data sources**

Test software and robustness

Monitor solutions performance **off site**

Evaluate **scalability** of new location

Federated learning

Increase AI performance by training on **multiple data sources**

Continual learning and federated predictions **without data transfer**

Monitor solutions performance **onsite on real data**

Encrypted computations

Access encrypted data and training models on encrypted data for **higher security**

Increase AI performance by training on **multiple encrypted data sources**

Add a layer of **security** in AI models

Continual learning and centralised predictions **with encrypted data transfer**

What is synthetic data?

Synthetic data simulates a given population rather than individuals – which is a key differentiator compared with conventional anonymisation techniques. Conventional anonymisation significantly lowers data quality and thus utility. In contrast, synthetic data aims to generate anonymised data sets without sacrificing utility.

Definition of synthetic data

Synthetic data can be generated in two main ways: Either by using real data to create a synthetic copy, or by using prior knowledge or aggregated statistics to simulate data, all requiring an algorithm using a source of randomness for the synthetic data generation[4].

Anonymisation of data

Due to privacy concerns and restrictions on data rights, synthetic data has gained increased focus as a way to anonymise sensitive data sets. The strength of synthetic data is that there is no connection between the synthetic patients and the individual real patients, making reidentification a much more complex task than with conventional anonymisation techniques. Additionally, synthetically generated data can often maintain most of the deep statistical properties that conventional anonymisation can not.

A few use cases enabled by synthetic data

- *Accelerate software development and research*
- *Centralise data from multiple data sources*
- *Increase AI performance by augmenting the real data with synthetic data*
- *Accelerate hypothesis testing*
- *Secondary purpose research as it is fully anonymised*
- *Test software pipelines*
- *Share data for educational purposes, hackathons, etc.*

We distinguish between four types of data

Real patient data

Jacob 23 years
Josefine 55 years
Jesper 67 years

The real patient data with sensitive health information.

Pseudonymisation

#1 23 years
#2 55 years
#3 77 years

Names and personal information removed or encrypted
Still subject to GDPR

Conventional anonymisation

Mean value 51±22 years

Aggregated information and statistics
**Excluded from GDPR*

Synthetic data

Morten 45 years
Stine 20 years
Thomas 74 years

Generated fake patient populations who simulate real ones
**Excluded from GDPR*

**) When constructed and evaluated correctly.*

How can useful synthetic data be created?

Synthetic data can be generated in many ways, both with and without available real data. The most promising and fastest-evolving fields are generative adversarial networks and variational autoencoders. A downside to these methods is that they are less straight forward to use than the two more simple approaches.

Statistical sampling

A simple way to generate synthetic data is by sampling from distributions observed in the real data. A downside is that correlations between variables can easily be lost. Public Health England has released a synthetic-data version of the national cancer registry called Simulacrum https://simulacrum.healthdatainsight.org.uk/.

Bayesian Networks

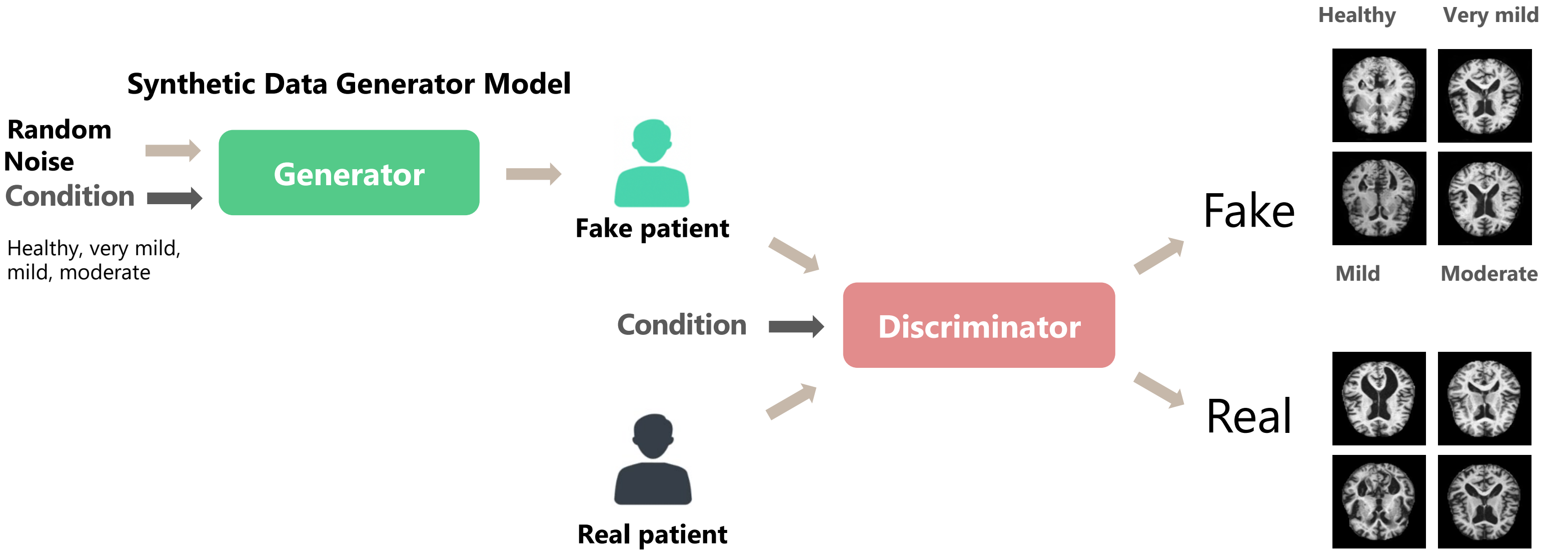
Bayesian networks is a strong approach to generate synthetic data as it models the relationship between variables[5]. However, it is not suitable for more complex data types, such as images.

Generative models

Novel **generative adversarial networks (GANs)** and **variational autoencoders (VAE)** are state-of-the-art generative methods for generating synthetic data, which exceed Bayesian approaches, and have proven to work for many data types, specifically images[6,7,8].

GANs

- GANs consist of two neural networks competing against each other, a **generator** and a **discriminator.**
- The generator's task is to fool the discriminator by generating fake data.
- The discriminator's task is to discriminate whether a data point is a real or fake (generated) one.
- The generator learns to generate very realistic data points which can be used as synthetic data.
- A GAN can be conditioned to generate specific type data points.



Maturity of the synthetic-data field

Synthetic data is still a very novel concept. The technology itself is maturing at pace and has now reached a level where many uses are viable. However, adapting technology to a business or organisational context takes time, iteration and learning. Moreover, regulation is lagging, stressing the need to explore use cases and raise awareness.

Research and technology

Tabular data

1 2 ●

Estimated Maturity

Synthetic-data generation of tabular data is relatively mature and can often generate data with nearly the same statistical properties as the real data[5,6]. The many variables constitute a possible limitation when synthetising tabular data if the data-set size is insufficient. Recent generative models, such as CTGAN, indicate very robust generation of tabular data and is available as open-source software.

Image data

1 2 ●

Image generation is the most thoroughly investigated field of generative models for synthetic images. These investigations have led to groundbreaking models for synthetising fake images even with limited amount of data and high resolution, such as StyleGAN2 ADA, which uses a residual outputs technique to achieve high-resolution images[7,8,9,10,11]. A downside to these complex models is that it can take weeks to train them.

Text, video and audio

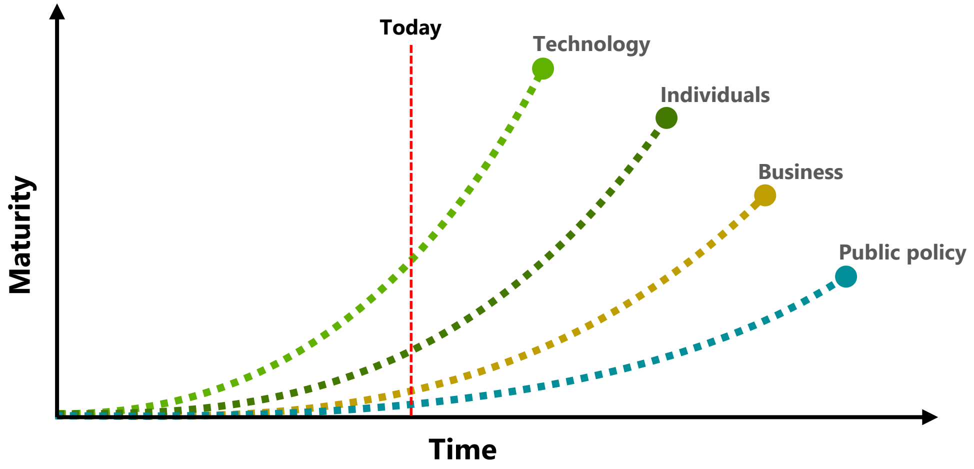
1 ● ●

More complex data types, such as text, video and audio, are also getting a lot of traction in research[12]. Danish AI company Corti.ai has an industrial PhD programme focused on researching generative models for uncertainty estimations and generation of audio and text. One of its recently published technologies is BIVA, a novel variational autoencoder which has also been tested for synthetic-text generation[13]. Corti.ai utilises anonymised transcriptions of medical phone conversations, transcribed by an automatic speech recognition model, to implement predictive machine learning models used in production for decision support.

Business, policy and regulation

Though synthetic data technologies are evolving fast, very few companies use them widely. Awareness is low, and regulations are still very immature. A comprehensive review of the regulatory aspects of synthetic data was recently conducted by Stanford Law School, highlighting the advantages, limitations and risks of using synthetic data to share sensitive data. The conclusion of the review was that synthetic data should not be considered risk-free, but also that this technology offers genuine progress and should be recommended as a valid solution to share useful private data without the otherwise common risks of direct reidentification[14].

Technology generally matures faster than organisational adoption, business use and public policy or regulation. Synthetic data has matured to a level where the regulatory aspects now need to be addressed for the method to be able to generate value outside limited settings.



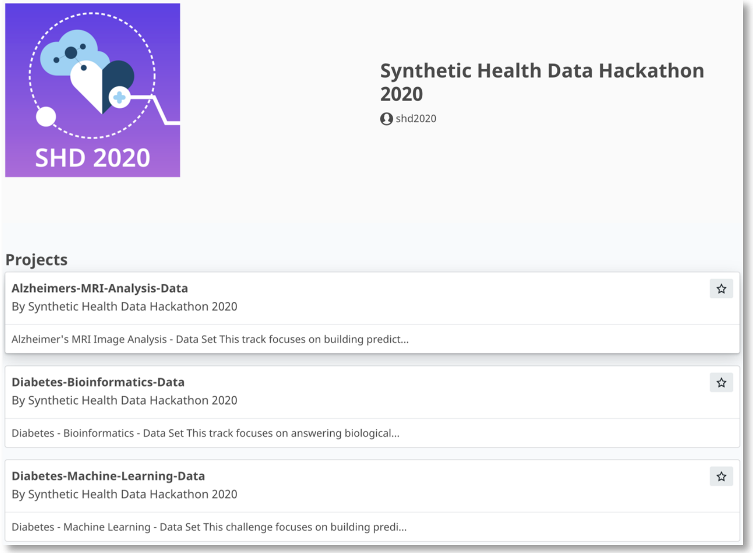
Source: 2017 Deloitte Global Human Capital trends

The hackathon

Participants' and judges' evaluations and survey results

What was the purpose of the hackathon?

The purpose of the hackathon was to bring together students, researchers, entrepreneurs and biopharma companies to work on new approaches to data related to diabetes and Alzheimer's. The hackathon was based on the exploration of synthetic data, including a comparing the use of synthetic patient data to the use of real patient data.



Synthetic Health Data Hackathon 2020

shd2020

Projects

Alzheimers-MRI-Analysis-Data
By Synthetic Health Data Hackathon 2020
Alzheimer's MRI Image Analysis - Data Set This track focuses on building predict...

Diabetes-Bioinformatics-Data
By Synthetic Health Data Hackathon 2020
Diabetes - Bioinformatics - Data Set This track focuses on answering biological...

Diabetes-Machine-Learning-Data
By Synthetic Health Data Hackathon 2020
Diabetes - Machine Learning - Data Set This challenge focuses on building predi...

SHARED

SHARED is an active research project exploring opportunities in, developing methods for and raising awareness of synthetic data in healthcare. SHARED is funded by Novo Nordisk Foundation.

Rigshospitalet

The hackathon was hosted by Rigshospitalet and its CIO, professor Henning Langberg.

Three challenges on diabetes and Alzheimer's

The teams could choose between three different challenges: The **first challenge** consisted of an Alzheimer's MRI image data set from Kaggle[15] (prepared by Deloitte) with a focus on building predictive machine learning models. The **second challenge** consisted of a diabetes data set from UCI[16] (prepared by Helsinki University) that focused on answering biological questions by identifying patterns in the data. The **third challenge** consisted of a diabetes data set with a focus on building predictive machine learning models. Both the diabetes data set and Alzheimer's MRI images were generated using conditional GANs.

A diverse set of participants

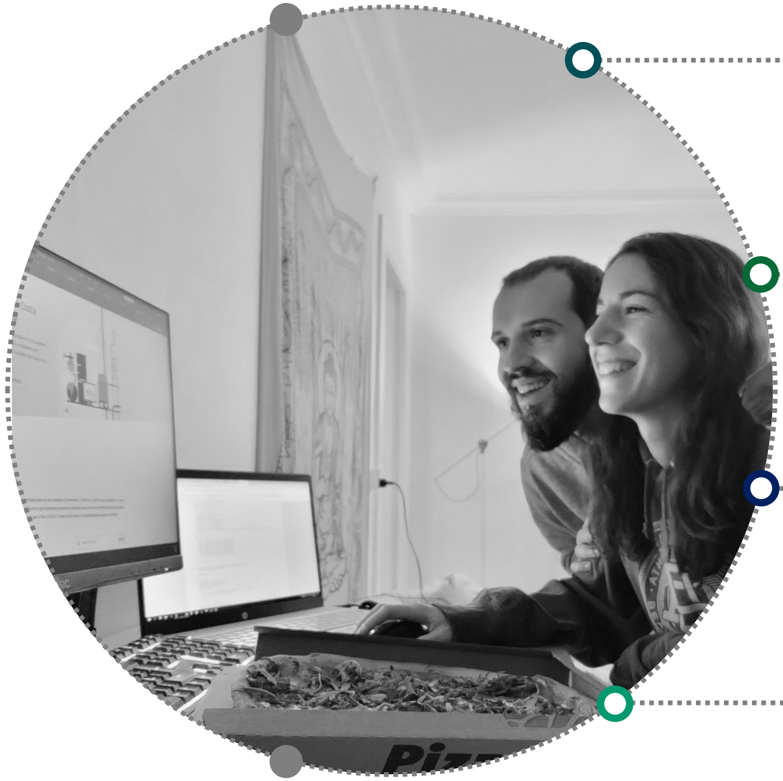
The event brought together students, researchers, entrepreneurs and biopharma companies to work on new approaches to data related to diabetes and Alzheimer's. Event attendees were characterised by possessing different levels of programming and experience.

biolib

BioLib is a Danish bioinformatics startup with a mission to make analysis of biological data secure and accessible. BioLib's digital platform empowers researchers to build, share and run bioinformatics applications and algorithms in a way that is accessible and ensures the protection of sensitive biomedical data when working collaboratively across organisations. BioLib provided the platform for our hackathon. Read more about the BioLib at https://biolib.com/.

The hackathon was evaluated with participants, organisers and judges

The hackathon Synthetic health data 2020 took place on the weekend of 27-29 November. This summary report of the hackathon rests on four main data sources: a survey, two sets of interviews and participant presentations .



Participant survey

A survey was sent to the **79 participants** after the hackathon of which **36 responded**. Results are presented on the following page.

Interviews with participants

Participants representing each of the three challenges were interviewed after the hackathon. The interviews focused on the relevance of using synthetic data compared to using real data.

Interviews with judges & organizers

After facilitating the hackathon, judges and organisers were asked to offer their reflections on the value of synthetic data and on the hackathon as a format.

Teams & participant presentations

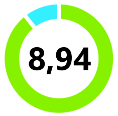
Each of the participating teams were asked to create a presentation with the findings from their work on the selected challenge.

Survey responses

Survey scores and selected quotes from participant survey

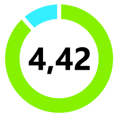
'In general, how happy are you that you attended the hackathon?'

Score 0-10, average:



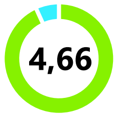
'How good/bad did you find: the process for finding a team?'

Score 0-5, average:



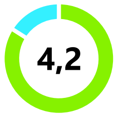
'How good/bad did you find: the communication from the organisers before and during the hackathon?'

Score 0-5, average:



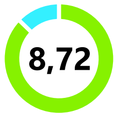
'How good/bad did you find: the adviser sessions?'

Score 0-5, average:



'Overall, did you like this challenge?'

Score 0-10, average:



How did you find out about the hackathon?	22 Friend or colleague	8 Facebook	4 LinkedIn	2 Twitter	2 Mailing list	2 Other

Experience with working with synthetic data compared to working with real data	"Synthetic data is a good tool **however it cannot 100% mimic** the real data and can be sometimes bit confusing."	"It is **challenging to generate** synthetic data which can be used to represent the exact distribution of real data."	"It seemed that the model trained on the synthetic data did not carry over to the real data as strongly as one could have hoped. Otherwise, it shows great potential regarding data privacy, and **I hope it becomes viable soon**."	"I think that synthetic data has **more simple structure** comparing to real data. Moreover, the synthetic MRI data didn't represent tiny elements of the brain structure."
Do you think synthetic data could be useful as a way to allow people to more freely share and work with health data?	"Yes. But we need **better data** than what we got"	"Many machine learning professionals are quick to arrive at conclusions without much **understanding of the medical nature** of the data."	"Yes. In the field of biomedical sciences, to **obtain real data is one of the most challenging things**. And here, synthetic data come into play."	"Yes, but synthetic data is **only as good as the data used** to generate it. We need to be careful!"
Praise and criticism	"**Not** enough time" "**More** descriptive dataset" "**Well** organized"	"**Learning** by doing" "**Challenging** task" "**Getting** to know colleagues"	"**Missing** meaningful discussion and conclusion between the participants" "**Learning** about synthetic data and its use"	'**Virtual** hackathon – hard to socialise and stay motivated' "**Experience** how others work"

Interviews with participants

Selected extracts from the participant interviews

Vajira
Representing the group TeamSyntheticMRI

'Access to synthetic data **while waiting for the real** data provides valuable insights.'

'It was a really **well organised** event with a good flow. However, with only two-three days, it is not easy to produce a good result.'

Ella
Representing the group TEMB

'I can see value in the ability to begin the analysis before the real data is available. However, one should be **aware of flaws in the data set:** The synthetic data did not have the same attributes as real data – and without a higher educational background in biological, you would not be able to spot this difference.'

'I have become **motivated** to become better at working with data.'

Marius
Representing the group Tandioli

'I definitely see the value in providing synthetic data for privacy preservation, as well as **accelerating development and augmenting scarce data** sources.'

'I learnt a lot about health data and **how important domain knowledge is** in order to validate the quality of the data.'

Interviews with judges and organisers

Selected extracts from the judge and organiser interviews

Lasse Westergaard Folkersen
Genetics expert at the Danish National Genome Center

‘Though I have worked a lot with data, I was probably among those who learnt the most about synthetic data. I really see the usability of it and **hope to see synthetic genomic data soon**.’

‘The ability to generate synthetic genomes **could break a lot of regulatory barriers** for genome centres around the world.’

Lars Lau Raket
Data science lead specialist at Lundbeck

‘It was interesting to see the different **innovative ideas**, and I learnt new things with regard to synthetic data.’

‘In clinical trials, the ability to generate **synthetic control groups** could be extremely valuable.’

‘Synthetic data can **enrich data** by imputing missing variables or increasing the amount of data.’

Professor Henning Langberg
CIO at Rigshospitalet

‘I was impressed by the **quality of what the participants achieved** in such as short time.’

‘Providing Danish health data as synthetic data could be used as a showcase to **attract research and industry to Denmark**.’

‘It was interesting to see that synthetic data could be used to **supplement the real data** to improve AI predictive performance.’

Deloitte. Rigshospitalet

Teams and participant presentations

79 participants were divided into 22 teams and one of three challenges. The participants had diverse backgrounds, encompassing biology and software engineering.

🏆 Alzheimer's MRI panel winner

🏆 Diabetes Machine Learning panel winner

🏆 Diabetes Bioinformatics panel winner

❤ People's choice

Aryubia 🏆 ❤
Diabetes – Machine Learning
Alzheimer's – MRI analysis

Spaghetti-Vector-Monster 🏆
Diabetes – Machine Learning
Diabetes – Bioinformatics

BitsPlease 🏆
Diabetes – Bioinformatics

Candy-Crush
Diabetes – Bioinformatics

Pandavas
Diabetes – Machine Learning
Diabetes – Bioinformatics

Gardariki-Hack
Diabetes – Machine Learning
Alzheimer's – MRI analysis

DJANGO-UNCHAINED
Diabetes – Machine Learning
Diabetes – Bioinformatics

TEMB
Diabetes – Bioinformatics

TeamSyntheticMRI
Alzheimer's – MRI analysis

PanicTeam
Alzheimer's – MRI analysis

The-dia-beaters
Diabetes – Machine Learning
Diabetes – Bioinformatics

Data-Wizards
Diabetes – Machine Learning

12monkeys
Diabetes – Machine Learning
Diabetes – Bioinformatics

Tandioli
Diabetes – Machine Learning
Diabetes – Bioinformatics

D-E-E-P
Diabetes – Machine Learning
Diabetes – Bioinformatics

Quarantine-Statistics
Diabetes – Machine Learning
Diabetes – Bioinformatics

Biohackers
Diabetes – Machine Learning
Diabetes – Bioinformatics

The-Classyfiers
Diabetes – Machine Learning

yellowBugs
Diabetes – Machine Learning
Diabetes – Bioinformatics

Team-Einstein
Diabetes – Machine Learning
Diabetes – Bioinformatics

chotbullar-warriors
Diabetes – Bioinformatics

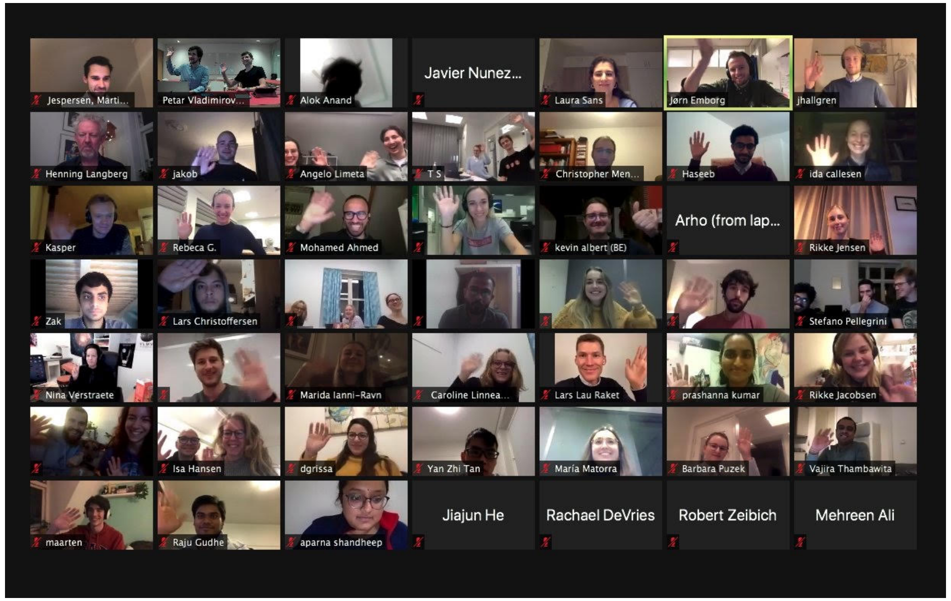
Camel-socks
Diabetes – Machine Learning
Diabetes – Bioinformatics

Introductory talks on synthetic data

By **Arho Virkki**, chief data officer at Turku University Hospital, and **Martin Jespersen**, PhD and machine learning expert at Deloitte, can be found here:
https://www.youtube.com/watch?v=n3CIQyoO2wg

Participant presentations

All presentations can be accessed here:
https://drive.google.com/drive/folders/1uCQInwfGJhZR6BM3eY7bhOWsxICpGujK



Findings and perspectives

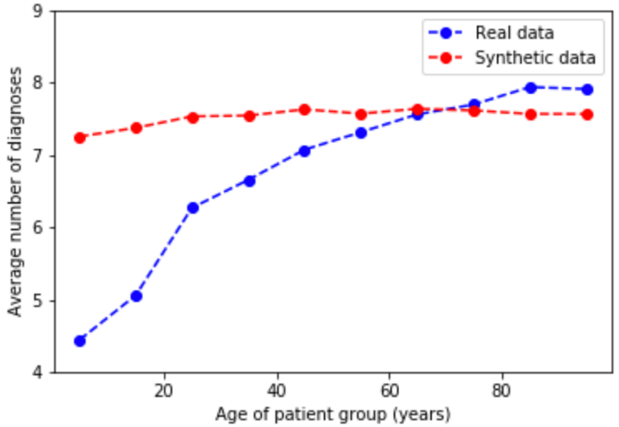
Learnings, findings and perspectives

Diabetes challenges: summary of participants’ findings

The synthetic diabetes data set was a tabular data set generated using a generative adversarial neural network. Two streams were defined for these data: develop AI models and do bioinformatics to investigate the biological relevance of the data.

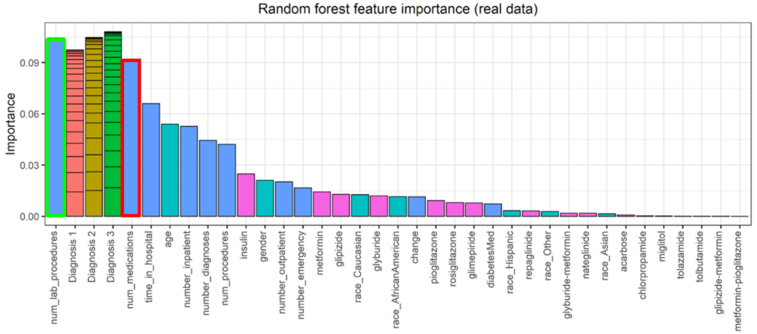
1. Data distribution

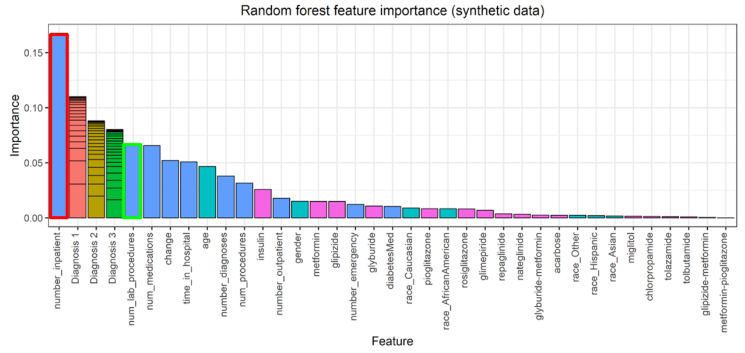
The data simulated independent data distributions quite well, however, correlations across variables and deep correlations were very limited. This indicated that the generative model used to generate the synthetic data could be improved by new methods or possibly longer training.



2. Biology

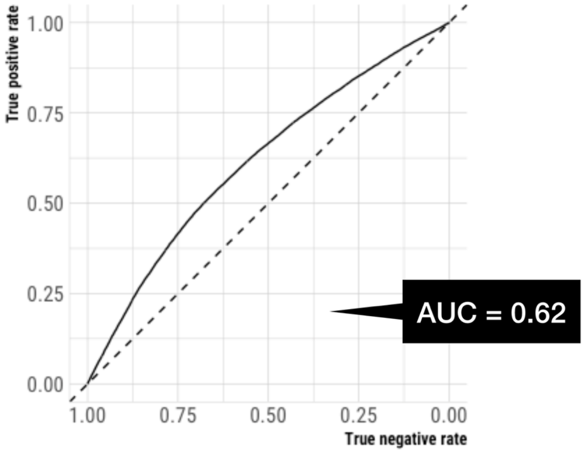
Models built on synthetic and real data, respectively, used the same variables to make the predictions, however, with varying importance. The real data showed a higher degree of biological relevance in its top-prioritised variables, indicating some loss of statistical important relationships.





3. Useability

As the quality of the data was relatively low, the predictive power of the synthetic data suffered when predicting the real data. A new synthetic data set of higher quality will be needed in order to build more robust and trustworthy AI models from these synthetic data.



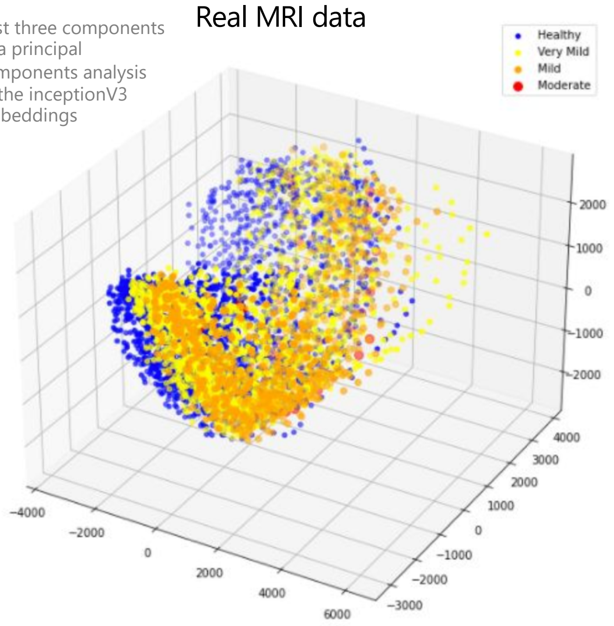
Alzheimer’s MRI challenge: Summary of participants findings

The synthetic Alzheimer’s MRI images were generated using a custom conditional generative adversarial network.

1. Data distribution

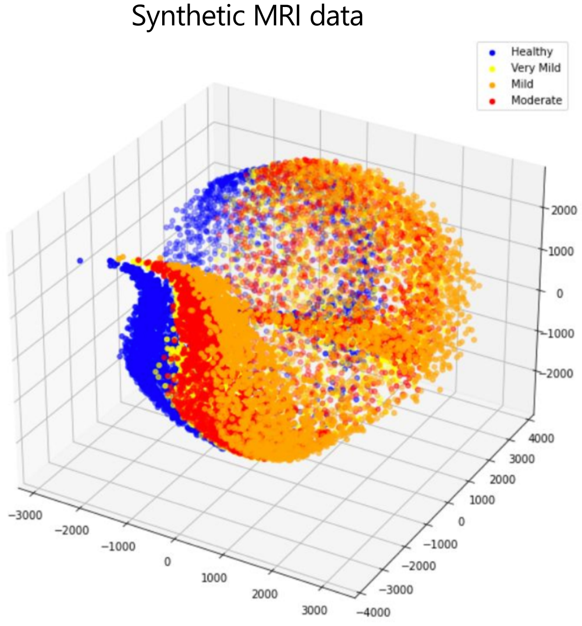
The synthetic MRI images had maintained similar distribution as the real MRI images and were hard to distinguish from real images. Less variability was observed in the synthetic MRI images. This indicates that the generator could be improved as it tends to generate the more common images (modes) with higher frequency than observed in the real data, resulting in lower quality and less diversity.

First three components of a principal components analysis of the inceptionV3 embeddings


Real MRI data

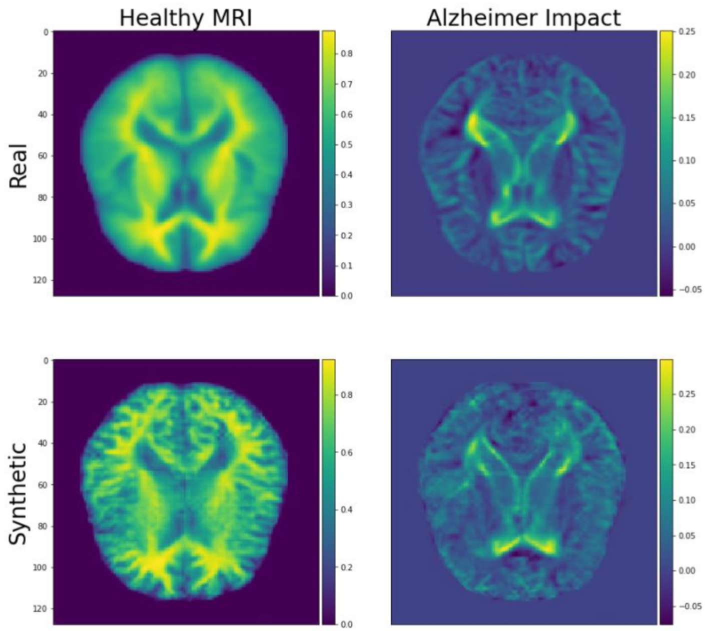
2. Biology

The impact of Alzheimer’s seemed to have been represented in the synthetic MRI images, but to a lower degree than in the real MRI images. This is seen in the figure on the right, which shows similar impact of real and synthetic data on Alzheimer’s. This similar impact indicates that the synthetic images potentially could be used to build AI models.


Synthetic MRI data

3. Useability

The synthetic MRI images had a lower predictive power than the real MRI images. However, combining synthetic and real MRI images increased overall predictive power. The fact that predictive power was observed is promising in terms of enriching scarce data sets or combining multiple sources for building robust AI models.


Healthy MRI
Alzheimer Impact

Insights into generating synthetic data

Documenting the utility of the synthetic data while generating them is essential for researchers to trust the synthetic data and to understand their limitations.

Example documentation for synthetic tabular data

Univariate distributions

This is the first level of tests to ensure that the generation has learnt each individual variable's distribution, i.e. distribution of age.

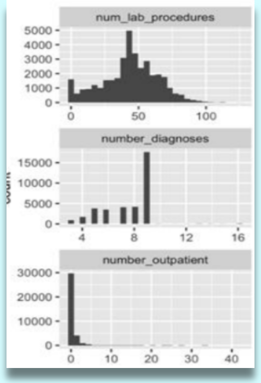
Bivariate correlations

The second level, this evaluates if relationships between variables are maintained, i.e. age positively correlated with number of hospitalisations.

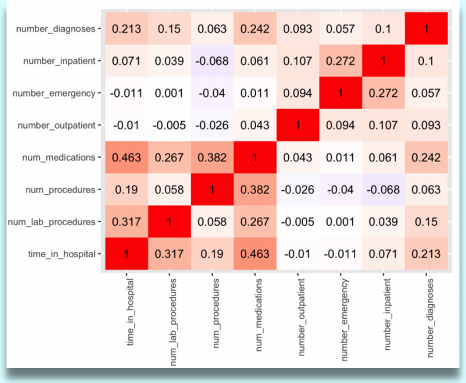
Multivariate correlations

The third level ensures that deep correlations are maintained by creating machine learning models on the synthetic data and evaluating real data with the relative prediction error compared to models trained on the real data. This could be done by leaving out one variable at the time to evaluate the predictive power of each individual variable.

Univariate

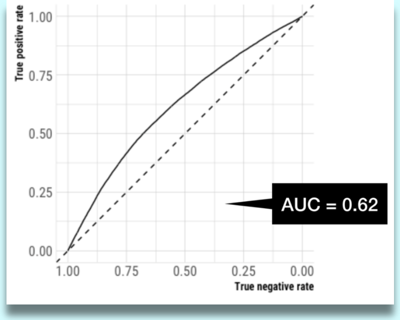


Bivariate



	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	number_diagnoses
number_diagnoses	0.213	0.15	0.063	0.242	0.093	0.057	0.1	1
number_inpatient	0.071	0.039	-0.068	0.061	0.107	0.272	1	0.1
number_emergency	-0.011	0.001	-0.04	0.011	0.094	1	0.272	0.057
number_outpatient	-0.01	-0.005	-0.026	0.043	1	0.094	0.107	0.093
num_medications	0.463	0.267	0.382	1	0.043	0.011	0.061	0.242
num_procedures	0.19	0.058	1	0.382	-0.026	-0.04	-0.068	0.063
num_lab_procedures	0.317	1	0.058	0.267	-0.005	0.001	0.039	0.15
time_in_hospital	1	0.317	0.19	0.463	-0.01	-0.011	0.071	0.213

Multivariate and relative prediction error



Example documentation for synthetic image data

Embedded distributions

Use pretrained image models to embed and calculate the difference in distribution of synthetic and real images (i.e. Fréchet inception distance (FID)).

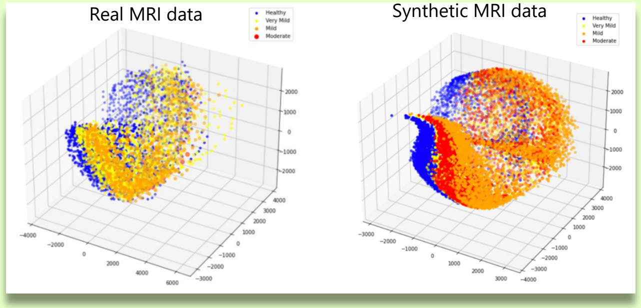
Mean images

Allow quick visual understanding of diversity of the images using the mean images (possibly over each type of the disease).

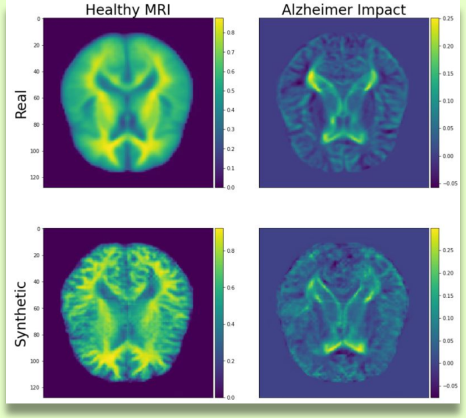
Relative prediction error

Similarly to tabular data checks, create machine learning models on the synthetic images and evaluate on the real data to calculate the relative prediction error compared to models trained on the real images.

Embedded distributions



Mean images



Insights into working with synthetic data

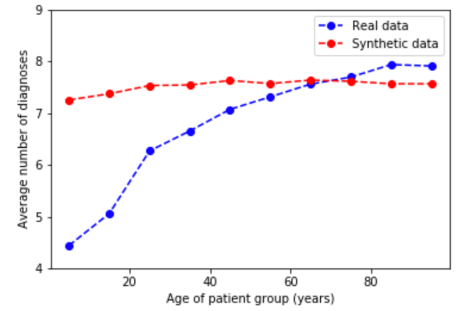
When working with synthetic data, it is important to understand potential limitations and to identify if any variable of the synthetic data is less reliable than other variables and should be neglected.

Domain knowledge

Participants showed that domain knowledge made it much easier to identify faults in the synthetic data. Documentation from the generation of synthetic data can enable a faster check for researchers to evaluate if the synthetic data were valuable to them or not.

Known correlations based on domain knowledge

Use domain knowledge for fast analysis of known correlations



What to use and what not to use

If the synthetic data consist of a lot of variables, some statistical properties may be better maintained than others. Researchers could therefore use documentation to understand which variables they could trust and which they could not.

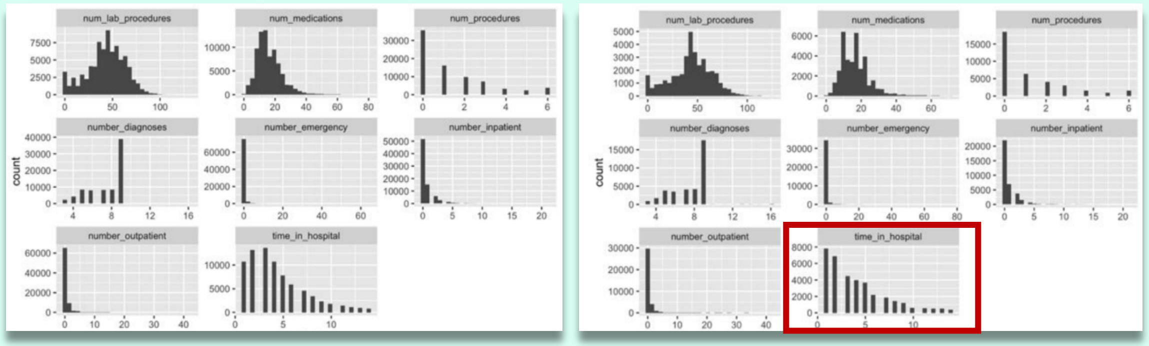
Examples of checks to evaluate what to use:

- Identify if the univariate distribution of the desired target variable is maintained in the synthetic data.
- Identify bivariate correlations which are maintained in the synthetic data.
- Evaluate feature importance of simple models and select strong features.

Univariate distributions

Avoid using poorly generated features not maintaining their distributions and correlations such as the highlighted region

Real data **Synthetic data**



Summary of findings

General summary

This fully virtual synthetic health data hackathon focused on the comparison of real data sets with their synthetic copies and sought to explore possible use cases within Alzheimer's and diabetes, respectively. 79 students and researchers with 15 different nationalities were divided into 22 teams. Each team chose one of three challenges and had 40 hours to work with the related data set. The task was to develop innovative solutions using the synthetic data and analyses to compare synthetic data to the real data.

Feedback from both participants and advisers reveals a shared view of synthetic data as a valuable new addition to the toolbox for working with sensitive patient data. The data sets proved to have a series of different potential uses, where some use cases are less vulnerable to limitations than others. Groups with domain knowledge quickly identified limitations in the diabetes synthetic data set as some expected correlations were lost. This loss highlights the importance of evaluating the synthetic data's quality.

Participants also pointed to the value of using synthetic data in combination with the real data in order to increase data set size and thereby strengthen the participants developed AI models' performances.

Six key findings are highlighted on the right-hand side of this page.

Advisers were present from Lundbeck, Novo Nordisk, the Danish National Genome Center, BioLib, Max Planck and Deloitte. The event was organised by BioLib and the project SHARED in a partnership with Digital Hub Denmark. The event was hosted by Rigshospitalet.

Data quality and biological insight

Synthetic data related to Alzheimer's and diabetes, respectively, were able to capture some of the same biological characteristics as the real data sets, however, there was also room for improvement of the methods. Future work should include investigation of the use of more recent state-of-the-art methods. Teams with knowledge of the biology were much more efficient at testing the data quality of the synthetic data sets.

Testing the synthetic data

When sharing synthetic data as a substitute to real data, an assessment of the similarity between the synthetic and real data is essential. Investigating and highlighting the similarity of the variables and correlations can shed light on the quality of the synthetic data for researchers.

Software development

Developing software and testing its functionality is one of the major, time-consuming components of developing a full solution. The pipelines built on the synthetic data by the participants were easily applied on the real data and appeared promising in terms of software development and testing.

Variety of analyses

The hackathon showed that even using synthetic data of a moderate quality could still lead to many different types of relevant analyses highlighting different biological research questions and comparisons between real and synthetic data.

Innovation

Several teams developed innovative analytical platforms and digital simulation games using the synthetic data. An example was 'Let's play doctor', where the team created a website for doctors to simulate treatment outcomes. This is an example of how synthetic data can be used to innovate.

Further work needed

Using synthetic data appears promising in terms of accelerating research and innovation. Further research on the methodology of generating synthetic data as well as evaluating data utility and privacy is needed. Along with awareness, this research can help pave the way for regulatory guidelines.

Next steps and perspectives

The journey of synthetic data is just beginning, and it will be essential to learn more by exploring potentials and limitations on more use cases and more types of data sets. Spreading the word about synthetic data's great potential among businesses and governments will help accelerate organisational adoption and regulatory paradigms.

How can we create useful synthetic data?

Can we create a toolbox for generating useful synthetic data requiring minimal human intervention?

Identify methodologies through researchers and partners for solutions for other datatypes and combine methods in a holistic synthetic data generation solution.

Where can synthetic data provide value?

Can synthetic data provide a temporary data set for researchers while waiting for access to the real data?

Can synthetic data enable combining data from multiple sources where it currently is very difficult?

Synthetic data could be used for educational purposes for students to strengthen innovation and education of digital health solutions.

How do we measure privacy in synthetic data?

Can we identify methods to evaluate the level of privacy-related risk in the synthetic data?

Identify where and if differential privacy can be utilised in synthetic-data generation and whether it is necessary, or if doing privacy checks, such as simulated model interference attacks on models built with synthetic data, is sufficient.

Can synthetic data have synergy with other privacy preserving techniques?

Can it be used as a precursor to explore the data before setting up a federated learning AI model?

A use case catalogue covering all the main privacy preserving techniques across a lifecycle (cf. page 7) would be valuable both for researchers and regulators.

Synthetic data is expected to be an integrated part of the toolbox in privacy preserving techniques and the true value is to identify the synergy between the different privacy preserving techniques.

Appendix

References, event collaborators and snapshots from the event site

References

1. Eliana Biundo, Andrew Pease, Koen Segers, Michael de Groote, Thibault d'Argent and Edouard de Schaetzen: "The socio-economic impact of AI on European health systems" (comissioned by MedTech Europe, authored by Deloitte), 2020 - https://www2.deloitte.com/be/en/pages/life-sciences-and-healthcare/articles/the-socio-economic-impact-of-AI-on-healthcare.html

2. Nordic Innovation (authored by Deloitte Legal): "Bridging Nordic Data", Nordic Innovation, 2020 - https://norden.diva-portal.org/smash/get/diva2:1441471/FULLTEXT04

3. NVIDIA: "Triaging COVID-19 Patients: 20 Hospitals in 20 Days Build AI Model that Predicts Oxygen Needs", NVIDIA blogposts, 2020 - https://blogs.nvidia.com/blog/2020/10/05/federated-learning-covid-oxygen-needs/

4. Mehreen Ali 1, Katariina Perkonoja, Javier Nunez-Fontarnau, Kasper Marstal, Henning Langberg, Janna Saarela, Arho Virkki and Timo Miettinen: "Synthetic data for the future of health technology", being submitted 2021

5. Zhang, Jun, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao: "Privbayes: Private data release via bayesian networks", in ACM Transactions on Database Systems (TODS) 42, no. 4 (2017): 1-41 - http://dimacs.rutgers.edu/~graham/pubs/papers/privbayes-tods.pdf

6. Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni: "Modeling tabular data using conditional gan", in Advances in Neural Information Processing Systems, pp. 7335-7345 (2019) - https://proceedings.neurips.cc/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf

7. Kovalev, Vassili and Siarhei Kazlouski: "Examining the Capability of GANs to Replace Real Biomedical Images in Classification Models Training", in International Conference on Pattern Recognition and Information Processing, pp. 98-107. Springer, Cham, 2019 - https://arxiv.org/pdf/1904.08688.pdf

8. Menon, Sumeet, Joshua Galita, David Chapman, Aryya Gangopadhyay, Jayalakshmi Mangalagiri, Phuong Nguyen, Yaacov Yesha, Yelena Yesha, Babak Saboury, and Michael Morris: "Generating Realistic COVID19 X-rays with a Mean Teacher+ Transfer Learning GAN", arXiv preprint arXiv:2009.12478 (2020) - https://arxiv.org/pdf/2009.12478.pdf

9. Salehinejad, Hojjat, Errol Colak, Tim Dowdell, Joseph Barfett, and Shahrokh Valaee: "Synthesizing chest x-ray pathology for training deep convolutional neural networks", in IEEE transactions on medical imaging 38, no. 5 (2018): 1197-1206 - https://www.researchgate.net/publication/328945795_Synthesizing_Chest_X-Ray_Pathology_for_Training_Deep_Convolutional_Neural_Networks

10. Karras, Tero, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila: "Training generative adversarial networks with limited data", arXiv preprint arXiv:2006.06676 (2020) - https://arxiv.org/abs/2006.06676

11. Razavi, Ali, Aaron van den Oord, and Oriol Vinyals: "Generating diverse high-fidelity images with vq-vae-2", in *Advances in Neural Information Processing Systems*, pp. 14866-14876 (2019) - https://arxiv.org/abs/1906.00446

12. Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu: "Wavenet: A generative model for raw audio", arXiv preprint arXiv:1609.03499 (2016) - https://arxiv.org/abs/1609.03499

13. Maaløe, Lars, Marco Fraccaro, Valentin Liévin, and Ole Winther: "Biva: A very deep hierarchy of latent variables for generative modeling", in Advances in neural information processing systems, pp. 6551-6562 (2019) - https://papers.nips.cc/paper/2019/file/9bdb8b1faffa4b3d41779bb495d79fb9-Paper.pdf

14. Bellovin, Steven M., Preetam K. Dutta, and Nathan Reitinger: "Privacy and synthetic datasets", in Stan. Tech. L. Rev. 22 (2019): 1 - https://www-cdn.law.stanford.edu/wp-content/uploads/2019/01/Bellovin_20190129-1.pdf

15. Dubey, Sarvesh, "Alzheimer's Dataset (4 classes of Images)", Kaggle datasets, 2019 - https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images

16. Strack, Beata, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. "Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records." BioMed research international (2014) - https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

Event collaborators and partners

Hosted by

Rigshospitalet

Rigshospitalet is Copenhagen's main university hospital. Located in the heart of the Danish capital, the hospital plays a crucial role in the Danish health-care system and has biomedical research as a key priority.

REGION H Rigshospitalet

Collaborators and Mentors

UNIVERSITY OF HELSINKI

TYKS Turku University Hospital

Teal Medical

Deloitte.

CSC

biolib

Lundbeck

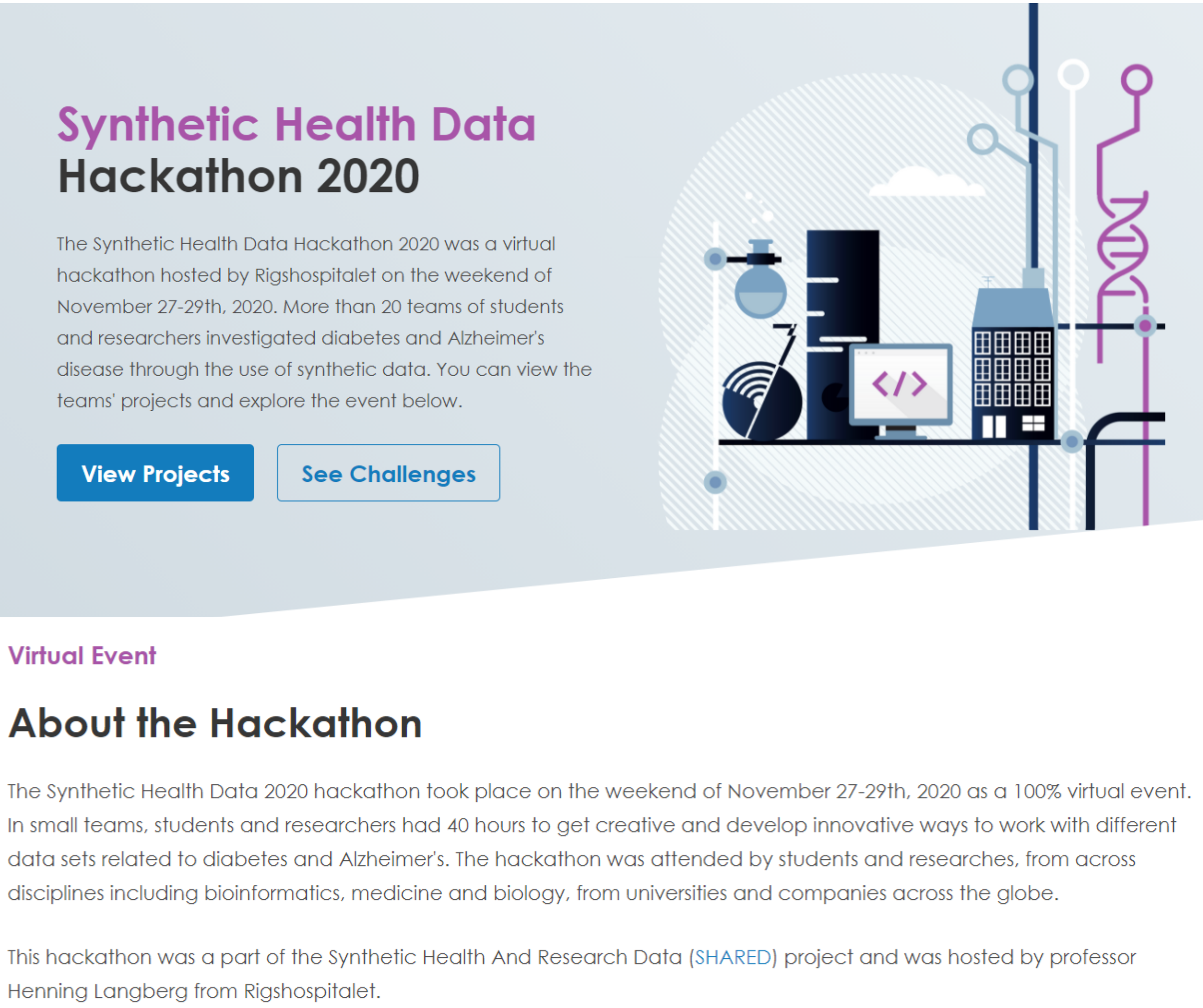
novo nordisk®

Partners and Sponsors

Digital Hub Denmark

ERHVERVSMINISTERIET

Snapshots from the event site (1/3)



Synthetic Health Data Hackathon 2020

The Synthetic Health Data Hackathon 2020 was a virtual hackathon hosted by Rigshospitalet on the weekend of November 27-29th, 2020. More than 20 teams of students and researchers investigated diabetes and Alzheimer's disease through the use of synthetic data. You can view the teams' projects and explore the event below.

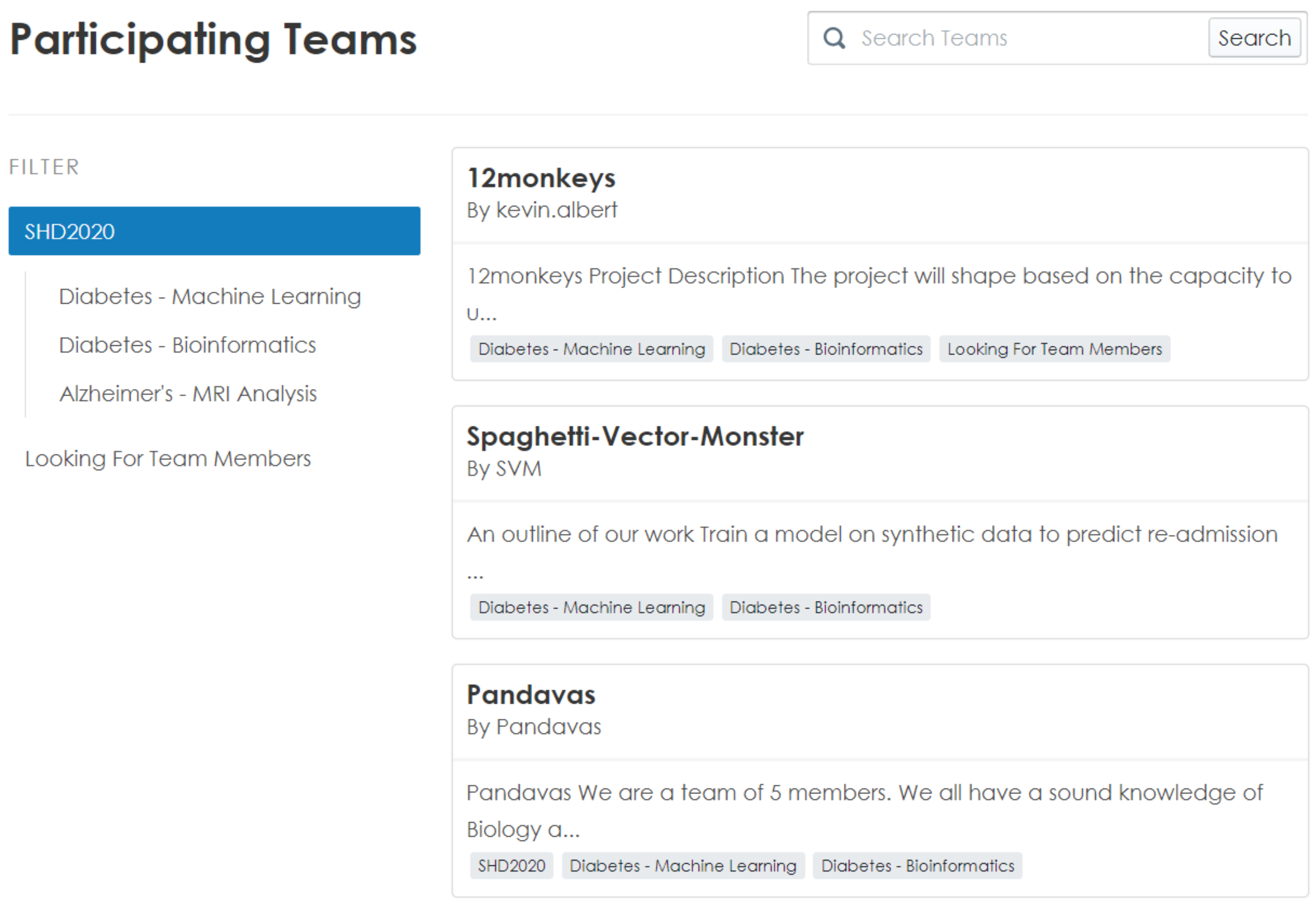
Virtual Event

About the Hackathon

The Synthetic Health Data 2020 hackathon took place on the weekend of November 27-29th, 2020 as a 100% virtual event. In small teams, students and researchers had 40 hours to get creative and develop innovative ways to work with different data sets related to diabetes and Alzheimer's. The hackathon was attended by students and researches, from across disciplines including bioinformatics, medicine and biology, from universities and companies across the globe.

This hackathon was a part of the Synthetic Health And Research Data (SHARED) project and was hosted by professor Henning Langberg from Rigshospitalet.

Snapshots from the event site (2/3)


Search Teams
Search
FILTER
SHD2020
Diabetes - Machine Learning
Diabetes - Bioinformatics
Alzheimer's - MRI Analysis
Looking For Team Members

Participating Teams

12monkeys
By kevin.albert

12monkeys Project Description The project will shape based on the capacity to u...

Diabetes - Machine Learning | Diabetes - Bioinformatics | Looking For Team Members

Spaghetti-Vector-Monster
By SVM

An outline of our work Train a model on synthetic data to predict re-admission ...

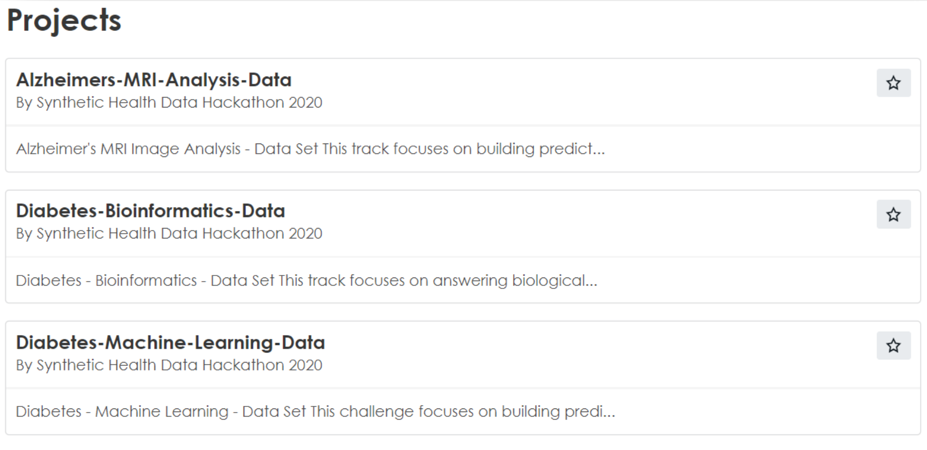
Diabetes - Machine Learning | Diabetes - Bioinformatics

Pandavas
By Pandavas

Pandavas We are a team of 5 members. We all have a sound knowledge of Biology a...

SHD2020 | Diabetes - Machine Learning | Diabetes - Bioinformatics

Snapshots from the event site (3/3)



Projects

Alzheimers-MRI-Analysis-Data ☆
By Synthetic Health Data Hackathon 2020

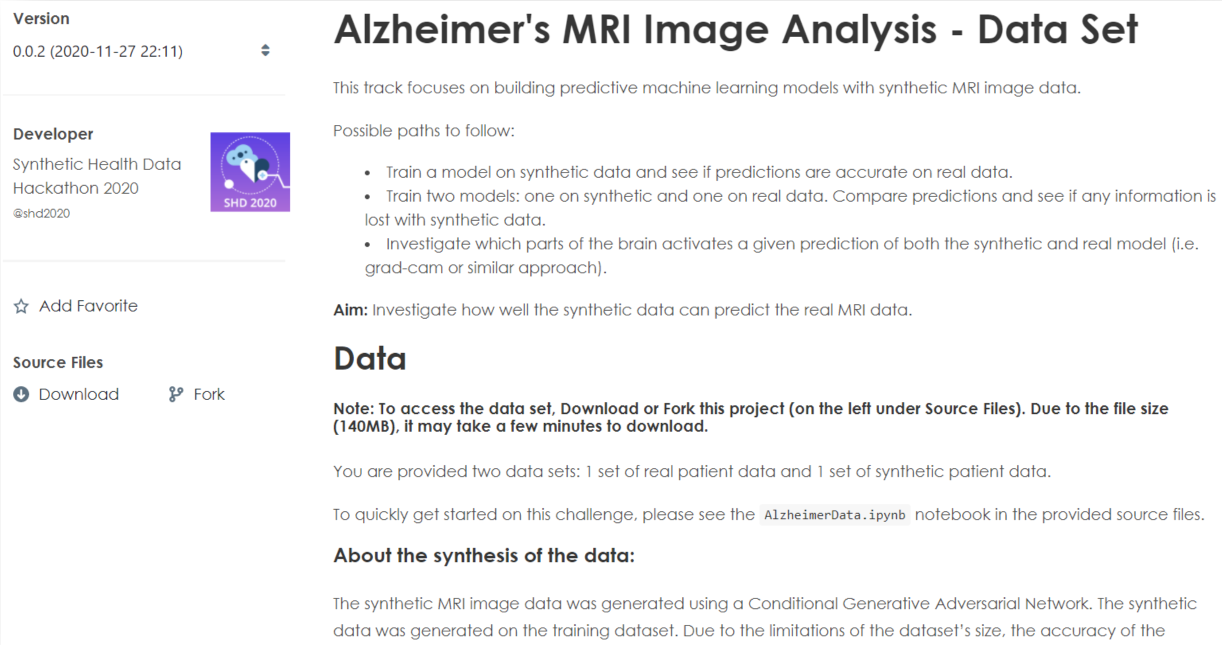
Alzheimer's MRI Image Analysis - Data Set This track focuses on building predict...

Diabetes-Bioinformatics-Data ☆
By Synthetic Health Data Hackathon 2020

Diabetes - Bioinformatics - Data Set This track focuses on answering biological...

Diabetes-Machine-Learning-Data ☆
By Synthetic Health Data Hackathon 2020

Diabetes - Machine Learning - Data Set This challenge focuses on building predi...



Version

0.0.2 (2020-11-27 22:11)

Developer

Synthetic Health Data Hackathon 2020
@shd2020

☆ Add Favorite

Source Files

Download Fork

Alzheimer's MRI Image Analysis - Data Set

This track focuses on building predictive machine learning models with synthetic MRI image data.

Possible paths to follow:

- Train a model on synthetic data and see if predictions are accurate on real data.
- Train two models: one on synthetic and one on real data. Compare predictions and see if any information is lost with synthetic data.
- Investigate which parts of the brain activates a given prediction of both the synthetic and real model (i.e. grad-cam or similar approach).

Aim: Investigate how well the synthetic data can predict the real MRI data.

Data

Note: To access the data set, Download or Fork this project (on the left under Source Files). Due to the file size (140MB), it may take a few minutes to download.

You are provided two data sets: 1 set of real patient data and 1 set of synthetic patient data.

To quickly get started on this challenge, please see the `AlzheimerData.ipynb` notebook in the provided source files.

About the synthesis of the data:

The synthetic MRI image data was generated using a Conditional Generative Adversarial Network. The synthetic data was generated on the training dataset. Due to the limitations of the dataset's size, the accuracy of the

Deloitte. REGION H Rigshospitalet

Digital Hub Denmark

novo nordisk fonden

In partnership with:

https://digitalhubdenmark.dk/
https://novonordiskfonden.dk/

Follow SHARED's journey on:

https://shared.landen.co/

This publication contains general information only, and none of the member firms of Deloitte Touche Tohmatsu Limited, its member firms, or their related entities (collective, the "Deloitte Network") is, by means of this publication, rendering professional advice or services. Before making any decision or taking any action that may affect your business, you should consult a qualified professional adviser. No entity in the Deloitte Network shall be responsible for any loss whatsoever sustained by any person who relies on this publication.

As used in this document, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of the legal structure of Deloitte USA LLP, Deloitte LLP and their respective subsidiaries. Certain services may not be available to attest clients under the rules and regulations of public accounting.