Discover Artificial Intelligence

Discover

Review

Synthetic data use: exploring use cases to optimise data utility

Stefanie James[1] · Chris Harbron[2] · Janice Branson[3] · Mimmi Sundler[4]

Received: 12 November 2021 / Accepted: 7 December 2021
Published online: 13 December 2021

Abstract
Synthetic data is a rapidly evolving field with growing interest from multiple industry stakeholders and European bodies. In particular, the pharmaceutical industry is starting to realise the value of synthetic data which is being utilised more prevalently as a method to optimise data utility and sharing, ultimately as an innovative response to the growing demand for improved privacy. Synthetic data is data generated by simulation, based upon and mirroring properties of an original dataset. Here, with supporting viewpoints from across the pharmaceutical industry, we set out to explore use cases for synthetic data across seven key but relatable areas for optimising data utility for improved data privacy and protection. We also discuss the various methods which can be used to produce a synthetic dataset and availability of metrics to ensure robust quality of generated synthetic datasets. Lastly, we discuss the potential merits, challenges and future direction of synthetic data within the pharmaceutical industry and the considerations for this privacy enhancing technology.

Keywords Synthetic data · Artificial intelligence · Privacy · Privacy enhancing technology · Pharmaceuticals · Software testing · Hackathons · Training · Internal secondary use · External sharing · Data retention · Vendor assessment · Machine learning · Data augmentation · Anonymisation · Pseudoanonymisation

1 Introduction

Increasingly, there is a need to optimize data utility and share data using innovative privacy enhancing technologies to prevent privacy related concerns. The Pharmaceutical industry has set up systems and processes to share data internally and externally which has matured throughout the years [1, 2].

There is a wealth of data to unlock for a range of different use cases, with the aim of optimizing the utility of data from clinical trials. The need to share data for wide ranging use cases requires the application of emerging privacy enhancing techniques which broadly fall into two categories:

- Data transformation, which removes potentially identifying features from the data but which may impact the utility of the data and;
- Behavioural measures which reduce the likelihood of an attempt to identify subjects.

Using innovative methods to adhere to patient privacy can enable data to be harnessed depending on the use case.

✉ Stefanie James, stef.james@astrazeneca.com | [1]Data Policy Director, Data Office, Data Science and Artificial Intelligence, Biopharmaceuticals, Research and Development, AstraZeneca, Academy House, 136 Hills Road, Cambridge CB2 8PA, UK. [2]Expert Statistical Scientist, Roche Pharmaceuticals, Welwyn Garden City, UK. [3]Global Head of Advanced Methodology and Data Science, Novartis, Basel, Switzerland. [4]Head of Data and AI Governance and Policy R&D, Data Office, Data Science and Artificial Intelligence, Biopharmaceuticals, Research and Development, AstraZeneca, Cambridge, UK.

Discover Artificial Intelligence (2021) 1:15 | https://doi.org/10.1007/s44163-021-00016-y

Springer

This paper will focus on the exploration of use cases for synthetic data, which is a topic generating interest throughout the industry and with European bodies [3]. Synthetic data is 'data that is artificially created/simulated rather than being generated by actual events.' [4] Synthetic data is modelled on actual data which mirrors the statistical properties of the original dataset. The goal of generating synthetic data is to enable faster access to fictional but useful datasets. To produce a quality synthetic dataset the complexities of the original data need to be captured, particularly where there is a focus on accelerating research and innovation.

The techniques to generate synthetic data in a way that can bridge the gap between privacy and utility are maturing rapidly in this space. However, with any type of technology and innovation it can be difficult to understand the complexities of synthetic data generation. Creating quality synthetic data requires a specialised skillset and the right technical and organisational measures in place therefore an investment is required to really optimise data utility through synthetic data use.

In this paper, we will consider the following key areas:

- Assessment on synthetic use cases to optimize data utility,
- Application of method(s) to produce a synthetic dataset,
- Utility measurements and parameters to ensure quality of the synthetic dataset,
- Potential challenges for the use of synthetic data.

2 Synthetic data use cases for optimising data utility

There are seven key use cases which this paper will explore to optimize data utility using synthetic data (Fig. 1) shows the relationships that can exist between the different use cases. For example, synthetic data can be used for machine learning purposes, which is also related to training due to the use case for training machine learning algorithms. Training is a separate use case, as training can also encompass the training of employees on realistic data. The seven use cases illustrate the potential of a synthetic data ecosystem that can be inter-connected throughout an organization to collect, store, analyze, and leverage data. Forrester Research identified synthetic data as a 'critical technology' [5] which will advance AI possibilities under 'AI 2.0' [5] as an emerging ecosystem.

A summary is provided (Table 1) of the use cases that this paper will explore with a high level observation on the financial, technical and organizational or reputational impact synthetic data could have on adoption within the Pharmaceutical industry.

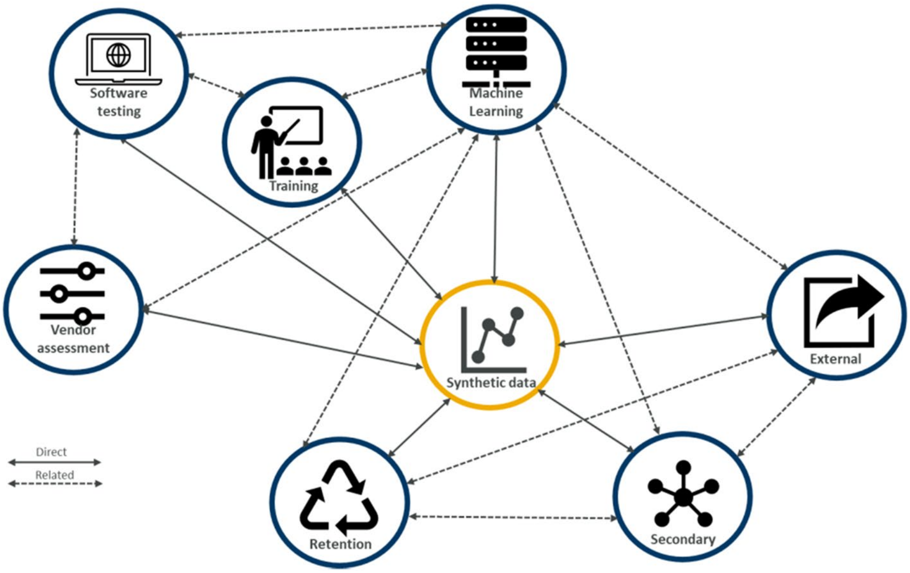


Fig. 1 The use cases to be explored for synthetic data use and the direct/relationships between the use cases

2.1 Requirements

One important question which will vary by use case is, how well does synthetic data need to capture all the properties within the original data? For all use cases there is a baseline requirement to ensure that no patients from the original dataset can be re-identified from the synthetic data or that any inferences could be made about them.

For some of the use cases: Machine learning—evaluation, Internal software testing, Education, training and hackathons and vendor assessment, higher level agreement between the original and synthetic datasets may be less critical. It will be important that each variable in the synthetic data covers the same range and distribution of values, including the same missing values as the original, and maintains key correlations e.g. zero cells in cross-tabulations, but for these applications capturing higher level data structure may be less important.

For the other use cases as well as the basic requirements described above, it may also be important that the synthetic data also captures higher level data structure from the original data. The requirements may vary depending upon the detailed specifics of the use case, for example if synthetic data is being used to develop a machine learning model, if the performance of the model is going to be evaluated on a test set consisting entirely of real data, then the agreement of the synthetic data may not matter as the model has been shown to work on real data. If, however, the model was also being tested on synthetic data then robustly showing that synthetic data reflected the statistical properties of real data would be critical.

2.1.1 Medical and Healthcare products Regulatory Agency (MHRA)

The MHRA are utilizing synthetic data. This can be seen by CPRD [6] real-world evidence. CPRD has generated 'high-fidelity synthetic datasets' [5]. The synthetic data generation and the evaluation framework used to generate the synthetic data are owned by the MHRA. The synthetic datasets replicate the 'complex clinical relationships in real primary care patient data while protecting patient privacy.' In the request form to access synthetic data, the use cases in (Table 2) are suggested. This would mean that the requirements to capture the properties of the original data would be higher.

2.1.2 Integraal Kankercentrum Nederland (IKNL)

The IKNL announced the first release of a synthetic version of the Netherlands Cancer Registry [7] which is available for scientific research and exploration. IKNL stipulate that synthetic data can be used for software development or to explore analysis methods but should not be used for clinical decision-making. This illustrates the variability in requirements for the use of synthetic data where a lower baseline could be adopted.

2.2 Machine learning

2.2.1 Use case 1.0: evaluation and comparison of machine learning techniques

New algorithms for machine learning are being developed and it can be challenging to understand whether these give a substantial benefit over existing algorithms using realistic data. Artificially generated data can be used but this often favours the algorithm making assumptions closest to those from which the data was generated, leading to contradictory results in different publications.

Generating synthetic data for the training, validation and test data from the same generative model can be an efficient way of making these comparisons. As well as removing privacy concerns, using synthetic data has the additional advantage that unlimited sample sizes of data can be generated. This allows for exploration of the impact of sample size on the performance of the model, and by evaluating the performance of any fitted model on a very large test data set removes the sampling variability that can be associated with evaluating a model on a small test data set that can be observed when the total amount of data available is limited.

2.2.2 Use case 1.1: data augmentation

Data augmentation is a technique used in machine learning where the amount of data used to train machine learning models is increased by the addition of either slightly modified copies of existing data or newly created synthetic data and can also increase regularization to help reduce overfitting.

Table 1 Summary exploration of use cases and the considerations on financial, technical, organisational and reputational impact

Number	Optimizing data utility through synthetic data Use cases:	Financial impact	Technical and organizational impact	Reputational impact
1.0	Machine learning I. Evaluation and comparison of machine learning techniques	☑ Generation of big data without the cost associated	☑ Benefits training, validation and testing	☑ Removes privacy and AI concerns
1.1	II. Data augmentation		☑ Increase model performance	
1.2	III. Prevent privacy attacks on machine learning models	☑ Increase exploration	☑ Benefits training, development and problem solving	☑ Removes privacy and AI concerns
2	Internal software testing	☑ Reduction of defects using sub-standard data	☑ Benefits software engineering, reliability and testing	☑ Removes privacy concerns
3	Education, training and hackathons	☑ Talent retention	☑ Benefits training, development and problem solving	☑ Avoids poor personal data handling
4	Data retention	☑ Avoids the need for expensive personal data storage and resourcing	☑ Generation of synthetic data requires processes and people	
5	Vendor assessment	☑ Avoid lengthy negotiation and increase exploration	☑ Ease of sharing synthetic data	☑ Removes privacy concerns
6	Internal secondary use		☑ Enables secondary analysis, exploration and investigation into the variables collected	☑ Demonstrates good practice in personal data handling
7.0	External sharing I. Alleviate regulatory challenges		☑ Enables potential ease of sharing across jurisdiction	☑ Demonstrates good practice in personal data handling
7.1	II. Accelerate access to data		☑ Enables faster data sharing	☑ Removes privacy concerns

Table 2 Suggested purpose for use of synthetic data from request form used by MHRA-CPRD database

What is your intended purpose for accessing the synthetic data? Please tick all that apply
Training of machine learning algorithms
Testing/validation of machine learning algorithms
Developing, populating or testing models
Methodological research
Sample size boosting
Feasibility counts
To support regulatory submissions
Building medical software applications
Trend simulation/analysis
Other (please state below)

Table adapted from cprd.com 2021 Synthetic data request form

A particular application of this may be to address under-sampling within a dataset, that is where a population subgroup, for example a particular ethnicity, is underrepresented within the dataset and so any fitted models may not perform well in that subpopulation. Adding synthetic data to boost that subpopulation may improve the performance of the model in that subpopulation and generate an overall more inclusive and fair model.

Whilst this can improve the predictive performance of algorithms, care is required in interpretation, both in any form of inference and in evaluating the performance of the fitted model by ensuring complete separation of the training and evaluation datasets with no data leakage.

2.2.3 Use case 1.2: prevent privacy attacks on machine learning models

There is a data security risk of privacy attacks on machine learning models to uncover the data used for training. It is becoming easier to attack machine learning models and recover their training data, using white box and black box attacks [8] which could lead to the leak of personal data. Synthetic data can be used to develop machine learning models with the view of protecting the training data used. If machine learning models are constructed using synthetic data then this risk would no longer be an issue of consideration. If the attacker recovers the training data, they would only recover the synthetic variants used. This would enable models to be shared in the scientific community and optimise utility and learning in this space. This use case illustrates the ability to use synthetic data to strengthen security, optimise sharing in the scientific community and reduce the risk of unveiling real data which could contain personal data.

2.3 Internal software testing

2.3.1 Use case: data engineering

There is an emerging need to test on realistic individual patient level data without the requirement to have access to the real dataset. Particularly, in the development of data products or more generally in data engineering and the creation of data pipelines. There are many examples of requiring representative data to conduct accurate software testing, such as data products that deal with adverse event reporting, clinical trial diversity and the use of natural language processing.

2.3.2 Example of GDPR fines for using personal data for testing

There was a recent fine in Norway where the Norwegian data protection authority (‘Datatilsynet’) fined the Norwegian Olympic and Paralympic Committee and Confederation of Sports [9]. The fine was for testing on personal data before a risk assessment had been performed or means to secure the information on a cloud solution. Datatilsynet found that testing could have been carried out by processing synthetic data, or using fewer personal data attributes.

Ultimately, there was no legal basis for testing and principles such as legality, data minimization and confidentiality were breached. By enabling data engineering through the use of synthetic data, this reduces the unnecessary processing of personal data for this use case and the need to utilize a legal basis of consent.

2.4 Education, training and hackathons

2.4.1 Use case: Onboarding and development

Synthetic data can be used as an effective training and onboarding tool, particularly for employees who need to have an understanding of how to handle personal data. Synthetic data enables employees to explore datasets that represent the same structural properties of clinical datasets without the privacy controls required.

Synthetic data can be used as a teaching aid, to develop skills in a range of different areas such as biometrics, data science and data engineering. The ability to use synthetic data for this purpose will nurture career development within the organization and allow for the cross-pollination of innovation, advancement of knowledge and skills whilst keeping the patient at the heart of development. The ability to show that employees are trained adequately is an important consideration within GDPR 'by design and default' [10]. To illustrate compliance, it is necessary to demonstrate privacy awareness.

2.4.2 Use case: problem solving

Synthetic data can be used to aid problem solving, a common technique for this is through hackathons. Hackathons provide the ability to concentrate on a problem or challenge in the real-world, and come together with colleagues either internally or externally to come up with a solution.

The ability to use synthetic data for hackathons means that all attributes in the artificially generated dataset can be used, rather than an aggregated dataset. It may also permit wider participation within the hackathon by allowing external participants who may not otherwise be allowed access to the real data to join, thus enhancing collaboration. Therefore, the value of the solution is not hampered by missing personal attributes which could be vital to the overall delivery of the solution and considerations for adoption of the solution.

2.5 Data retention

2.5.1 Use case: enhance data storage considerations

In the General Data Protection Regulation (GDPR) there is a principle of storage limitation, where personal data must not be kept for longer than required [11]. This means that policies need to be developed by companies to provide retention periods for data that contains personal information. Retention of personal data is costly for an organization, in terms of management of resource, infrastructure and processes.

If the data is of significant utility and to enhance reuse of data, it is possible to apply synthesis techniques to retain data utility without privacy concerns. The generation of synthetic datasets, where privacy is no longer a concern, could significantly reduce the financial burden to organizations. There is also the option of anonymizing datasets which would also adhere to the principle of storage limitation. The benefits need to be considered for both synthetic data use and anonymization depending on the use case and potential future usability of the dataset which is considered later in this paper. This would mean that pharmaceutical companies could still derive value from data and adhere to the principle of storage limitation.

Good data management principles should also be considered when retaining data for longer. There are implications in terms of storage costs and the potential of retaining data sets that are unused and provide no value for retention- instilling a 'data-hoard' mentality. Therefore, there is a need for robust organizational processes to effectively determine the utility of a synthetic dataset.

2.6 Vendor assessment and sharing data with third party services

2.6.1 Use case: transfer of data

Vendor evaluation, assessments and sharing data with third party services can be time consuming. Contractual obligations and data transfer agreements can increase the resource, timelines and involve complex business processes across organizations such as legal, procurement, IT etc. The ability to use synthetic data to provide data to vendors and third party services could expedite the process and avoid unnecessary use of personal data.

Synthetic data would enable the ability to assess options and work with vendors to conduct pilot studies or fine tune developments before going through contractual considerations including complex privacy agreements. Key areas where synthetic data can advance procurement include speed in contractual negotiation, assessment and agility in decision making by sharing synthetic data for vendor assessments and third party services.

2.7 Internal secondary use

2.7.1 Use case: conduct further research beyond primary purpose

The best practice for secondary data is to utilize a privacy enhancing technology, such as anonymization or the production of synthetic data. Synthetic data and anonymization of data solve the same problem and involve a transformation of data.

Synthetic data can aid with internal data selection before using personal data. Synthetic data as a prelude to access to personal datasets means that Scientists can explore data, assess the data that is available internally and whether this could answer their scientific questions and exploration. Scientists can then determine the data they require based on synthetic data.

2.7.2 Example in the literature

Synthetic data generated from clinical trial data has been used for secondary analyses, with a focus of analyzing the synthetic data as an objective to determine whether analyses using real versus synthetic datasets were similar. In this scenario, there was a 'high concordance between the analytical results and conclusions' [12] from synthetic and real data suggesting synthetic data use could broadly 'serve as a proxy for real data.'

2.8 External sharing

2.8.1 Use case 1.0: alleviate regulatory challenges

Synthetic data provides the possibility of increased transparency within the pharmaceutical industry. The ability to use synthetic data could alleviate any difficulties with data sharing between organizations and across jurisdictions, this can often be complex due to external laws and regulations and in particular, international data transfers. Synthetic data provides an alterative avenue of exploration that could remove this complexity. There is still a lot of work to do and more maturity required to roll this out effectively in the external domain. A comprehensive effort would be required in the industry to adhere to a common code of conduct and standards for generating synthetic data to share externally. The development of a common sharing platform focused on synthetic data and industry guidelines would greatly aid this effort.

2.8.2 Use case 1.1: accelerate access to data

Synthetic data can be used as a method of accelerating access to data which can be a lengthy process on current external data sharing platforms. This delay can be seen as discouraging the use of secondary data. Synthetic data can optimise this process by allowing rapid access to researchers and analysts to quickly understand the statistical properties of the dataset. This could inform whether a full data application should be pursued.

Similarily, this type of rapid access to synthetic data in the external domain could inform future operating models for sharing data. Researchers and analysts could build code in their preferred development language based on synthetic data which could be submitted and re-run on the real data. This way of operating would optimise data sharing and utility through the provision of rapid access to datasets. The ability to explore synthetic datasets and develop for execution of the real data provides a balance between data utility and adherence to privacy.

2.8.3 Example of sharing synthetic data externally

There are already instances of using synthetic data in the external environment. The Simulacrum [13] provides a dataset that contains artificial patient-like cancer data to aid scientific insight. The Simulacrum imitates data held by Public Health England's National Cancer Registration and Analysis Service. Scientists get access to Simulacrum synthetic data, once the scientific query is refined scientists are able to submit a request to Public Health England to run queries on the real data. Public Health England will provide aggregate and anonymous data back to the scientist. Scientists are able to publish results based on the synthetic data.

3 Application of methods to produce synthetic data

There are various methods that can be used to produce synthetic data. A synthetic data set can be produced using one or several methods [14]. Typical methods used to produce synthetic data are provided in (Fig. 2).

3.1 Data perturbation

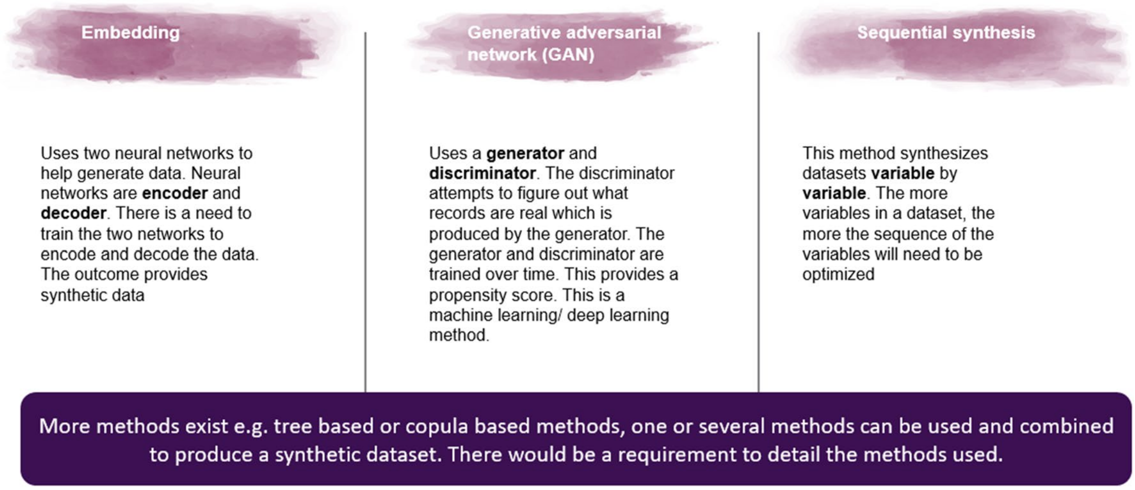
Data perturbation is a data security technique [15] which seeks to add noise to the original data set to produce more diverse data sets. The challenge with data collected during the conduct of a clinical trial is that often, hidden knowledge is within the data set. Data perturbation can assist with this difficulty. There are additional considerations with data perturbation, as with any method of producing synthetic data and this evolves around the introduction of bias.

The area of synthetic data generation is a highly active research area where we would anticipate seeing the development of multiple new methods and improvement of existing methods over the coming years.

4 Synthetic data quality metrics

The application of utility measurements and examining the distance between the original data set and the synthetic data set that is produced, is a key factor to uphold privacy and accuracy of use for the synthetic data set.

Which metrics are relevant is dependent upon the use case being considered. Other features may also be considered, for example for the use cases of vendor assessment, software testing or training, we may wish the synthetic data to have similar data-missingness patterns to the real data in terms of the proportions of missing values in each variable and the



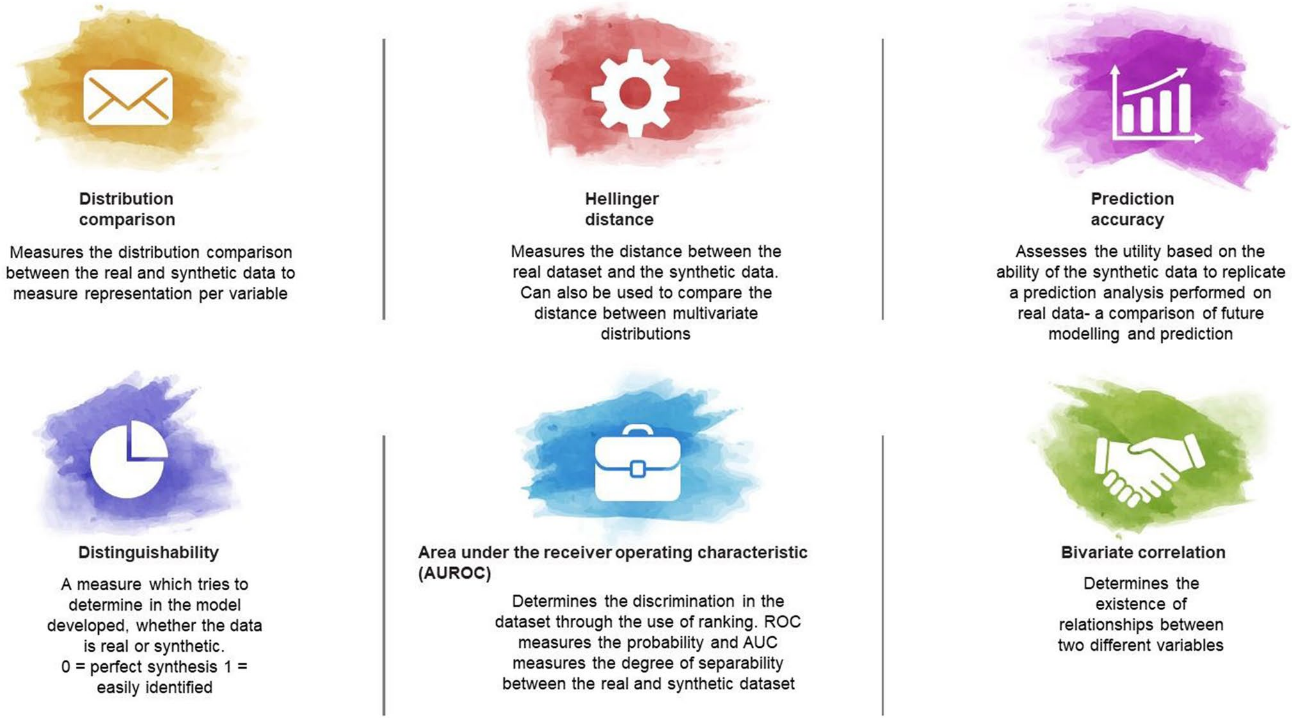
Uses two neural networks to help generate data. Neural networks are **encoder** and **decoder**. There is a need to train the two networks to encode and decode the data. The outcome provides synthetic data

Uses a **generator** and **discriminator**. The discriminator attempts to figure out what records are real which is produced by the generator. The generator and discriminator are trained over time. This provides a propensity score. This is a machine learning/ deep learning method.

This method synthesizes datasets **variable** by **variable**. The more variables in a dataset, the more the sequence of the variables will need to be optimized

More methods exist e.g. tree based or copula based methods, one or several methods can be used and combined to produce a synthetic dataset. There would be a requirement to detail the methods used.

Fig. 2 Example methods that can be used to produce synthetic data



Distribution comparison

Measures the distribution comparison between the real and synthetic data to measure representation per variable

Hellinger distance

Measures the distance between the real dataset and the synthetic data. Can also be used to compare the distance between multivariate distributions

Prediction accuracy

Assesses the utility based on the ability of the synthetic data to replicate a prediction analysis performed on real data- a comparison of future modelling and prediction

Distinguishability

A measure which tries to determine in the model developed, whether the data is real or synthetic. 0 = perfect synthesis 1 = easily identified

Area under the receiver operating characteristic (AUROC)

Determines the discrimination in the dataset through the use of ranking. ROC measures the probability and AUC measures the degree of separability between the real and synthetic dataset

Bivariate correlation

Determines the existence of relationships between two different variables

Fig. 3 Typical data quality metrics

relationship of a value being missing to the values of other variables. Whilst for other applications where the synthetic data will be analyzed there is little value in artificially introducing missing values into the data. The typical utility measurements that can be applied are displayed in (Fig. 3).

5 Synthetic data and anonymisation

There are several techniques to enhance privacy, each technique can be used for different purposes (Fig. 4). The external regulations applicable to each technique must be considered. Synthetic data is a complimentary privacy enhancing technique and not a replacement.

The merits of each technique [16] needs to be considered for the use case under exploration. If there is a requirement to use data without privacy concerns, then the options reduce to anonymization or synthetic data use.

Anonymization is applied to real data however during the application of anonymization techniques, there can be substantial data loss which can impact use cases- especially for scientific exploration and utility. The data loss experienced through anonymization is through the direct result of adherence to the requirement that data is no longer 'in a form which permits identification of data subjects.' [17] Typical strategies for anonymisation which would result in a loss of data utility include grouping variables which loses data resolution, omitting variables, or adding noise. Synthetic data can offset some of the challenges experienced through anonymization. Synthetic data enables all the statistical properties of the data to be shared but on data that has been artificially generated. It will depend on the researcher or analysts requirements to determine whether anonymized data or synthetic data can better meet the needs of the use case, with consideration on real versus artificially generated data.

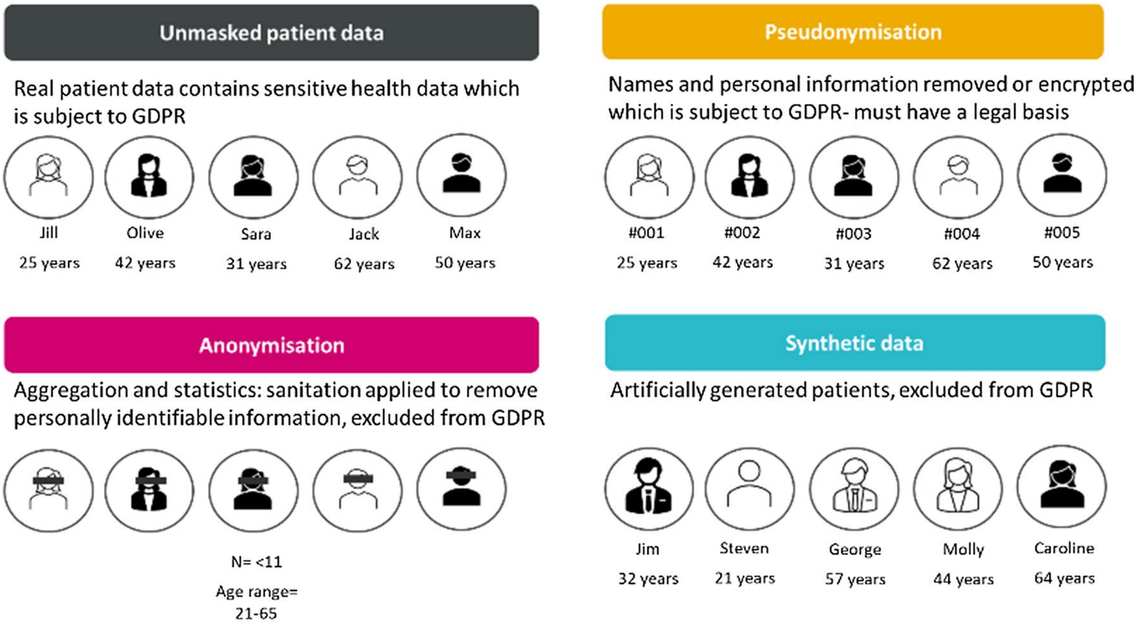

Unmasked patient data
Real patient data contains sensitive health data which is subject to GDPR
Jill
25 years
Olive
42 years
Sara
31 years
Jack
62 years
Max
50 years
Pseudonymisation
Names and personal information removed or encrypted which is subject to GDPR- must have a legal basis
#001
25 years
#002
42 years
#003
31 years
#004
62 years
#005
50 years
Anonymisation
Aggregation and statistics: sanitation applied to remove personally identifiable information, excluded from GDPR
N= <11
Age range= 21-65
Synthetic data
Artificially generated patients, excluded from GDPR
Jim
32 years
Steven
21 years
George
57 years
Molly
44 years
Caroline
64 years

Fig. 4 The different privacy enhancing techniques that can be used

6 Challenges with the use of synthetic data

The challenges include how the synthetic data is produced and how the application of the method is measured to ensure the characteristics between the original and synthetic data set are comparable. This can be difficult to illustrate due to the diverse data sets that can be collected during the conduct of a clinical trial and would require statistical analysis to compare.

The application of the method to produce synthetic data can also be time consuming. If the Pharmaceutical industry will look to generate synthetic data in the future, then it is likely that new technical and organizational measures will need to be adopted- including business processes and skilled individuals to produce synthetic data. This becomes particularly important due to the need to avoid overfitting or underfitting the models [18, 19] that are applied to produce synthetic data.

As with other privacy enhancing technologies [20], there is also the challenge of ensuring that a patient cannot be re-identified. This also relates to the need for robust business processes, method application and skilled individuals who are aware of the regulations and privacy that need to be adhered to.

Industry and regulatory acceptance of synthetic data for the different use cases will need to be considered, especially if synthetic data is used for internal or external data sharing practices.

6.1 Adoption of technical and organizational measures

There are several use cases where synthetic data could be an advantage. This needs to be balanced with the effort and potential expense in producing a representative synthetic dataset. To effectively use synthetic data, there is a requirement to adopt technical and organization measures (TOMs) for change and adoption. Gartner have predicted that by 2024 [21], 60% of data used for AI and machine learning will be synthetic data, overshadowing the use of real data for this purpose. There would be several challenges for organisations to overcome to really make this prediction a reality, primarily centered around adoption of TOMs.

These measures need to consider:

- Who is best placed to synthesize data, with a profile that includes the technical ability to generate models for data synthesis, balanced with the privacy knowledge to remain aware of the need to uphold privacy and risk of re-identification.

- Where in terms of how data synthesis is applied in reference to the infrastructure of the organization and processes around data sharing.
- What is used to synthesize, such as systems that are either applied in house, or off-the-shelf commercial offerings.
- When data synthesis is applied, to leverage the adoption of end to end data pipelines, where like for like datasets can be used for multiple purposes without privacy concerns.
- How the application of the model is applied, which closely relates to who synthesizes the data. To produce a trusted synthetic dataset, documentation around the utility of synthetic data needs to be embedded into the data generation pipeline. To make this feasible for organizations to run as a service, the burden of synthetic data needs to be reduced by automation.
- Why synthetic data has been produced, dependent on the use case that is being explored. Audit trails and transparency of data production should be developed to increase organizational trust in the generation of the datasets.

There will be some level of change for the organization to adopt synthetic data, but this organizational change is similar to the application of other privacy enhancing technologies such as anonymization.

6.2 The re-identification risk

There is a re-identification risk where a patient may be deemed an outlier [22] such as a rare disease or demographic. There could be a potential solution such as removing these patients before producing the synthetic dataset however, this could harm the future utility and introduce bias.

6.3 Scientific and medical acceptance

It is unclear to what extent conclusions generated from synthetic data will be accepted by the scientific and medical community in, for example, peer reviewed publications. It may be expected that until greater confidence has been established in the robustness of synthetic data, that journals will expect to see results directly demonstrated on real subjects. Synthetic data may still have a role, for example a machine learning algorithm may be generated using synthetic data as a training set, but then its utility is demonstrated using a real dataset.

In the future, it could be possible to apply synthetic data to an interim cut of data from a clinical trial. This would enlarge the overall cohort size of the dataset to replicate the final expected recruitment number. This type of utility would require scientific and medical acceptance before it is used in everyday ongoing clinical trial management. The benefit of applying synthetic data, could be to aid recruitment in under-represented regions/populations to increase trial diversity. This provides an ability to augment trial recruitment during the conduct of the clinical trial rather than when a final cut of data is available. This use of predictive analytics could mean that protocols are amended to identify vulnerabilities [23] in the design of the trial. It should be noted that in this example, applying synthetic data generation to clinical trials should not be confused external control arms [24] which sometimes confusingly get referred to as synthetic control arms. An external control arm is a term used to describe the generation of a cohort of real patients which mimics a clinical trial, usually to act as a control arm and enhance data from single-arm clinical trials, using data source external to the clinical trial, typically either real world data sources or historical clinical trials.

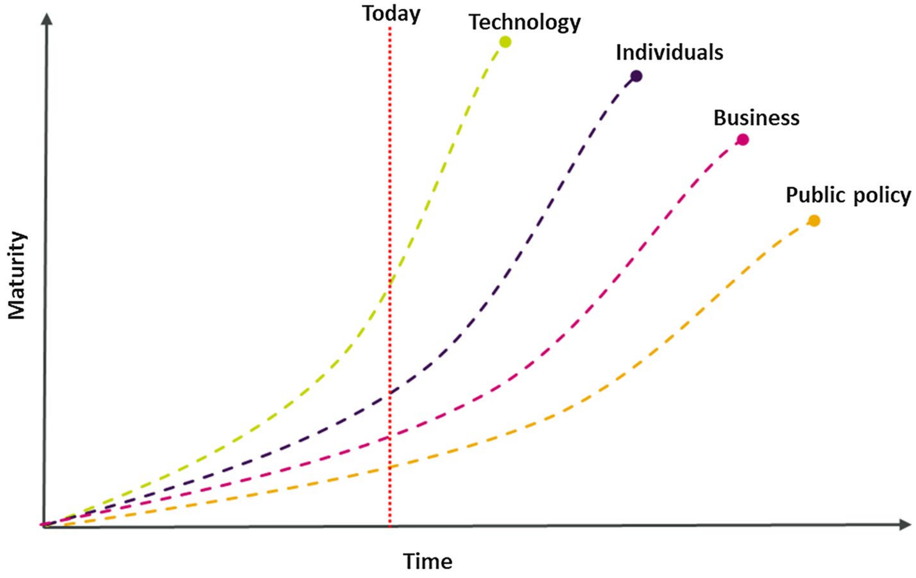
6.4 Industry acceptance

Ultimately, there is a challenge around the acceptance of synthetic data use and where it will be deemed fit for purpose. It is an emerging area and more work is needed to understand how the methods can be applied systematically and tailored around each use case.

6.5 Innovation

The technology for creating synthetic data has matured at a faster pace than the level of adoption within organisations [25] as illustrated by (Fig. 5).

Fig. 5 The maturity and time for the full adoption of synthetic data, with technology, individuals, business and public policy as key factors



Graphic adapted from (Global Human Capital Trends 2017, 2021)

We have addressed several use cases in this paper where the pharmaceutical industry could benefit from synthetic data adoption. Synthetic data maturity within the regulatory or policy environment now needs to be addressed [26] so that the gap between technology, adoption and utility can be fulfilled with regulatory requirements built in.

The following considerations should be built into an organizational approach to synthetic data generation. These considerations are:

- The development of a privacy assurance report to remove concerns of re-identification.
- Utility measurements and assessment of the quality of synthetic data to enable generation.

These considerations in combination with technical and organizational measures will increase the trust in synthetic datasets and in time, aid acceptance of this privacy enhancing technology.

Acknowledgements The authors acknowledge the review and contributions of the PSI Data Sharing Special Interest Group, Rebecca Sudlow from Roche, Katherine Tucker from Roche, Lucy Frith from GlaxoSmithKline, Christoph Gerlinger from Bayer, Hoi Shen Radcliffe from Amgen, Mahumood Hameed from Eisai, Tamsin Sargood from Johnson and Johnson, Catrin Tudor Smith from University of Liverpool, Carrol Gable from University of Liverpool, Matt Sydes from MRC Clinical Trials Unit at UCL and Michael Robling from Cardiff University.

Authors' contributions All authors contributed to the article. The literature search and analysis was conducted by SJ and CH. The first draft of the article was written by SJ. SJ, CH, JB and MS commented on previous versions of the article. SJ, CH, JB and MS approved the final article. All authors read and approved the final manuscript.

Funding No funds, grants, or other support was received.

Data availability (data transparency) Not applicable.

Code availability (software application or custom code) Not applicable.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Vivli. About Vivli: Overview—Vivli; 2021. https://vivli.org/about/overview/. Accessed 4 Oct 2021.
2. Sdv.dev. The synthetic data vault. Put synthetic data to work!; 2021. https://sdv.dev/. Accessed 4 Oct 2021.
3. European Data Protection Supervisor. Is the future of privacy synthetic; 2021. https://edps.europa.eu/press-publications/press-news/blog/future-privacy-synthetic_en. Accessed 4 Oct 2021.
4. AIMultiple. The ultimate guide to synthetic data: uses, benefits & tools; 2021. https://research.aimultiple.com/synthetic-data/. Accessed 4 Oct 2021.
5. Forrester. AI 2.0: upgrade your enterprise with five next-generation ai advances; 2021. https://www.forrester.com/report/AI-20-Upgrade-Your-Enterprise-With-Five-NextGeneration-AI-Advances/RES163520?objectid=RES163520. Accessed 29 Nov 2021.
6. Cprd.com. Synthetic data | CPRD; 2021. https://www.cprd.com/content/synthetic-data. Accessed 4 Oct 2021.
7. Iknl.nl. Synthetische dataset NKR beschikbaar voor onderzoekers; 2021. https://iknl.nl/nieuws/2021/synthetische-data-nkr-beschikbaar-voor-onderzoeker. Accessed 22 Oct 2021.
8. Ico.org.uk. Privacy attacks on AI models; 2021. https://ico.org.uk/about-the-ico/news-and-events/ai-blog-privacy-attacks-on-ai-models/. Accessed 29 Nov 2021.
9. OneTrust Data Guidance. Norway: datatilsynet fines NIF NOK 1.2M for disclosing personal data of 3.2M individuals; 2021. https://www.dataguidance.com/news/norway-datatilsynet-fines-nif-nok-12m-disclosing. Accessed 4 Oct 2021.
10. General Data Protection Regulation (GDPR). Chapter 2—principles—General Data Protection Regulation (GDPR). https://gdpr-info.eu/chapter-2/. Accessed 3 Nov 2021.
11. General Data Protection Regulation (GDPR). Art. 5 GDPR—principles relating to processing of personal data—General Data Protection Regulation (GDPR). https://gdpr-info.eu/art-5-gdpr/. Accessed 3 Nov 2021.
12. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open. 2021;11(4):e043497.
13. healthdatainsight.org.uk. The Simulacrum—healthdatainsight.org.uk; 2021. https://healthdatainsight.org.uk/project/the-simulacrum/. Accessed 4 Oct 2021.
14. Replica-analytics.com. Replica analytics | resources | knowledgebase; 2021. https://replica-analytics.com/knowledgebase. Accessed 4 Oct 2021.
15. Wilson R, Rosen P. Protecting data through perturbation techniques: the impact on knowledge discovery in databases. J Database Manag. 2003;14:14–26. https://doi.org/10.4018/978-1-59140-471-2.ch003.
16. Tucker A, Wang Z, Rotalinti Y, et al. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit Med. 2020;3:147. https://doi.org/10.1038/s41746-020-00353-9.
17. Legislation.gov.uk. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance); 2021. https://www.legislation.gov.uk/eur/2016/679/contents. Accessed 4 Oct 2021.
18. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66(3):411–21.
19. Langberg H, Hvidbak T, Closter Jespersen M. Synthetic health data Hackathon. Deloitte and Rigshospitalet, 2021, p. 8.
20. Bamford S. Applications of privacy-enhancing technology to data sharing at a global pharmaceutical company. J Data Protect Privacy. 2021;3(3):281–90.
21. Andrew White. By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated—Andrew White; 2021. https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/. Accessed 22 Oct 2021.
22. Zimek A, Filzmoser P. There and back again: outlier detection between statistical reasoning and data mining algorithms. Wiley Interdiscip Rev Data Mining Knowl Discov. 2018;8:e1280.
23. Peachey J, Li G, Chew P, Manak D. Faster and cheaper clinical trials. The benefit of synthetic data. [ebook] Accenture; 2021. https://www.accenture.com/_acnmedia/PDF-148/Accenture-Insilico-Faster-And-Cheaper.pdf. Accessed 7 Oct 2021.
24. Burger H, Gerlinger C, Harbron C, Koch A, Posch M, Rochon J, Schiel A. The use of external controls: to what extent can it currently be recommended? Pharm Stat. 2021;20(6):1002–16.
25. Deloitte Insights. Global Human Capital Trends 2017; 2021. https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2017.html. Accessed 15 Oct 2021.
26. van der Schaar Lab. Synthetic data: breaking the data logjam in machine learning for healthcare // van der Schaar Lab; 2021. https://www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/. Accessed 4 Oct 2021.