SESSION 4: APPLICATIONS OF SYNTHETIC DATA IN THE LIFE SCIENCES INDUSTRY II

uOttawa

Replica Analytics
AN AETION COMPANY

APPLICATIONS OF PRIVACY ENHANCING TECHNOLOGY TO DATA SHARING AT A GLOBAL PHARMACEUTICAL COMPANY

Presented by:

Stephen Bamford,
Senior Director,
Clinical Data Standards & Transparency,
IDAR, J&J Innovative Medicine

Applications of Privacy-Enhancing Technology to Data Sharing

At a Global Pharmaceutical Company

Stephen BAMFORD
Head of Clinical Data Standards & Transparency; IDAR, Global Development
Janssen Research & Development
November 29th, 2023

The opinions in this presentation are my own and do not necessarily reflect the views and policies of J&J

The Background to Data Sharing

Major influencers that help define data sharing processes, policy & compliance

Legal considerations, regulatory commitments and ethical rights are considered before data is made available to external researchers for scientific purposes

In certain situations, we are unable to provide external access to clinical trial data

STUDY TIMELINES

The trial is ongoing or completed too recently. We have an 18-month moratorium after study completion to allow the disseminating of clinical trials results in the peer reviewed biomedical literature.

INFORMED CONSENT

May specify that data cannot be used for research if not directly related to the product / condition / disease state studied in the trial. In cases where conduct of research could benefit public health, it may be permissible to share de-identified data sets.

PRIVACY & CONFIDENTIALITY

Phase 1 trials & studies of rare diseases cannot be always fully de-identified and redacted in such a way as to guarantee the anonymity of the research participants according to US & EU standards.

LEGAL AUTHORITY

Compound has been obtained from, or has been developed in collaboration, with an external partner under a contract that does not permit, or specifically prohibits, external access to the data.

PRACTICAL CONSTRAINTS

Legacy request may only be available in paper formats and not readily accessible or may be collected in a foreign language.

REGULATORY APPROVAL

Clinical trial data will not be made available until after regulatory approval has been granted in the U.S. & European Union.

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

The handling of data has changed dramatically in a relatively short period of time

2005
Data under "lock & key" & hardly ever used post-trial

2015
Open sharing between, within & outside companies

2020
Privacy, legal & ethical considerations control the space

Raiders of the Lost Ark, Lucasfilm Ltd
A Night at the Museum, Twentieth Century Fox

Privacy-Enhancing Technologies

There are different approaches that can be used to share data, each with unique characteristics

Key-Coded Data[1]

- Clinical data from an internal data base (e.g., CDISC® SDTM files)

Pseudonymized Data[2]

- A level of de-identification is done to ensure that there are no unique patients (demographics) or exact event dates in the data, coupled with stronger administrative controls

Anonymized Data[3]

- The industry agreed standard (EMA, Health Canada) for the anonymization of patient data

Clinical Data Synthesis[4]

- Create a synthetic model that is then used to generate artificial, realistic study data

[1] Highly Restricted, Personal Data Type 3 Information
[2] Restricted, Personal Data Type 2 Information
[3] Restricted **or** Confidential **or** Public (once published) Information (depending on the business risk / context / level of anonymization / DUA)
[4] Confidential Information – (No privacy risk, just business risk)

One needs to consider different aspects of the data usage for each of the different approaches

	Key-Coded	GDPR Pseudonymization	Risk-Based De-identification	Clinical Data Synthesis
Governance[4]	maximum	medium	low	minimal
Privacy Risk	very high	medium	low	none
Data Utility	maximum	high	medium	medium
Adherence to the Primary Use agreement	maximum	high[1,2]	medium[1]	minimal[1]
Data Minimization (i.e., providing partial data sets or variables to enable the analysis)	essential[3]	highly recommended	no issue	no issue

[1] Still needs to adhere to corporate data sharing guidelines (e.g., non-commercial use)
[2] The data is anonymized considering the context (i.e., this would not be adequate for public disclosure, but is enough for limited (internal) disclosure with appropriate governance)
[3] In case of data sharing only the data necessary for the purpose should be shared
[4] This could include a data use agreement but does include documentation

The following rules need to be followed depending on each individual use case[1]

	Key-Coded	GDPR Pseudonymization	Risk-Based De-identification	Clinical Data Synthesis
Software Testing (internal)	no[2]	no[2]	yes[3]	yes
Software Testing (external)	no	no	yes[3]	yes
Primary Re-use	yes	yes	yes	yes
Secondary research (internal)	no	yes[4]	yes	yes
Secondary research (external)	no	no	yes	yes

[1] Even if sharing is permissible, there still needs to be ethical governance, including legal considerations
[2] Unless necessary to verify data integrity and there is no possibility to use anonymized or synthetic data
[3] Although this is permissible, it is still preferred to use synthetic data
[4] The data is anonymized considering the context (i.e., this would not be adequate for public disclosure, but is enough for limited (internal/partner) disclosure with appropriate governance)

A careful balance between RISK and DATA UTILITY

GDPR Pseudonymization

- Removal of direct identifiers (e.g., names, ID numbers) while leaving indirectly identifying info (e.g. age, gender, race)
- Used for internal purposes only
- Considered personal information under regulations
- Additional safeguards required when using this data

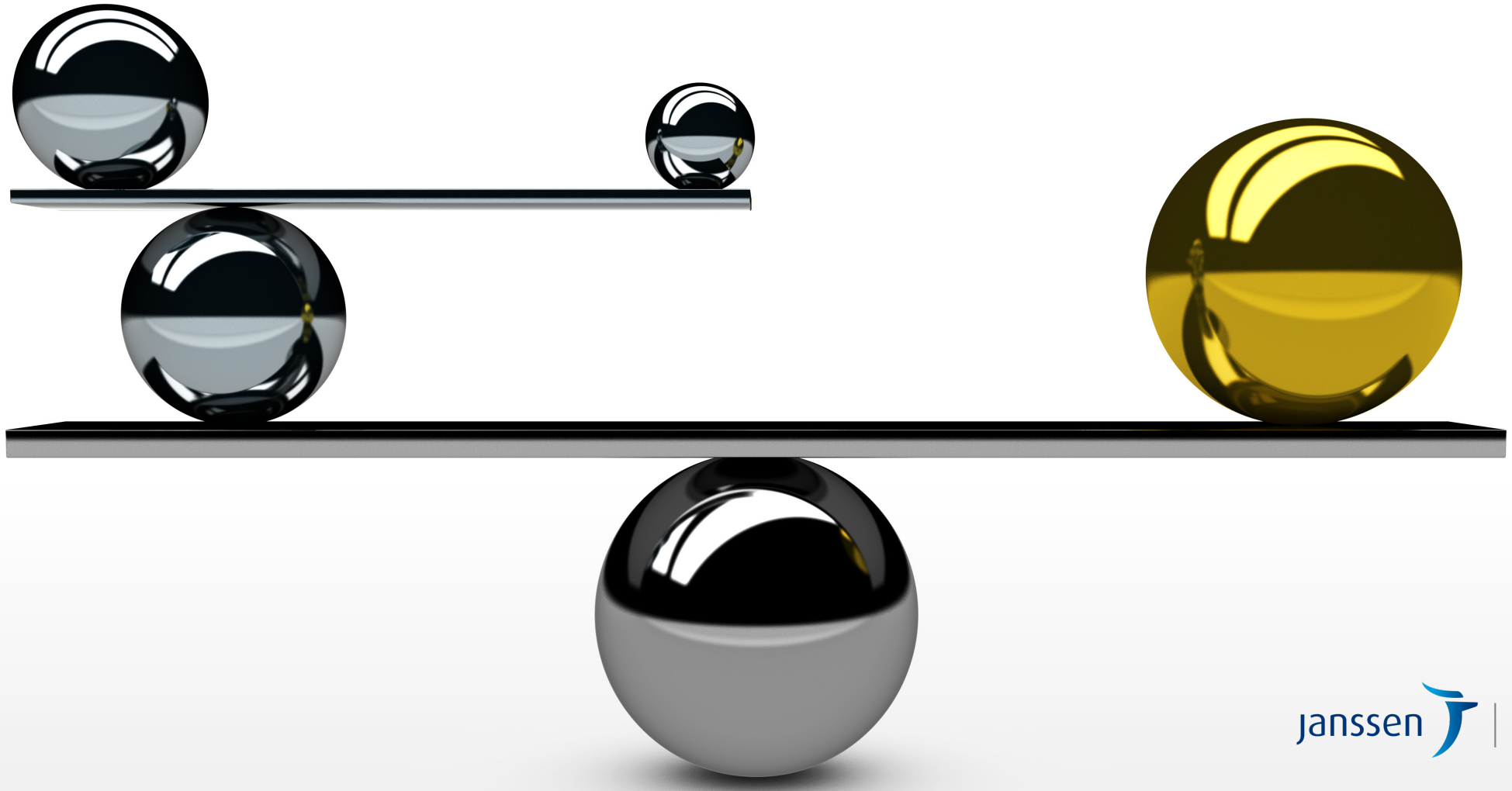
Risk-Based De-identification

- Also called de-identification
- Addresses risk from Indirect Identifiers
- No longer considered personal information
- Anonymized data shared for research purposes

Data Synthesis

- Uses characteristics of a real data set to generate "fake" data
- Models statistical distributions and structure of clinical trial data set
- Generates synthetic data records like the original
- Not considered personal information because data is not linked to actual individuals

Data Controllers have the responsibility to carefully balance RISK with DATA UTILITY



There are three general technical approaches to data synthesis

Statistical Methods

Based on imputation and multiple imputation techniques

Machine Learning Methods

Such as decision trees and random forests

Deep learning methods

Such as autoencoders and generative adversarial networks

There are several ways to synthesize data

Method	Works with small datasets?	Computational demands?
Statistical	Yes	Not computationally demanding
Machine learning	Yes	Computationally demanding
Deep learning	Not very well – *requires a minimal size to be effective; a minimum of 10k to 20k observations (patients) would be typical*	Very computationally demanding; requires high powered machines

NOTE: For all methods, there also needs to be a high ratio between observations (patients) and variables.

Two types of synthetic data can be generated

Partially Synthetic

- Some of the records are synthesized

  OR

- Some of the variables are synthesized

Fully Synthetic

- All of the variables are synthesized

Our approach at Janssen allows us to synthesize the key quasi-identifiers in the dataset

Method	Works with small datasets ?	Computational demands
Statistical	Yes	Not computationally demanding
Machine learning	Yes	Computationally demanding
Deep learning	Not very well – *requires a minimal size to be effective; a minimum of 10k to 20k observations (patients) would be typical*	Very computationally demanding; requires high powered machines

Partially Synthetic

- Some of the records are synthesized

OR

- Some of the variables are synthesized

Fully Synthetic

- All of the variables are synthesized

- We are using:
  - **a machine learning approach**
    - so that we can synthesis small datasets (clinical trials)
  - **a partial synthesis method**
    - to be able to maintain a reasonable observations to variables ratio

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
RANKING				
SCORE[3]				

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY				
PATIENT TRUST				
OPERATIONAL COST				
DATA UTILITY				
RANKING				
SCORE[3]				

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45			
PATIENT TRUST	0.40			
OPERATIONAL COST	0.05			
DATA UTILITY	0.10			
RANKING				
SCORE[3]				

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40			
OPERATIONAL COST	0.05			
DATA UTILITY	0.10			
RANKING				
SCORE[3]				

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05			
DATA UTILITY	0.10			
RANKING				
SCORE[3]				

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05	3	2	1
DATA UTILITY	0.10			
RANKING				
SCORE[3]				

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05	3	2	1
DATA UTILITY	0.10	1	2	2
RANKING				
SCORE[3]				

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05	3	2	1
DATA UTILITY	0.10	1	2	2
RANKING				
SCORE[3]				

$$\frac{(0.45 \times 3) + (0.40 \times 2) + (0.05 \times 3) + (0.10 \times 1)}{4}$$

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05	3	2	1
DATA UTILITY	0.10	1	2	2
RANKING		0.6		
SCORE[3]				

$$\frac{(0.45 \times 3) + (0.40 \times 2) + (0.05 \times 3) + (0.10 \times 1)}{4}$$

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05	3	2	1
DATA UTILITY	0.10	1	2	2
RANKING		0.6	0.39	0.28
SCORE[3]				

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05	3	2	1
DATA UTILITY	0.10	1	2	2
RANKING		0.6	0.39	0.28
SCORE[3]				

Inversely “level−set” the ranges and then proportionally position any middle rankings.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05	3	2	1
DATA UTILITY	0.10	1	2	2
RANKING		0.6	0.39	0.28
SCORE[3]				1

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05	3	2	1
DATA UTILITY	0.10	1	2	2
RANKING		0.6	0.39	0.28
SCORE[3]		0		1

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Privacy Enhancing Technologies

A weighted ranking method

	Weight[1]	Rankings[2]		
		GDPR Pseudonymization	Risk-Based De-identification	Data Synthesis
PRIVACY	0.45	3	1	1
PATIENT TRUST	0.40	2	2	1
OPERATIONAL COST	0.05	3	2	1
DATA UTILITY	0.10	1	2	2
RANKING		0.6	0.39	0.28
SCORE[3]		0	0.65	1

[1]The weights (column) reflect the priority attached to the four criteria that Janssen uses: (a) the extent of privacy protection for the individual, (b) the utility of the data after it has been transformed by the PET, (c) maintaining consumer trust, and (d) the cost sensitivity of the organization. The weights must add up to one.

[2] A rank of one means that a particular PET better satisfies a criterion than a rank of two or three.

[3] The score is a normalized average rank that has a higher value when a particular PET satisfies the four criteria.

Frequently Asked Questions about the use of synthetic data

Is synthetic data utility good enough?

- The weight of evidence is growing rapidly that it works extremely well
- The model accuracy is between 95 – 97% due to the privacy concern

What are the privacy risks with synthetic data?

- Evaluations show that synthetic data is below acceptable thresholds and below that of DEID clinical trial data

Do drug and device regulators accept synthetic data as a surrogate to clinical data?

- There is interest but they are reviewing the evidence as it accumulates

Do privacy regulators accept synthetic data is not personal information?

- This area is very new, but the responses have been positive as it removes a lot of practical problems compared with anonymization

Difference between synthesis and anonymization

Identity disclosure risks for synthetic data are generally lower than identity disclosure risks for anonymization

- Fewer controls needed to share synthetic data = less business and economic burden

In principle synthesis can be highly automated / less labor intensive

- Once all of the automated pipelines are developed

Fewer skills needed to synthesize compared to anonymization

- This requires appropriate automation, but that is necessary in any case
- Makes it easier to scale synthesis

There is an increasingly negative narrative around anonymization because of the frequency of publicized attacks:

- Reduced public trust and reduced regulator confidence
- Initial response from regulators regarding synthetic data has been positive

Can potentially use generative models to perform “simulations” (not applicable to anonymized data)

There are specific use cases for which synthetic data provides an ideal solution

Hackathons and data competitions / challenges

- These require data sets that can be distributed widely with minimal demands on the entrants

Proof of concept and technology evaluations

- Often times technology developers or technology acquirers need to quickly evaluate whether a new technology works well in practice and they need realistic data with which to work, with minimal constraints

Algorithm testing

- One of the biggest challenges when developing AI and machine learning algorithms is getting a sufficient number of data sets, that are large enough, and that are sufficiently realistic on which to test the algorithms

Software testing

- Testing data-driven applications requires realistic data for functional and performance testing. Random data cannot replicate what will happen when a system goes into production

Open data

- Sharing complex data sets publicly is challenging because of privacy concerns. This can now be achieved by sharing synthetic data instead

Data exploration

- Organizations that want to maximize the use of their data can make synthetic versions available for exploration and initial assessment by potential users, and if the exploration yields positive results, the users would go through the process to obtain access to the de- identified data

Algorithm development

- Data analysis programs can be developed on synthetic data and then submitted to the data custodian for execution on the real data – this brings the verified code to the data rather than sharing the data itself

Simple statistics

- When the desired analytics require only a handful of variables, it is possible to use synthetic data as a proxy for real data and to produce more or less the same results

Education and training

- Synthetic data can be used for teaching practical courses on data analysis and for software training


janssen
PHARMACEUTICAL COMPANIES OF
Johnson&Johnson