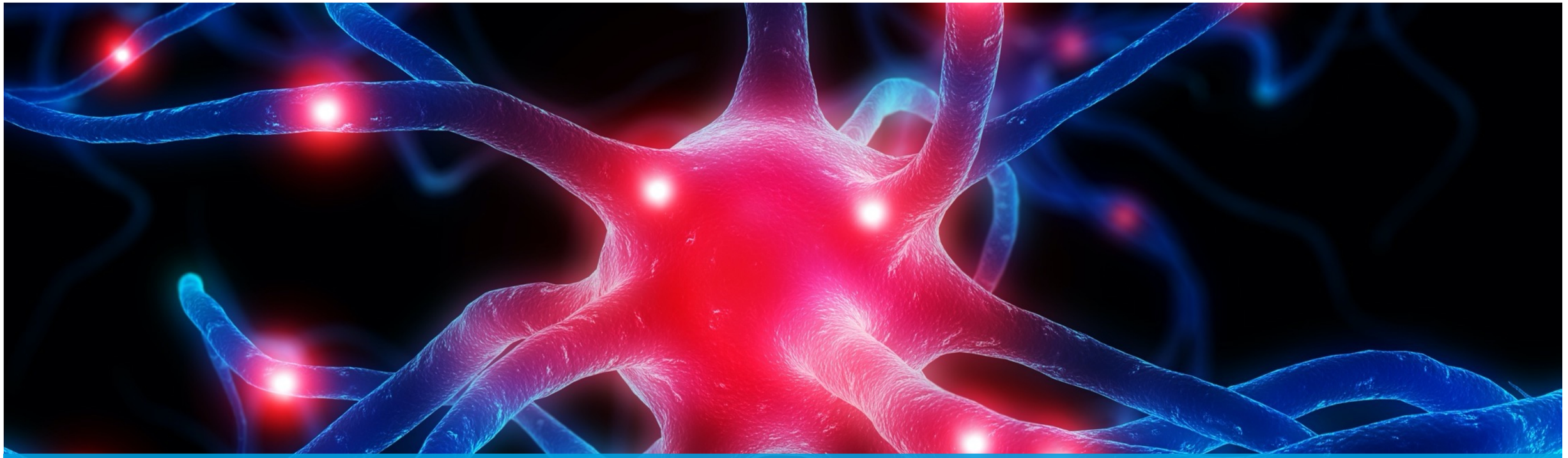


Synthetic Data and Privacy

Experiences Implementing Data Synthesis in a Global Life Sciences Company

Stephen BAMFORD
Head of Clinical Data Standards & Transparency; IDAR, Global Development
Janssen Research & Development
June 16th, 2021

Active neuron

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

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

Stephen Bamford

Also answers to "Bamf".
For almost 25 years, has successfully managed companies, departments and business units. Experienced in the strategic development and delivery of industry innovations. A global operational leader with a pragmatic managerial style that is results-orientated. Experienced in planning and executing business transformations. Focussed on both corporate objectives and meeting stakeholder/customer needs.

Hobbies:

phuse

Alongside my working career, in 2004 I founded the PHUSE organisation. From inception, I have driven, and continue to drive PHUSE that now has over 10,000 global members. It runs over 25 events globally each year, including a data innovation symposium in partnership with the FDA. Our annual "Connect" conferences in both Europe and America regularly attract more than 600 delegates to each event and feature over 125 industry presentations.

- **2004** Founded PHUSE
- **2005** Conference Chair (Heidelberg)
- **2006** Conference Chair (Dublin)
- **2011–2013** PHUSE President
- **2014–2016** PHUSE Chairman of the Board

1991 – Education, *Coventry University*

- BSc (Hons) 2:1 in Statistics and Operational Research

1993 – IGER, *Aberystwyth UK*

- Statistician (1993–1994)

1995 – Pfizer, *Sandwich UK*

- Statistical SAS Programmer – Statistical Analysis and Reporting Group (1995–1998)
- Associate Director – Statistical Analysis and Reporting Group (1998–2002)
- Director – Statistical Analysis and Reporting Group (2002–2005)
- European Head of Phase I Clinical Programming & Medical Writing (2005–2007)

2004 – Quanticate, *Canterbury UK*

- Head of Programming and Medical Writing (2007–2008)
- General Manager (2009–2011)

2011 – Cerafor, *Broadstairs UK*

- Executive Director (2011–2014)

2014 – Business & Decision Life Sciences, *Brussels BE*

- Director of Biometrics Consultancy (2014–2016)

2016 – Janssen (J&J), *High Wycombe UK*

- Director of Information & Knowledge Management (2016–2017)
- Head of Data Transparency (2018–2020)
- Head of Clinical Data Standards & Transparency

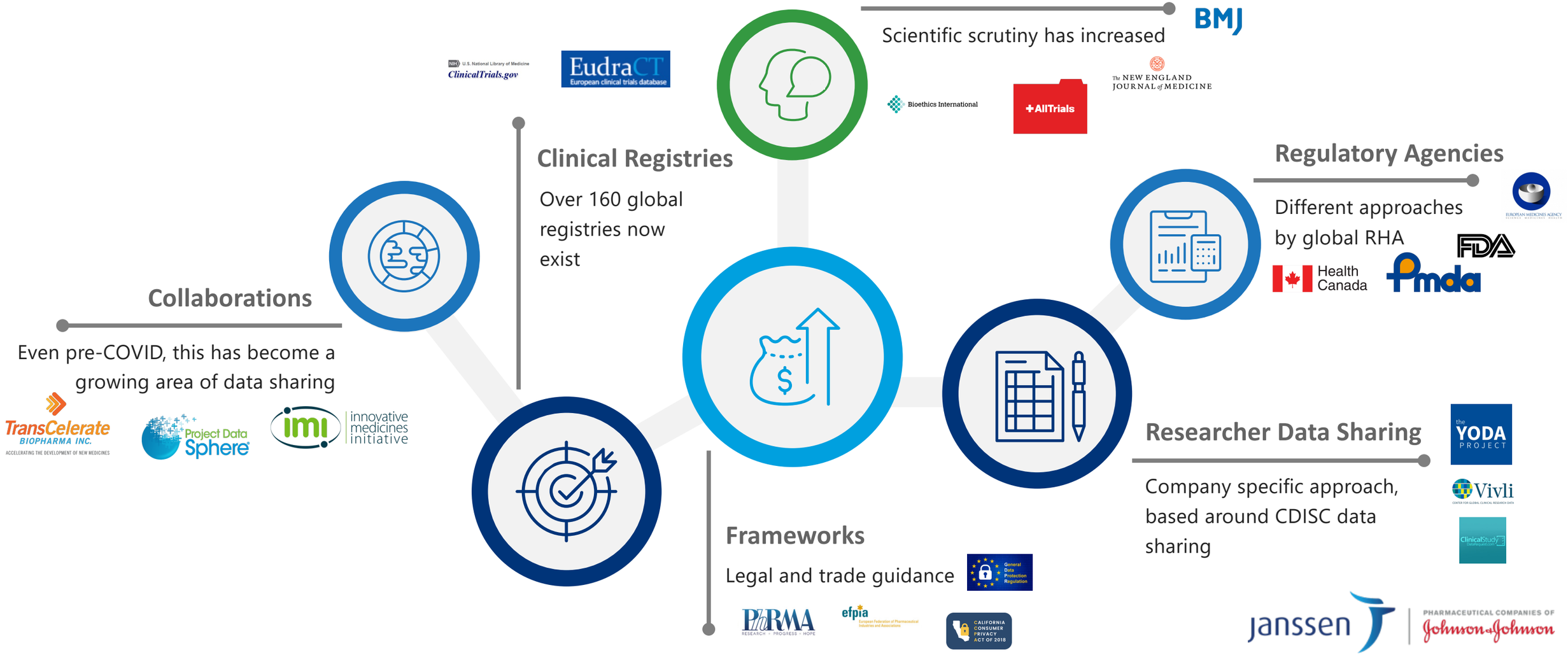
Today’s Agenda and Purpose

1	**Background of Sharing**	What has led us to where we are today & what do others do?
2	**Privacy-Enhancing Technologies (PETs)**	Overview of the different approaches to protect personal information
3	**Synthetic Data within Janssen**	The journey so far and what progress has been made
4	**Thoughts Around the Future of Synthetic Data**	Reflections on how/where synthetic healthcare data can progress
5	**Q&A**	Open discussion

The Background to Data Sharing

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

Enhanced transparency obligations have evolved quickly over the last few years



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

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

GDPR Pseudonymization

- 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

Risk-Based De-identification

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

Clinical Data Synthesis

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

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

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	very low
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

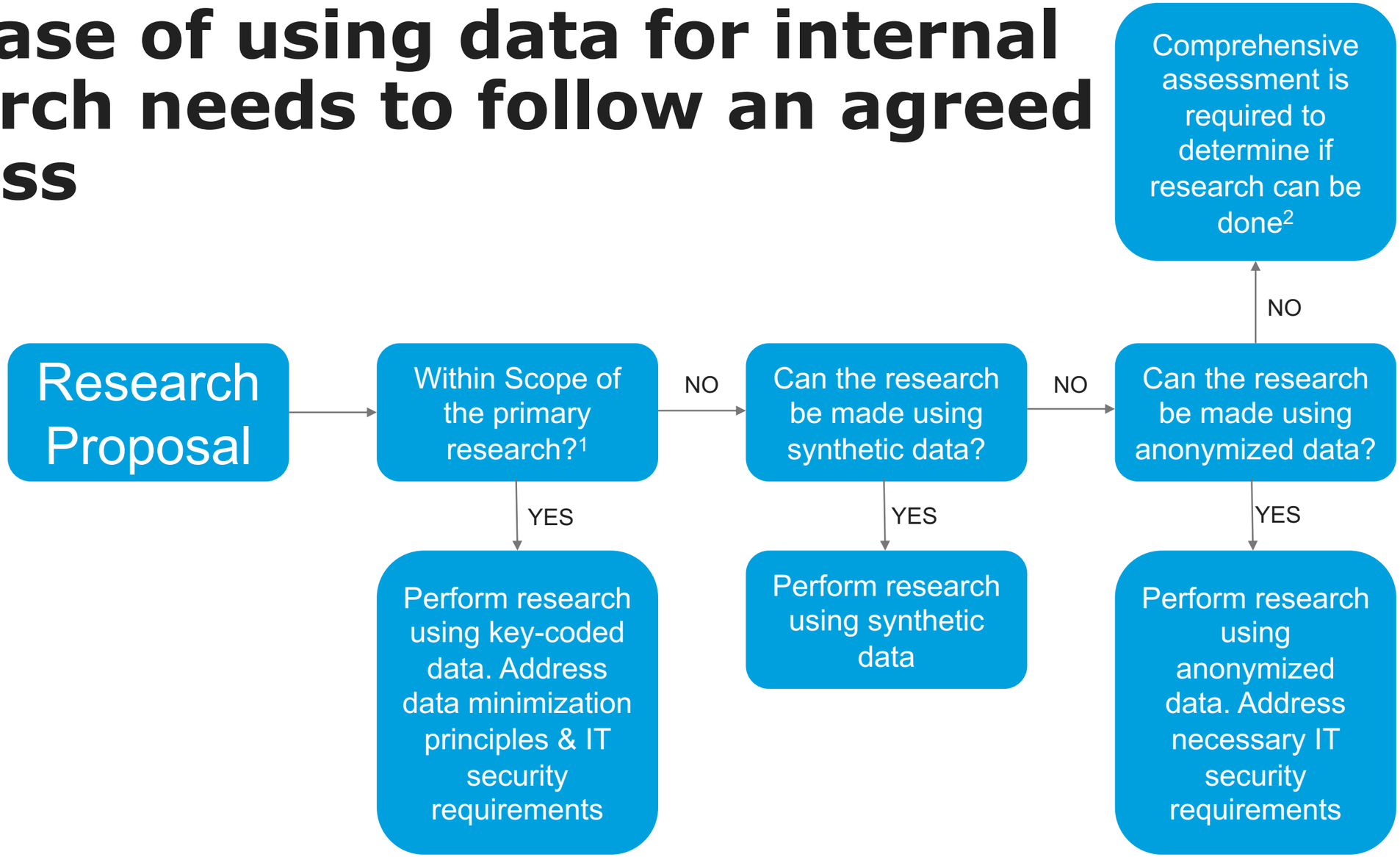
The following rules need to be followed depending on each individual use case[1]

	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)

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

The case of using data for internal research needs to follow an agreed process

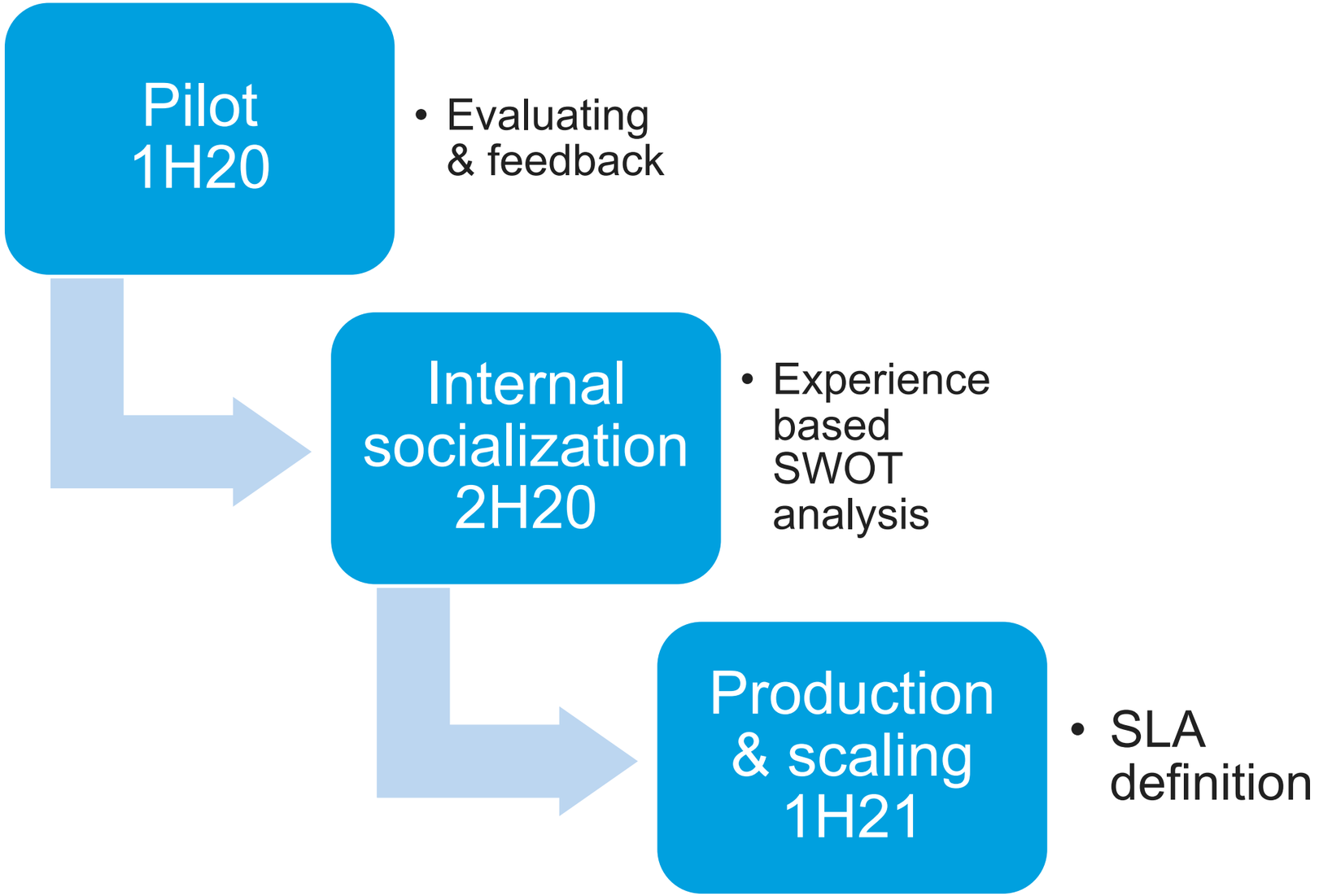


[1]Primary research is all scientific research activities to learn about the pharmaceutical product, get permission to introduce and keep it on the market, monitor its safety and get it covered by health insurances

[2]The scope of the research and any associated prerequisite to be considered

Synthetic Data within Janssen

To date, 33 clinical studies have been synthesized



janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

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

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

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?

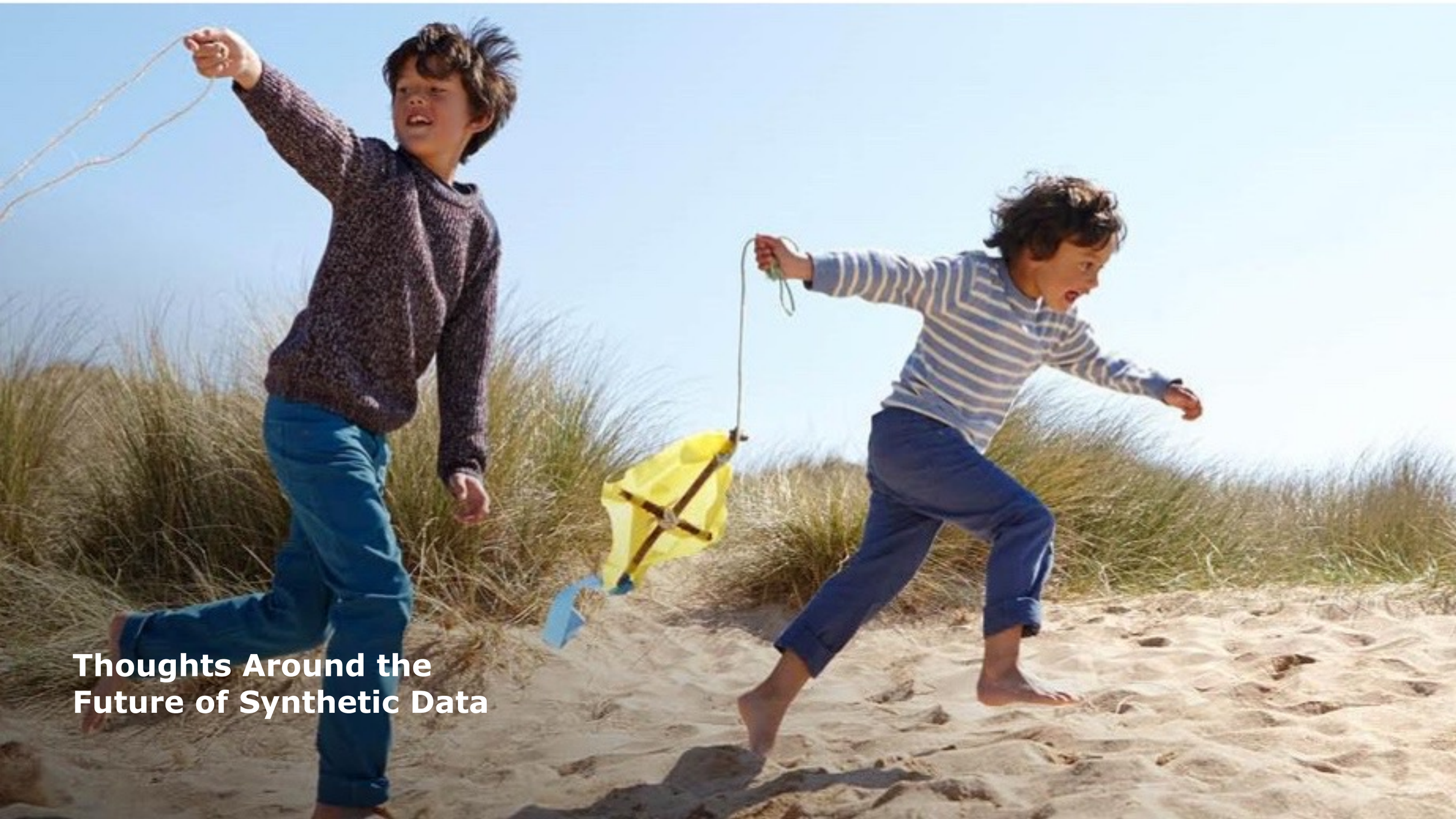
- Evaluations show that synthetic data is below acceptable thresholds and below that of deidentified clinical trial data

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


Thoughts Around the
Future of Synthetic Data

Acceptance of synthetic data within Janssen

examples

- 95% of statistical modeling can be done on synthetic data
- The data uses have been growing over time from testing to data science

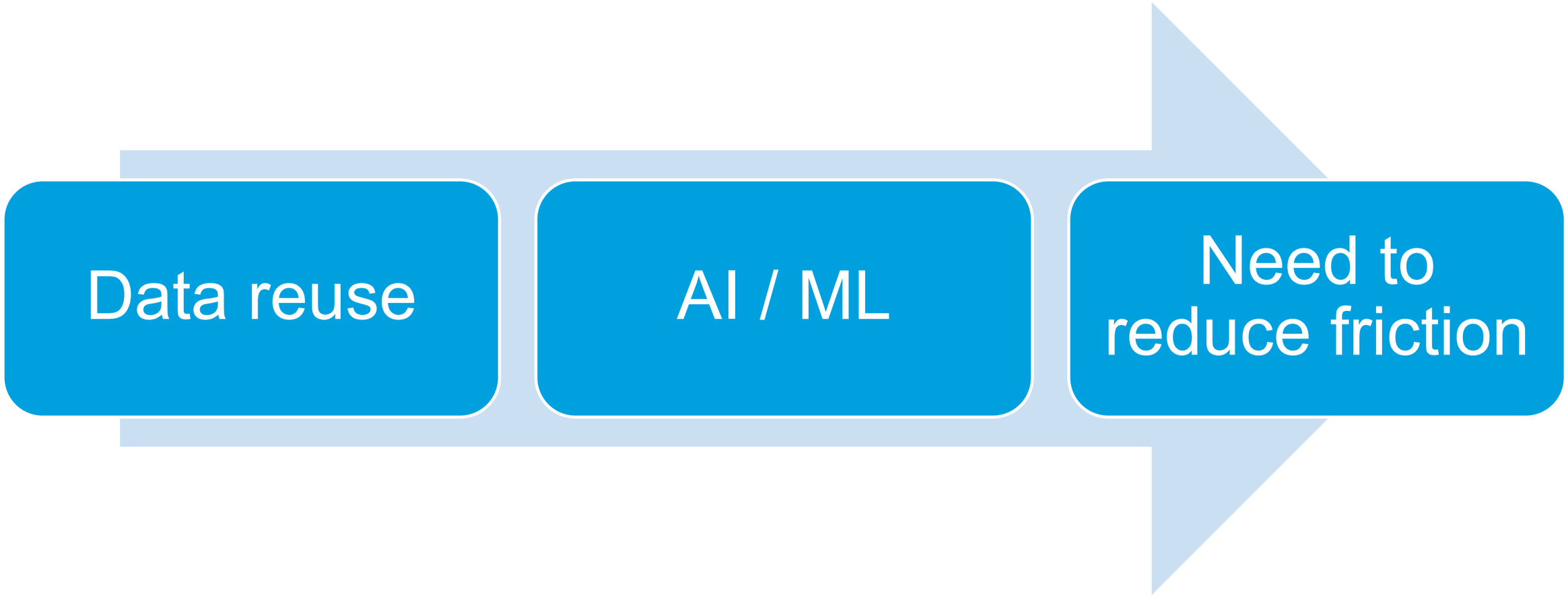
methods

- Access to synthetic data is easier than alternative methods
- Future use of simulators can be explored

regulators

- Regulator perspectives will be very important for internal adoption
- Release a set of parameters for when synthetic data would be considered non-identifiable

Acceptance of synthetic data more broadly in the life sciences industry



Closing Remarks

clear guidance and frameworks for evaluating privacy risks

reduce uncertainty, easier to adopt, increase use of synthetic data


janssen
PHARMACEUTICAL COMPANIES OF
Johnson&Johnson