Replica Analytics

AN AETION COMPANY

Generating & Evaluating Synthetic Clinical Trial Data in a Pharmaceutical Company

Mark Baillie & Lucy Mosquera

Agenda

- Overview of the requirements and use cases for synthetic clinical trial data
- Clinical trial datasets used in this case study
- Synthesis process for clinical trial data, utility, and privacy results
- Impact of the synthetic clinical trial datasets

Replica Analytics
AN AETION COMPANY

AMDS, Analytics
Global Drug Development

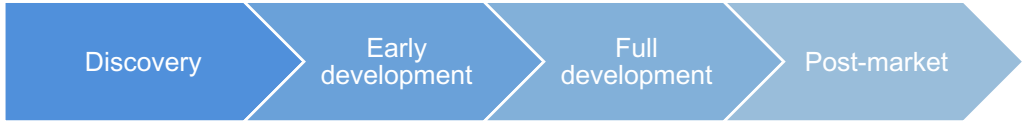
Synthetic data and clinical drug development

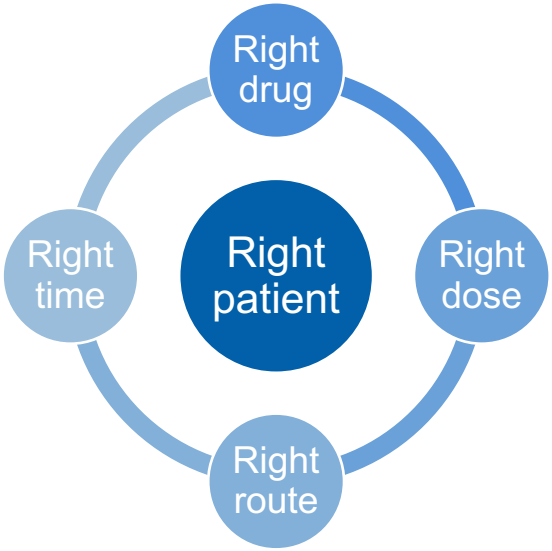
Mark Baillie
Director, Data Science
January 18th, 2023

NOVARTIS | Reimagining Medicine

Why synthetic data?

The potential value for Novartis is to facilitate timely access to realistic data to drive both internal and external projects and programs.





Synthetic data may enable **learning from existing and future data** using advances in statistics, machine learning, data science and AI:

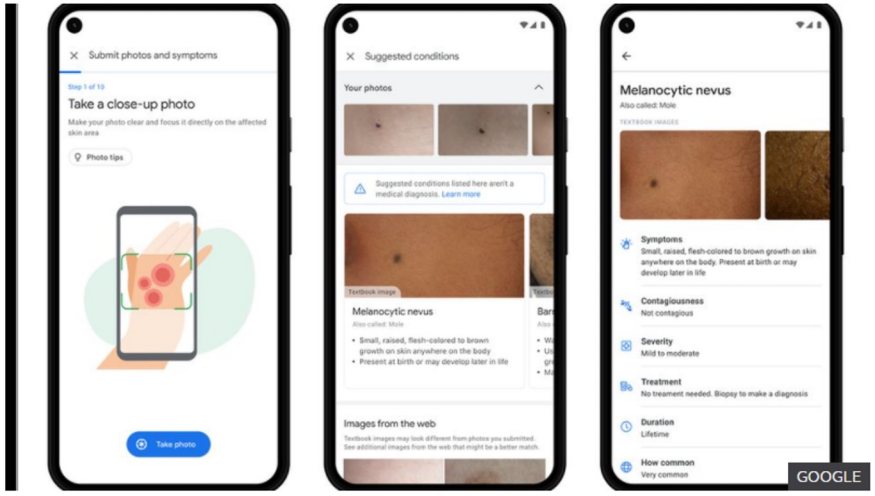
- increase understanding of drug, disease and patients,
- accelerate and improve development projects, and
- inform decision making.

We want to leverage advances in data science and AI

Google AI tool can help patients identify skin conditions

By Zoe Kleinman
Technology reporter

20 hours ago

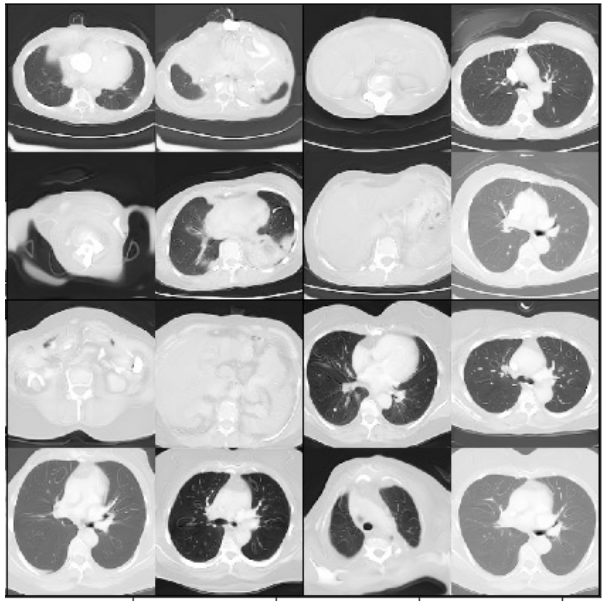


Google has unveiled a tool that uses artificial intelligence to help spot skin, hair and nail conditions, based on images uploaded by patients.

Source: https://www.bbc.com/news/technology-57157566

NOVARTIS | Reimagining Medicine

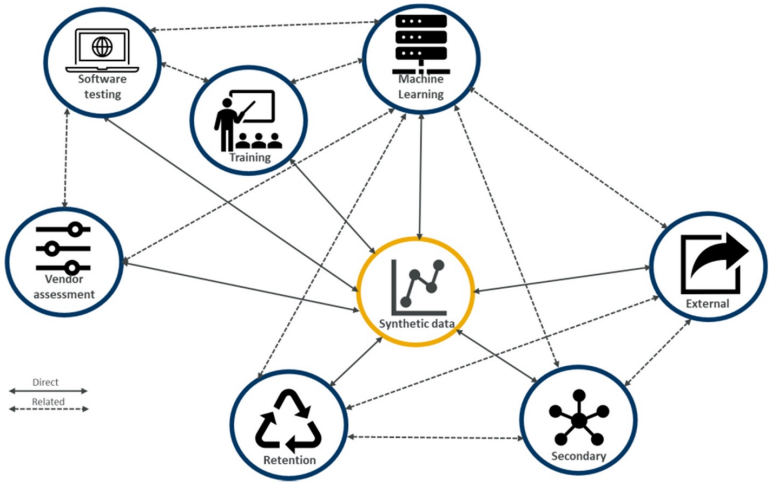
We view synthetic data as an enabler



Source: Multi-Modal Conditional GAN: Data Synthesis in the Medical Domain. Jonathan Ziegler et al. https://openreview.net/pdf?id=8PI7W3bCTI

- Synthetic data is an important and visible research topic with growing interest from multiple industry, academic and regulation stakeholders
- Data sharing is a driver:
  - An alternative to anonymization for data sharing (i.e., improved privacy and data utility)
  - Generate **privacy preserving** datasets with robust utility that can be used for **collaborations** and knowledge generation (tools and publications**)**
- Synthetic data solutions may also have intrinsic value beyond data sharing
  - Use conditional generation to augment sparse datasets (e.g., 2D/3D images)

What are the potential use cases?



Source: James, S., Harbron, C., Branson, J. et al. Synthetic data use: exploring use cases to optimise data utility. Discov Artif Intell 1, 15 (2021). https://doi.org/10.1007/s44163-021-00016-y

- Statistical/machine learning methodology development and benchmarking
- **Internal (external) software development**
- **Education, training, data challenges, and hackathons**
- Internal secondary use
- Data retention
- **Vendor assessments and engagements**
- External sharing

Requirements for the synthetic clinical trial data case study

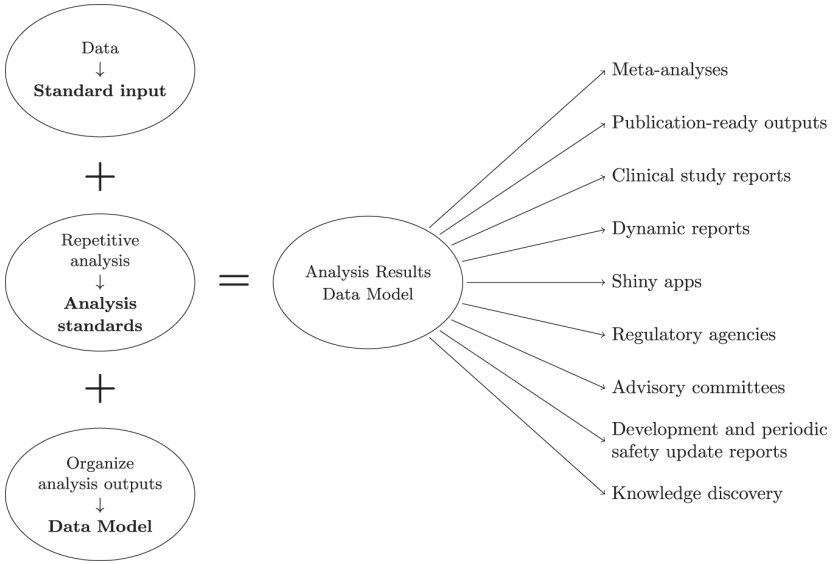
- **The value for Novartis** is to facilitate timely access to realistic data to drive both internal and external projects and programs and avoid downtime as we wait for anonymization or approval to share data.
- **Case study purpose –** to support the internal focused use cases of tool development (for trial reporting), and methodology development.
- **Task definition** - synthesis of six complete Phase 3/4 clinical trials (CDISC ADaM)
- **Success criteria:**
  - **Privacy**
  - **Utility**

Six clinical trials covering various therapeutic areas and study designs

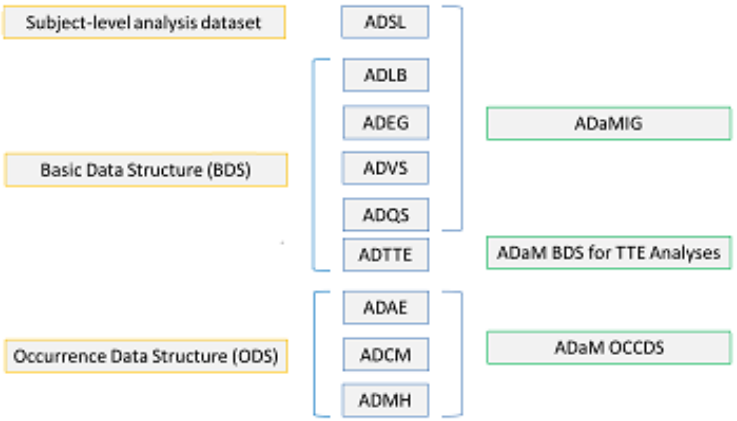
	Indication	Design/phase	#patients
Immunology	Plaque Psoriasis	52-week, Randomized, Double-blind Phase III Study	1,114
Cardiovascular	Recurrent Major CV Disease	Randomized, Double-blind, Placebo-controlled, Event-driven Phase III Trial	10,066
	Acute heart failure	Phase IIIb outcome study in AHF patients was designed as a multicenter, randomized, double-blind, placebo-controlled, event-driven study	6,600
Renal	Renal Transplantation	2-year, randomized, multicenter, open-label, 2-arm Phase IV study	2,037
Respiratory	Asthma	Multicenter, Randomized, 52-week, Double-blind, Parallel group, Active Controlled Phase III Study	3,092
Oncology	Breast Cancer	Multi-center, randomized, double-blind, placebo controlled Phase III study	1,147

Overview of the clinical trial data

- All available study data, stored in CDISC ADaM format - focus on the analysis and reporting
- Synthesis of all measurement domains including:
  - demographics and baseline characteristics, (ADSL),
  - adverse events (ADAE),
  - laboratory measurements (ADLB),
  - time to event (ADTTE) and,
  - efficacy (ADEFF).
- Many linked, curated domains with varying data structures
  - Business logic added to support trial reporting i.e., imputations, aggregations, derivations, etc.
  - The goal is to ensure measurements are be linked together to retain coherent, consistent and logical patients



Source: Barros, J.M., Widmer, L.A., Baillie, M. *et al.* Rethinking clinical study data: why we should respect analysis results as data. *Sci Data* **9**, 686 (2022). https://doi.org/10.1038/s41597-022-01789-2



Source: http://pharma-sas.com/introduction-on-how-to-create-adam/

Requirements for the synthetic clinical trial data case study

- **The value for Novartis** is to facilitate timely access to realistic data to drive both internal and external projects and programs and avoid downtime as we wait for anonymization or approval to share data.
- **Case study purpose –** to support the internal focused use cases of tool development (for trial reporting), and methodology development.
- **Task definition** - synthesis of six complete Phase 3/4 clinical trials (CDISC ADaM)
- **Success criteria:**
  - **Privacy**:
    - facilitate internal data sharing – clinical study data has restricted access
  - **Utility**:
    - The synthetic data is a 1:1 replication of the original data provided in terms of structure
    - Variables that are numerical, binary, or categorical (ordinal or non-ordinal) remain the same with similar distributions and min/max characteristics
    - The synthetic data should have similar characteristics to the original data but not be identical i.e., primary and secondary research follow similar trends

Introduction to synthetic clinical trial data

Replica Analytics

AN AETION COMPANY

Synthetic data

WHAT IT IS

Synthetic data is **generated from real data**, but is not real data.

WHY IT MATTERS

It has the **same patterns and statistical properties** as real data.

Real

COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	3	1	0	1	25.44585
United States	3	1	1	0	24.09375
United States	3	1	1	1	33.07829
United States	2	1	1	0	33.64845
United States	3	1	1	0	25.66958
United States	3	1	1	0	25.85938
United States	2	1	1	0	24.7357
United States	5	0	0	0	27.75276
United States	5	0	1	1	28.07632

Synthetic

COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

HOW IT CAN BE USED

For certain use cases it **can act as a proxy for real data**.

Replica Analytics

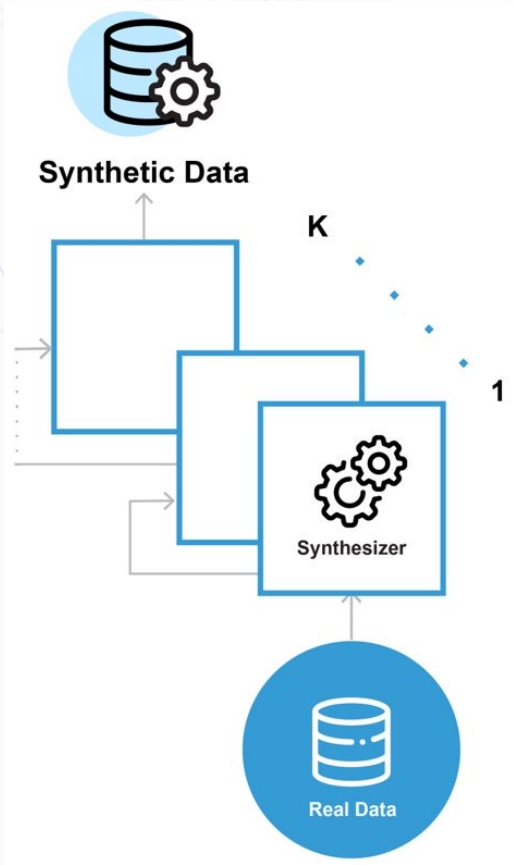
AN AETION COMPANY

Synthetic data generation

Machine learning or deep learning models **capture patterns** in the real data, and then **generate new data** from that model.

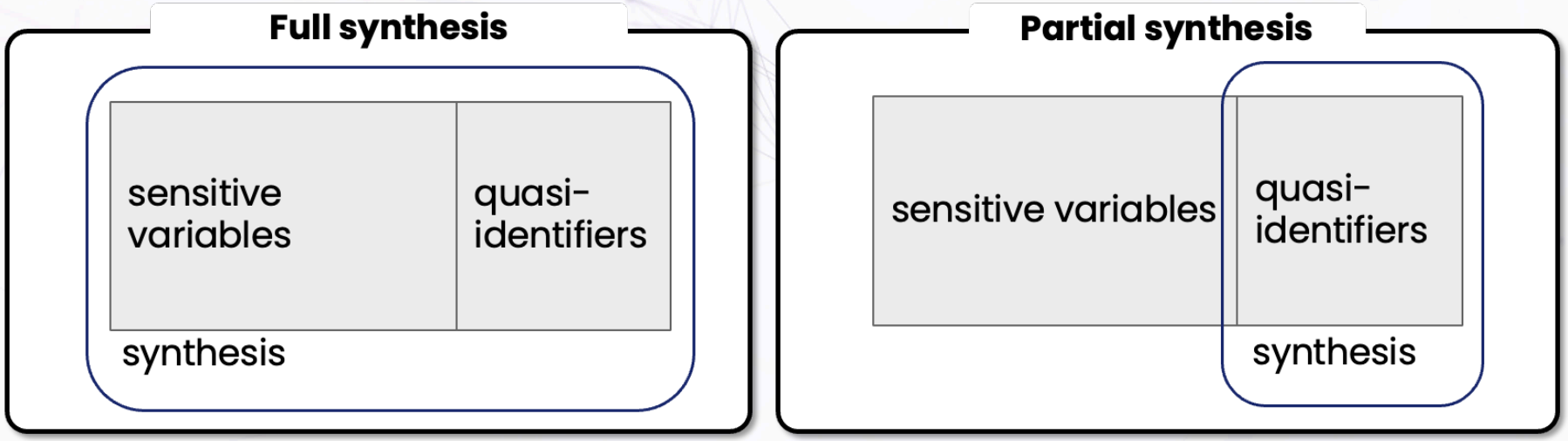
Replica Analytics
AN AETION COMPANY

Sequential synthesis utilizes multiple machine learning methods in a sequence



Replica Analytics
AN AETION COMPANY

Two types of synthesis



Necessary for **maximum privacy** or when amplifying **smaller datasets**

Sometimes used **when the data is complex** or **for small datasets** with very large number of variables

Replica Analytics
AN AETION COMPANY

Partial synthesis for clinical trial data

- Synthesis of quasi-identifiers and key baseline measures
- Mitigates privacy risks while producing high utility data
- Allows for synthesis of all domains in a clinical trial dataset
- Avoids the challenges seen in clinical trial data posed by modelling rare or unique patterns in data (e.g., an adverse event that leads to additional concomitant medications or hospitalizations)
- Additional strategies to mitigate risk such as date shifting and suppression of free text

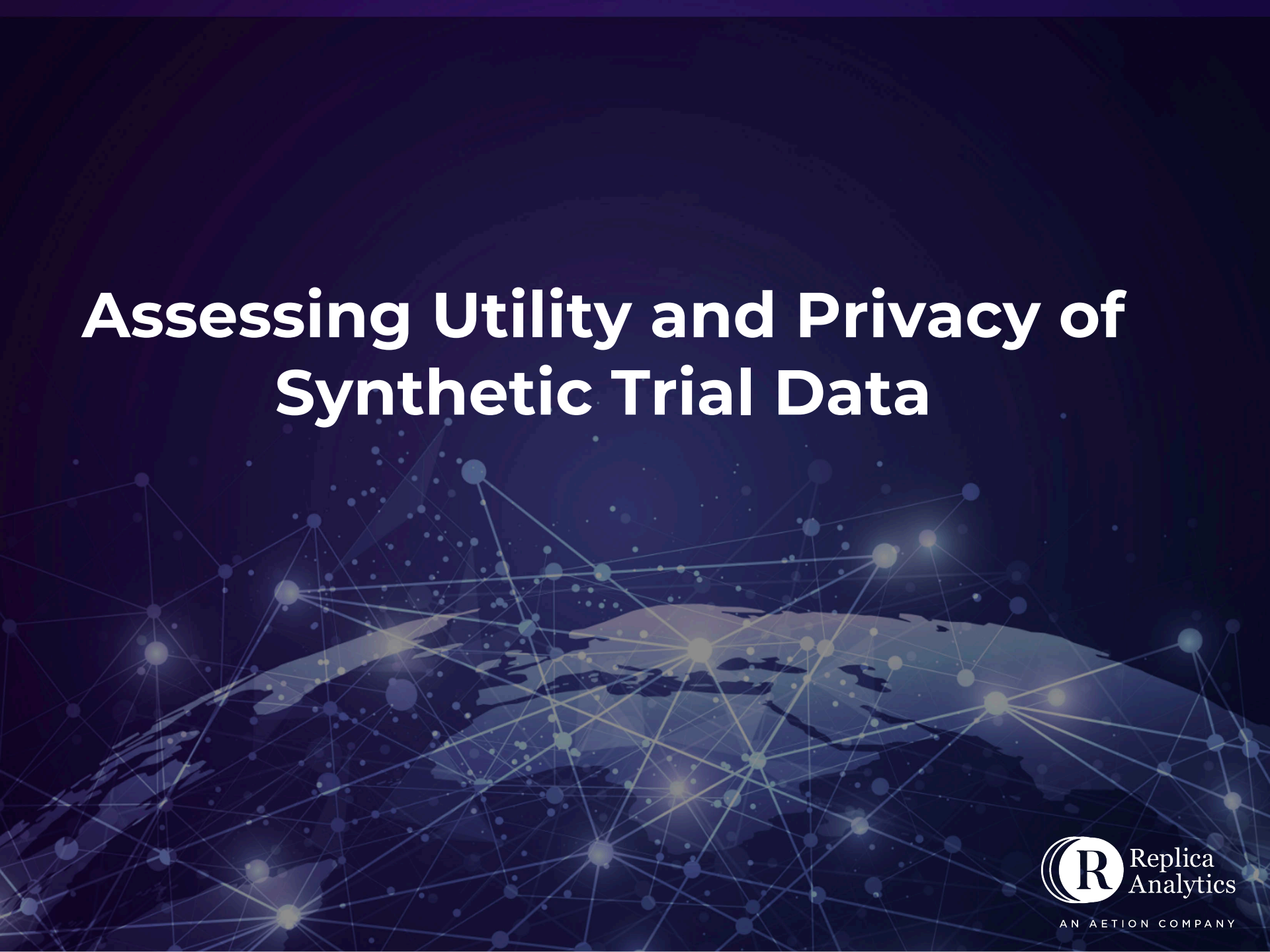
Replica Analytics
AN AETION COMPANY

Partial synthesis vs traditional de-identification for clinical trial

- Synthesis maintains correlations between variables while traditional de-identification transforms each variable independently
- Synthesis can scale to complex datasets with many quasi-identifier variables while traditional de-identification may require shortcuts such as examining only a subset of quasi-identifiers to scale to complex datasets
- Less suppression required meaning key variables will be present in the output data

Replica Analytics
AN AETION COMPANY

Assessing Utility and Privacy of Synthetic Trial Data



Replica Analytics

AN AETION COMPANY

Utility assessments

- Generic or broad utility assessments show how similar synthetic data is to the real data it was generated from without referencing a specific analysis
- Our utility assessment framework compares real and synthetic data using a range of metrics aimed to assess similarity at different levels of complexity

Replica Analytics
AN AETION COMPANY

Univariate comparison

- Hellinger distance compares the distribution of a variable in the real data to the distribution seen in the synthetic data
  - Hellinger distance = 0 → the distributions are identical
  - Hellinger distance = 1 → the distributions are completely different and non-overlapping

Trial	Hellinger distance - Median (IQR)	Trial	Hellinger distance - Median (IQR)
CACZ885M2301	0.000 (0.002)	CQVM149B2302	0.002 (0.008)
CAIN457A2326	0.000 (0.014)	CRAD001A2433	0.002 (0.014)
CBKM120F2302	0.000 (0.009)	CRLX030A2301	0.000 (0.004)

Replica Analytics
AN AETION COMPANY

Bivariate comparison

- Assesses the absolute difference in bivariate correlation between all pairs of variables seen in the dataset
  - Uses correlation metrics specific to the data types of the variables (e.g., Pearson's correlation for pairs of continuous variables or Cramer's V for pairs of categorical variables)

Trial	Absolute difference in bivariate correlation - Median (IQR)	Trial	Absolute difference in bivariate correlation - Median (IQR)
CACZ885M2301	0.001 (0.008)	CQVM149B2302	0.004 (0.015)
CAIN457A2326	0.003 (0.019)	CRAD001A2433	0.005 (0.017)
CBKM120F2302	0.004 (0.019)	CRLX030A2301	0.001 (0.008)

Replica Analytics
AN AETION COMPANY

Multivariate comparison

- Assesses the absolute difference in predictive ability seen for a model trained on the real data compared to a model trained on the synthetic data, iterating over every variable in the dataset as the 'target' variable

Trial	AUROC difference - Median (IQR)	AUPRC difference - Median (IQR)	Trial	AUROC difference - Median (IQR)	AUPRC difference - Median (IQR)
CACZ885M2301	0.0016 (0.094)	0.028 (0.132)	CQVM149B2302	0.028 (0.132)	0.0003 (0.048)
CAIN457A2326	0.002 (0.014)	0.002 (0.025)	CRAD001A2433	0.002 (0.025)	0.010 (0.044)
CBKM120F2302	0.001 (0.020)	0.002 (0.024)	CRLX030A2301	0.002 (0.024)	0.0001 (0.056)

Replica Analytics
AN AETION COMPANY

Privacy concerns with synthetic data

In general, identity disclosure is not the main type that is of concern

- Unless the generative model has been overfit, in which case many records would just be replicated; but that should not be a common occurrence

We are concerned with other types of inferences from the dataset:

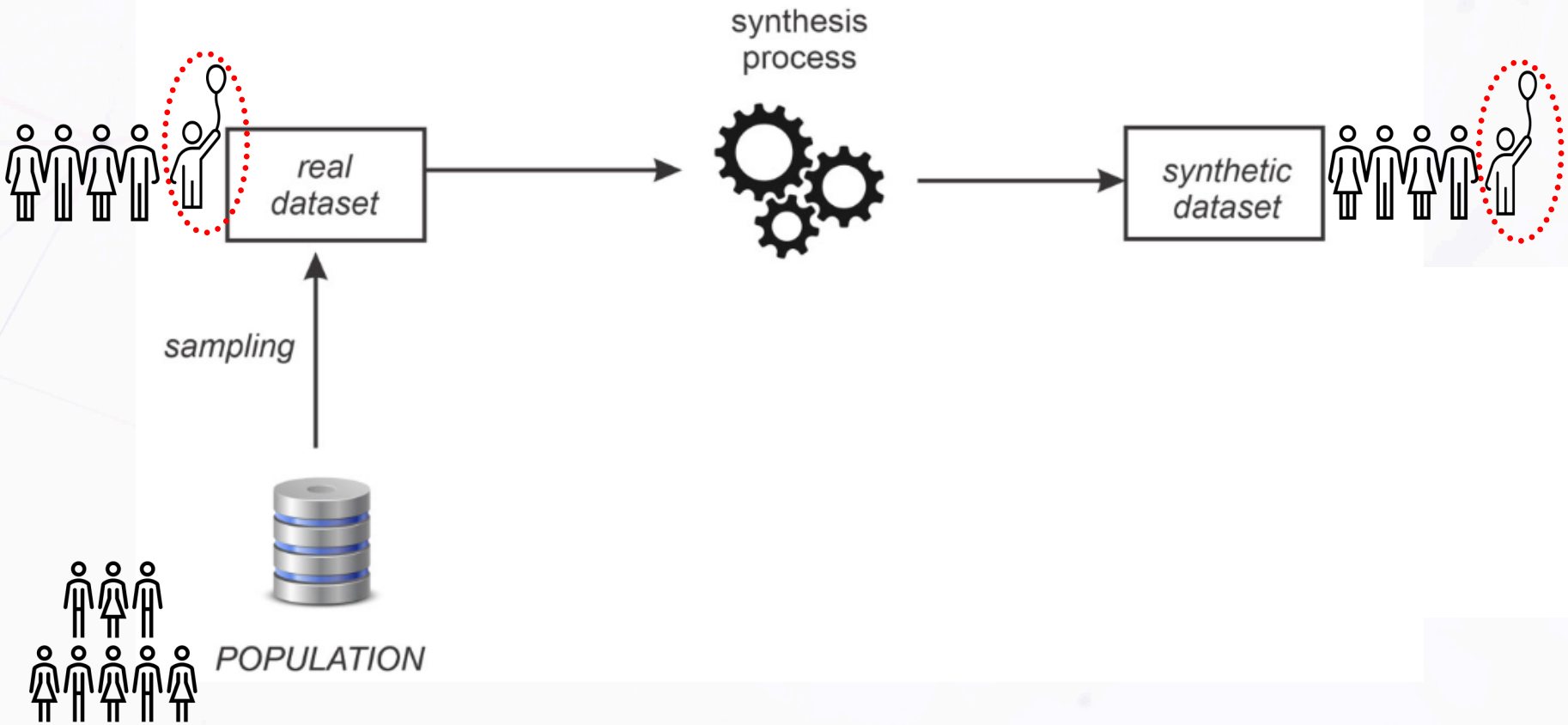
- Attribution disclosure
- Membership disclosure

Replica Analytics
AN AETION COMPANY

Membership disclosure

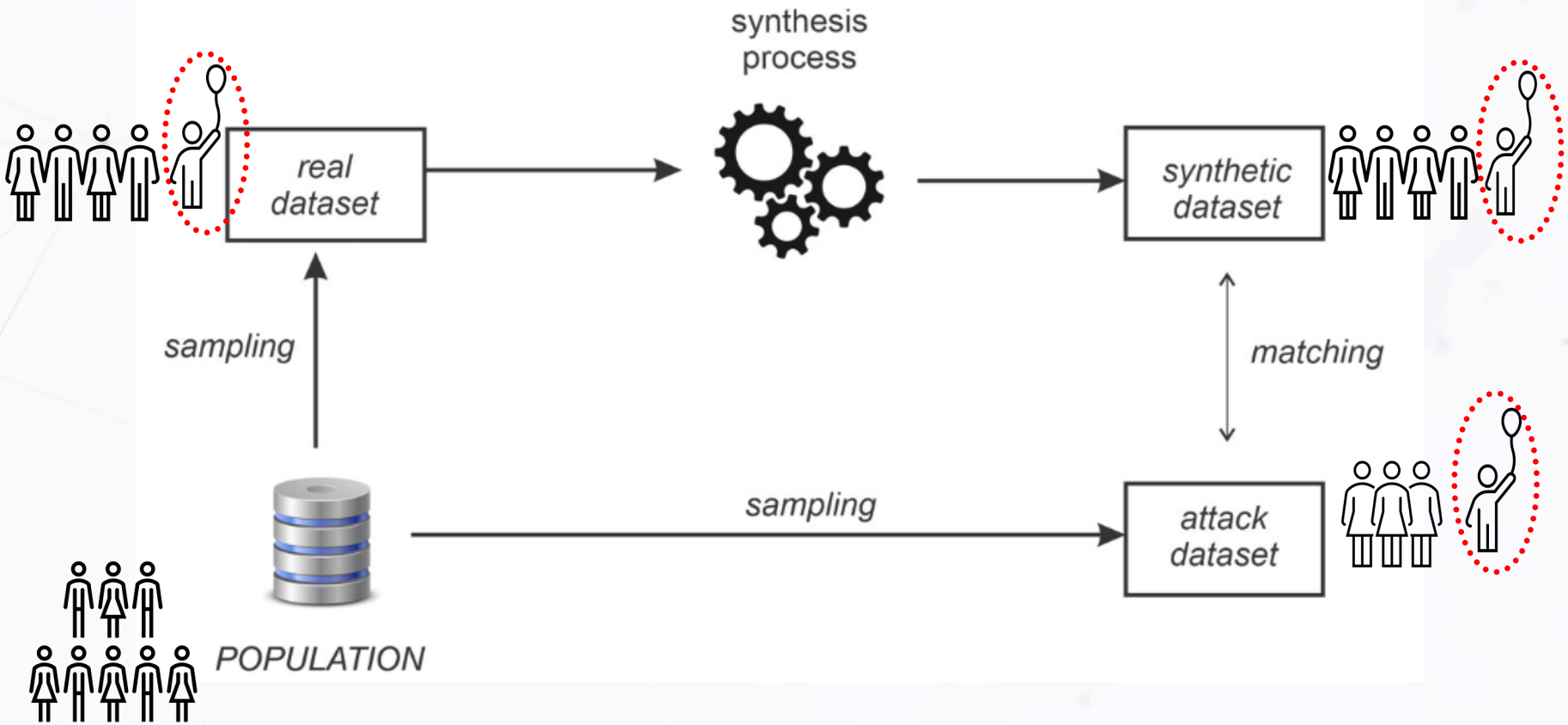
- To what extent an adversary could determine that a target individual is in the training data that was used for training the generative model
- Knowing that someone is in the training dataset may reveal sensitive information about them, for example, if the dataset was about individuals who participated in an HIV study

The process for a membership disclosure attack



Replica Analytics
AN AETION COMPANY

The (ground truth) process for a membership disclosure attack



Replica Analytics
AN AETION COMPANY

Probability of membership disclosure results

Trial	Max Risk	Max Adjusted Risk
CACZ885M2301	0.0057	<0.0001
CAIN457A2326	0.0018	<0.0001
CBKM120F2302	0.0035	<0.0001
CQVM149B2302	0.0002	<0.0001
CRAD001A2433	0.0117	<0.0001
CRLX030A2301	0.0076	0.000275

Replica Analytics
AN AETION COMPANY

Probability of membership disclosure results

Trial	Max Risk	Max Adjusted Risk
CACZ885M2301	0.0057	<0.0001
CAIN457A2326	0.0018	<0.0001
CBKM120F2302	0.0035	<0.0001
CQVM149B2302	0.0002	<0.0001
CRAD001A2433	0.0117	<0.0001
CRLX030A2301	0.0076	0.000275

All below the commonly used acceptable risk threshold of 0.09

Replica Analytics
AN AETION COMPANY

Utility and privacy conclusions

- Across all 6 datasets the synthetic data produced retains the same patterns and relationships seen in the real data
- Within a dataset there is a range of utility values, some patterns or relationships may be reproduced better than others
- Across all 6 datasets the synthetic data produced has low privacy risks in terms of membership disclosure. Without adjustment the max risk values seen are all below acceptable thresholds

Replica Analytics
AN AETION COMPANY

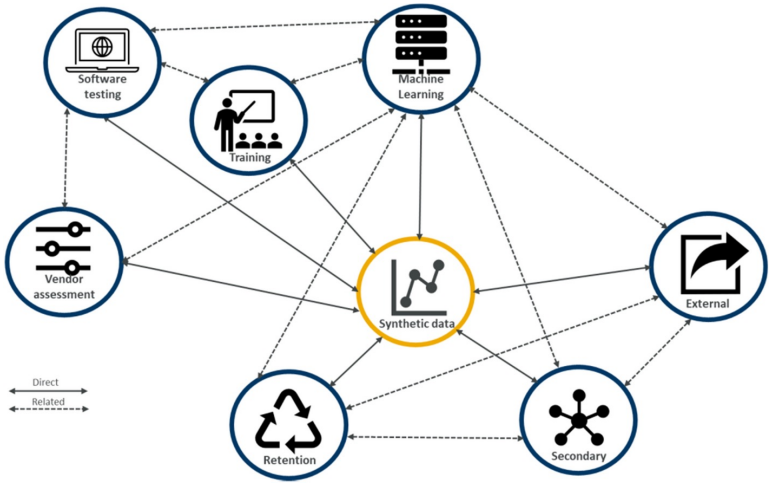
Impact of the synthetic clinical trial datasets?

Mark Baillie

Requirements for the synthetic clinical trial data case study

- **The value for Novartis** is to facilitate timely access to realistic data to drive both internal and external projects and programs and avoid downtime as we wait for anonymization or approval to share data.
- **Case study purpose –** to support the internal focused use cases of tool development (for trial reporting), and methodology development.
- **Task definition** - synthesis of six complete Phase 3/4 clinical trials (CDISC ADaM)
- **Success criteria:**
  - **Privacy**:
    - facilitate internal data sharing – clinical study data has restricted access
  - **Utility**:
    - The synthetic data is a 1:1 replication of the original data provided in terms of structure
    - Variables that are numerical, binary, or categorical (ordinal or non-ordinal) remain the same with similar distributions and min/max characteristics
    - The synthetic data should have similar characteristics to the original data but not be identical i.e., primary and secondary research follow similar trends

Impact of the synthetic clinical trial datasets?



Source: James, S., Harbron, C., Branson, J. et al. Synthetic data use: exploring use cases to optimise data utility. Discov Artif Intell 1, 15 (2021). https://doi.org/10.1007/s44163-021-00016-y

- Statistical/machine learning methodology development and benchmarking
- **Internal (external) software development**
- **Education, training, data challenges, and hackathons**
- Internal secondary use
- Data retention
- **Vendor assessments and engagements**
- External sharing

Impact of the synthetic clinical trial datasets?

Outcomes:

- we have six fully synthetic clinical trials i.e., immediate value
- ongoing plans to make this resource available for future use cases
- track usage to better understand and assess demand and understand better the benefits of synthetic data

The future:

- Policy and guidelines for internal and external use of synthetic data
- Explore further the value proposition for synthetic data value
  - Data augmentation (e.g., for clinical trial design, prior / evidence synthesis)
  - The balance between privacy/utility for external data sharing
- Further evaluation of synthetic data and data generators (i.e., models) for further use

Questions?

Replica Analytics

AN AETION COMPANY

Thank you!

Replica Analytics

AN AETION COMPANY