SESSION 3: APPLICATIONS OF SYNTHETIC DATA IN THE LIFE SCIENCES INDUSTRY I

uOttawa

Replica Analytics
AN AETION COMPANY

APPLICATIONS OF SYNTHETIC DATA IN SUPPORTING DATA PRODUCTS AND ANALYTICS FOR REAL WORLD EVIDENCE

Presented by:

Leonardo D’Ambrosi,
Senior Lead Data Scientist,
Bayer

Shared learnings from the field

Synthetic Data for Real World Data Analytics

//////////

London, 30.11.2023

Leonardo D’Ambrosi

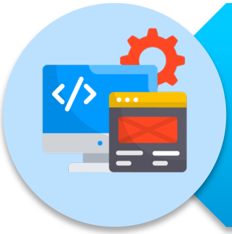
image: DALL-E 3

Disclaimer

The considerations and opinions are my own and do not necessarily reflect the views of Bayer

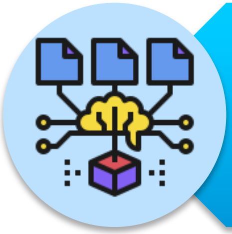
Why Synthetic Data?

Why synthetic data?



Software development

Privacy / Anonymization



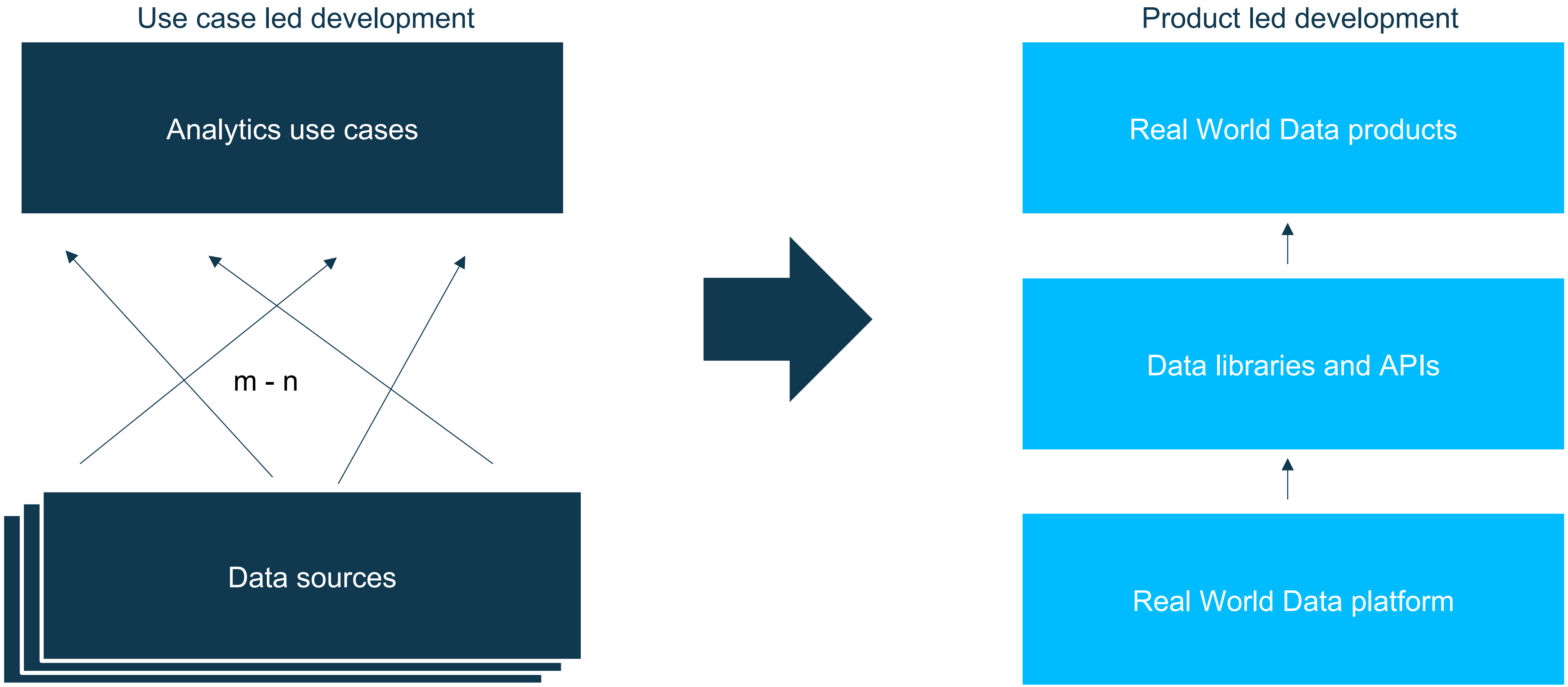
AI/ML training models

Image: flaticon.com

Where are Synthetic Data Used?

Modern software development practices in real-world data analytics

Benefit: improve reuse, accuracy and efficiency with modular and reusable components



Bespoke analyses creates use-bespoke data assets, rework and duplication

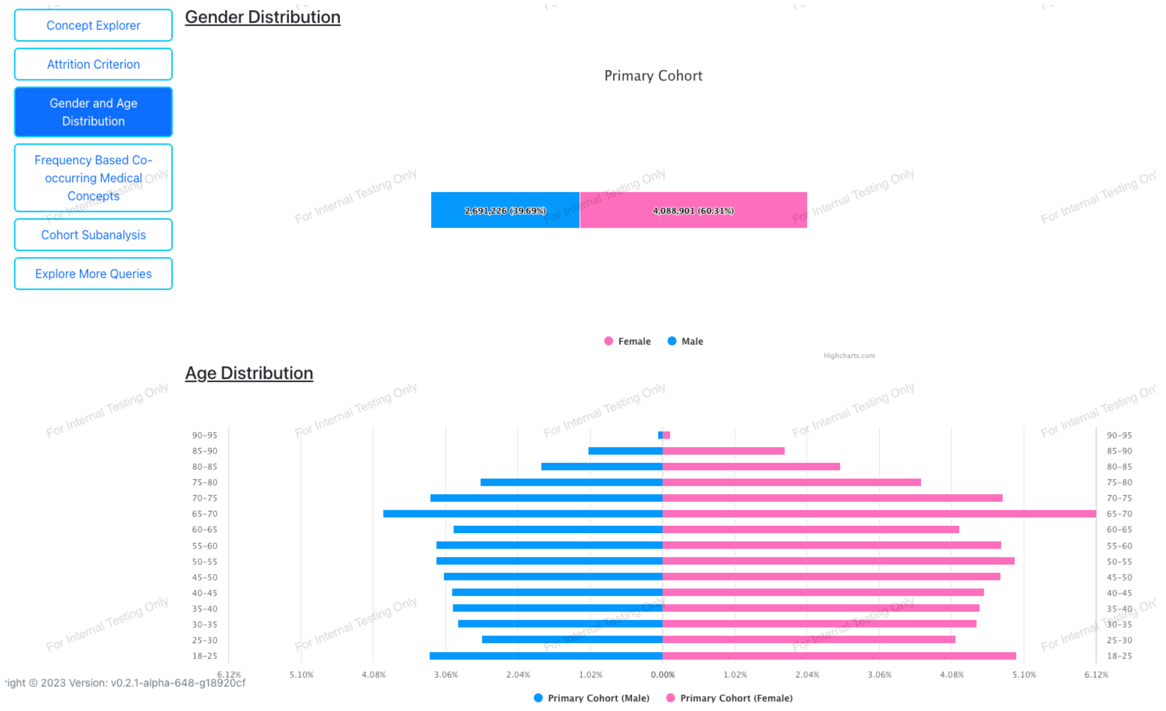
Execute standardized analysis libraries for an identified set of product features

Software development

Creation of data products and analytics for real-world evidence

1

- **Situation:**
  - Protect healthcare data during data analytics software development and testing
- **Complication:**
  - External developers need data access but are restricted from accessing licensed healthcare data
- **Resolution:**
  - Use synthetic data sets that mirror the complexity, missing values, schema of real healthcare data for development and testing

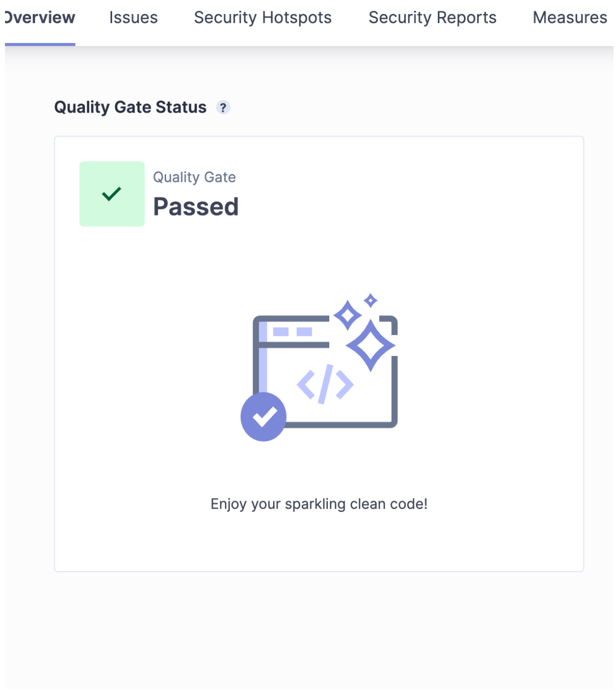


ight © 2023 Version: v0.2.1-alpha-648-g18920cf

Enhancing data analytics with unit testing

2

- **Situation:**
  - Ensuring software reliability is critical. Often testing methods rely on real-world data
- **Complication:**
  - Real-world data can be scarce, restricted, and may not adequately represent all scenarios, leading to biased or incomplete tests
- **Resolution:**
  - Unit tests with synthetic data allows comprehensive testing in controlled environments, overcoming privacy issues



Optimizing user journey with synthetic data and AI-generated questions

- **Situation:**
  - Users have unique needs when using analytical products, generating valuable data on preferences and needs
- **Complication:**
  - Data collection is limited and slow, often missing key user interactions (cold start problem)
- **Resolution:**
  - Employ AI to generate synthetic data simulating real user queries and preferences, enabling quicker and more effective user experience optimization

Image: flaticon.com

Synthetic data for algorithm development

4

An evolutionary algorithm for covariate balance between non-randomized populations

- **Situation:**
  - Developing and publishing innovative healthcare algorithms requires rigorous validation and value demonstration for scientific understanding
- **Complication:**
  - Real World Evidence data is sensitive, licensed, and restricted, impeding sharing and access
- **Resolution:**
  - Use synthetic or simulated data sets that mirror the complexity of real healthcare data helps us understand the algorithm with controllable inputs and understandable outputs

Parameter	Patient pool	Target population
height	uniform distribution min = 125, max = 195	normal distribution $\mu = 150, \sigma = 20$
weight	normal distribution $\mu = 90, \sigma = 20$ min = 50, max = 120	uniform distribution min = 50, max = 120
age	normal distribution $\mu = 65, \sigma = 20$ min = 18, max = 75	normal distribution $\mu = 50, \sigma = 20$ min = 18, max = 75
gender	0.6 m, 0.4 w	0.5 m, 0.5 w
country	0 A, 0.1 B, 0.2 C 0.3 D, 0.3 E, 0.1 F	0.1 A, 0.2 B, 0.2 C 0.1 D, 0.2 E, 0.2 F
hair color	0.4 fair, 0.3 medium, 0.3 black	0.2 fair, 0.4 medium, 0.4 black
(height, weight)	correlated $\sigma_{hw} = 0.5$	correlated $\sigma_{hw} = -0.8$
(age, gender)	correlated $\sigma_{ag} = -0.5$	correlated $\sigma_{ag} = 0.5$
binary 0 – 3	$p = [0.1, 0.3, 0.5, 0.8]$	$p = [0.3, 0.5, 0.3, 0.5]$

TABLE 2 Parameters governing the statistical distribution of the parameters in the simulated dataset.

Contact: Stephen Privitera

Privacy / Anonymization

Synthetic data from clinical trial data and real-world data

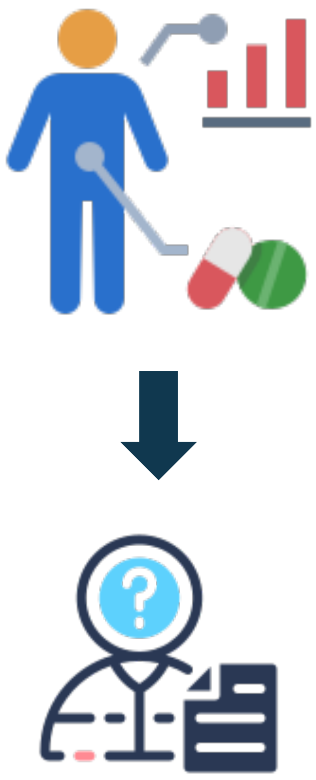
5

2 examples of data anonymization

1. Clinical Trial Data

- Multiple vendors performed proof of concepts to generate synthetic data from clinical trial data
- Outcome:
  - High quality synthetic data generation from small data point (~100s) turned out to be challenging

Contact: Christoph Gerlinger



2. Real-World data

- The Bayer “Future Clinical Trials” project aims to speed up drug development with advanced data anonymization
  - The Finnish use case

Mehtälä *et al.*
BMC Medical Research Methodology (2023) 23:258
https://doi.org/10.1186/s12874-023-02082-5

BMC Medical Research Methodology

RESEARCH — Open Access

Utilization of anonymization techniques to create an external control arm for clinical trial data

Juha Mehtälä[1*†], Mehreen Ali[2,3†], Timo Miettinen[2,3], Liisa Partanen[4], Kaisa Laapas[4], Petri T. Niemelä[4], Igor Khorlo[5], Sanna Ström[4], Samu Kurki[4], Jarno Vapalahti[4], Khaled Abdelgawwad[5] and Jussi V. Leinonen[4]

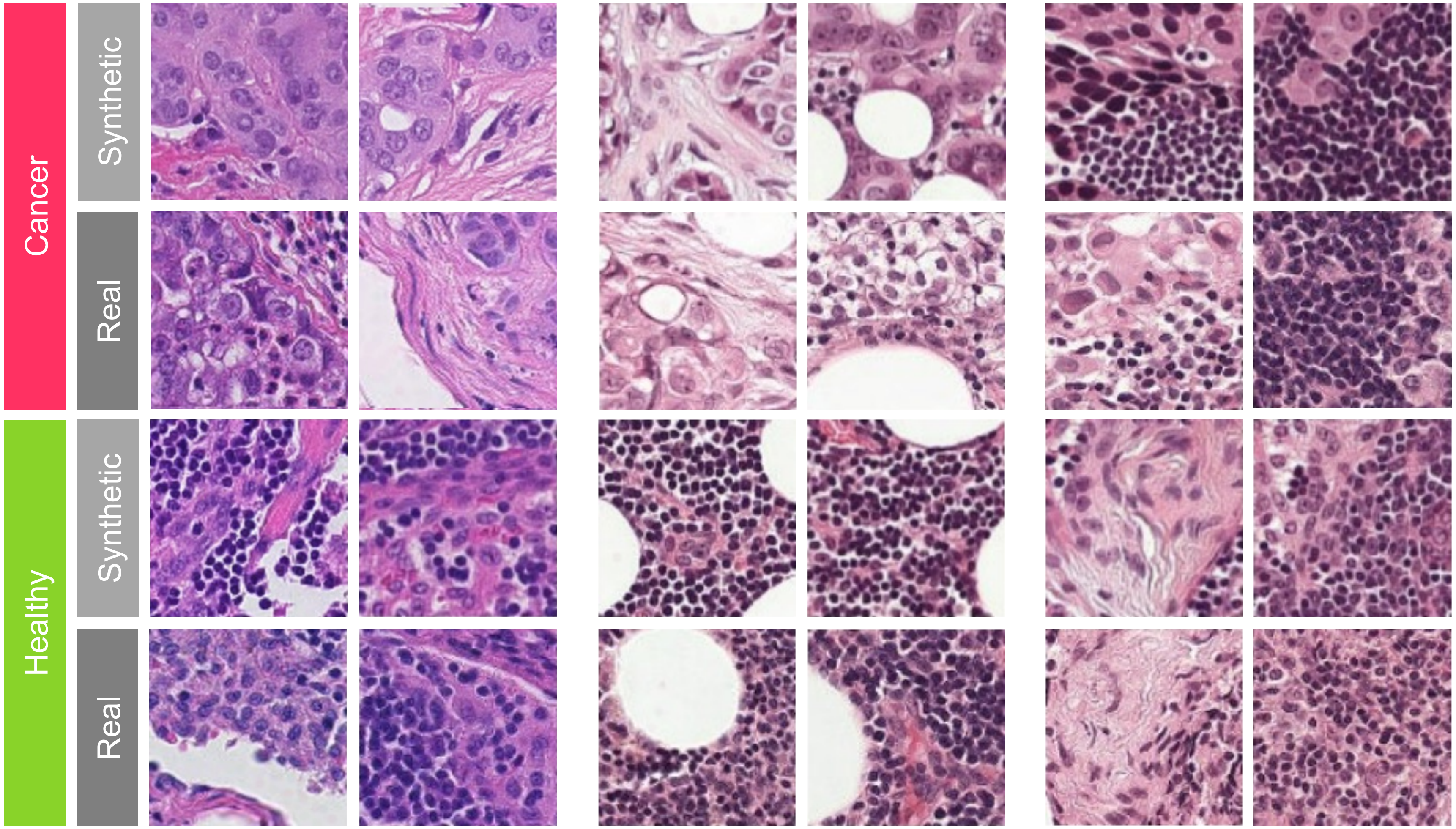
Source: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-023-02082-5

Contact: Jussi Leinnonen

Image: flaticon.com

AI generated synthetic images for histopathology

Synthetic generated data exhibits a quality that matches that of real data, a confirmation provided by in-house pathologists



Contact: Sadegh Mohammadi

Some learnings from the field

Synthetic data holds great promise…

Synthetic data generation impacts data sharing & AI development within the organization in various ways

- **Synthetic data is currently successfully used to develop and test software programs:**
  - It may be also useful for training purposes
- **Enable data sharing & AI models development:**
  - Enable quickly, safely and efficiently share data with external partners to accelerate scientific findings
  - Synthetic data generation will result in a lower barrier for data access
- **Generative AI explosion pushes enthusiasm and awareness:**
  - Faster and more efficient data creation, reducing the time and resources required for manual data input
- **A trending topic, very active area:**
  - Numerous publications, rapid methodological improvements
  - High quality open-source code to generate synthetic data

…but important challenges remain

- **Lack of shared definition:**
  - No common understanding across stakeholders about what synthetic data are
  - Synthetic vs simulated
- **Resistance to adoption:**
  - Real data “vs” synthetic data
- **“Small” data sets:**
  - More variables than data points (wide data)
  - Data for rare diseases
- **Data utility loss:**
  - If the signal in the real data is weak, it can be lost during the synthetization process

Thank you!