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

APPLICATIONS OF SYNTHETIC DATA TO ACCELERATE DATA ACCESS AND INSIGHT GENERATION

Presented by:

George Kafatos,
Director,
Data & Analytics Int'l team lead, Amgen

Applications of synthetic data to accelerate data access and insight generation

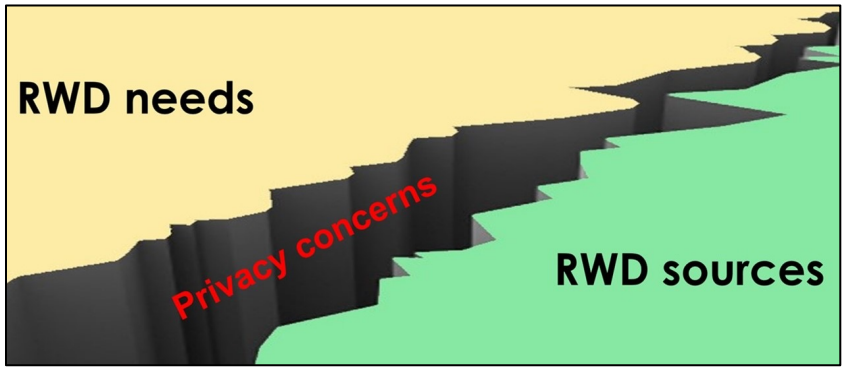
George Kafatos

Synthetic Data Summit 2023

AMGEN

Limited access to data remains one of the biggest hurdles in Real-World Data (RWD) use

- RWD are increasingly recognized as playing an important role in guiding drug development and understanding healthcare (HC) delivery
- This is in part due to the growing availability of RWD sources
- However, an important barrier, especially outside the USA, is the restricted access to patient-level data due to patient privacy concerns (e.g. GDPR rules implemented in the European Union in 2018[1]).



BRIDGING THE GAP BETWEEN DATA NEEDS AND DATA AVAILABILITY WHILST PROTECTING PATIENT PRIVACY IS FUNDAMENTAL IN HC RESEARCH

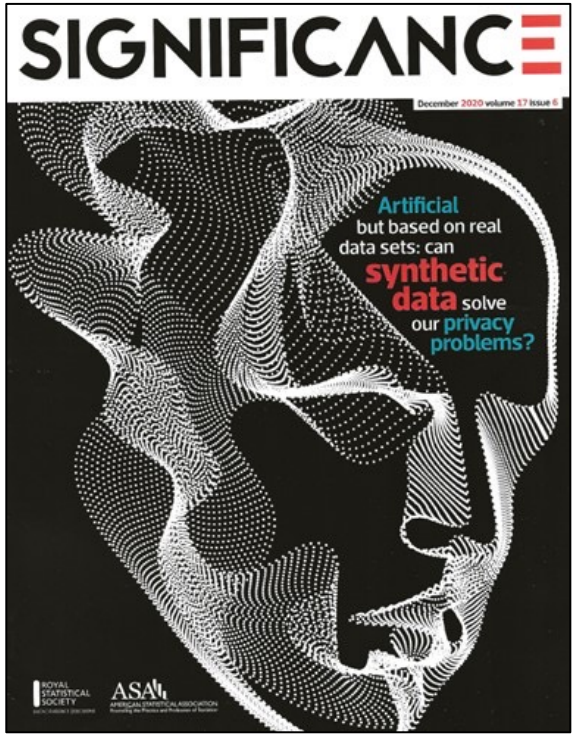
[1] https://www.clinicalleader.com/doc/how-to-take-full-advantage-of-rwd-without-jeopardizing-privacy-laws-0001

Synthetic data can be used to accelerate data access and insights generation

ONE SOLUTION THAT CAN BE USED TO ENABLE DATA INSIGHTS AND ACCELERATE ACCESS TO RESTRICTED PATIENT-LEVEL DATA IS SYNTHETIC DATA

Specifically, in terms of data access, synthetic data can be used to:

- Develop programming code
- Carry out feasibility analysis (e.g. hypotheses generation, sample size calculations, understanding missing data, evaluation of different methodologies)
- Produce teaching/training material
- Build complex models on synthetic data that can be subsequently run on real data[1].



"*Synthetic data are artificially generated data that are modelled on real data, … except they don't contain any real … information about individuals*"[2]

[1] Kokosi et al (2022) An overview of synthetic administrative data for research *Int J Popul Data Sci* **7**(1):1727
[2] Kaloskampis et al (2020) Synthetic Data in The Civil Service *Significance* **17**(6): 8–23

AMGEN

The choice of method for generating synthetic data should depend on their intended use

DIFFERENT DATA GENERATION METHODS (PREDICTION-BASED, SAMPLING-BASED, GENERATIVE MODELS)

When creating synthetic datasets for data access purposes different factors can be accounted for such as:

- Privacy risk
- Fidelity-level*
- Utility level
- Computational limitations (scale up within data sources)
- Standardizing synthetic data creation approach (scale up between data sources).

- **NOT ALWAYS BEST TO BE STRIVING FOR HIGHEST-FIDELITY SYNTHETIC DATA**
- **INSTEAD, THE TYPE OF SYNTHETIC DATA GENERATED SHOULD BE BASED ON THEIR INTENDED USE.**

* Fidelity is defined as the degree the synthetic data resemble the real data

AMGEN

There are synthetic data available for general use but there are limited use cases within the literature

Some examples include[3]:

- The NIH National COVID Cohort Collaborative
- The CMS Data Entrepreneur's Synthetic Public Use Files
- Various synthetic datasets available from UK CPRD
- A&E data from NHS England
- England Cancer Analysis System (CAS)
- A synthetic dataset from the Dutch cancer registry
- Synthetic variants of the French public health system claims data source (SNDS)
- South Korean Health data from Health Insurance Review and Assessment service (the national health insurer)[1].

NEED TO BETTER UNDERSTAND OF HOW SYNTHETIC DATA CAN BE USED IN PRACTICE IN RELATION TO DATA ACCESS

[1] Mosquera et al (2023) A methodology for generating synthetic longitudinal health data *BMC Med Res Methodol* **23**(1): 67

AMGEN

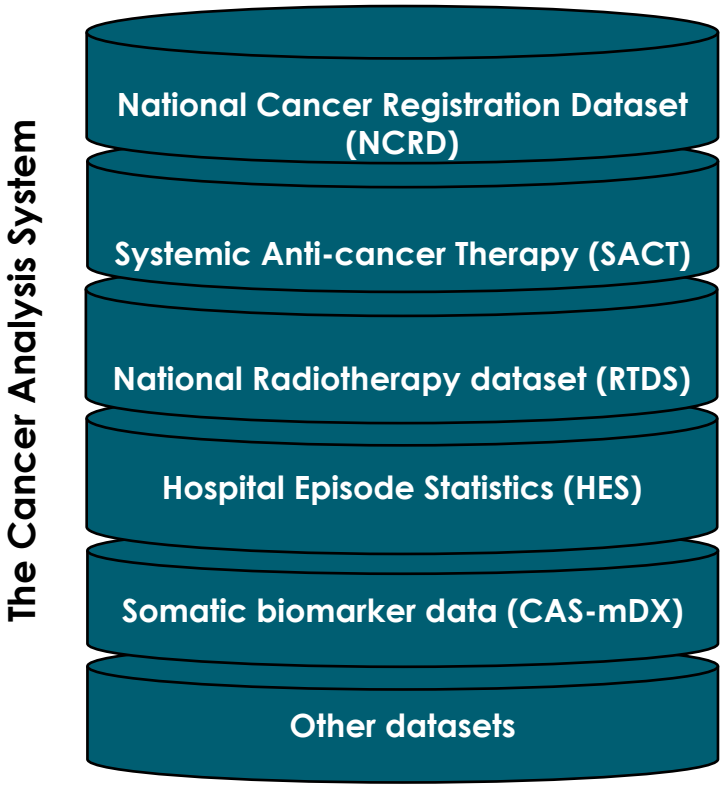
Leveraging synthetic data to facilitate data access: An example using the CAS data source in England

AMGEN

The Cancer Analysis System (CAS) data source

NDRS
NATIONAL DISEASE REGISTRATION SERVICE

THE CAS, COLLECTED BY NATIONAL DISEASE REGISTRATION SERVICE (NDRS), NHS ENGLAND COMPRISES OF SEVERAL LINKED DATA SOURCES FROM CANCER PATIENTS DIAGOSED AND TREATED IN ENGLAND AT POPULATION-LEVEL[1].



LARGE PATIENT POPULATION: ~5 MILION CANCER PATIENTS (10-YEAR PERIOD)

- Patient demographics
- Clinical characteristics
- ONS Death status / date
- Administration date /dosing / cycle
- Treatment intent
- Regimen modification
- Paediatric patients
- Hospitalisations
- Adverse events
- Somatic biomarkers in routine clinical practice.

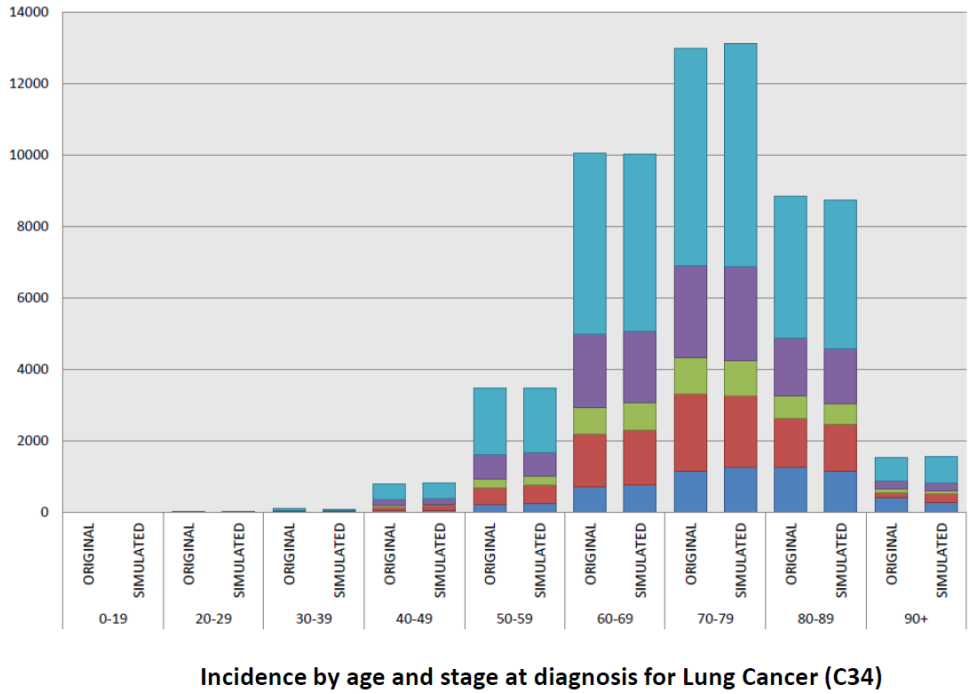
Note: The CAS data are available within the Genomics England environment

[1]Bright et al (2020) Data Resource Profile: The Systemic Anti-Cancer Therapy (SACT) dataset Int *J Epidemiol* **49**(1):15-15l

AMGEN

The Simulacrum is a synthetic dataset that resembles the CAS data

- **THE SIMULACRUM DATASET WAS DEVELOPED BY HEALTH DATA INSIGHT (HDI) IN COLLABORATION WITH IQVIA AND AZ[1]**
- **IT WAS INITIALLY RELEASED IN 2018 AND SIMULACRUM v2 BECAME AVAILABLE IN APRIL 2022**



Incidence by age and stage at diagnosis for Lung Cancer (C34)

Source: Vernon S & Chen C. The Simulacrum. NAACCR 2017

Simulacrum v2:

- High-fidelity dataset
- (Some) multivariate distributional properties reflected on synthetic data
- Bayesian network approach with privacy measures applied
- Patients diagnosed in years 2016-19
- Includes NCRD, SACT, RTDS and somatic biomarker datasets (but not HES)[2]
- Includes most data variables in these datasets

- **WORKS REASOBALY WELL FOR ONE AND TWO DIMENTIONAL COUNTS**
- **NOT GUARANTEED TO GIVE RIGHT ANSWERS FOR MULTIPLE VARIABLES / SUBGROUPS AND RARER TUMOUR TYPES.**

[1] https://simulacrum.healthdatainsight.org.uk/

[2] NCRD: National Cancer Registration Dataset; SACT: Systemic Anti-Cancer Therapy; RTDS: National Radiotherapy Dataset; HES: Hospital Episode Statistics

The Simulacrum synthetic dataset can be used to facilitate access to the CAS data

THE SIMULACRUM SYNTHETIC DATASET WAS DESIGNED TO GENERATE HYPOTHESES, FEASIBILITY ANALYSIS AND DEVELOP PROGRAMMING CODE

EXAMPLES OF USES OF SIMULACRUM DATA

- Estimating prevalence of patients with hematological and solid tumours (children/adults) and % of those on therapy by year
- Developing descriptive tables with demographic and clinical information by subgroups of patients/characteristics
- Support development of Line Of Treatment (LOT) algorithms.

THE OUTPUT OF THE SYNTHETIC DATA ANALYSIS CAN HELP INFORM

THE BASIS OF THE STUDY PROTOCOL AND STATISTICAL ANALYSIS PLAN (SAP) DOCUMENT.

In addition to Simulacrum use, the collaboration model established has been key for accessing CAS

AMGEN ROLE

- Leverage Simulacrum to form research questions
- Outline analyses required
- Define intended study design, study population (e.g. ICD10 codes), exposure and outcomes
- Study period, required datasets
- Patient eligibility criteria
- For certain (less complex) analyses, Amgen contributes by developing programming code using the Simulacrum data.

IQVIA ROLE

- Leverage the Simulacrum data to develop the SAP document
- Generate programming code using the Simulacrum data

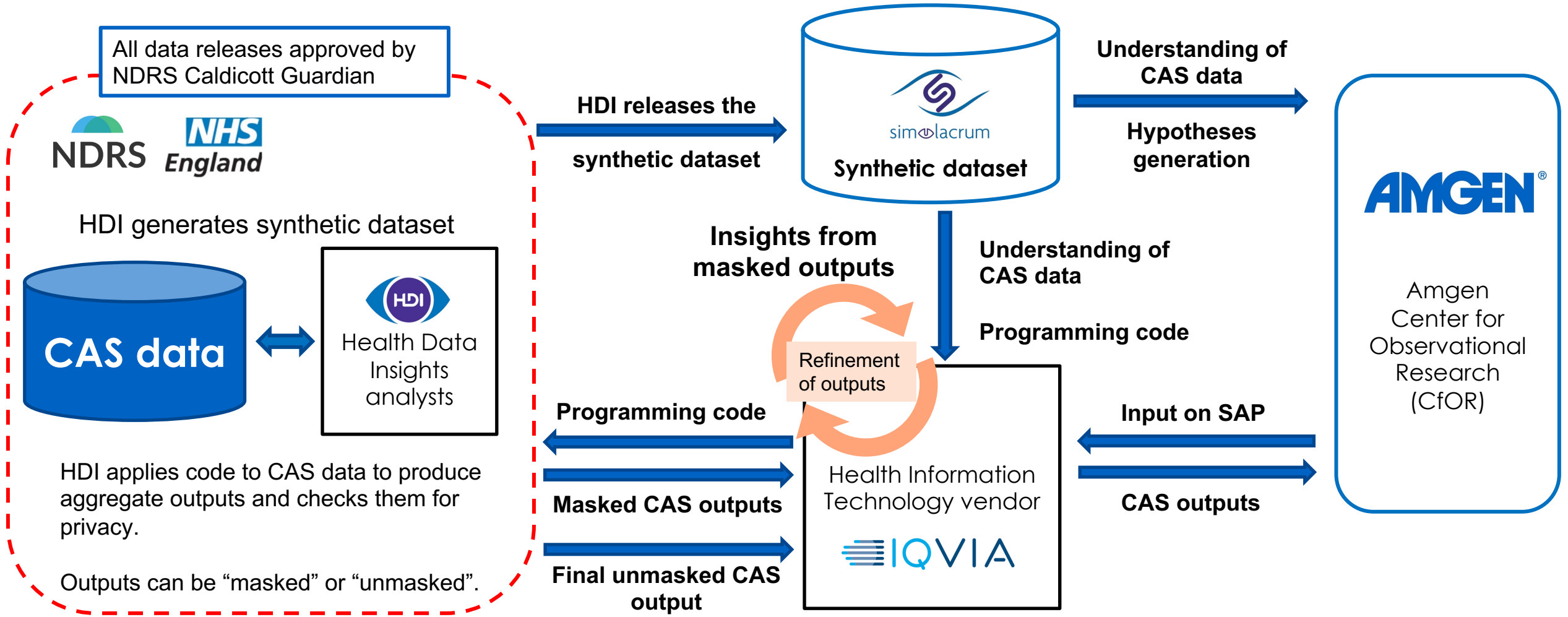
In particular:

- Loading, linking, cleaning and analysing data
- Implementation of eligibility criteria
- Derivation of study variables
- Feasibility analysis
- Production of table shell outputs
- Debugging of programming code in collaboration with HDI
- Review and delivery of "masked" results (i.e., rounded patient numbers)
- Refinement of analyses based on "masked" results
- Review and delivery of "unmasked" results (i.e., exact patient numbers)
- Review of Amgen's programming code.

HDI ROLE

- Executes programming code
- Discuss programming issues with IQVIA and adjusts code accordingly
- Produces the aggregated results and apply privacy checks
- Obtains approval for release of results from the NDRS Caldicott guardian
- Following approval, delivers the aggregated results.

Collaboration workflow leveraging the Simulacrum synthetic data to gain insights and enable quick analyses to the CAS data



RECENT EXAMPLE OF AMGEN STUDY WITH ~6 MONTH TURN-AROUND TIME FROM PROTOCOL APPROVAL TO DELIVERY OF RESULTS

Features and benefits of the CAS collaboration

FEATURES:

- Ability to set up a flexible contract that allows multiple analyses
- Good working relationship and regular engagements between the different parties i.e. data owners and users
- Focus on producing evidence that will have public health benefits
- Understanding and commitment by all parties of the importance to protect patient privacy
- Development of process that can streamline and accelerate analyses (e.g. data guides, processes, repository of algorithms, refinement of programming code and final analysis outputs).

BENEFITS:

- Speed of analyses delivery
- Optimal use of resources / cost efficiencies
- Transparency of analysis steps (visibility of programming code / algorithms).

Use of synthetic data as part of regulatory filing

AMGEN

Health Insurance Research Database (HIRD) in Taiwan

National Health Insurance Research Database
NHIRD

POPULATION-LEVEL DATA SOURCE (~23M)

Information includes:

- Registry for beneficiaries
- Ambulatory care claims
- Impatient claims
- Prescriptions dispenses at pharmacies
- Registry for medical facilities
- Registry for board-certified specialists

Additional information by linking with other registries: deaths, lab measurements, cancer stage, socio-economic factors.

RESTRICTED AMOUNT OF DATA PROVIDED TO RESEARCHERS (≤10% OF TAIWAN POPULATION)[1]

[1] Hsieh et al (2019) Taiwan's National Health Insurance Research Database: past and future *Clin Epidemiol* **11**: 349-358

AMGEN

Synthetic dataset was submitted to China Center of Drug Evaluation (CDE) as supporting material

Source: https://www.encepp.eu/encepp/viewResource.htm?id=47635

- Regulatory submission in China for Denosumab for male osteoporosis indication*
- Requirement for information from Chinese patients
- Study to assess the safety and clinical effectiveness of denosumab among Chinese men with osteoporosis**
- Proposal to use of Taiwan HIRD database
- CDE guidelines require accessibility of data so they can evaluate if needed

- **A LOW FIDELITY TRAINING DATASET WAS CREATED BASED UPON THE HIRD DATA SOURCE (SAME DATA STRUCTURE; ~5,000 PATIENTS)**
- **PROVIDED CDE WITH ALL THE STUDY MATERIAL NEEDED TO FACILITATE VDE EVALUATION OF EXTERNAL DATABASE.**

* Submitted April 2022; Approved in February 2023

** Study report was included within the filing package

Discussion

Discussion points

- Creation of synthetic dataset should be based on data governance specifications provided by the data owners
- There are limited examples of how synthetic data can be used in practice to accelerate data access
- Potential data owners' concerns in releasing synthetic data may be due to lack of understanding of a successful implementation of a business model using synthetic data.

COMMUNICATION OF USE CASES SUCH AS THE CAS COLLABORATION COULD BE KEY FOR THE WIDER ADOPTION OF SYNTHETIC DATA FOR DATA ACCESS PURPOSES

AMGEN

Acknowledgments

AMGEN: OLIA ARCHANGELIDI, ZHENNA HUANG
IQVIA: JULIA LEVY, POOJA HINDOCHA
HDI: LORA FRAYLING

Disclosure

GEORGE KAFATOS IS AN EMPLOYEE OF AMGEN LTD AND OWNS AMGEN INC SHARES