

# PHUSE SDE – Utrecht

Standards and Databases



## **Standardising Biopharma Data through Conformed Data Models**

James McDermott  
Marten Krijgsman  
Wilhelmina Lucinescu



# Agenda



- Introduction
- Case study: A tale of two databases
- What is a Conformed Data Model?
- Future-state Clinical Data Platform (CDP) architecture

# Agenda



[ 5 mins ]

Overview



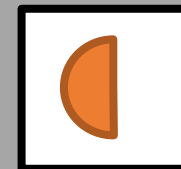
[ 10 mins ]

Case Study



[ 10 mins ]

Conformed  
Data Models



[ 5 mins ]

Discussion

# Today's goals



- Understanding how **public sector** data models currently function
- Understanding the **underlying principles** that complicate unified data platforms
- Outlining the key principles of a **conformed data model**
- Starting the conversation on designing a **future-state conformed data platform**

# A Tale of Two Databases



## Brief introduction

MRes student at University of Oxford, studying genetics of fibromyalgia and endometriosis

MSci in Biomedical Research from Newcastle University

My research interests:

- Genetics of common diseases
- Chronic pain
- Computational biology

Conflicts of interests: none

# A Tale of Two Databases



## Topic introduction

- **Advantages** and **challenges** with large, unified databases
- Bringing together different clinical codes
- Data cleaning
- Making sense of historical data
- Future directions

# A Tale of Two Databases



## Worldwide initiatives

### UK Biobank

deCODE Genetics (Iceland)

Estonian Biobank

Biobank Japan

China Kadoori Biobank

### FinnGen (Finland)

Lifelines (Holland)

Million Veteran Program (USA)

All of US Research Program (US)

# A Tale of Two Databases



## UK BIOBANK at a glance

- ~500,000 participants recruited between 2006 and 2010
- in depth genetic and health information
- between 40 and 69 years old
- Health information and clinical diagnoses brought together by a combination of health records and self reports

# A Tale of Two Databases



## What health information is available?

- COVID 19 Registry
- Primary care (read codes - now obsolete)
- Hospital inpatient (ICD9/ICD10 codes)
- Cancer register

# A Tale of Two Databases



## Primary care data

| Condition     | ICD10     | ICD9      | Read_ctv3                                                                                                                                                                                                                                                                                     | Read_v2                                                                                                                                                                                                                 | Self |
|---------------|-----------|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| Endometriosis | N800:N809 | 6170:6179 | X101N, X408N, X408O, X408R, X408T, X408V, XE0eW, XE0eX, XaEWA, K5000, K5001, K5002, K500z, K501., K502., K503., K5030, K5031, K5032, K5033, K503z, K504., K5040, K5041, K504z, K505., K5050, K5051, K5052, K505z, K506., K50y., K50y0, K50y1, K50y2, K50y3, K50yz, K50z., Kyu9., Kyu90, K500. | K5000, K5001, K5002, K500z, K501., K502., K503., K5030, K5031, K5032, K5033, K503z, K504., K5040, K5041, K504z, K505., K5050, K5051, K5052, K505z, K506., K50y., K50y0, K50y1, K50y2, K50y3, K50yz, K50z., Kyu9., Kyu90 | 1402 |

41 read ctv3 codes!

# A Tale of Two Databases



## Primary care data

Different data suppliers

Combination of different codes and free text (manually analysed)

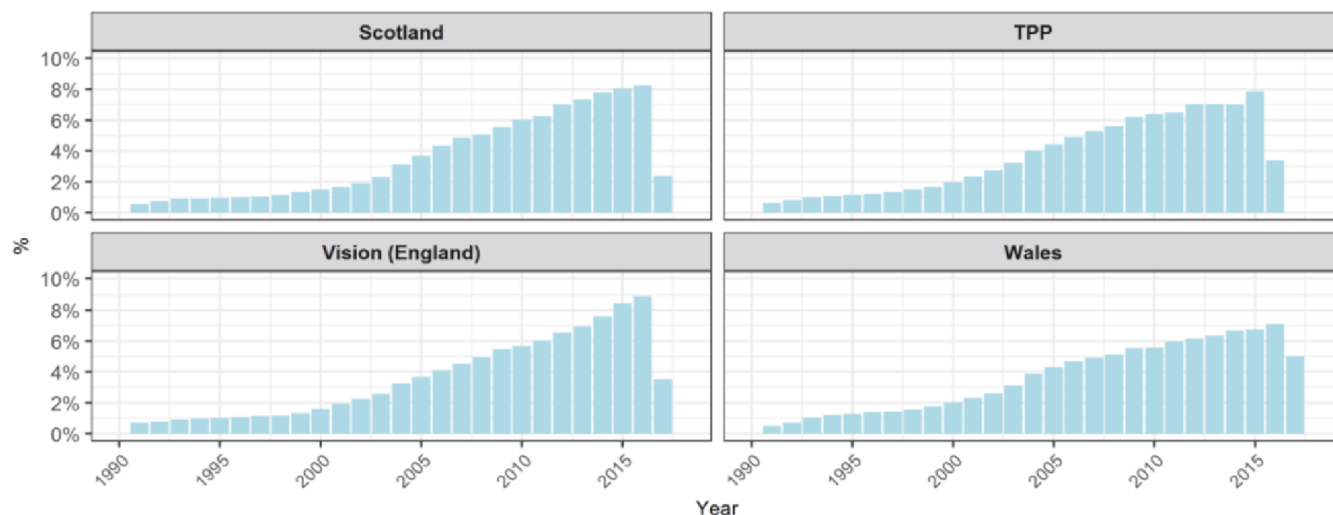


Figure 1. Proportion of participants with clinical data by year and source.<sup>16</sup>

# A Tale of Two Databases



## Advantages

- Research trends are about bringing together data: as **much** as possible, from as **many** people as possible
- This has enabled important advances (large **GWAS**)
- Easier collaboration

## Pitfalls

- Different data suppliers
- Different coding systems
- Evolving coding systems (**ICD10** vs **ICD11**)

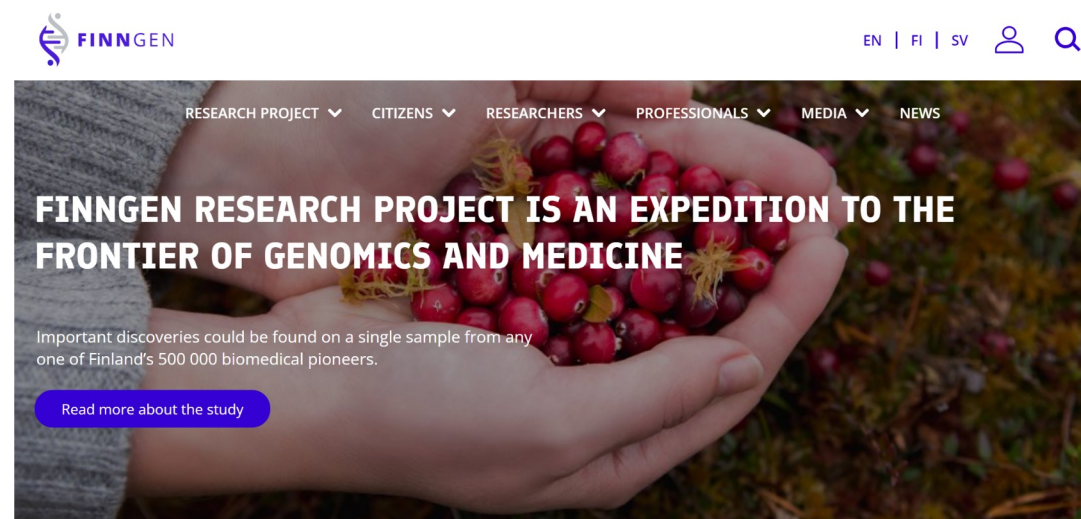
# A Tale of Two Databases



## FinnGen

### Private-public collaboration

- National institutes, hospitals, GPs, health service providers, insurance, pharmaceutical companies, biobanks, other registries
- Diagnoses coded in **ICD10**
- Longitudinal data on 500,000 participants



# A Tale of Two Databases



## Future directions

### Private-public collaboration

- Collaboration improves **study power**
- Private-public collaborations are ideal
- Biobanks provide a **unified platform** for research
- Clinical terms will evolve - **we have to keep up!**



Menu ≡

[NHS Digital](#) > [Services](#) > [Terminology and Classifications](#) > [SNOMED CT](#)

## SNOMED CT

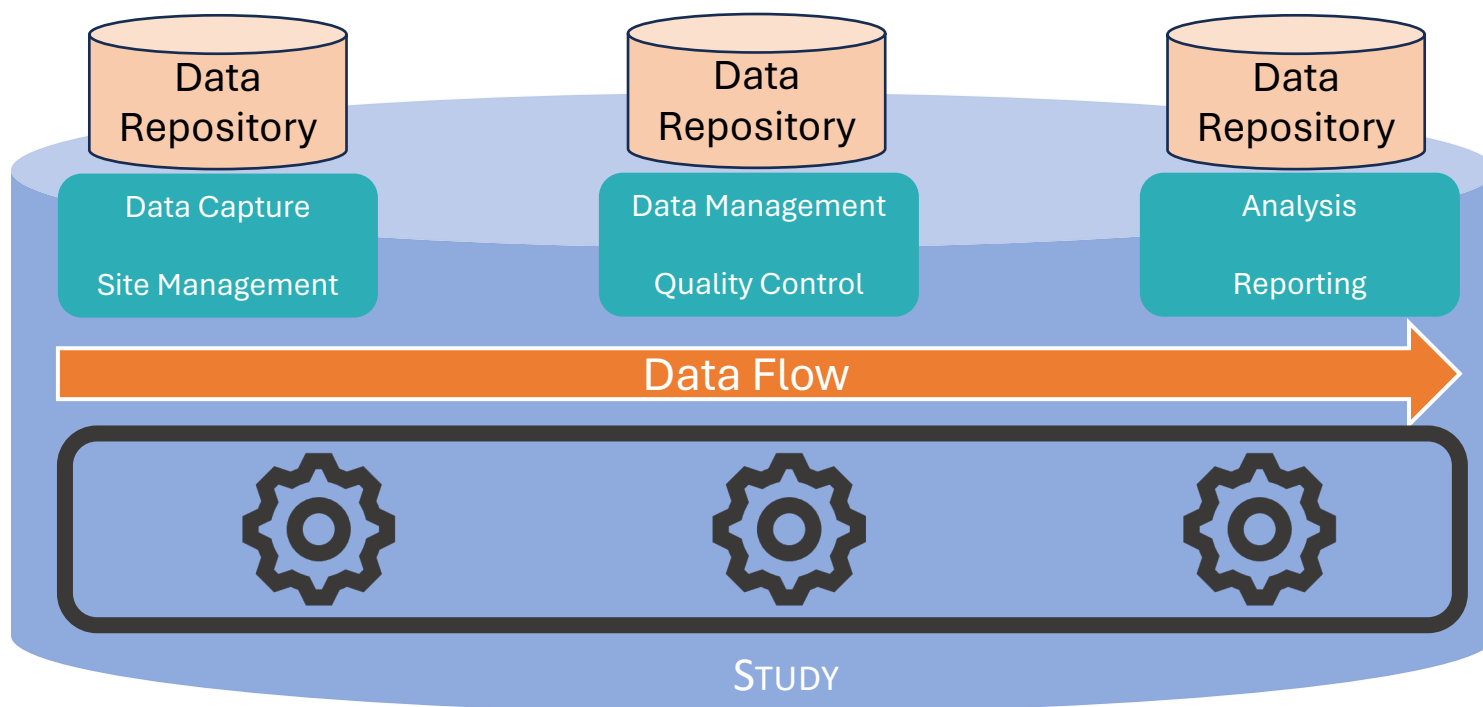
SNOMED CT is a structured clinical vocabulary for use in an electronic health record. It is the most comprehensive and precise clinical health terminology product in the world.



# Towards Conformed Data Models



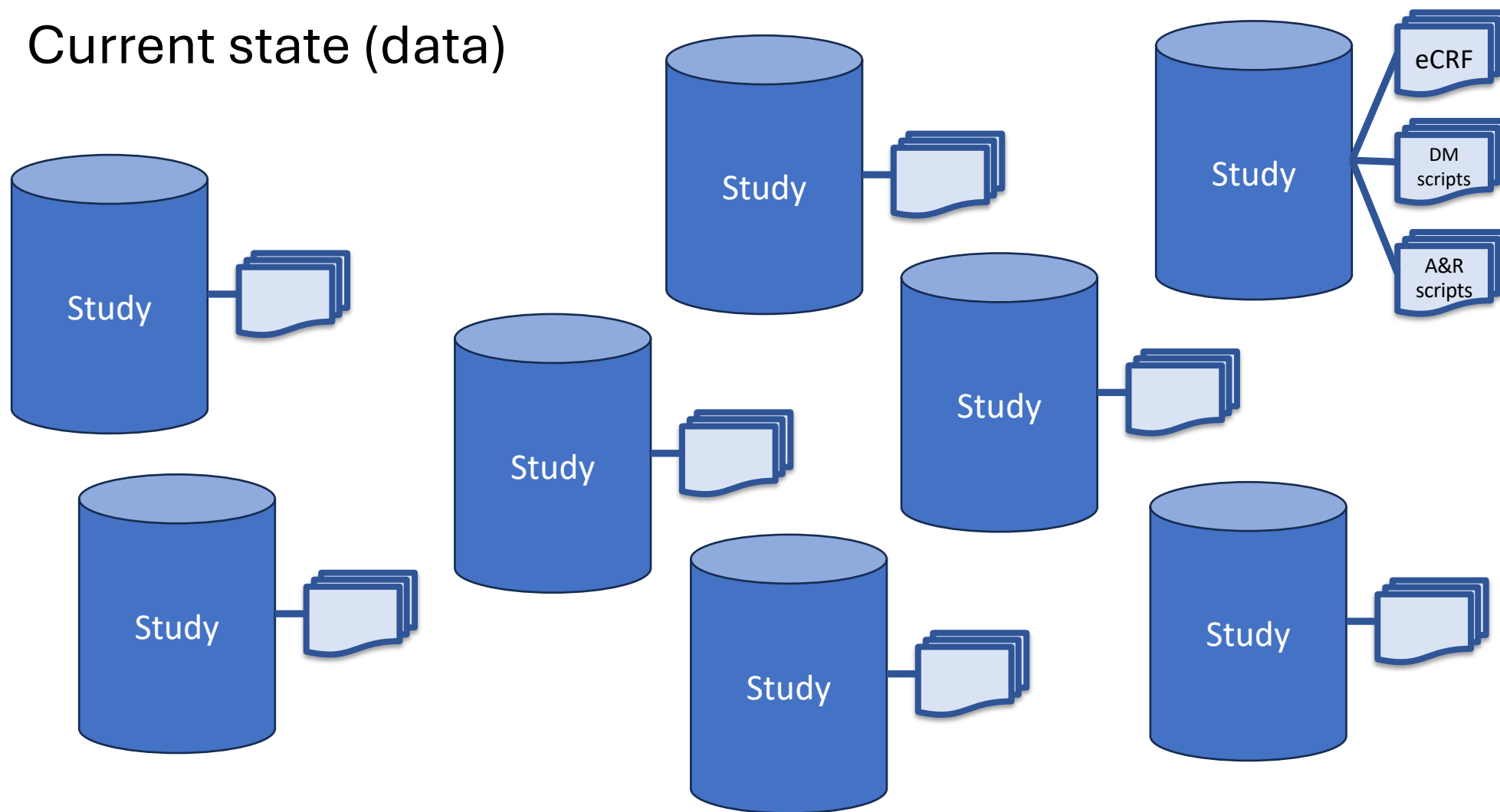
Current state (process)



# Towards Conformed Data Models



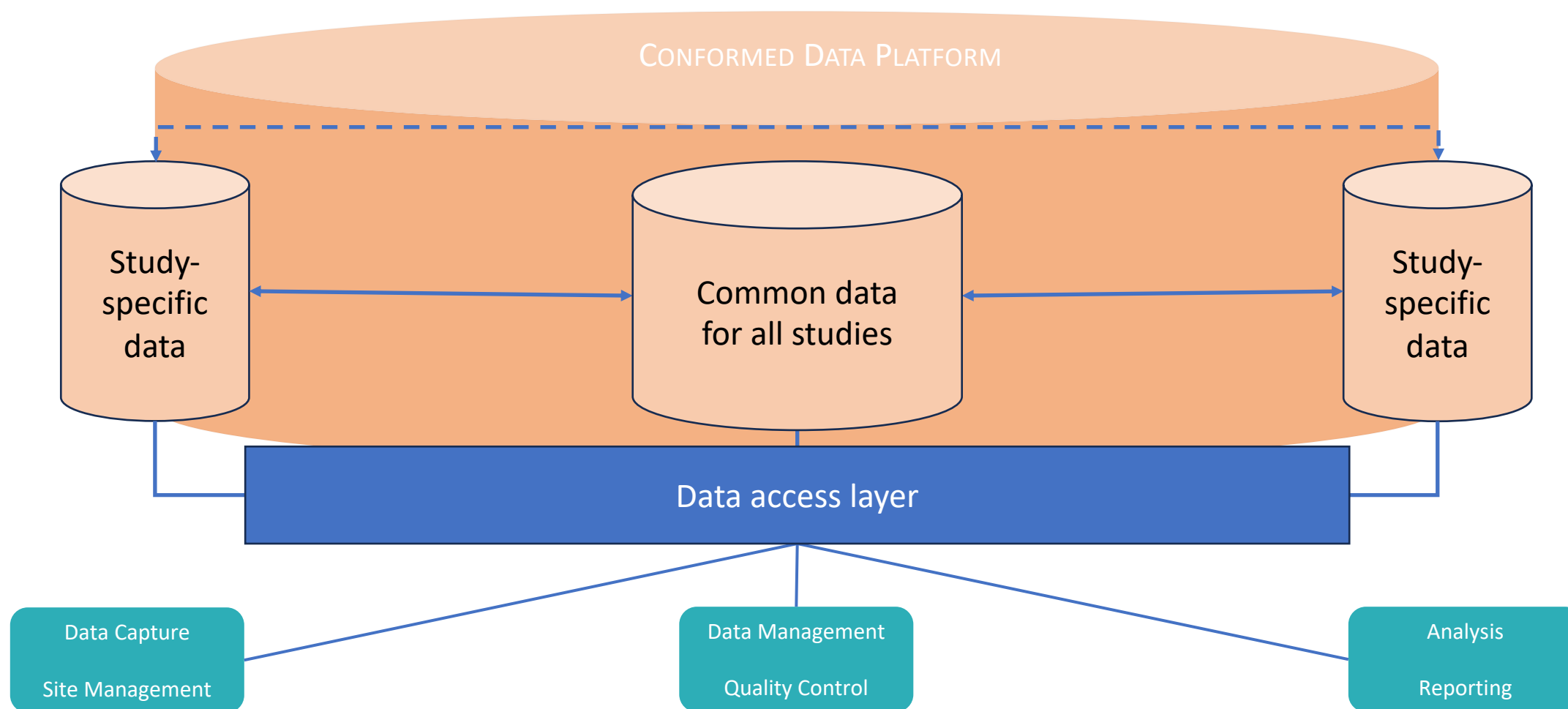
Current state (data)



# Towards Conformed Data Models



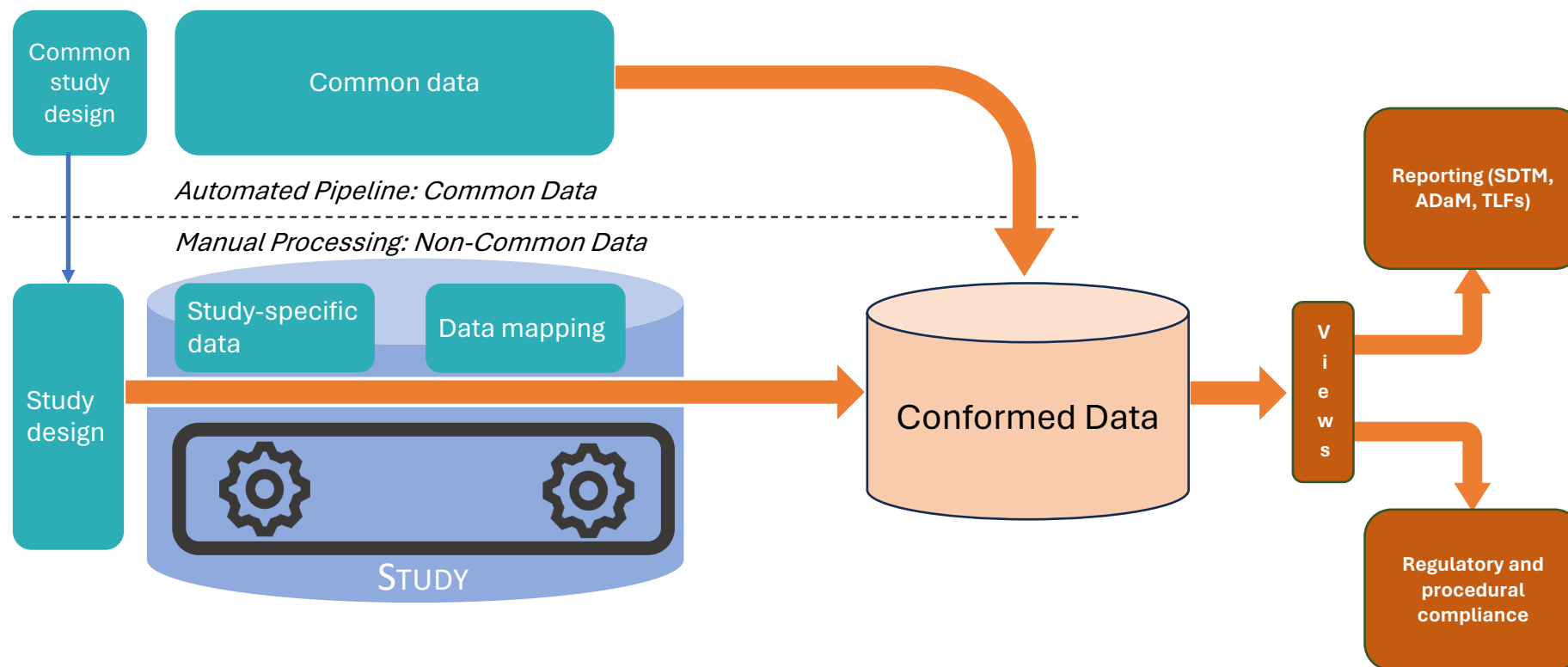
Future state (data)



# Towards Conformed Data Models



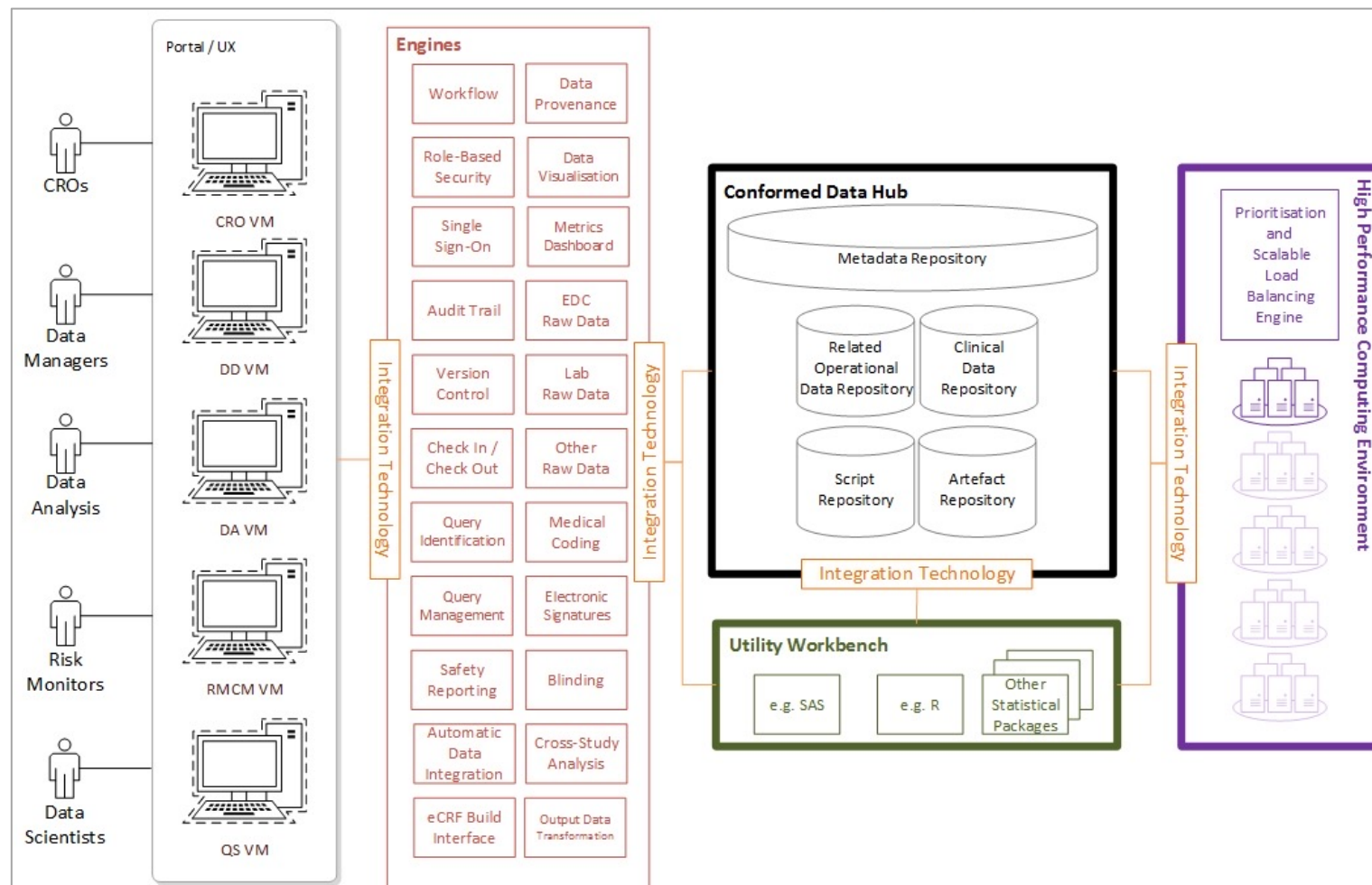
Future state (process)



# Towards Conformed Data Models



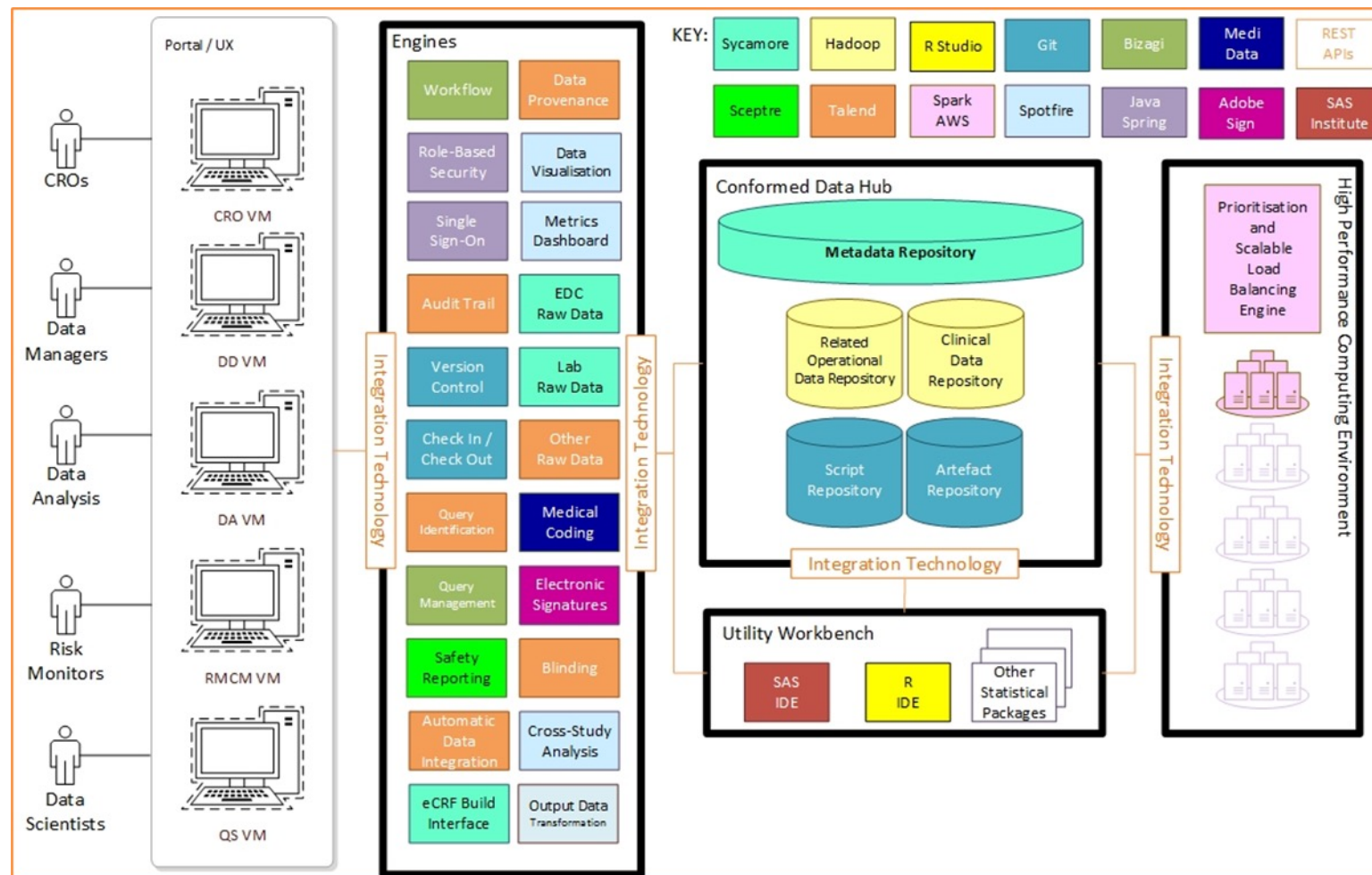
## Future-state architecture (generic)



# Towards Conformed Data Models



## Future-state architecture (applied example)





# Towards Conformed Data Models

## Future directions

- A unified platform for each company
  - Conformed Data Model
  - Data Governance of the CDM
  - A platform with API interfaces for the required engines
  - Cultural change within companies

# Today's goals



- What can we learn from **public sector** data models?
- What gains can be made in **industry** data models?
- How can we set up **conformed data models**?
- How do we take the next step towards a **conformed clinical data platform**?



**Thank you for joining us today!**

**Contact details for further information and questions:**

**James McDermott**

**Marten Krijgsman**

**Wilhelmina Lucinescu**

