



MRC-NIHR Trials Methodology Research Partnership: Webinar recording

## **Data Integrity**

***Presented by Macey Murray (UCL)***

19 October 2022

On behalf of Health Data Research UK



The slides are also available below.

For any queries, please contact [uktmn@nottingham.ac.uk](mailto:uktmn@nottingham.ac.uk)

<https://www.youtube.com/watch?v=m-2z1QkcT7>



MRC  
Clinical  
Trials Unit



# Assessing data provenance and integrity of NHS Digital datasets for clinical trials

Dr Macey Murray,

Senior Research Fellow in Trial Conduct Methodology;

Clinical Trials secondee, NHS DigiTrials

19 Oct 2022 | TMRP webinar series 5

Smarter Studies  
Global Impact  
Better Health

# Acknowledgements

## Healthcare Systems Data for Clinical Trials Collaborative Group

**MRCCTU at UCL** Matthew Sydes, Sharon Love, James Carpenter, Mahesh Parmar.

**University of Oxford** Marion Mafham, Martin Landray.

**NHS Digital** Heather Pinches (DigiTrials), Suzanne Hartley (DARS, formerly Leeds CTRU).

## Thanks also to

**NHS Digital** Michael Chapman, Laura Sato and Jaspal Panesar (Metadata).

**MHRA** who provided constructive feedback during development.

# Overview

- Background
- Purpose of the position paper\*
- Assessment process
- Data integrity of the two selected datasets
- Conclusions and Recommendations
- Follow-on project: Demonstrating Data Integrity of routine health data in Clinical Trials (DEDICaTe)

# NHS Digital



- Statutory role
  - Collect, analyse, and publish health data
  - Provide technical infrastructure to support health and social care
- NHSD and NHSX to merge with NHS England and Improvement
- Hosts >200 data assets, 51 currently available with more in the near future
  - via Data Access Request Service (DARS) or DigiTrials (clinical trials service).
- Most widely used in trials are HES APC and CRD (Lensen *et al.* 2020)
  - HES APC: Admissions data from NHS Trusts in England
  - CRD: Death registration data from the Office for National Statistics (ONS) for England and Wales

# Advantages of healthcare data...

...being recognised as high-quality data suitable for use in trials:

- ✓ Sponsors: Healthcare data as trial data, simpler data collection, more complete, more efficient, less costly, especially for large multicentre trials.
- ✓ Investigators: Reduced burden on NHS site staff, can focus on patient care and collection of data such as patient-reported outcomes.
- ✓ Data providers: Documentation of provenance demonstrating integrity of their datasets enables their use in clinical trials.
- ✓ Public: Trials run more efficiently through use of centralised national datasets, supports innovation, research and development of better treatments, in a timely manner.

# MHRA real-world data guideline series

- Published in Dec 2021
- Recognises value of real-world data (RWD)
- Encourages sponsors to use RWD in trials that support regulatory decisions.
- Says that RWD must be demonstrated to be “of sufficient quality”



## Contents

1. Scope
2. Introduction to MHRA RWD guidelines
3. Data Quality
4. Advice

 [Print this page](#)

## 1. Scope

This document provides an introduction to the MHRA’s real-world data (RWD) guideline series, and points to consider when evaluating whether a RWD source is of sufficient quality for the intended use.

Sponsors interested in the use of RWD in their development programmes are encouraged to [engage with the MHRA](#) for further advice on specific proposals.

## 2. Introduction to MHRA RWD guidelines

There are vast amounts of data being collected on patients, for example, in electronic health records (EHR), and disease and patient registries. Such data are commonly

# MHRA RWD guidance: Data quality

Sponsors should include in the study protocol a **description of the tools and methods for selection, extraction, transfer, and handling** of data and how they have been or will be validated. It is essential that **processes are established to ensure the integrity of the data** from acquisition through to archiving and sufficient detail captured to allow for the verification of these activities. It is expected that the validity of the RWD that are intended to be used in the study is formally documented and approved by the sponsor before the study protocol is published or submitted to the MHRA.



# Revision of ICH GCP (E6)



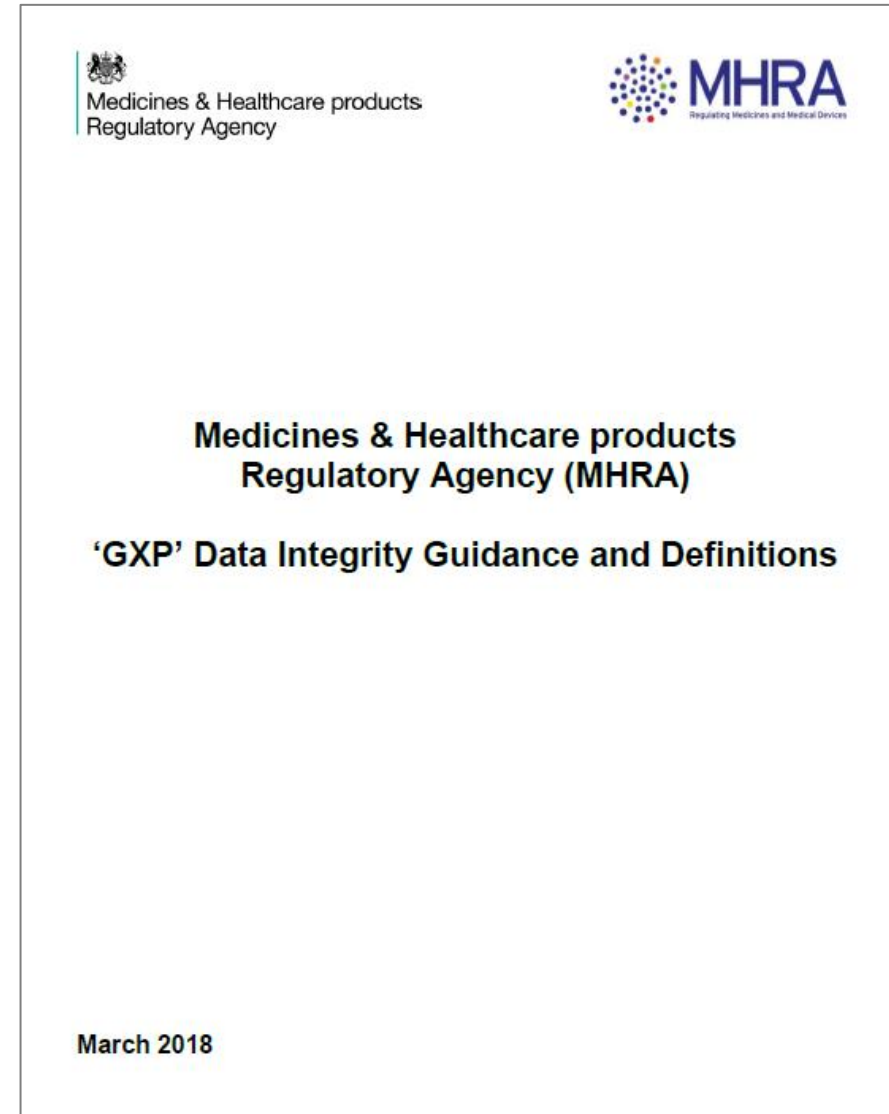
- ICH E6 R3 draft principles published in April 2021
- Flexibility to allow for innovation, but key principles endure: protection of trial participants and reliability of trial results
- intends to be “media neutral to enable the use of different technologies for the purposes of documentation”
- acknowledges “the use of a variety of relevant data sources in clinical trials”, including patient/disease registries and electronic medical records
- expects sponsors to demonstrate data integrity and reliability of trial data
- More detail on secondary use of healthcare data expected in forthcoming Annexes.

# Regulatory definitions

- Trial sponsors need to demonstrate to regulatory authorities that all data, including healthcare systems data, are integral, reliable, and complete.
  - Provenance = detailed record of the origins of the data, the processes, and the methods by which it is produced.
  - Integrity = extent to which all data are complete, consistent, accurate, and reliable throughout the data lifecycle
- Sponsors review source data for quality control and reliability of trial data
- Source data = Original or certified copy, necessary for evaluation and reconstruction
  - ALCOA: Accurate, Legible, Contemporaneous, Original, Attributable, Complete (ICH E6 R2 section 4.9.0)
  - ALCOA+ (used by CDISC eSource standard): includes consistent, enduring and available
- ALCOA and ALCOA+ have limited use for centrally curated healthcare data like HES APC (Appendix 1 of Position paper)

# MHRA GXP Data Integrity guidance

- Risk-based approach to data management
- Covers data integrity risk, criticality and data lifecycle.
- Documented system providing acceptable state of control based on data integrity risk
- Use of ALCOA+
- Also, emphasis on data governance measures to ensure data are complete, consistent, enduring and available through the lifecycle.



February 11, 2022

Report Open Access

# Use of NHS Digital datasets as trial data in the UK: a position paper

Murray, Macey Lee; Pinches, Heather; Mafham, Marion; Hartley, Suzanne; Carpenter, James R; Landray, Martin; Love, Sharon B; Parmar, Mahesh KB; Sydes, Matthew R

**Background:** Clinical trial teams increasingly use digital data, particularly to enhance recruitment and retention, and to reduce effort. However, there is continued concern about the provenance, quality and reliability of such data.

**Objectives:** There were two key objectives: (a) to assess whether NHS Digital (NHS) datasets can be similarly evaluated to clinical trial data.

**Method:** The data lifecycles of the datasets were compared to the healthcare provider's databases to assess the evidence of whether the datasets are reliable transcribed copies of source data.

**Result:** The assessment method was applied to (a) the Civil Registration dataset and (b) the Civil Registration dataset held by the originating provider.

**Conclusion:** Based on these findings, the datasets are reliable transcribed copies of source data.

**Implications:** First, these datasets can be used to identify algorithms and processes to identify data. Second, an assessment approach should be used to assess the reliability of transcribed copies of source data.

On behalf of the Healthcare System (HCS), JRC, MKBP, and MRS are funded by the Department of Health and Social Care.

Preview

903

views

635

downloads

See more details...

All versions This version

COMMENT | VOLUME 4, ISSUE 8, E567-E568, AUGUST 01, 2022

## Data provenance and integrity of health-care systems data for clinical trials

Macey L Murray  Sharon B Love James R Carpenter Suzanne Hartley Martin J Landray Marion Mafham et al. [Show all authors](#)

Open Access Published: August, 2022 DOI: [https://doi.org/10.1016/S2589-7500\(22\)00122-4](https://doi.org/10.1016/S2589-7500(22)00122-4)

 Check for updates

The need to run clinical trials quickly and efficiently is well recognised, none more so than during the pandemic. The need for rapid access to treatments and preventative measures for COVID-19. Late phase clinical trials can take many years and immense efforts necessary to deliver them. There are various ways in which the conduct of clinical trials can be improved through judicious use of data already collected in health-care interactions. These data might be known as routinely collected health-care data (RCHD), or real-world data. We describe here how one key roadblock to clinical trials can be removed.

Murray ML *et al.* 2022.  
Zenodo.org.  
<https://doi.org/10.5281/zenodo.6047155>

Murray *et al.* 2022.  
Lancet Digital Health  
[https://doi.org/10.1016/S2589-7500\(22\)00122-4](https://doi.org/10.1016/S2589-7500(22)00122-4)

# Purpose

To demonstrate:

1. The integrity of two sets of healthcare systems data, held by NHS Digital, so they are suitable for trial use.
  - Hospital Episode Statistics Admitted Patient Care (HES APC)
  - Civil Registration of Deaths (CRD)
2. Our approach to ascertain data integrity can be similarly applied to other healthcare data.

# Assessment process

- Based on ICH E6 R3 and GXP Data Integrity guidance to show the integrity and reliability of trial data, including healthcare data
- Requires documentation of tools, systems, controls, lineage, access.
- As no established procedure for assessing healthcare datasets, three key stages of data lifecycle assessed:
  - 1) **collection** of data from healthcare systems: *submission*
  - 2) centralised **processing and curation** to form the validated data: *production*
  - 3) and **linkage and extraction** for the end user: *releaseable*.

# Key stages of data lifecycle

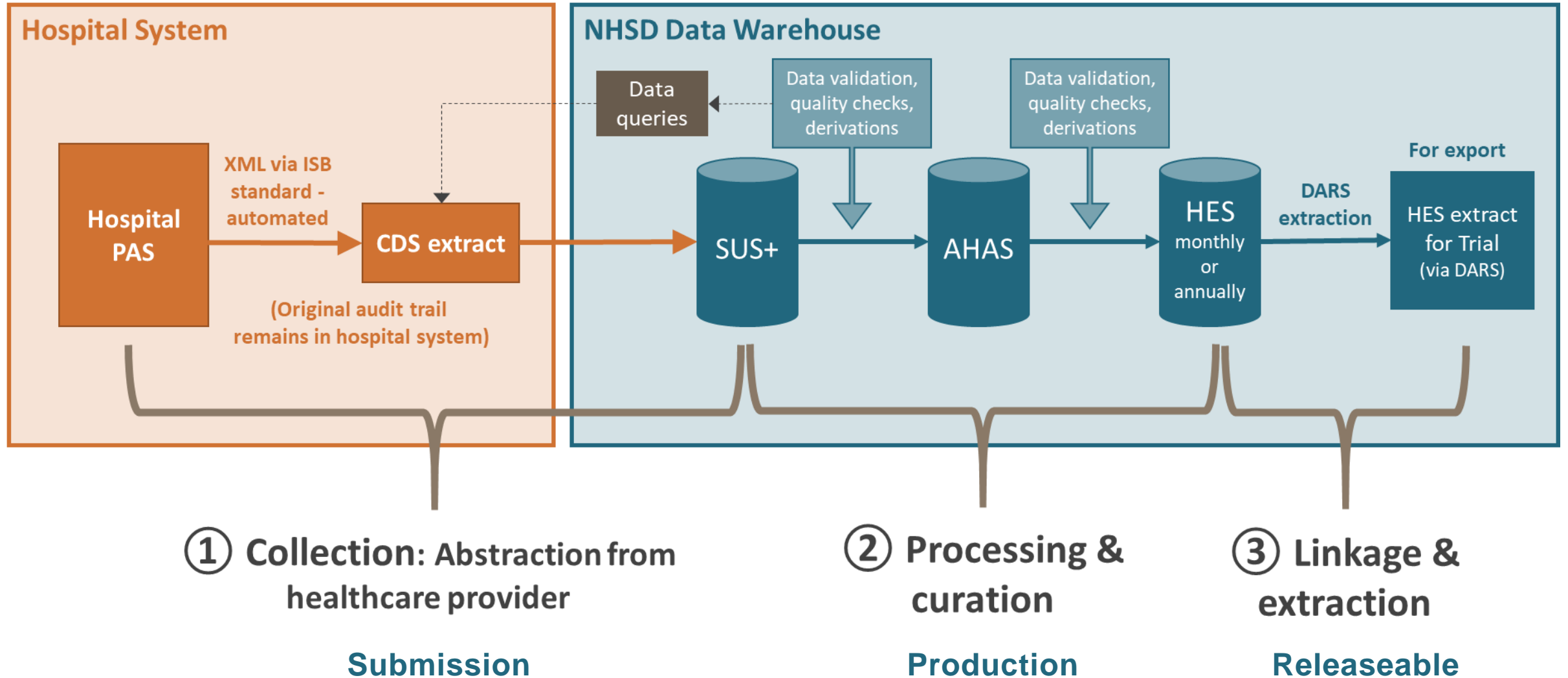
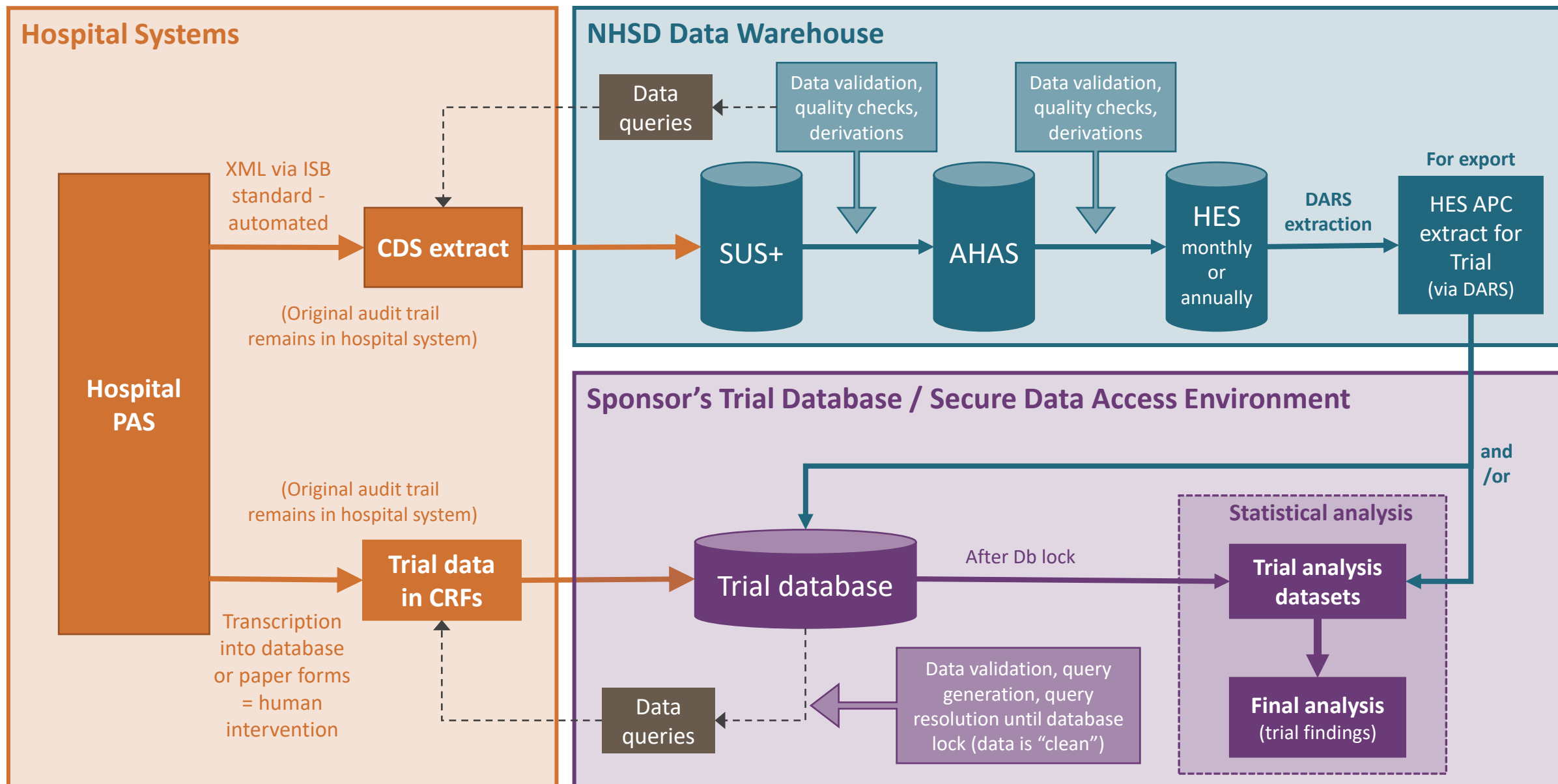
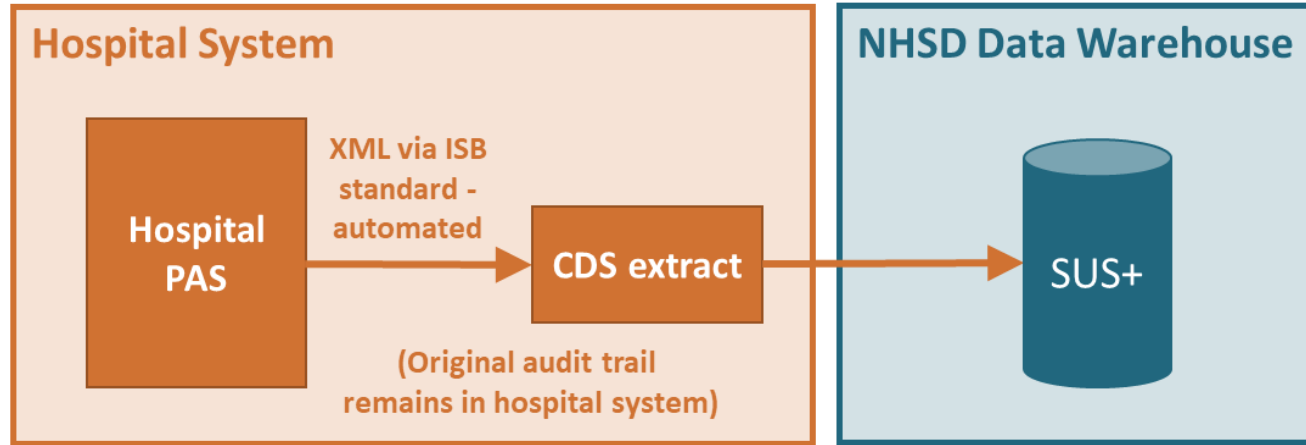


Fig 1: Schematic showing data flows in a clinical trial using both trial-specific data collection & NHSD datasets





# Stage 1 Collection from acute NHS Trusts

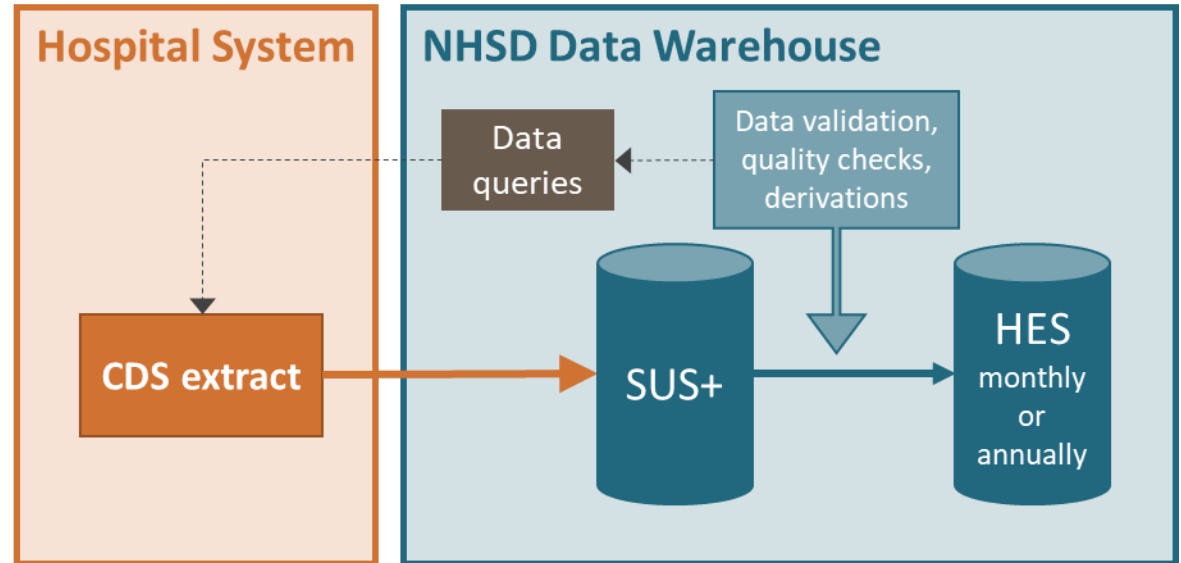


- Twice monthly submission
- XML schema: Information standard (ISB 0092) and CDS Business rules
- Coded with ICD-10 and OPCS-4 pre-submission
- Hospital audit trail not submitted
- Data Quality Maturity Index (DQMI) feedback to Trusts.

- **Accurate, copy** from PAS (note: diagnostic accuracy out-of-scope)
- **Reliable and complete** due to XML schema, Information standard, Business rules.
- Recording accuracy aided by Payment by Results data assurance framework, Care Act 2014, and NHSD's DQMI.
- But primary purpose of PAS for healthcare/patient administration, not research.

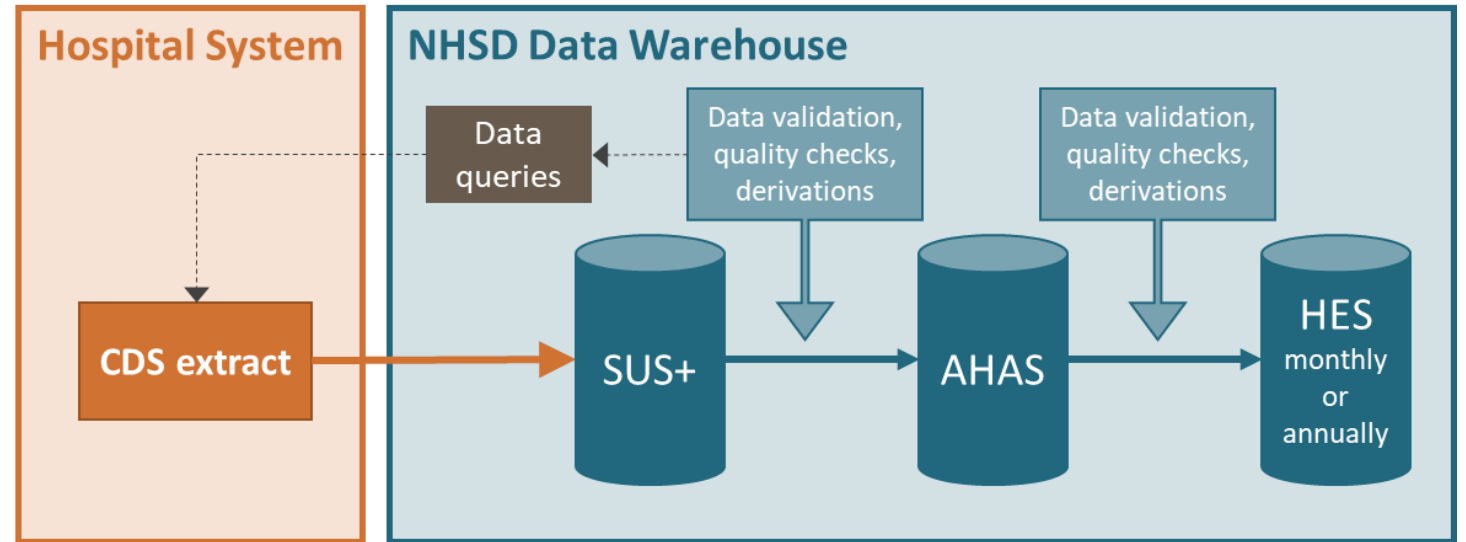
# Stage 2 Processing and curation

- Legacy flow from SUS+ to HES in Oracle data warehouse; decommissioned in July 2021.
- Processing, data quality, and removal of duplicate records described in NHSD publications
- Autocleans dictionary of rules for data cleaning and derivations
- Person-matching using HES identifier (HESID).
- Provisional monthly HES datasets generated, and one annual update in March when data records become static.
- 40-45 days to generate HES APC from CDS submission.
- Data Quality notes and DQMI highlight known issues to Trusts and users, including organisational coverage and completeness of data fields.



# Stage 2 Processing and curation

- DPS Platform initiated 2018, standardises and automates processing
- Used for collection, person-matching, controlled linkage, de-identification and re-identification
- **Improves data integrity, better governance of datasets.**



- Data flows daily from SUS+ to AHAS cumulatively, and person-matched via the Master Person Service (MPS) using MPS identifier (MPSID).
- AHAS daily delta feed available <16 hours of receipt from hospitals as CDS, so **contemporaneous**.
- Provisional monthly HES and annual update generated from AHAS, using same rules as legacy system.
- Changes to the processing of HES allows handling of larger data volumes, **increasing accuracy and timeliness**.

# Stage 3 Linkage and extraction

---

- Linkage via person identifiers HESID (legacy) or MPSID (DPS) which use algorithms for person-matching records.
- Three access routes after successful data access request:
  1. Data Access Environment to DPS hosted by NHSD.
  2. Data extracts from Data Production team at NHSD, securely transferred to user (256-bit AES encryption).
  3. Through the newly developing Secure Environment (aka Trusted Research Environment) hosted by NHSD.
- Trial-specific procedures needed (out-of-scope):
  - to verify record matching to HES APC and,
  - to minimise recording errors of linkage information (NHS number, demographics).

# Conclusions on HES APC and CRD

- ✓ HES APC and CRD are integral datasets based on evidence collated on data management within NHSD.
- ✓ Development of NHSD's DPS platform has improved timeliness and accuracy.
- ✓ Handled robustly, appropriate controls and automation to assure quality for secondary uses, including trials.

# Recommendations

## For HES APC & CRD

- Data lifecycles are documented, and their integrity confirmed, so should be considered **suitable for use as trial data, and equivalent to a transcribed copy of the original source data.**
- Trial teams should prioritise processes that support access and use of healthcare data
- Trial teams must document reasons for using specific healthcare data in the Trial Master File.

## For other healthcare datasets

- Characterisation of the data lifecycle provides a template for other healthcare data providers to follow.
- Assessments of other datasets should be made publicly available, and data providers must take steps to rectify issues with data integrity where possible.

(Murray *et al.*, 2022 [https://doi.org/10.1016/S2589-7500\(22\)00122-4](https://doi.org/10.1016/S2589-7500(22)00122-4))

# Influence on MHRA guideline for RCTs\*

Section 5 states:

*“An example of a suitable scenario for a RWD based trial and an appropriate design would be...an objective endpoint routinely and consistently collected in the EHR database(s) for the patient population considered of interest. For example, **all-cause mortality and inpatient hospitalisations are known to be well recorded in the UK general population.**”*

\*using RWD to support regulatory decisions



MRC  
Clinical  
Trials Unit



HDRUK  
Health Data Research UK

## Demonstrating the Data Integrity of routinely collected healthcare systems data for Clinical Trials (DEDICaTe)

**Lead:** Macey Murray, Senior Research Fellow in Trial Conduct Methodology;  
Clinical Trials secondee, NHS DigiTrials

### Collaborators:

**MRCCTU at UCL** Matthew Sydes, Sharon Love, James Carpenter, Mahesh Parmar.

**University of Oxford** Marion Mafham, Martin Landray.

**NHS Digital** Michael Chapman, Heather Pinches, Laura Sato, Jaspal Panesar, Annie Walker.

Smarter Studies  
Global Impact  
Better Health



# Background

- Trial sponsors need to demonstrate to regulatory authorities that all data, including healthcare systems data, are integral, reliable, and complete.

Provenance = detailed record of the origins of data, processes, & methods by which it is produced.

Integrity = extent to which all data are complete, consistent, accurate, & reliable throughout the data lifecycle

- Provenance and integrity of two NHS Digital's (NHSD) datasets recently assessed & documented: **Suitable for use in trials, equivalent to a transcribed copy of original source data\***.

# Aims

1. To use a data intelligence platform (Collibra) to record provenance & integrity of NHSD datasets including:
  - Hospital Episode Statistics: Admitted Patient Care (HES APC)
  - Civil Registration of Deaths (CRD)
  - Hospital Episode Statistics: Outpatients, Critical Care
2. To semi-automate the process in NHSD's Central Metastore aka "single source of truth"
3. To develop an operating manual



# Methods

## Gather dataset information within NHSD

Logical data dictionary, rules at each stage of submission, production, releaseable

## Use Collibra to ingest information into the Central Metastore to form complete data models

Includes business rules, processing, data quality and validation rules.

## Develop data lineage diagrams in the Metastore

Provide information on data provenance and integrity to NHSD and users (trials community, regulators).

# Central Metastore – key capabilities and vision

## National governance / Burden reduction

- Analytical tool to help determine whether a new data collection is needed or whether a new product can be created from existing data
- Approval workflow(s) for new data collections and data standards

Working with national standards governance

## Data catalogue views

- Provide rich information about our data products to data consumers – feeding future open external and internal data catalogue views of data at NHS Digital + platform-wide business glossary and data dictionary

Internal and external catalogues rollouts

## Corporate governance

- *Approval workflows* for business terms and logical data definitions,
- pipelines specifications and automated communication of designs to production environment

Quality assuring designs, supporting info management / governance audits

## Data design re-use

- Re-use existing business terms, logical and physical data definitions and data processing rules in new product design
- Hold and maintain all specifications in one place – central source of all info - no more spreadsheets!
- Sync with other metadata sources - e.g. NHS DD, DARS, IAOs, DSDS, IOPS, data platform, etc.

Adopting standards, avoiding re-invention

## Metadata-driven processing

- Using metadata to expose pipelines design to business approvers and to define what happens in the data production

Avoiding handcrafted code in production platform



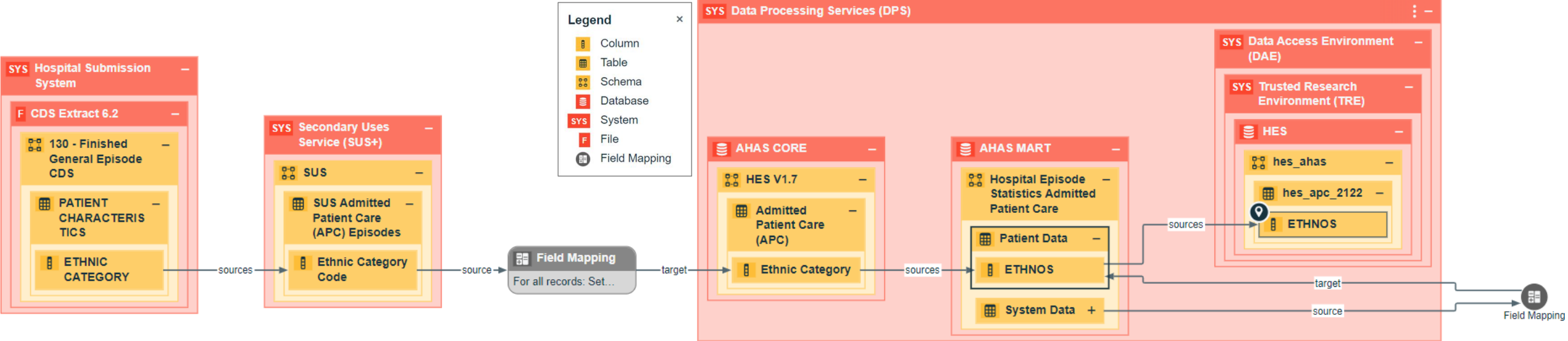










# Results so far

1. Project on track to deliver next Spring (Mar 2023)
2. Complete ingestion of HES APC information in the Central Metastore
  - Data flow diagram (lineage)
3. Ingestion of CRD almost complete
4. Operating manual in development; draft version already in use.







# Simplified data flow of HES APC



**Legend** ×

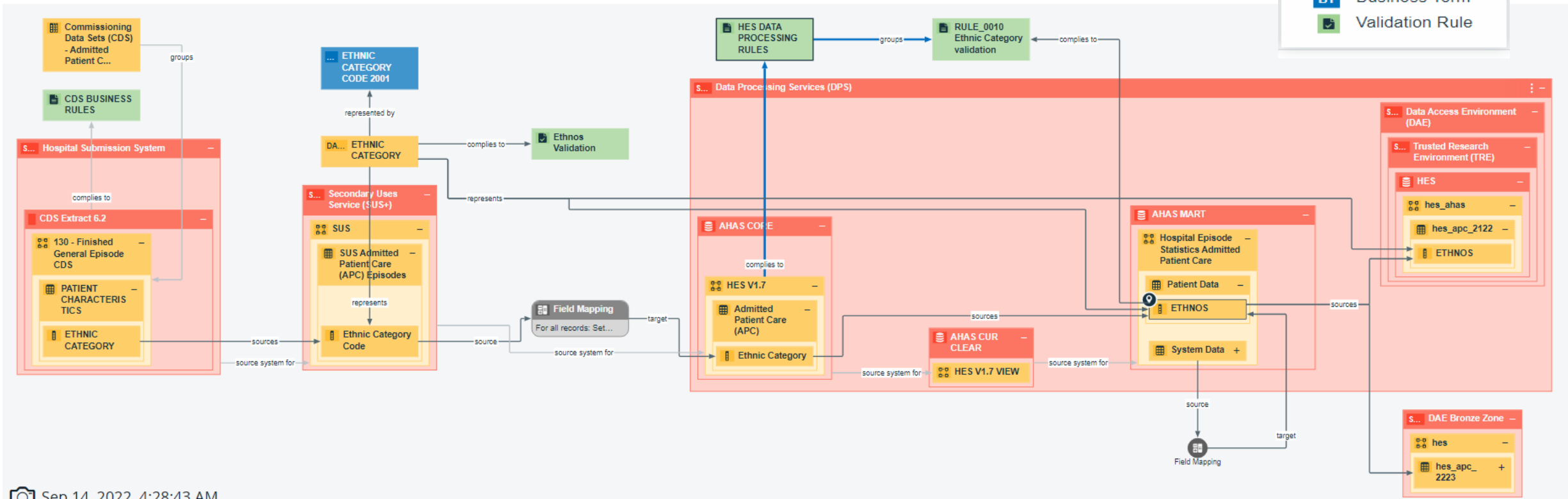
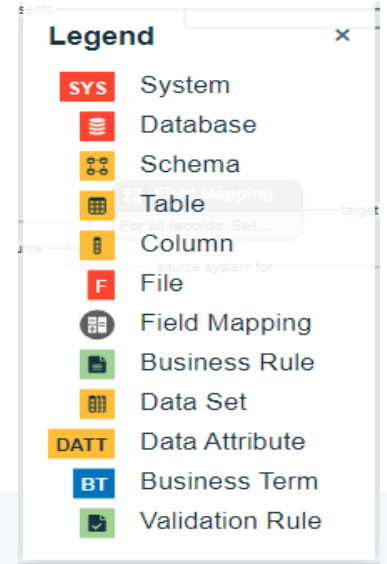
-  Column
-  Table
-  Schema
-  Database
- SYS** System
-  File
-  Field Mapping

**Legend** ×

-  Column
-  Table
-  Schema
-  Database
- SYS** System
-  File
-  Field Mapping



# ...with rules



# Discussion

## Automation

- Initial process is only semi-automated (e.g. ingesting info from spreadsheets).
- Lineage/provenance data relationships may be graphically generated
- Data definitions and standard processing rules may be re-used in building new extracts and pipelines.

**Huge number of NHSD assets, but once information is ingested in Metastore, improves data governance & aids transparency to users.**

# Discussion

## Consistency

- Information gathering phase is most time-consuming, so important to start early and in parallel with other tasks.
- Started early for next datasets.

**Consistency in record keeping is vital**

# Next steps

- Complete ingestion of more data assets including HES OP data
- **User-test content:** Is this what trialists need to know about NHSD data? What do regulators want to see when they audit?
- **Make Central Metastore information accessible to users,** e.g. HDR innovation gateway, NHS Digital webpages.
- Agree a set of minimum requirements to document data integrity and provenance

# Acknowledgments

[macey.murray@ucl.ac.uk](mailto:macey.murray@ucl.ac.uk)   
[@drmakerbaker](https://twitter.com/drmakerbaker) 

- My collaborators and CTU colleagues
- Special thanks to NHSD's Corporate Metadata Team

Laura Sato, Jaspal Panesar, Esha Sandhane,  
Maitree Singh, Jo Simpson, Alex Zhao

**FUNDING:** Health Data Research UK,  
Director's Discretionary Fund 2022.