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 uktnm@nottingham.ac.uk

[https://www.youtube.com/watch?v=m-2z1QkcT7](https://www.youtube.com/watch?v=m-2z1QkcT7)
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
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”
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.
Use of NHS Digital datasets as trial data in the UK: a position paper


Murray et al. 2022. Lancet Digital Health 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**

**1. Collection:** Abstraction from healthcare provider

**2. Processing & curation:** Production

**3. Linkage & extraction:** Releaseable
Fig 1: Schematic showing data flows in a clinical trial using both trial-specific data collection & NHSD datasets

**Hospital Systems**

- Hospital PAS
  - Trial data in CRFs
    - Transcription into database or paper forms = human intervention

- CDS extract
  - XML via ISB standard - automated
  - (Original audit trail remains in hospital system)

**NHSD Data Warehouse**

- SUS+
  - Data queries
  - Data validation, quality checks, derivations

- AHAS
  - Data validation, quality checks, derivations

- HES monthly or annually
  - Data validation, quality checks, derivations
  - DARS extraction
  - HES APC extract for Trial (via DARS)
  - For export

**Sponsor’s Trial Database / Secure Data Access Environment**

- Trial database
  - Data queries
  - After Db lock
  - Statistical analysis
  - Trial analysis datasets
  - Final analysis (trial findings)

- Data queries
  - Data validation, query generation, query resolution until database lock (data is “clean”)
  - (Original audit trail remains in hospital system)
Stage 1 Collection from acute NHS 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.

- 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.
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)
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
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.
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

**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

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

**Data design re-use**
- Re-use existing business terms, logical and physical data definitions and data processing rules in new product designs
- 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.

**Metadata-driven processing**
- Using metadata to expose pipelines design to business approvers and to define what happens in the data production platform

- Working with national standards governance
- Internal and external catalogues rollouts
- Quality assuring designs, supporting info management / governance audits
- Adopting standards, avoiding re-invention
- Avoiding handcrafted code in production platform
Metastore logical data dictionary view

<table>
<thead>
<tr>
<th>Description</th>
<th>Status</th>
<th>Tags</th>
<th>Asset Type</th>
<th>represents</th>
<th>complies to</th>
</tr>
</thead>
<tbody>
<tr>
<td>AREA OF WORK NAME</td>
<td>Candidate</td>
<td>IHS ED</td>
<td>Data Attribute</td>
<td>Hospital Episode Statistics Outpatients, Hospital Episode Statistics Accident and Emergency</td>
<td></td>
</tr>
<tr>
<td>AREA OF WORK NAME (CLINICAL SUB SPECIALTY)</td>
<td>Candidate</td>
<td>IHS_ED</td>
<td>Data Attribute</td>
<td>Hospital Episode Statistics Outpatients, Hospital Episode Statistics Accident and Emergency</td>
<td></td>
</tr>
<tr>
<td>AREA TEAM OF GP PRACTICE</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AREA TEAM OF RESIDENCE</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AREA TEAM OF TREATMENT</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CENSUS OUTPUT AREA, 2001</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CENSUS OUTPUT AREA, 2001 (6 CHARACTER)</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Metastore logical data dictionary view

<table>
<thead>
<tr>
<th>Filters</th>
<th>Description</th>
<th>Status</th>
<th>Tags</th>
<th>Asset Type</th>
<th>Represents</th>
<th>Complies to</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AREA OF WORK NAME</td>
<td>Candidate</td>
<td>NHS_ED</td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREA OF WORK NAME (CLINICAL SUB SPECIALITY)</td>
<td>Candidate</td>
<td>NHS_ED</td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREA TEAM OF GP PRACTICE</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREA TEAM OF RESIDENCE</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREA TEAM OF TREATMENT</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CENSUS OUTPUT AREA, 2001</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CENSUS OUTPUT AREA, 2001 (6 CHARACTER)</td>
<td>Candidate</td>
<td></td>
<td>Data Attribute</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Datasets including this data:
- COVID-19 UK Non-hospital Antibody Testing Results
- Hospital Episode Statistics Outpatients, Hospital Episode Statistics Admitted Patient Care
- Hospital Episode Statistics Accident and Emergency

Physical columns mapped to this item are logically equivalent:
(Multiple releases for the same dataset often have the same physical column name in different tables)

Tagged if in the NHS Data Dictionary

Searchable
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
- 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

- 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.