CPRD data quality work
Assuring quality of datasets
Lessons from CPRD

Dr Tim William, Head of IR
5th December 2023
What is CPRD?

UK Government health data research service supporting **observational** and **interventional** public health and clinical studies by academics, industry and regulators worldwide

Services based on > 30 years of collecting longitudinal primary care EHR across UK

- **>67 million** Patients for observational studies
- **18.7 million** Patients for trials & clinical studies
- **Median 10 years follow-up** 25% 20 years follow-up
- **GP Network** 1 in every 4 GP practices in UK

Daily data collection

28% UK population coverage

Representative data
Impact of CPRD data on public health

- Drug safety
- Drug use
- Disease epidemiology
- Incidence/prevalence
- Care delivery

3000 peer-reviewed publications using CPRD data
CPRD Interventional research services

CPRD SPRINT (Speedy Patient Recruitment INto Trials): supports commercial organisations rapidly recruit high quality patients living with chronic conditions in the community, into phase 2 and 3 trials.

Patient referrals: supports patient invites into academic led studies following location via EHR searches and clinical review.

CPRD PROVE (PRoviding Online Verification of EHR): verification and provision of supplementary information of coded records for selected patients from observational studies.

Cluster trials: supports recruitment of GP practices which are randomised to either intervention or control arms.

Data enabled clinical trials support: data enabled clinical trials support for low-intervention, randomised, phase IV clinical effectiveness trials taking place in a primary care setting.
Central search of CPRD primary care database for eligible patients → CPRD network of 2000 GP practices → GP reviews eligible patients for suitability → GP invites only suitable patients → High quality patients contact study centre
DaRe2THINK EHR-embedded approach

**SCREENING**
- Automated screening for selection criteria across >13 million NHS patients
- Primary care sites invited based on the number of patients meeting inclusion and exclusion criteria
- Weekly updates to notify each General Practice of potentially eligible participants

**ENROLMENT**
- Individual eligibility confirmed by local Primary Care staff
- Potential participants texted/mailed information and invited to discuss their enrolment
- Primary Care or central team go through trial information and obtain informed e-consent
- Counter-signature of consent completed by local Investigator or delegate

**RANDOMISATION**
- Randomisation 1:1 within CPRD portal
- **Intervention arm:** DOAC therapy prescribed using local clinical systems
- **Control arm:** Usual care (no anticoagulant therapy)

**FOLLOW-UP**
- Technology-supported patient reported cognitive function (yearly) and quality of life assessment (6-monthly)
- Key secondary & additional secondary outcomes
- Adverse events acquired from routine clinical records across all primary and secondary NHS care (yearly)
- Primary & additional secondary outcomes

*Auto SAE reporting*

Indicates data-driven automated process
Technology supported to reduce burden

*Courtesy of University of Birmingham: D Kotecha*
CPRD Data quality checks

- CPRD collects data from contributing practices on a daily basis and integrates this with existing data to create releases for research.
- Before the data is made available for research, checks are carried out covering the integrity, structure and format of the data.
- Issues highlighted by the checks are reviewed and addressed before data is incorporated into the data release for researchers.

CPRD checks:
- the volume of data downloaded against that supplied
- data volumes are in the expected range
- all data elements received are of the correct type, length and format
Validation and quality checks

Collection-level validation:
• ensures integrity by checking that data received from practices contain only expected data files and ensures that all data elements are of the correct type, length and format. Duplicate records are identified and removed.

Transformation-level validation:
• checks for referential integrity between records ensure that there are no orphan records included in the database (for example, that all event records link to a patient).

Research-quality-level validation:
• covers the actual content of the data. CPRD provides a patient-level data quality metric in the form of a binary ‘acceptability’ flag. This is based on recording and internal consistency of key variables including date of birth, practice registration date and transfer out date.
Data Quality (DQ) Framework

**Level 1:** Conducted at both a collection level and at a record level, the data is verified to ensure that the data conforms to an agreed specification. Ensures that the collection contains only expected data files, that the collection files are from a known source, and that the collections’ files are in the correct sequence. Structural in nature and are applied to every data element. These checks include that the data item is of the correct type, that fields that should have values do indeed contain a value, and that, where specified, the format and length of the data is accurate.

**Level 2:** Validate the actual content of the data, considering any existing content already collected. Contextual in nature and consider the values of multiple data items within the master database. Data is not corrected, but patient, practice, and database level data quality metrics and markers are put in place. The only exception are opt-outs where a patient's data is removed from future builds.

**Level 3:** Study-specific tests that are either undertaken by researchers themselves or as part of good data management.
Conceptual Process with Data Quality

1. **Capture**
   - Data is pulled from source.

2. **Process**
   - New data is checked, integrated, and processed.

3. **Store**
   - Processed data is stored to reflect the current state.

4. **Build**
   - Database is transformed into static research version with value added fields created and opt-outs applied.

5. **Release**
   - Databases are released along with metadata, release notes, and DOIs.

6. **Use**
   - Databases used on analytics platforms and tools.

**Level 1 DQ**

**Level 2 DQ**

**Level 3 DQ**
CPRD data quality metrics (1)

Acceptability flag (GOLD and Aurum):
- Patient must pass all:
  - *Plausibility*: A patient passes the test if their age is less than or equal to 115 and their gender is male, female, or indeterminate.
  - *Completeness*: A patient passes the test if they have a valid date of birth, valid first registration date and current registration date.

Approximately 98% of CPRD GOLD and 93% of CPRD Aurum patients are acceptable.

‘up to standard’ (UTS) date (GOLD):
- the date at which data in the practice is considered to have continuous high-quality data fit for use in research.
- assurance of continuity in data recording (gap analysis)
- avoidance of use of data for which transferred out and dead patients have been removed (death recording)
CPRD data quality metrics (2)

CPRD death date:

- Source data for CPRD may not always correspond to the date of occurrence. For instance, it may reflect the date of notification of the death to the GP, or when the deceased patients’ registration record was updated.
- CPRD provide an estimate of the date of death (cprd_ddate), based on an algorithm which uses additional information in the patient record.
- Gallagher et al. found that for censoring follow-up and calculating mortality rates, the derived death dates in the CPRD GOLD database are likely to be sufficient as the CPRD mortality rates is comparable with the ONS rate.
CPRD data strategy

- CPRD have implemented a CPRD Research Data Quality Strategy
- Set up an internal CPRD Data Quality Working Group (CPRD DQWG) on behalf of the CPRD Senior Management Team (CPRD SMT).
- Set up a Data Quality Advisory Group (CPRD DQAG) composed of members of the CPRD DQWG and external stakeholders representing sectors that CPRD operate in will act as advisers on development and delivery of the strategy.
- Developing and publishing a Data Quality Strategy which will be a detailed implementation plan and standard operating procedures.
- Developing a series of data quality projects
CPRD Research Data Quality Strategy principles:

1. All staff recognise the need for high standards of data quality and understand their individual roles in achieving this.
2. All members of staff who liaise with CPRD data users will communicate information about data quality clearly, openly, and effectively.
3. All data supplied by CPRD for research are accurate (i.e., reflect the source data), valid, reliable, timely, relevant, and complete, where possible.
4. All validation and verification processes are fit for purpose and transparent, that is, relevant stakeholders have appropriate visibility of data quality processes, procedures, and outputs.
5. All data quality standards and procedures are proportionate for a data custodian and are actively monitored and reviewed.
CPRD Research Data Quality Strategic Objectives

1. Develop and publish data quality standards and metadata that are proportionate, fit for purpose, and that maintain public and stakeholder trust.
2. Implement and maintain robust data quality validation, verification, and monitoring processes to ensure that all CPRD research data is right the first time and made available in a time frame appropriate for its intended purpose.
3. Leverage partnerships with data suppliers/controllers and the CPRD research community to assure the quality of research data.
4. Ensure internal adherence to and external awareness of the roles and responsibilities for high data quality.
5. Annually review and update data quality procedures to ensure that these maintain high quality standards in data collection, processing, and release.
Other CPRD Data Quality Projects

- **Ethnicity record**
- **Exploration of death rates**: GP practices in CPRD Aurum - consistency and comparisons between CPRD Aurum, GOLD and ONS mortality rates
- **National Data Opt-outs Impact**: Explore data sources as external comparators for CPRD data
- **Census comparability**: Comparisons of demography, location/region, ethnicity, urban/rural, long-term health problems, deprivation
- **Primary care resource use**: Completeness and variability between practices, by patient demographics, and over time.
- **Tools development**: Scoping development of effective and efficient tools for the internal/external use and delivery of research data
- **PROVE services**: provide a form of source data verification
- **Prescription cost analysis look-up**: Quality assurance and documentation