



Data Harmonisation : Enabling a federated model for research collaboration

Professor Simon Lovestone, University of Oxford; Mr. Mike Denis, Oxford Academic Health Science Network; Mr. David Newton, University of Oxford; and Mrs. Tanya Smith, Oxford Health NHS Foundation Trust

Background

Specialist or secondary care for Mental Health and Dementia in both hospitals and in the community is provided by dedicated NHS Trusts in the UK, of which there are 54 in England. These Trusts have complete penetration of Electronic Medical Records (EMRs) having been effectively paperless for at least 5 years. The entire clinical records is contained in these records and includes a range of structured and unstructured data such as narrative contemporaneous notes from clinical staff, letters to primary care and other services, Recognising the potential value of these data we established the Clinical Records Interactive Search (CRIS) system, first at the South London and Maudsley NHS Trust and then others¹. CRIS facilitates research access to the entire EMR dataset enabling research use of both structured data and use of natural language processing for derivation of information from the unstructured narrative text.

UK-CRIS and DPUK



CRIS is now established in 14 NHS Trusts, covering approximately one third of those in England. Of these, 11 are now linked allowing research queries to be implemented through federation. In order to capitalize on this capability, we needed to attempt harmonisation to a common standard, establish a governance model to safeguard and secure patient data, and provide an accountable system for sharing de-identified data.

References

1 Stewart, R. *et al.* The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* **9**, 51, doi:10.1186/1471-244X-9-51 (2009).

Methods

CRIS OMOP CDM Mapping and Extract Transform Load (ETL) Development

- Data mapping workshops carried out to map electronic health record entities to OMOP CDM
- Mapping and interface specification developed for RiO and Carenotes system
- Specification used to inform and develop CRIS ETL pipeline
- Enable CRIS OMOP Database to be built as part of each NHS Trust Extract Transform and Load process.



Federated Governance Model

- NHS Workshops to establish information governance principles
- Patient and public engagement to obtain feedback and input to the model
- Iterative testing of the model with each Trust to refine and improve model
- Update technical system to support agreed governance processes
- Establish local and national committees that will conduct project review and approval process
- Develop effective privacy notices to inform the patient and public population
- Design, develop and sign off data sharing agreement



Results

The following was achieved through the data mapping process and governance workshops:

- **54 Concepts** across **6 domains** were mapped between OMOP and individual field level items within the RiO and Carenotes Electronic Health Record Systems.
- Domains: Demographics, Diagnosis, Measurements, Referral, GP Details, and Clinical Notes.
- Total number of de-identified patient records in UK-CRIS: **2.21 million**
- Total number of data points across CRIS OMOP Databases: **100 million+**
- **Federated Governance Model** established across 11 NHS Mental Health Trusts with a data sharing agreement executed between all organisations to support the sharing of de-identified data.
- Online application, review and approval process through <https://ukcris.nhs.uk>
- Several test queries executed successfully across the network of CRIS OMOP databases.



Conclusions

This work illustrates a process whereby disparate localised datasets can be harmonised into a national resource, that enables research and collaboration to take place, whilst providing local control. It offers a case study of how a technical solution (the OMOP CDM) and procedural solution (CRIS governance model) can come together to provide a functional and scalable research platform.

The dataset within the CRIS OMOP CDM can be expanded, and future work will look into adding further mappings to enrich the data available for federated queries. In addition, a dedicated remote analytics space will be built to support analysis of federated datasets (in a secure and contained environment), run NLP programmes and manage derived data.



Our goals in the coming year are to build the CRIS Network with additional NHS Trusts, establish the first set of federated projects utilising this national federated dataset and link to other data sources such as population research cohorts, primary care datasets and other national and regional datasets. We are open for collaboration.

Acknowledgements: We are grateful to funding from the NIHR and MRC to establish and expand the CRIS network as well as to on-going support from CRIS Network member NHS Trusts.