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

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

In the UK, 50% of the National Institute for Health and Care Research-funded trials (to 2019) were planning to access and use RCHD.[1] However, looking at the data given to trials from registries between 2013 and 2018, fewer than 5% of UK-based randomised clinical trials obtained RCHD.[2] As notable examples, the RECOVERY[3] and PRINCIPLE[4] trials of potential treatments for COVID-19 have each successfully harnessed benefits of RCHD, which has simplified multi-site data collection through centralised collation and aided identification of potential trial participants.[5] There is an intention by trialists more widely to make use of RCHD for study design and recruitment through to outcome ascertainment and post-trial follow-up;[1,2,5,6] successful high-profile example trials will encourage further uptake.

Regulatory issues are a major challenge to wider use of RCHD in trials. The Medicines and Healthcare products Regulatory Agency in the UK and the US Food and Drug Administration recognise the potential value of RCHD for clinical trials that support regulatory decisions and each published draft guidance.

Trial sponsors nevertheless need to demonstrate that all the data used in the trial, including the RCHD, are reliable, complete, and relevant. This involves assessing data provenance and integrity, and the validity (diagnostic value) and suitability of routine datasets for trial measures (eg, outcomes, exposures, and covariates).[7] Data provenance is the detailed record of the origins of the data, the processes, and the methods by which it is produced. Data integrity is defined as the extent to which all data are complete, consistent, accurate, and reliable throughout the data lifecycle.[8] Although there is a standard endorsed by regulators for assessing systems that create or capture electronic clinical data as source (that is, original records),[9] there has been insufficient guidance on how to assess centrally curated RCHD.

Therefore, we developed a process to ascertain and document the provenance and integrity of RCHD. Our in-depth report has recently been made publicly available.[9] This report sets out the methods which we applied to the two NHS Digital data assets most requested by trialists: the Admitted Patient Care dataset of Hospital Episode Statistics (HES APC) and the Civil Registration of Deaths (CRD).[9]

The provenance and integrity of these two datasets were evaluated in three key stages: first, collection and transfer of data from health-care systems to NHS Digital's systems; second, centralised processing and curation to form the validated dataset; and, finally, linkage and extraction for trialists and the sponsor. At each stage, we reviewed the tools and systems used, and the controlled processes for managing data, data lineage, and access arrangements. Advice about the level of detail required for documentation was sought throughout from the Medicines and Healthcare products Regulatory Agency, who provided helpful feedback through the development process.

By investigating the data lifecycles of HES APC and CRD, we have demonstrated that their curation is robust, and handled with appropriate controls and automation. We are confident that the data can be considered as equivalent to high-quality transcribed versions of the original source data, and so are sufficiently reliable for use in clinical trials.

Our detailed approach has clear implications for the design, conduct, and analysis of clinical trials. We have demonstrated that these two key health-care systems datasets have the provenance and data integrity for use in clinical trials that would be suitable in regulatory submissions. This approach is relevant to industry as well as academia. Greater use of RCHD should change

For more on the **guidance from the Medicines and Healthcare products Regulatory Agency** see https://www.gov.uk/government/publications/mhra-guidance-on-the-use-of-real-world-data-in-clinical-studies-to-support-regulatory-decisions/mhra-guidance-on-the-use-of-real-world-data-in-clinical-studies-to-support-regulatory-decisions

For more on the **draft guidance from the US Food and Drug Administration** see https://www.fda.gov/media/152503/download

many aspects of trial conduct, including the way trials are monitored and in particular, probably decreasing their carbon footprint.

Our work on two health-care systems datasets is only the initial step. The integrity and provenance of each routinely collected dataset that might be used in clinical trials should now be systematically assessed and clearly documented using the same approach. We call upon, and strongly encourage, all data collators to share and maintain the necessary documentation in a similar manner that we have started for HES APC and CRD.

Trialists must also record the relevance of RCHD (validity and suitability) in their trial protocol and Trial Master File, and we suggest a process of curation and documentation of these choices.[10] Further work is important to assess the use of RCHD against traditional trial-specific data collection methods so trialists can choose which approach to use for data items, accounting for availability, completeness, timeliness, latency, and cost. Such assessments can be achieved through studies-within-a-trial in existing trials.

In conclusion, RCHD has the potential to transform the conduct of clinical trials, but their sponsors need confirmation of their integrity and provenance to satisfy regulatory-grade standards. We have demonstrated the process for two important datasets and now urge data providers to take the necessary steps to facilitate this for all relevant datasets. These steps will make trials more efficient and consequently lead to faster improvements in health care for all.

MM declares research grants from Novartis and Novo Nordisk unrelated to this manuscript. MJL declares grants from the Industrial Strategy Challenge Fund, HDR UK, National Institute for Health and Care Research Oxford Biomedical Research Centre, and Medical Research Council (MRC) Population Health Research Unit unrelated to this manuscript. MRS declares research grants from Astellas, Clovis Oncology, Janssen, Pfizer, Novartis, and Sanofi-Aventis unrelated to this manuscript; and speaker fees from Lilly Oncology and Janssen unrelated to this manuscript. SBL, MLM, JRC, SH, MKBP, and HP declare no competing interests. MLM, SBL, JRC, MKBP, and MRS acknowledge funding from HDR UK (HDR-9005) and MRC (MC_UU_00004/07, MC_UU_00004/08, and MC_UU_00004/09).

*Macey L Murray, Sharon B Love, James R Carpenter, Suzanne Hartley, Martin J Landray, Marion Mafham, Mahesh K B Parmar, Heather Pinches, Matthew R Sydes, on behalf of The Healthcare Systems Data for Clinical Trials Collaborative Group

macey.murray@ucl.ac.uk

Institute of Clinical Trials and Methodology, University College London, London WC1V 6LJ, UK (MLM, SBL, JRC, MKBP, MRS); Health Data Research UK, London, UK (MLM, SBL, JRC, MJL, MM, MKBP, MRS); NHS DigiTrials Programme (MLM, MJL, MM, HP) and Data Services Directorate (SH), NHS Digital, Leeds, UK; Medical Statistics, London School of Hygiene and Tropical Medicine, University of London, London, UK (JRC); Clinical Trial Service Unit and Epidemiological Studies Unit (MM), Nuffield Department of Population Health (MJL), University of Oxford, Oxford, UK; British Heart Foundation Data Science Centre, Health Data Research UK, London, UK (MRS)

1 McKay AJ, Jones AP, Gamble CL, Farmer AJ, Williamson PR. Use of routinely collected data in a UK cohort of publicly funded randomised clinical trials. *F1000 Res* 2020; **9:** 323.

2 Lensen S, Macnair A, Love SB, et al. Access to routinely collected health data for clinical trials – review of successful data requests to UK registries. *Trials* 2020; **21:** 398.

3 Horby P, Lim WS, Emberson JR, et al. Dexamethasone in hospitalized patients with Covid-19. *N Engl J Med* 2021; **384:** 693–704.

4 Yu LM, Bafadhel M, Dorward J, et al. Inhaled budesonide for COVID-19 in people at high risk of complications in the community in the UK (PRINCIPLE): a randomised, controlled, open-label, adaptive platform trial. *Lancet* 2021; **398:** 843–55.

5 Pinches H. NHS DigiTrials: how the search for COVID-19 treatments is revolutionising clinical trials. *The Pharmaceutical Journal* 2021; **306:** DOI:10.1211/PJ.2021.1.75436.

6 Sydes MR, Barbachano Y, Bowman L, *et al.* Realising the full potential of data-enabled trials in the UK: a call for action. *BMJ Open* 2021; **11:** e043906.

7 Medicines and Healthcare products Regulatory Agency. 'GXP' Data Integrity Guidance and Definitions. 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/687246/MHRA_GxP_data_integrity_guide_March_edited_Final.pdf (accessed Feb 10, 2021).

8 Clinical Data Interchange Standards Consortium Electronic Source Data Interchange Group. Leveraging the CDISC standards to facilitate the use of electronic source data within clinical trials. 2006. https://www.cdisc.org/system/files/all/reference_material_category/application/pdf/esdi.pdf (accessed Sept 29, 2020).

9 Murray ML, Pinches H, Mafham M, *et al.* Use of NHS Digital datasets as trial data in the UK: a position paper. *Zenodo* 2022; published online Feb 11. https://www.doi.org/10.5281/zenodo.6047155.

10 The Healthcare Systems Data for Clinical Trials Collaborative Group. Routine dataset justification template. *Zenodo* 2022; published online Feb 15. https://doi.org/10.5281/zenodo.6047938.