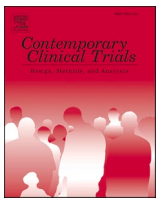
Contemporary Clinical Trials 141 (2024) 107514

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial



Getting our ducks in a row: The need for data utility comparisons of healthcare systems data for clinical trials

Matthew R. Sydes [a,b,c,*], Macey L. Murray [a,b], Saiam Ahmed [a,s], Sophia Apostolidou [a], Judith M. Bliss [d], Claire Bloomfield [e,f], Rebecca Cannings-John [g], James Carpenter [a,h], Tim Clayton [i], Madeleine Clout [j], Rebecca Cosgriff [f], Amanda J. Farrin [k], Aleksandra Gentry-Maharaj [a,r], Duncan C. Gilbert [a], Charlie Harper [l], Nicholas D. James [m], Ruth E. Langley [a], Sarah Lessels [c], Fiona Lugg-Widger [g], Isla S. Mackenzie [n], Marion Mafham [b,l,o], Usha Menon [a], Harriet Mintz [a], Heather Pinches [p], Michael Robling [g], Alexandra Wright-Hughes [k], Victoria Yorke-Edwards [a,q], Sharon B. Love [a]

[a] MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, University College London, London, UK
[b] Health Data Research UK (HDR UK), London, UK
[c] BHF Data Science Centre, Health Data Research UK (HDR UK), London, UK
[d] Clinical Trials and Statistics Unit, Division of Clinical Studies, The Institute of Cancer Research, London, UK
[e] Insitro Inc, San Francisco, CA, USA
[f] NHS Transformation Directorate, NHS England & NHS Improvement, London, UK
[g] Centre for Trials Research, Cardiff University, Cardiff, UK
[h] London School of Hygiene and Tropical Medicine, London, UK
[i] Department of Medical Statistics and Clinical Trials Unit, London School of Hygiene and Tropical Medicine (LSHTM), London, UK
[j] Bristol Trials Centre, University of Bristol, Bristol, UK
[k] Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, UK
[l] Nuffield Department of Population Health, University of Oxford, Oxford, UK
[m] The Institute of Cancer Research, London, UK
[n] MEMO Research, Division of Molecular and Clinical Medicine, University of Dundee, Dundee, UK
[o] Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), NDPH, University of Oxford, Oxford, UK
[p] NHS DigiTrials, NHS England, Leeds, UK
[q] Centre for Advanced Research Computing, University College London, London, UK
[r] Department of Women's Cancer, UCL Elizabeth Garrett Anderson Institute for Women's Health, University College London, London, UK
[s] UCL Comprehensive Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London, UK

Abbreviations: CDM, Common Data Model; CRFs, Case Report Forms; DUCkS, Data Utility Comparison Study; EHDEN, European Health Data and Evidence Network; FDA, Food and Drug Administration; HDR UK, Health Data Research UK; HSD, Healthcare Systems Data; ICD10, International Classification of Diseases v10; MedDRA, Medical Dictionary for Regulatory Activities; MHRA, Medicines and Healthcare products Regulatory Agency; MRC, Medical Research Council; NHS, UK National Health Service; NIHR, National Institute of Health and Care Research; NPV, Negative Predictive Vale; OMOP, Observational Medical Outcomes Partnership; ONS, Office for National Statistics; OPCS-4, Office of Population Censuses and Surveys Classification of Interventions and Procedures version 4; PPV, Positive Predictive Value; RCHD, routinely-collected healthcare data; RCT, Randomised controlled trials; SCORE-CVD, Standardising Clinical Outcome measures in Routinely-collected Electronic healthcare systems data; SDE, Secure Data Environment; SNOMED CT, Systematized Nomenclature of Medical terms – Clinical Terms; SWAT, Study Within A Trial; TRE, Trusted Research Environment.

* Corresponding author.
E-mail addresses: m.sydes@ucl.ac.uk (M.R. Sydes), macey.murray@ucl.ac.uk (M.L. Murray), saiam.ahmed@ucl.ac.uk (S. Ahmed), s.apostolidou@ucl.ac.uk (S. Apostolidou), judith.bliss@icr.ac.uk (J.M. Bliss), cbloomfield@insitro.com (C. Bloomfield), canningsrl@cardiff.ac.uk (R. Cannings-John), j.carpenter@ucl.ac.uk (J. Carpenter), Tim.Clayton@lshtm.ac.uk (T. Clayton), madeleine.clout@bristol.ac.uk (M. Clout), r.cosgriff@nhs.net (R. Cosgriff), A.J.Farrin@leeds.ac.uk (A.J. Farrin), a.gentry-maharaj@ucl.ac.uk (A. Gentry-Maharaj), duncan.gilbert@ucl.ac.uk (D.C. Gilbert), charlie.harper@ndph.ox.ac.uk (C. Harper), nick.james@icr.ac.uk (N.D. James), ruth.langley@ucl.ac.uk (R.E. Langley), Sarah.Lessels@hdruk.ac.uk (S. Lessels), LuggFV@cardiff.ac.uk (F. Lugg-Widger), i.s.mackenzie@dundee.ac.uk (I.S. Mackenzie), marion.mafham@ndph.ox.ac.uk (M. Mafham), u.menon@ucl.ac.uk (U. Menon), HXM074@student.bham.ac.uk (H. Mintz), h.pinches@nhs.net (H. Pinches), Roblingmr@cardiff.ac.uk (M. Robling), A.Wright-Hughes@leeds.ac.uk (A. Wright-Hughes), v.yorke-edwards@ucl.ac.uk (V. Yorke-Edwards), s.love@ucl.ac.uk (S.B. Love).

https://doi.org/10.1016/j.cct.2024.107514
Received 13 October 2023; Received in revised form 23 February 2024; Accepted 24 March 2024
Available online 26 March 2024

ARTICLE INFO

Keywords:
Healthcare systems data
Health policy
Routinely-collected healthcare data
Electronic health records
Real world data
Routinely-collected data
RCTs
Data utility
Data utility comparison studies
DUCkS

ABSTRACT

Background: Better use of healthcare systems data, collected as part of interactions between patients and the healthcare system, could transform planning and conduct of randomised controlled trials. Multiple challenges to widespread use include whether healthcare systems data captures sufficiently well the data traditionally captured on case report forms. "Data Utility Comparison Studies" (DUCkS) assess the utility of healthcare systems data for RCTs by comparison to data collected by the trial. Despite their importance, there are few published UK examples of DUCkS.

Methods-and-Results: Building from ongoing and selected recent examples of UK-led DUCkS in the literature, we set out experience-based considerations for the conduct of future DUCkS. Developed through informal iterative discussions in many forums, considerations are offered for planning, protocol development, data, analysis and reporting, with comparisons at "patient-level" or "trial-level", depending on the item of interest and trial status.

Discussion: DUCkS could be a valuable tool in assessing where healthcare systems data can be used for trials and in which trial teams can play a leading role. There is a pressing need for trials to be more efficient in their delivery and research waste must be reduced. Trials have been making inconsistent use of healthcare systems data, not least because of an absence of evidence of utility. DUCkS can also help to identify challenges in using healthcare systems data, such as linkage (access and timing) and data quality. We encourage trial teams to incorporate and report DUCkS in trials and funders and data providers to support them.

1. Introduction

Randomised controlled trials (RCTs) are essential in providing reliable evidence of the efficacy and safety of healthcare interventions intended to treat or prevent disease, and remain the most reliable method to assess new interventions. However, they are commonly expensive in absolute terms, complex to implement and can take many years. Healthcare systems data, also known as routinely-collected healthcare data (RCHD), are collected routinely during interactions between patients and the healthcare system, whether or not they are part of research. Judicious use of these data has the potential to transform the design and conduct of future trials, [1] although many challenges need to be addressed before the wider trials community, including regulators, funders and guideline developers, embrace this approach. Here, we set out the rationale for using such data and summarise these challenges, before focusing on how trial teams can contribute to assessment of "data utility" of healthcare systems data.

1.1. Rationale for using healthcare systems data in trials

The daily pressures in public healthcare settings, like the National Health Service (NHS) in the UK, are immense. Reducing the data collection burden imposed by research, even partially, should provide notable relief to sites and remove a limit on the amount of research sites can deliver. *Re*-use of data already collected by the healthcare system should potentially increase quality (e.g. trial-relevant events are identified even when they occur outside the trial site, such as hospitalisations) and lead to a substantial increase in efficiency, avoiding wasteful, duplicative trial-specific data collection systems. Healthcare systems data could widely support retrospective provisioning of participants' baseline characteristics and pertinent medical history, and healthcare systems datasets could further support prospective collection of follow-up information including trial-specific outcome measures and facilitate efficient and complete long-term follow-up.

Many trial teams have been considering the use of healthcare systems data from primary care (GPs), secondary care (hospitals), and other healthcare delivery settings to support identification of trial sites or even to identify and directly approach potential participants. Such routes to recruitment may be more efficient and inclusive than efforts through clinics and/or advertisements, thus, avoiding research waste from trials that struggle to recruit. This approach may particularly suit trials involving rare/uncommon diseases, patient groups not being seen routinely at research-active sites, and screening, vaccine or public health interventions trials where there is need for recruitment from the general population. Such an approach may also help reduce inequalities of access, given the well-documented under-representation in many trials of participants from ethnic minority, participants from deprived backgrounds, women, people with co-morbidities and hard-to-reach populations. Up-to-date institutional healthcare systems datasets have shown value in terms of assessing site feasibility and set-up and to monitor recruitment eg 3C (NCT01120028) and RECOVERY (NCT04381936) trials, [2,3] and can allow (near-) real-time checking of how reflective recruited participants are of the intended, underlying populations.

Box 1
Broad challenges in using healthcare systems data for clinical trials.

Area		Expansion
Knowledge of and access to appropriate, timely data	→	Data structure and demonstrable evidence of the availability, structure and contemporaneity of dataset; route to approvals, application times, contracts, technical requirements; and transparent costs
Integrity and provenance of the data	→	Is the healthcare systems dataset a reliable, transcribed copy of the original source data?
Accurate linkage through identifiers	→	Appropriate collection, checking and matching required for accurate linkage
Utility of the healthcare systems data to replace trial-specific data collection	→	Area in which Data Utility Comparisons Studies (DUCkS) are needed
Archiving and retention of data according to good practice and regulations	→	Ensuring consistent between legal agreements and regulations
Onward sharing of the trial dataset	→	Issues of consent, anonymisation and data ownership

1.2. Challenges to using healthcare systems data in trials

Box 1 lists some of the numerous challenges to using healthcare systems data in clinical trials.[1,2] [1,4–11] We focus here specifically on the challenge of "utility", because this challenge can be directly addressed by trial teams; others challenges require evaluations to be led by the data provider. A healthcare systems dataset may be considered appropriate and useful for aspects of a given trial — to have utility — if it

[1] BHF Data Science Centre. 2023. "Navigating Health Care Systems Data for Clinical Trials". 2023. https://www.hdruk.ac.uk/wp-content/uploads/2023/02/Datasets-website-3.pdf

[2] NHS England. 2023. "Demonstrating the Data Integrity of routinely collected healthcare systems data for Clinical Trials (DEDICaTe)". https://dedicate.healthandcaremetadata.uk

Table 1
Considerations in developing and delivering data utility comparison studies for outcome measures.

Issue	Participant-level[1]	Trial-level
Broad question	Is there agreement of fact of data item (e.g. did event happen) and, if so, is there agreement on its timing and definition?	Is there broad agreement on the comparative treatment effect in terms of clinical relevance and statistical certainty?
Data requirement	Data items or outcome measures available in both trial-specific data collection and healthcare systems data. Any algorithms or code lists for deriving items or outcome measures must be clear.	Data item or outcome measure available in both trial-specific data collection and healthcare systems data. Any algorithms or code lists for deriving items or outcome measures must be clear.
Analysis setting	Participant level data for both trial data and healthcare systems data must be stored and available for analysis in the same location.	Participant level data can be stored and available for analysis in separate locations (provided treatment allocation is present in both locations).
Timing of comparison	Can be done at any time in a trial's lifecycle, including pilot assessments, without disclosing accumulating comparative treatment effects of the trial. If done early, need to have sufficient events to be reliable.	Can only be done after comparative data from the main trial have already been disclosed, else confidential findings would be revealed.
Which participants to include?	Comparison of timing and nature of events can be done in participants appearing in both datasets. In time-to-event trials that are planned according to a number of control arm events, it may be judicious to focus on only control arm participants, whereas in trials planned according to number of events in total, it may be safer to focus on the arms combined without reference to allocated treatment. [28]	Comparison of timing and nature of events can be done in participants appearing in both datasets.
Which source is the reference for the comparison?	May be appropriate to assume neither is the "gold standard". Exploratory analyses can be done with both trial-specific data collection as the reference and again with healthcare system data as the reference.	May be appropriate to use trial-specific data collection as reference point as the primary analysis may already have been done and published.
Sample size and power	What power is available for the assessment? Follow justification for most SWATs where sample size is defined by the number of participants in the underpinning trial, but need to know that the effort is worthwhile. Calculation would be separate for each source of healthcare systems data.	What power is available for the comparison of treatment effects? Consider whether heterogeneity of treatment effect could be presented. Follow justification for most SWATs where sample size is defined by the number of participants in the underpinning trial, but need to know that the effort is worthwhile. Calculation would be separate for each source of healthcare systems data. Power for any comparison of treatment effect by data collection approach will be very limited: looking for similarity in the treatment effect rather than testing whether the treatment effect is different by data source.
Analysis method	Analysis of agreement. Cohen's kappa can be used as a measure of agreement for binary analyses, or suitable alternatives for categorical or continuous data. Note that sensitivity, specificity, PPV and NPV assume one source to be a gold standard, but both may have events not (yet) reported in the other source.	Summarise proportionate and absolute effects from both methods, noting also number of events, width of confidence intervals and, where appropriate, median follow-up time. May require separate analyses by data source if data are from across national borders; important that the data are interpreted for the methodology implications rather than the clinical implications.
Potential implications	Inform future trials and potentially this trial: could this provoke a point of choice between two planned data collection methods during an ongoing trial i.e. switch some or all future data collection to only healthcare systems data or confirm to use only trial-specific data collection?	Inform future trials
Other considerations	Allows evaluation and exploration of discrepant events.	If there is complete agreement on the patient-level comparison, there should also be agreement on the trial level comparison already and trial-level comparison is not needed. Use of healthcare systems data may also improve censoring time (time known to be event-free) in patients without events, giving more information. Recognise that there will be very limited statistical power to detect small but potentially important differences in treatment effect. Therefore, should be combined with a participant-level assessment of agreement wherever possible. Focus is not on further clinical dissection of the findings by potential subgroup, only on assessing use of healthcare systems data. Draft FDA guidance[2] suggests trial-level treatment effect comparisons.

[1] (For recurrent events, consider at event-level rather than participant-level)
[2] US FDA. 2021. "Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products: Guidance for Industry." https://www.fda.gov/media/152503/download

sufficiently captures the relevant required data.

Traditional means of data collection in trials includes transcription of data at sites from healthcare records (source data) onto paper or electronic case report forms (CRFs). CRFs can be prone to both transcription errors, misclassification errors and inadvertent omissions, such as missed admissions in other healthcare settings. [12] Trial teams, particularly in industry-led trials, spend huge amounts of time and money sending staff physically, or increasingly virtually, to trial sites to laboriously check that the data recorded on the CRFs match the source data. This places a burden on sites as well as the sponsor's team and has a considerable carbon cost. [13] The Medicines and Healthcare products Regulatory Agency (MHRA) has referred to trials that undertake partial-replacement or supplementation of site data collection with HSD as

"hybrid trials".[3] This could fundamentally shift how data in trials are collected, processed and checked. This efficiency would increase with detailed national-level data over regional-level or site-level data.

Trial teams are increasingly supplementing CRFs with capture of data from participants e.g. subjective patient-reported outcomes (PROs), research use and wearables data. Such data are not yet regularly included in the healthcare systems data and may come into scope in the future.

1.3. What are DUCkS?

Assessing whether relevant data could be sufficiently well-collected from healthcare systems datasets can be done through a series of "Data Utility Comparison Studies" (DUCkS). In these studies, the relevant data collected through trial-specific data collection is compared to data acquired from the healthcare systems. This ensures that new algorithms or phenotypes [14] that describe how categorised data can be drawn together from healthcare systems data capture what they intend to for trial outcomes.

The considerations for DUCkS were developed through a series of informal, multi-way and multi-disciplinary discussions in various forums, where the authors iteratively shared lessons learned from their experiences of using healthcare systems data in trials and from developing, delivering and supporting data utility comparisons.

Table 1 sets out considerations in developing and delivering DUCkS for trial outcome measures. These can be done on a participant-level and/or a trial-level. Participant-level comparisons would focus on assessing agreement within the records of each participant, such as fact of event, timing of event, participant characteristic or treatment detail. Trial-level comparison would focus on assessing consistency of treatment effect estimates by randomised comparison. Thus, some DUCkS may only be done on a completed trial, while others can be done during an ongoing trial.

The approach can also apply beyond outcome measures to assessing the use of other data points for trials. For example, in a randomised trial, baseline concomitant medications should likely be balanced across the groups, so the focus should be on participant-level agreement. In contrast, for data on those additional treatments used subsequent to initiation of the primary treatment period in a trial, both participant-level and trial-level agreement might be useful.

There appear to be few data utility comparisons from the UK in the published literature [15] and, with a fuller review of the literature ongoing, [16] those that do exist vary in the methods used to quantify utility and focus on patient-level comparisons, trial-level comparisons or both. For example, the DUCkS of death data in the BOSS trial (ISRCTN54190466) of screening for Barrett's oesophagus compared death records in trial-specific data collection with those from the Office for National Statistics (ONS). [17] In patient-level comparisons, at each of the yearly data snapshots, the ONS dataset contained more deaths than trial-specific data collection contained, but there were also some deaths known to the trial teams through CRFs which were not yet available in the ONS datasets. Trial-level comparison of the treatment effect was rightly avoided in that DUCkS because BOSS, an open–label trial, was still "blinded" in terms of comparative treatment effect – the accumulating, comparative data had not yet been disclosed. Such careful considerations show how DUCkS can be safely done in an ongoing trial.

Another example is from the ASCEND trial (NCT00135226) in people with diabetes. That DUCkS looked retrospectively at both patient-level and trial-level comparisons of cardiovascular and bleeding events expected to be detectable within healthcare systems datasets and those reported by participants on a mail-based questionnaire and adjudicated by a committee. [18] In that case, the trial results had already been published, so comparisons could be done at both participant-level and trial-level. The DUCkS approach can also be used to assess datasets outside of the trial's primary outcome measure and across multiple healthcare systems datasets. For example, further studies within the UK-wide ovarian cancer screening trial, UKCTOCS (NCT00058032), were able to assess the concordance of common cancers other than ovarian cancer across death registries, cancer registries and national hospital data, [19,20] and the PATCH trial (NCT00303784) of transdermal oestrogens for men with prostate cancer compared cardiac toxicities between registries and CRF data. [21]

The UKCTOCs main results paper also include trial-level comparison as a sensitivity analysis [22], as did the Building Blocks trial. [23]

1.4. Conducting DUCkS

Table 2 sets out experience-based considerations for DUCkS in five broad steps, structured in terms of planning, protocol development, data, analysis and reporting. We expand on three considerations.

First, ensuring patients are correctly linked to the necessary healthcare systems datasets in line with participant consent and with appropriate ethical approvals. This has been challenging, anecdotally. Inaccurate linkage can result from errors in the identifiers collected by the trial teams or by mis-linkage by data custodians. Therefore, trial teams need systems for checking either source of error. Data providers must confirm whether linkage has been achieved: a "null return" could mean the participant has not yet contributed to that healthcare systems dataset (e.g. perhaps had not yet been hospitalised) rather than had not been linked. Accurate linkage [24] is a separate challenge to utility and the two should not be confused: DUCkS can be done successfully by focusing only on those people for whom linkage has been reassuringly achieved.

Second, selecting the appropriate times at which snapshots of datasets are taken ("data-freeze") for DUCkS is critical. Researchers need to compare like with like. This may be straightforward for baseline and historical data. For prospective and outcome data, it is appropriate to find as close a data-freeze date as possible for the two sources. Administrative censoring (ignoring data after an agreed date) should be applied to the dataset with the later data-freeze and accounting for delays in routine data reaching certain healthcare systems datasets. It is common in some trial settings to use adjudication committees or Endpoint Review Committees to determine, often retrospectively, whether a participant had a trial event. Any DUCkS must consider whether the healthcare systems data would be compared to adjudicated or unadjudicated trial data and whether it would also be adjudicated. The need for adjudication is likely to continue in trials even with a move to greater use of healthcare systems data.

Third, trial teams must consider carefully whether it is appropriate to conduct DUCkS at the participant-level, trial-level or both. To reiterate, ongoing trials must not disclose information that is usually kept confidential, such as the treatment effect. Therefore, ongoing trials can likely do only participant-level DUCkS whereas completed trials can additionally do trial-level DUCkS. Participant-level comparisons in an ongoing trial may provide evidence to support the choice data collection method for the latter parts of the trial, as done in the SHIFT trial [25], and could be incorporated within pilot studies.

1.5. Disseminating the findings of DUCkS

The findings of new DUCkS should be published promptly so the community can consider findings while they are current. Health Data Research UK has developed a forum for key stakeholders to discuss how to support appropriate use of healthcare systems data, including the results of DUCkS. Its aim is for the research community to discuss the findings

[3] MHRA. 2021. "MHRA guideline on randomised controlled trials using real-world data to support regulatory decisions". https://www.gov.uk/government/publications/mhra-guidance-on-the-use-of-real-world-data-in-clinical-studies-to-support-regulatory-decisions/mhra-guideline-on-randomised-controlled-trials-using-real-world-data-to-support-regulatory-decisions

Table 2
Experience-based considerations for data utility comparisons.

Step 1: Planning
- Is a new DUCkS required for the relevant baseline characteristics, treatment adherence data or outcome measures? No need to do if there is already sufficient evidence in the literature.
- Plan as early as possible for comparison of trial-specific data collection and healthcare systems data, ideally into trial protocol and build consent to access healthcare systems datasets into participant information sheets from the outset. Where healthcare systems data has not yet been acquired, it could be brought in for the purposes of these assessments and may require re-consent of participants or permitted work-around.[1] Before a new trial, may seek to do in a previous trial.
- Engage data provider(s) early in the process for support, because demonstrating utility for the datasets they hold should encourage other research teams to work closely with them in the future.
- Ensure sufficient support in place from funders (e.g. trial funders) to access the healthcare systems data and to complete the project.
- If the outcomes required are present in the healthcare systems data, this information can be directly extracted using the coding. However, where outcomes are not directly available, e.g. progression, algorithms need to be designed to indirectly identify outcomes from patterns of healthcare interactions. Such extracted outcomes also need to be validated via participant and/or trial-level DUCkS.
- Consider implications of doing a DUCkS and how the findings would (or would not) impact the trial
- Determine who will be involved in the DUCkS (trial team or independent people) and the route to approval
- Plan to complete the work in a timely fashion so the findings are contemporaneous and relevant and do not only apply to the past.
- Undertake comparisons of trial-specific data against various computable phenotypes and algorithms (code lists with application rules) to help researchers better choose those which are most information-rich and most sensitive to capturing treatment effects.
- Determine where the DUCkS will be done: secure network, secure data environment (SDE) or trusted research environment (TRE), or multiple locations.

Step 2: Protocol and/or Statistical Analysis Plan
- Follow a pre-published SWAT protocol where possible or publish one if a suitable protocol does not yet exist.[2] Incorporate into trial protocol if possible. Develop and follow tailored Statistical Analysis Plan.
- Confirm the code lists / phenotyping algorithm used, including classification system being used (e.g. MedDRA, SNOMED, ICD10, OPCS-4, etc.) and the version. Note the "correct" list or algorithm may not be known at the outset but can follow precedent where possible.
- Where data are derived from multiple codes and sources, make these clear in appendices and, ideally, link to precedent such as the HDR UK Phenotype Library.

Step 3: Data
- Access healthcare systems data from each of the providers in an appropriate environment.
- Check the linkage of the participants. Recognise that accurate linkage is a separate problem to utility. Utility comparisons should focus on those participants that were successfully linked. The number of patients not successfully linked should be clearly listed in a flow diagram.
- Has the trial previously connected to healthcare systems data and used it check or update the dataset? The cleanest comparison would involve trial data and healthcare systems datasets being connected for the first time.
- Be clear on the data freeze date for each dataset. These should be closely aligned, ideally the same date. If this is not possible, use administrative censoring to effectively ignore data after the data freeze date in the dataset with the longer follow-up period.

Step 4: Analysis
- Calculate agreement at the participant-level and, or (for recurrent events) the event-level.
- Evaluate time or value differences where there is agreement in fact of event.
- Evaluate the discrepancies: what might have caused them? This could be: calendar time; geography or participating site; participant characteristic; healthcare system dataset artefact; trial-specific data collection artefact.
- Undertake trial-level comparison of treatment effect when appropriate.[3] Report on number of events, treatment effect (relative and absolutely), confidence interval width and follow-up compliance.
- Report separately by source of healthcare systems data and also by sources that would be combined in practice (i.e. including by nations or regions separately and together).
- Report separately for each code list or phenotyping algorithm used, allowing others to decide on "broad" or "targeted" approaches.

Step 5: Reporting
- Promptly and transparently make the findings publicly available; flag the work with systematic reviewers while the work is ongoing.
- Reporting must be done in a way that does not impact an ongoing trial or inappropriately disclose accumulating, comparative data.
- Notify data providers of the findings so that they can address any issues identified.
- Upload findings to a suitable repository e.g. an element of HDR UK Gateway.
- Notify any systematic reviewers that have pledged to collate accumulating information in relevant areas or provide context in manuscript e.g. [16]
- If following, or starting from, a standard SWAT protocol, notify the SWAT protocol author and upload results to the SWAT repository for the relevant protocol.

[1] In some nations, laws may permit temporary lifting of the duty of confidentiality e.g. for England and Wales this can be done appropriately under the Health and Social Care Act via application to use Section 251.

[2] SWAT / SWAR repository: https://www.qub.ac.uk/sites/TheNorthernIrelandNetworkforTrialsMethodologyResearch/SWATSWARInformation/Repositories/SWARStore/

[3] Unlikely to be appropriate during an unreported trial

and to determine whether particular datasets are ready for use in this way, and to encourage data holders to take action if required. Forum members will also prioritise future data utility comparisons, contextualised by removing the other barriers to using healthcare systems data. Supporting development of this evidence-base should be of interest to data holders, not least because it may drive researchers to seek their services, or better still their collaboration, more often in the future.

Understanding where onerous trial-specific data collection could be replaced by healthcare systems data should be of interest to all funders in relative terms and of particular interest to industry sponsors where trial-specific data collection may currently be of higher volume in absolute terms and where the reduction in effort would, therefore, be more pronounced. It is understood that regulators are becoming increasingly accustomed to seeing healthcare systems data being used *around* trials as supporting evidence, but have less often seen use of healthcare systems data *within* trials. The findings of these comparisons should be of relevance to all parties in these discussions.

It is not clear just how much evidence would be required from DUCkS to significantly shift data collection practices for future trials. Positive findings from just one DUCkS should not be sufficiently reassuring that a particular healthcare systems dataset can definitively replace trial-specific data collection; similarly, negative findings from just one DUCkS should not deter researchers from considering that data source in the future. The wider trials community has not swiftly embraced other approaches improving trials efficiency so the evidence-base may need to be quite extensive to drive change. It is through the conduct, sharing and

collation of such comparisons that trial teams can decide whether to have confidence in choosing to use data from particular healthcare systems datasets rather than trial-specific data collection. It is also important that interpreters of trial outputs (e.g. funders, healthcare professionals and healthcare regulators) can have confidence in results where healthcare systems data has been used within trials. When choosing to use a healthcare systems dataset in a trial, the trial should reference relevant DUCkS as justification, in their Trial Master File, alongside other evidence of integrity and provenance (see Supplement of Murray et al). [7] Once confidence in the utility of healthcare systems has been assessed, and each of the challenges addressed, the general need for DUCkS should diminish, except for outcome measures where utility has not yet been addressed.

1.6. Further considerations

Currently, the nationally-collated healthcare datasets in the UK do not have the same richness as local primary and secondary healthcare records. They also do not contain common trial outcome measures and those require calculation across multiple systems, if the components are even recorded. For example, it has been anecdotally difficult for cancer researchers to differentiate new metastases from new primary tumours in healthcare systems datasets. The BHF Data Science Centre-led SCORE-CVD project is looking at how to define key outcome measures for cardiovascular trials available from healthcare systems data. [26] The planned NHS Research Secure Data Environment (SDE) Network, comprising the NHS England and Sub-National SDEs in England are a step towards addressing the depth of data in collated systems.[4,5] DUCkS could usefully be performed in such an environment, provided systems are in place for linkage with research data and adequate protections are provided for the trial data. Such potential uses of proposed SDEs, and other methodological research, should be considered by the SDEs developers from the outset.

Some organisations are seeking to map data so trial-specific CRFs to be signed off by an authorised investigator can be auto-populated directly from local electronic healthcare records. This may be useful where the trial needs a richness of data that is not currently nationally-available in collated healthcare systems datasets e.g. laboratory data. Rich datasets often include free-text which would need careful processing. However, population of site-level CRFs would require considerable effort to set-up at each site and would still leave trial teams exposed to a risk that healthcare interactions and events may be missed or not validated if they happen elsewhere. This may change in the future when all healthcare sites have healthcare systems data that have been mapped onto an international standard, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).[6] Greater uptake of OMOP in the UK and Europe is being encouraged by HDR UK and the European Health Data and Evidence Network (EHDEN).[7]

Further thoughts on assessing the utility of "real-world data" more broadly than clinical trials are offered in Health Data Research UK's (HDR UK) Data Utility Framework, [27] and the US Food and Drug Administration's (FDA) draft guidelines,[8] the latter being explicitly targeted more narrowly at industry. Trial funders should consider how they can support DUCkS within the trials they fund which may help deliver ways to better funder future trials. Evaluations should be explored on the impact of trials' carbon-footprinting with a shift in data collection. The trials community must also look outside trials for evidence from other forms of research, including prospective cohort studies, which collect data on CRFs and could also be contributing to DUCkS of healthcare systems data.

1.7. Limitations

We have drawn together these considerations from informal discussions. They can serve as building blocks for future conduct of such DUCkS. We anticipate there may be a need to revisit these considerations and perhaps develop formal guidelines through an international Delphi process. in the future after more DUCkS are undertaken and published.

2. Summary

Planning the assessment of data utility involves lining up many ducks in a row, including getting the data (costing, consent) and appropriately defining outcome measures. Finding time for these activities, unless planned from the outset, can be difficult in busy trials, but more DUCkS are needed to build the wider body of evidence for using healthcare systems data. There are many organisations interested to engage, for example Health Data Research UK and the NIHR-MRC Trials Methodology Research Partnership. With further data utility comparisons completed and shared in the literature, and with necessary evidence addressing each of the key challenges, we can start to change how trials are conducted, with the aim of improved efficiency and quality.

Funding

AGM, DG, JC, MLM, MRSy, REL, SBL, UM acknowledge funding from UKRI's (UK Research and Innovation) MRC (Medical Research Council) to MRC CTU at UCL (MC_UU_00004/08).

MRSy and SLe acknowledge funding from the British Heart Foundation This work was supported by the British Heart Foundation Data Science Centre (grant SP/19/3/34678); awarded to Health Data Research UK.

ISM, MLM, MR and MRSy acknowledge funding from HDR UK to University of Cardiff, University of Dundee, University of Oxford and UCL (2023.0025).

MLM, MRSy and SBL acknowledgement funding from HDR UK to UCL (TF2022.28).

FLW and MR acknowledge funding from Health and Care Resarch Wales and Cancer Research UK.

CRediT authorship contribution statement

Matthew R. Sydes: Conceptualization, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing.

Macey L. Murray: Conceptualization, Investigation, Methodology, Writing – review & editing, Writing – original draft.

[4] NHS England blog (Claire Bloomfield). 2023. "Investing in the future of health research: secure, accessible and life saving". https://www.england.nhs.uk/blog/investing-in-the-future-of-health-research-secure-accessible-and-life-saving/#:~:text=The%20Sub%20National%20SDEs%20are,collaborations%20and%20successful%20research%20partnerships

[5] Dept of Health and Social Care. 2023. "Secure data environment for NHS health and social care data - policy guidelines". https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines/secure-data-environment-for-nhs-health-and-social-care-data-policy-guidelines

[6] Observational Health Data Sciences and Informatics. 2023. "Standardized Data: The OMOP Common Data Model". https://www.ohdsi.org/data-standardization

[7] HDRUK news article. 2022 "New data partners join cross-border effort to standardise data to the Observational Medical Outcomes Partnership (OMOP) common data model". https://www.hdruk.ac.uk/news/hdr-uk-with-ehden-to-announce-a-total-of-22-data-partners-have-been-selected-in-the-7th-and-last-ehden-data-partner-call

[8] US FDA. 2021. "Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products: Guidance for Industry". https://www.fda.gov/media/152503/download

Saiam Ahmed: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Sophia Apostolidou: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Judith M. Bliss: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Claire Bloomfield: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Rebecca Cannings-John: Investigation, Methodology, Writing – original draft, Writing – review & editing.

James Carpenter: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Tim Clayton: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Madeleine Clout: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Rebecca Cosgriff: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Amanda J. Farrin: Methodology, Writing – original draft, Writing – review & editing, Investigation.

Aleksandra Gentry-Maharaj: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Duncan C. Gilbert: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Charlie Harper: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Nicholas D. James: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Ruth E. Langley: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Sarah Lessels: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Fiona Lugg-Widger: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Isla S. Mackenzie: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Marion Mafham: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Usha Menon: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Harriet Mintz: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Heather Pinches: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Michael Robling: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Alexandra Wright-Hughes: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Victoria Yorke-Edwards: Writing – original draft, Writing – review & editing, Investigation, Methodology.

Sharon B. Love: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. CB is now employed at Insitro, South San Francisco, CA. Insitro had no involvement in the design or implementation of the work presented here.

Data availability

No data was used for the research described in the article.

References

[1] M.R. Sydes, Y. Barbachano, L. Bowman, T. Denwood, A. Farmer, S. Garfield-Birkbeck, et al., Realising the full potential of data-enabled trials in the UK: a call for action, BMJ Open 11 (6) (2021) e043906.

[2] R. Haynes, P. Harden, P. Judge, L. Blackwell, J. Emberson, et al., 3C Study Collaborative Group, Alemtuzumab-based induction treatment versus basiliximab-based induction treatment in kidney transplantation (the 3C study): a randomised trial, Lancet 384 (9955) (2014) 1684–1690.

[3] RECOVERY Collaborative Group, Higher dose corticosteroids in patients admitted to hospital with COVID-19 who are hypoxic but not requiring ventilatory support (RECOVERY): a randomised, controlled, open-label, platform trial, Lancet 401 (10387) (2023) 1499–1507.

[4] S. Lensen, A. Macnair, S.B. Love, V. Yorke-Edwards, N.M. Noor, M. Martyn, et al., Access to routinely collected health data for clinical trials - review of successful data requests to UK registries, Trials 21 (1) (2020) 398.

[5] A. Macnair, S.B. Love, M.L. Murray, D.C. Gilbert, M.K.B. Parmar, T. Denwood, et al., Accessing routinely collected health data to improve clinical trials: recent experience of access, Trials 22 (1) (2021) 340.

[6] M.L. Murray, S.B. Love, J.R. Carpenter, S. Hartley, M.J. Landray, M. Mafham, et al., Data provenance and integrity of health-care systems data for clinical trials, Lancet Digit. Health 4 (8) (2022) e567–e568.

[7] M.L. Murray, H. Pinches, M. Mafham, S. Hartley, J.R. Carpenter, M. Landray, et al., Use of NHS Digital Datasets as Trial Data in the UK: A Position Paper (2.0), Zenodo, 2022. https://doi.org/10.5281/zenodo.6047155.

[8] A.D.N. Williams, G. Davies, A.J. Farrin, M. Mafham, M. Robling, M.R. Sydes, F. V. Lugg-Widger, A DELPHI study priority setting the remaining challenges for the use of routinely collected data in trials: COMORANT-UK, Trials 24 (1) (2023) 243.

[9] H.P. Mintz, A.R.S. Dosanjh, H. Parsons, M. Sydes, R.T. Bryan, N.D. James, P. Patel, Making administrative healthcare systems clinical data the future of clinical trials: lessons from BladderPath, BMJ Oncol. 2 (1) (2023).

[10] J. MacArthur, L. Morrice, C. Sudlow, M. Sydes. BHF DSC Data-Enabled Trials Survey Report, 2021, https://doi.org/10.5281/zenodo.5079178.

[11] J. MacArthur, M. Sydes, B.D.S. Centre, How to facilitate the use of healthcare systems data in cardiovascular clinical trials, Report (2022), https://doi.org/10.5281/zenodo.6384888.

[12] R. Iyer, A. Gentry-Maharaj, A. Nordin, R. Liston, M. Burnell, N. Das, et al., Patient-reporting improves estimates of postoperative complication rates: a prospective cohort study in gynaecological oncology, Br. J. Cancer 109 (3) (2013) 623–632.

[13] F. Adshead, R. Al-Shahi Salman, S. Aumonier, M. Collins, K. Hood, C. McNamara, et al., A strategy to reduce the carbon footprint of clinical trials, Lancet 398 (10297) (2021) 281–282.

[14] S. Denaxas, J. MacArthur, S. Lessels, M.R. Sydes, J. Farrell, J. Nolan, et al.. Ensuring Phenotyping Algorithms Using National Electronic Health Records are FAIR: Meeting the Needs of the Cardiometabolic Research Community, Zenodo, 2023, https://doi.org/10.5281/zenodo.10209724.

[15] S. Ahmed, M.R. Sydes, S.B. Love, N.D. James, J. Carpenter, PS8C-01: Agreement and Completeness of Routine Versus Trial-Specific Patient Outcome Data , A Systematic Review, ICTMC2022, Zenodo 109 (3) (2023) https://doi.org/10.5281/zenodo.7741866.

[16] S. Ahmed, M. Sydes, S. Love, J. Carpenter, Assessing the Agreement of Routinely-Collected Health Data and Case Record Form Data in Randomised Controlled Trials: A Systematic Review (CRD42020186048), 2020.

[17] S.B. Love, A. Kilanowski, V. Yorke-Edwards, O. Old, H. Barr, C. Stokes, et al., Use of routinely collected health data in randomised clinical trials: comparison of trial-specific death data in the BOSS trial with NHS digital data, Trials 22 (1) (2021) 654.

[18] C. Harper, M. Mafham, W. Herrington, N. Staplin, W. Stevens, K. Wallendszus, et al., Comparison of the accuracy and completeness of Records of Serious Vascular Events in routinely collected Data vs clinical trial-adjudicated direct follow-up Data in the UK: secondary analysis of the ASCEND randomized clinical trial, JAMA Netw. Open 4 (12) (2021) e2139748.

[19] A. Gentry-Maharaj, E.O. Fourkala, M. Burnell, A. Ryan, S. Apostolidou, M. Habib, et al., Concordance of National Cancer Registration with self-reported breast, bowel and lung cancer in England and Wales: a prospective cohort study within the UK collaborative trial of ovarian Cancer screening, Br. J. Cancer 109 (11) (2013) 2875–2879.

[20] D.S. Thomas, A. Gentry-Maharaj, A. Ryan, E.O. Fourkala, S. Apostolidou, M. Burnell, et al., Colorectal cancer ascertainment through cancer registries, hospital episode statistics, and self-reporting compared to confirmation by clinician: a cohort study nested within the UK collaborative trial of ovarian Cancer screening (UKCTOCS), Cancer Epidemiol. 58 (2019) 167–174.

[21] A. Macnair, M. Nankivell, M.L. Murray, S.D. Rosen, S. Appleyard, M.R. Sydes, et al., Healthcare systems data in the context of clinical trials - a comparison of cardiovascular data from a clinical trial dataset with routinely collected data, Contemp. Clin. Trials 128 (2023) 107162.

[22] I.J. Jacobs, U. Menon, A. Ryan, A. Gentry-Maharaj, M. Burnell, J.K. Kalsi, et al., Ovarian cancer screening and mortality in the UK collaborative trial of ovarian Cancer screening (UKCTOCS): a randomised controlled trial, Lancet 387 (10022) (2016) 945–956.

[23] M. Robling, M.J. Bekkers, K. Bell, C.C. Butler, R. Cannings-John, S. Channon, et al., Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (building blocks): a pragmatic randomised controlled trial, Lancet 387 (10014) (2016) 146–155.

[24] K. Harron, Data linkage in medical research, BMJ Med. 1 (1) (2022) e000087.

[25] A. Wright-Hughes, E. Graham, D. Cottrell, A. Farrin, Routine hospital data - is it good enough for trials? An example using England's hospital episode statistics in the SHIFT trial of family therapy vs. treatment as usual in adolescents following self-harm, Clin. Trials 15 (2) (2018) 197–206.
[26] BHF Data Science Centre, M.R. Sydes, C. Sudlow, S. Denaxas, T. Chico, S. Lessels, et al., Standardising Clinical Outcome Measures in Routinely-Collected Electronic Healthcare Systems Data (SCORE-CVD) Initial Report, Zenodo, 2023 (BHF DSC reports). https://doi.org/10.5281/zenodo.8169092.
[27] B. Gordon, J. Barrett, C. Fennessy, C. Cake, A. Milward, C. Irwin, et al., Development of a data utility framework to support effective health data curation, BMJ Health Care Inform. 28 (1) (2021).
[28] Choodari-Oskooei B, Love S, Sydes M, White I, Parmar M. PS4B-04: Total or control events: choosing approach for timing of trial analyses. Zenodo (ICTMC2022 abstracts). 2023. https://doi.org/10.5281/zenodo.7741866.