

# Using healthcare systems data for outcomes in clinical trials: Issues to consider at the design stage

**Alice-Maria Toader** (✉ [Sgatoade@liverpool.ac.uk](mailto:Sgatoade@liverpool.ac.uk))

University of Liverpool <https://orcid.org/0000-0002-3801-3486>

**Marion K Campbell**

University of Aberdeen Health Services Research Unit

**Jennifer K Quint**

Imperial College London School of Public Health

**Michael Robling**

Cardiff University Centre for Trials Research

**Matthew R Sydes**

MRC Clinical Trials Unit at UCL: Medical Research Council Clinical Trials Unit at University College London

**Joanna Thorn**

University of Bristol Population Health Sciences

**Alexandra Wright-Hughes**

University of Leeds Clinical Trials Research Unit

**Ly-Mee Yu**

University of Oxford Department of Primary Care Health Sciences: University of Oxford Nuffield Department of Primary Care Health Sciences

**Tom E.F. Abbott**

William Harvey Research Institute: Queen Mary University of London William Harvey Research Institute

**Simon Bond**

Cambridge Clinical Trials Unit

**Fergus J Caskey**

University of Bristol Population Health Sciences

**Madeleine Clout**

University of Bristol Population Health Sciences

**Michelle Collinson**

University of Leeds Clinical Trials Research Unit

**Bethan Copsey**

University of Leeds Clinical Trials Research Unit

**Gwyneth Davies**

UCL GOS Institute of Child Health: University College London Great Ormond Street Institute of Child Health

**Timothy Driscoll**

Swansea University

**Carrol Gamble**

University of Liverpool

**Xavier L Griffin**

Barts and The London School of Medicine and Dentistry Institute of Dentistry: Queen Mary University of London Institute of Dentistry

**Thomas Hamborg**

Queen Mary University of London Wolfson Institute of Population Health

**Jessica Harris**

University of Bristol

**David A Harrison**

ICNARC: Intensive Care National Audit and Research Centre

**Deena Harji**

University of Leeds

**Emily J Henderson**

University of Bristol Population Health Sciences

**Pip Logan**

University of Nottingham

**Sharon B Love**

MRC CTU at UCL: Medical Research Council Clinical Trials Unit at University College London

**Laura A Magee**

King's College London

**Alastair O'Brien**

UCL Institute for Liver & Digestive Health: University College London Institute for Liver and Digestive Health

**Maria Pufulete**

University of Bristol Medical School

**Padmanabhan Ramnarayan**

Imperial College London

**Athanasios Saratzis**

University of Leicester

**Jo Smith**

Keele University

**Ivonne Solis-Trapala**

Keele University

**Clive Stubbs**

BCTU: University of Birmingham Clinical Trials Unit

**Amanda Farrin**

University of Leeds

**Paula Williamson**

University of Liverpool School of Health Sciences


---

## Research Article

**Keywords:** Healthcare systems data, outcomes, clinical trials, routinely collected data, data validity, registries

**Posted Date:** October 24th, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3351132/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Trials on January 29th, 2024. See the published version at <https://doi.org/10.1186/s13063-024-07926-z>.

# Abstract

## Background

Healthcare system data (HSD) are increasingly used in clinical trials, augmenting or replacing traditional methods of collecting outcome data. The PRIMORANT study set out to determine when HSD are of sufficient quality and utility to replace bespoke outcome data collection, a methodological question prioritised by the clinical trials community.

## Methods

The PRIMORANT study had three phases. First, an initial workshop was held to scope the issues faced by trialists when considering whether to use HSDs for trial outcomes. Second, a consultation exercise was undertaken with clinical trials unit (CTU) staff, trialists, methodologists, clinicians, funding panels and data providers. Third, a final discussion workshop was held, at which the results of the consultation were fed back, case studies presented, and issues considered in small breakout groups.

## Results

Key topics included in the consultation process were validity of outcome data, timeliness of data capture, internal pilots, data-sharing, practical issues, and decision-making. A majority of respondents (n = 78, 95%) considered the development of guidance for trialists to be feasible. Guidance was developed following the discussion workshop, for the five broad areas of terminology, feasibility, internal pilots, onward data sharing, and data archiving.

## Conclusions

We provide guidance to inform decisions about whether or not to use HSDs for outcomes, and if so, to assist trialists in working with registries and other HSD-providers to improve the design and delivery of trials.

## Background

Healthcare systems data (HSD) refers to health care information, gathered from providers including primary and secondary care, for the delivery of healthcare but not purposely designed for its use in research. Such data are sometimes referred to as routinely-collected health data (RCHD). These data may come from administrative, surveillance, registry or audit systems, and may facilitate research, with potential benefits such as a reduction in the burden on patients and health professionals of collecting research-specific data (1).

Between 2013 and 2018, less than 5% of all UK RCTs were granted HSD access from registries (2). As of 2019, 47% of the 216 in-progress clinical trials in the NIHR Journals library planned to use HSD (3). Recent estimates show that in 2022, this percentage has increased to 62%.

Methodological research priorities for use of HSD within trials have previously been established through a Delphi study (4). Stakeholders, including trialists, research funders, regulators, data-providers and the public, identified 40 unique research questions that were ranked in importance via a survey and a virtual consensus meeting. The top seven priorities, in order, relate to: data collection method; outcome selection; communication with participants; regulatory approvals; data access and receipt; data quality; and data analysis. A summary is available on the COMORANT study website (5), with full details published. (4)

The PRIMORANT study aimed to explore two of the COMORANT methodological research questions by 1) addressing a “real” challenge through methodology work and 2) addressing a “perceived” challenge through training. This paper describes the work undertaken to address the first of these and focuses on the COMORANT priority question relating to outcome selection at the trial design stage: *‘How should the trials community decide when routinely collected data for outcomes are of sufficient quality and utility to replace bespoke data collection?’*. The aim was to identify issues to be considered before the decision to use HSD for outcome data in a clinical trial is finalised.

## Methods

### i. Initial workshop

An initial workshop was hosted online on 28th September 2022, and comprised three presentations followed by breakout group discussions. Invitations were distributed in the UK among the COMORANT, Trial Methodology Research Partnership - Health Informatics Working Group (TMRP HI WG), NIHR Methodology Incubator HI subgroup, UK Clinical Research collaboration – Clinical Trials Unit (UKCRC CTU) Network Statistics group, and SPIRIT-Routine lists; 25 people attended. The presentations covered the use of HSD in trials, SPIRIT Extension for trials using routine data and terminology and data integrity. During the breakout groups, it was proposed to discuss, in the context of case studies, how the decision to use HSD was made, alongside lessons learned and relevant guidance. The aim was to identify existing relevant guidance on using HSD data for clinical trial outcomes and to explore areas to consider when using or deciding whether to use HSD.

### i. Consultation exercise

Based on the seven topics identified from the initial workshop, crosschecked for consistency against the existing Medicines and Healthcare products Regulatory Agency (MHRA) guidance (6) 16 questions were developed for the consultation.

The JISC Online Survey tool (7) was used to create, host, and distribute the consultation. All questions were optional, allowing the responder to engage with topics that aligned with their expertise. All

responses were provided anonymously. A copy of the consultation questions can be found in **Additional File 1**.

Between December 2022 and January 2023, the consultation was sent to over 200 individuals worldwide, including CTU staff, trialists, methodologists, clinicians, funding panels and data providers. Consultation recipients were identified and selected from initial workshop attendees, HTA funding committee members, Chief Investigators (CIs) of RCTs using HSD funded by NIHR and attendees of the SPIRIT Extension report meeting. Recipients were encouraged to distribute the consultation to others with relevant expertise.

## ii. Discussion workshop

The results from the consultation were used to identify issues to be discussed at a face-to-face workshop in March 2023. Respondents to the consultation were asked to note their interest in attending this second workshop, and whether they could present a case study. Findings from the first two stages were summarised descriptively, with free text responses grouped into topics (initially A-MT and verified by PRW and AF), and presented during the discussion workshop.

Case study presentations were selected by the study team from those offered, based on the range of issues they highlighted and ensuring a diversity of trial designs, trial populations, trial outcomes and data sources. The speakers were asked to prepare a short PowerPoint presentation describing the case study, and the issues related to HSD, alongside their recommendations.

The second part of the workshop focused on the list of issues to consider that arose from the consultation. The participants were divided into six break-out groups and discussed the completeness of the list and generated recommendations for trial teams about points to consider when deciding whether to use HSD for trial outcomes.

# Results

## i. Initial workshop

The initial workshop was attended by 26 participants. Key topics identified to include in the subsequent consultation process were validity of outcome data, timeliness of data capture, internal pilots, data sharing, practical issues, and decision-making. The conclusion from the meeting was that the development of practical guidance to be used when considering the use of HSD for outcomes would be helpful.

## ii. Consultation exercise

Responses were received from 82 individuals invited. A majority of responders (n = 70) considered that evidence from previous feasibility studies would be sufficient evidence to confirm the validity of outcome data from HSD. Most responders (n = 72, 89%) agreed that a Standard Operating Procedure (SOP) for data providers for the handling and resolution of discrepancies in the HSD would be helpful; but fifty

responders (61%) considered that such an SOP would not be feasible. In contrast, a template SOP or guidance for trialists was considered feasible by 78 (95%) responders. Further results are shown in **Additional File 2**.

Many responders (n = 64) suggested several elements specifically related to the use of HSD for outcomes to be appropriate for inclusion in trial progression criteria. These included: data availability and completeness; time to access the data; data quality; linkage; the potential for any bias or confounding. Details to be made public by trial teams included: cost of acquiring the data; time needed for each step of the data acquisition and linkage process; and information regarding the quality and validity of data.

Issues to be considered when deciding between using HSD, more traditional data collection through bespoke trial CRFs, or a hybrid approach for collecting outcome data were taken as the starting point for presentations and discussions at the subsequent workshop.

### iii. Discussion workshop

Invitations were issued to 45 individuals, including the PRIMORANT team, with 35 (78%) able to attend in person. Six case studies and further areas of investigation were presented in 7 talks, and these are summarised in **Additional File 3**. The key messages across the studies were as follows.

- The need for clear outcome definitions and use of validated code lists.
- Feasibility assessments can provide assurance about the validity and availability of HSD for outcomes.
- HSD can improve outcome data collection, but challenges include classification; subsequent changes to datasets and linkage; retention and archiving requirements of the clinical trial versus routine data provider; specialist knowledge and resource to analyse hospital episode statistics (HES) data; and adapting traditional data management processes to handle HSD.
- Assessing the utility of HSD against medical records, or through data linkage to other sources, is important in order to understand whether HSD is appropriate for both clinical and health economic outcomes in an individual trial.
- The impact of HSD availability on timing for trial reporting and interim analysis requirements should be considered.
- The volume and types of incomplete data within HSD should be assessed.
- The potential for delay in accessing HSD should be considered against trial timelines.
- Work on demonstrating the integrity and provenance of data is on-going through collaboration between NHS England (formerly NHS Digital) and HDR UK.

### iv. Feedback from breakout group discussions

**Box 1** provides a list of considerations for trialists at the design stage. The content was iteratively discussed and developed during the workshop break-out groups, and subsequently finalised by email.

# Discussion

To address an agreed methodological research evidence gap prioritised by the research community, we have systematically developed a comprehensive and easy-to-use list of issues to consider when deciding whether to use HSD for trial outcomes. Discussions emphasized the need for careful planning/exploration of the datasets before making the decision. Discussions with funders around phased approaches and contingency planning are recommended.

The list complements other resources available or planned for trialists using HSD for trial outcomes, including MHRA guidance, CTTI guidance (8), and the HDR UK 'Route Map' described above. Forthcoming SPIRIT-Routine guidance is anticipated to highlight some of these issues to be considered in trial protocols (9). Consideration of the issues described here will also allow trial teams to meet the reporting standards of the CONSORT-Routine guideline (10).

Several areas of potential concern, which are likely to be more commonly encountered, were discussed:

1. Finding data specialists with experience with HSD can be difficult. If unavailable, identifying appropriate training and funding for this should be built into grant applications, also recognising the increased risk on research delivery and time required.
2. Sample datasets are not always available. Early discussion with HSD controllers may be useful, both for them and for the users.
3. There are examples of registry trials which supplement core registry data with add-on modules which collect trial-specific outcomes. If this approach is used, data management processes require careful consideration in advance to ensure that the integrity of registry data are not compromised (11).
4. When choosing outcome measures, the potential limitations of choosing only those where HSD exists, which may exclude some agreed to be of core importance, e.g. in core outcome sets (12), needs to be considered.

Several areas were identified where further work would be helpful.

1. Validation studies to demonstrate HSD quality are needed in terms of integrity and provenance (Murray 2022) and utility (under review). One question raised was whether data providers should be responsible for providing information about the validity of the data they provide. Expansion of the work to demonstrate integrity and provenance of data (13, 14) to cover more providers will be useful.
2. Examples of helpful discussions with research funders were given, whereby phased feasibility studies to assess uncertainties related to HSD were agreed. A point for further discussion with funders is whether a different costing model should be applied to access data for feasibility and pilot studies. It was considered helpful to explore this concept with funders and HSD providers, to see how it might be potentially supported.

Strengths of this work include the range of stakeholders engaged, and the breadth of examples and case studies discussed. The responses to the consultation allowed exploration of a range of potential areas

for consideration that mapped onto issues across the lifecycle of the trial and covered topic areas that were likely to be relevant to the range of disciplines and roles involved in trial design. Limited representation of funders and data providers at the discussion workshop is recognised as a limitation, however planned dissemination activities will be aimed at greater engagement, with potential for future revisions to the list of issues to consider. The main focus of this work was on UK practice and datasets, although some of the findings may be considered to be generalisable outside the UK.

The focus of the PRIMORANT study was on issues to consider during the design phase of a clinical trial. It is important to note however that there are other aspects of conduct and reporting in relation to using HSD for trial outcomes. For example, algorithms used within trials should be well-documented to enhance reproducibility. Code list and data fields provided may change over time, so algorithms will need to change, and those changes will also need to be documented. If data are sourced from multiple providers, consistency of coding across the datasets should be checked and reconciliation clearly documented. Code lists and/or algorithms should be made publicly available to improve efficiency for future researchers, for example in the HDR UK phenotype library.

## Conclusion

In summary, the issues identified here should strengthen the decision-making process for trialists when considering the use of HSD for trial outcomes. The work should also inform discussions with funders to build in mitigation (e.g. include an option to supplement with data directly from participants or sites) and allow for additional costs that could be incurred or unanticipated workarounds required (e.g. for changes in legislation, delays in data release, periodic renewal of data sharing agreements), as well as discussions with HSD-providers about how to improve the design and delivery of trials using HSD.

## Abbreviations

A&E Accidents and Emergency

CAG Confidentiality Advisory Group

COPD Chronic Obstructive Pulmonary Disease

CTU Clinical Trials Unit

DfE Department for Education

DMC Data Monitoring Committee

HER Electronic Health Records

EMIS Egton Medical Information Systems

HDR Health Data Research

HES Hospital Episode Statistics

HRA Health Research Authority

HSD Healthcare Systems Data

ICD-10 International Classification of Diseases, 10th Revision

MHRA Medicines and Healthcare products Regulatory Agency

NHS National Health Service

NIHR National Institute for Health and Care Research

ONS Office for National Statistics

PROMs Patient-Reported Outcome Measures

PSA Prostate-Specific Antigen

QoL Quality of Life

RCHD Routinely Collected Health Data

RCT Randomised Controlled Trial

REC Research Ethics Committee

SOP Standard Operating Procedures

TARN Trauma Audit and Research Network

UKPDS UK Prospective Diabetes Study

## Declarations

### **Ethics approval and consent to participate**

Not applicable

### **Consent for publication**

Not applicable

## **Availability of data and materials**

Not applicable

## **Competing interests**

GD reports speaker honoraria from Chiesi Ltd and Vertex Pharmaceuticals outside the submitted work.

PR reports consultancy fees from Vyaire Medical and Sanofi outside the submitted work.

EJH reports honoraria and travel support from Kyowa Kirin; Abbvie; Ever; Bial, The Neurology Academy and CME Institute outside the submitted work.

JKQ has received grants from MRC, HDR UK, GSK, BI, asthma+lung UK, and AZ and personal fees for advisory board participation, consultancy or speaking fees from GlaxoSmithKline, Evidera, Chiesi, AstraZeneca, Insmed.

MKC was a member of the CONSORT-ROUTINE group. The Health Services Research Unit, where MKC works, receives core funding from the Scottish Government Health Directorates.

SBL reports no conflicts of interest.

MRSy reports speaker fees at clinical trial statistics training meeting for clinicians (no discussion of particular drugs) from Lilly Oncology; Speaker fees at clinical trial statistics training meeting for clinicians (no discussion of particular drugs) from Janssen; and Educational video on clinical trial statistics (no discussion of particular drugs) from Eisai.

MClout reports no conflicts of interest.

MC reports no conflicts of interest.

JThorn reports no conflicts of interest.

MR reports no conflicts of interest.

JH reports no conflicts of interest.

TJD reports no conflicts of interest.

AJF reports no conflicts of interest.

None of the other authors reported any conflicts of interest.

## **Funding statement**

This work was funded through an HDR UK award (TF2022.31 PRIMORANT). Alice-Maria Toader is funded by MRC Trials Methodology Research Partnership (TMRP) Doctoral Training Partnership (DTP). Grant Number MR/W006049/1. GD is funded by a UKRI FLF [MR/T041285/1]. MRS and SBL are supported by the Medical Research Council (MRC, part of UKRI) [grant number MC\_UU\_00004/08].

## Authors contributions

AF and PRW conceived the idea for the project. A-MT, AF and PRW organised the two meetings. A-MT conducted the consultation and analysed the results. A-MT wrote the first draft in collaboration with AF and PRW.

All authors commented on and approved the final manuscript.

## Acknowledgments

We are grateful to those completing the consultation. Fiona Lugg-Widger was a co-applicant for the PRIMORANT study, leading for the second prioritised project and led the COMORANT study that identified research priorities in using routine data.

## References

1. Sydes MR, Barbachano Y, Bowman L, Denwood T, Farmer A, Garfield-Birkbeck S, et al. Realising the full potential of data-enabled trials in the UK: a call for action. *BMJ open*. 2021;11(6):e043906.
2. Lensen S, Macnair A, Love SB, Yorke-Edwards V, Noor NM, Martyn M, et al. Access to routinely collected health data for clinical trials—review of successful data requests to UK registries. *Trials*. 2020;21(1):1-11.
3. 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. *F1000Research*. 2021;9:323.
4. Williams AD, Davies G, Farrin AJ, Mafham M, Robling M, Sydes MR, et al. A DELPHI study priority setting the remaining challenges for the use of routinely collected data in trials: COMORANT-UK. *Trials*. 2023;24(1):1-8.
5. COMORANT-UK Consensus on methodological opportunities for routine data and trials. [Available from: <https://www.cardiff.ac.uk/centre-for-trials-research/research/studies-and-trials/view/comorant-uk>.
6. MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions 16 December 2021 [Available from: <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>.
7. Jisc - Online surveys [Available from: <https://beta.jisc.ac.uk/online-surveys>.
8. (CTTI) CTTI. Recommendations for Registry of Clinical Trials June 2021 [Available from: [https://ctti-clinicaltrials.org/wp-content/uploads/2021/06/CTTI\\_Registry\\_Trials\\_Recs.pdf](https://ctti-clinicaltrials.org/wp-content/uploads/2021/06/CTTI_Registry_Trials_Recs.pdf).

9. McCarthy M, O'Keeffe L, Williamson PR, Sydes MR, Farrin A, Lugg-Widger F, et al. A study protocol for the development of a SPIRIT extension for trials conducted using cohorts and routinely collected data (SPIRIT-ROUTINE). HRB open research. 2021;4.
10. Kwakkenbos L, Imran M, McCall SJ, McCord KA, Fröbert O, Hemkens LG, et al. CONSORT extension for the reporting of randomised controlled trials conducted using cohorts and routinely collected data (CONSORT-ROUTINE): checklist with explanation and elaboration. *bmj*. 2021;373.
11. Brohi K. The trials of being a national trauma registry. *Emergency Medicine Journal*. 2015;32(12):909-10.
12. Core Outcome Measures in Effectiveness Trials (COMET Initiative) [Available from: <https://www.comet-initiative.org/>].
13. Murray ML, Love SB, Carpenter JR, Hartley S, Landray MJ, Mafham M, et al. Data provenance and integrity of health-care systems data for clinical trials. *The Lancet Digital Health*. 2022;4(8):e567-e8.
14. Murray ML, Pinches H, Mafham M, Hartley S, Carpenter J, Landray MJ, et al. Use of NHS Digital datasets as trial data in the UK: a position paper. 2022.

## Box 1: Issues to consider

Issues to be considered before the decision to use HSD for collecting outcomes in an RCT is finalised are described here. The aim is to help the trial team make an informed judgment based on an understanding of the suitability of HSD for outcome data in the context of the specific clinical trial, and to build in mitigation, for example including the option to supplement with data directly from participants or sites. Working through the items below may highlight ways trialists can work with HSD providers to improve how such trials are designed and delivered.

It is recommended that trialists consider additional costs that could be incurred or unanticipated workarounds required during the trial, such as changes in legislation, delays in data release and periodic renewal of data sharing agreements. Strategies to address these uncertainties might include building in a contingency fund or agreeing a phased project plan with the funder; researchers are encouraged to risk assess a broad range of possible scenarios and consider potential mitigation strategies.

### (1) Terminology

Be aware that terminology within data access applications will likely differ between providers; seek clarification or examples from the provider if available.

Ensure awareness of how terms can be interpreted by the different individuals involved across the multiple organisations.

### (2) Feasibility

#### 2.1 Team

Where possible, include trial operations professionals, data and health specialists with experience of completing data access forms and analysing the data from the provider/s for the relevant health research question, in the trial team. This ideally needs to include individuals who (1) understand the data, its structure, its interpretation, and its quality; (2) understand how and when the data are collected at source; (3) have the skills to handle the data when they are provided; and (4) will undertake the statistical and health economic analysis. Where knowledge gaps are identified, look to include funding for training and development activities.

## **2.2 Data**

Trialists should be aware of how HSD are entered, coded, the QA processes, how data are validated at the point of upload and then transferred. Data providers should be approached to provide this information. Trialists should justify the use of healthcare systems datasets in their Trial Master File. A suggested template form is in the Supplement of Murray (2022) Zenodo: <https://zenodo.org/record/6047938>

### **a. Does the HSD include what the trial needs?**

Using the data provider's data dictionary, where available, establish which outcomes are collected "routinely", and ascertain any cost of data provision and the data provider timelines for data verification/release. Consider the need for repeated data releases and costs relating to data retention. Discuss the process for data linkage if linking to a trial cohort and/or multiple data sources are sought. If time and resources permit, interrogate the dataset to understand any limitations prior to the decision to use HSD. The dataset may cover only a subset of the outcomes deemed relevant to the trial question. If this is the case, consider how the other outcome data will be collected, or whether the benefit of using a single approach to data collection outweighs the value of collecting data on all outcomes from multiple sources. For a registry-based trial, discuss whether the registry team could adapt or supplement routine HSD collection to meet the trial's needs without compromising the integrity of the registry.

HSD may be appropriate for aspects of reporting safety data depending on the risk profile of the clinical trial. This should be considered during trial design and clearly defined in the protocol. This is likely to be appropriate in low-risk trials where adverse events are not informing the emergent safety profile of the trial; timeliness of data provision should be considered in relation to safety monitoring plans.

Establish whether any precedent, or evidence of public support for accessing these data for research, exists, or alternatively whether issues have arisen previously. Consider trial participants' needs for understanding of the use of their HSD for outcomes in research and how that may vary according to study populations.

### **b. Data quality assurance**

Establish whether the provider can provide information regarding data provenance, integrity, and completeness. Understand the timeliness of the collection of the data held by the provider, for example whether there is a lag between site data collection and entry into the provider system, or whether data is

only released at a certain time of year. In addition, understand how the provider receives and processes the data, and how changes in processing and coding are handled and communicated.

Consider what is known, from previous literature, about the validity and completeness of the outcome data, which may include national audit reports. Assess whether it is realistic to be able to provide the funder with an accurate idea of HSD data quality at application, or whether it is possible to build in approaches to examine the uncertainty during the trial.

### **c. Time**

Ask the provider how long it will take from the point of request and then from the point of approval to supply a specified dataset to the trial team; determine if the contract includes binding timelines and decide what is an acceptable delay for delivery of data for the first occasion and subsequent deliveries. Establish whether this time will reduce if datasets are requested on multiple occasions during the trial. Consider this in relation to whether any interim analyses are planned or when using HSD for monitoring safety outcomes.

### **d. Algorithms for deriving outcomes**

Explore whether a validated algorithm for deriving outcomes from HSD exists. If not, consider whether to include time to develop and test the proposed algorithm, within a utility comparison.

#### **a. Considerations around missing data**

Be aware of the timing of data entry processes into the HSD resource by clinical teams and data entry clerks, and their subsequent availability or missingness, which may also vary across sites. For example, within registries outcomes may be entered on an annual basis or annual reviews may be delayed. Similarly, be aware of how long the data may take from local collection into a national or collated set, and how long it takes for the latter to be released.

Discuss whether it may be possible to go back to participating sites to collect missing data. Otherwise consider imputation from other available variables, or other HSD datasets, with the collection of extra variables to maximise the effectiveness of the imputation method. This may be where a contingency fund for unanticipated workarounds would be helpful.

#### **b. Consideration of potential reporting errors/discrepancies**

Discuss the mechanism and opportunity for resolution of discrepancies with the provider. Ask the provider whether they have any guidance on the range of possible solutions based on their experience (e.g., rules of precedence, windows for 'same dates', impossible events). Always cost for managing data queries – this could be part of contingency management.

### **Preparation of trial dataset**

A discussion with the provider about whether raw data or analysis-ready data will be provided may be appropriate. For example, it may be useful to consider whether the trial team will need to do additional analyses over the primary analysis. If so, the trial team may consider that raw data may be more appropriate. However, if the trial team has limited statistical support or only need one or two defined analyses, analysis-ready data might be more appropriate. Cost and time may also be a factor – access to analysis-ready data could be more costly or take longer to receive. Additional considerations might be the ability to verify the derivation of analysis-ready data undertaken by the third party. In this case raw data might be more appropriate, where the trial team can have complete control over the analysis steps provided there is local statistical expertise to do this.

### **(3) Internal pilot**

For an internal pilot to be undertaken to determine how use of HSD compares to collecting outcome data traditionally, for example in terms of sufficiency, timeliness, completeness, cost-effectiveness, the trial team needs to consider whether setting up the trial using both approaches can be justified in terms of cost and complexity, e.g., by providing added value for the health area more widely than the individual trial. If an internal pilot to assess this question is felt to be valuable and feasible, due consideration should be given to the progression criteria to be applied to the aspects related to the use of HSD.

### **(4) Onward data sharing**

In principle, onward data sharing can facilitate further research and extend the efficiency gains from using HSD. Discuss the funder's requirements for onward data sharing and whether the provider can approve this.

Onward sharing may not be permissible or subsequent access may not be straightforward (e.g., if access through a trusted environment is needed). Ensure these issues are considered in the data sharing agreement/contract as well as any resources involved.

It is important to consider prospectively who (in the broadest sense, e.g., trial oversight committees, trial team, industry partners, future meta-analysts) will need to see HSD, as raw or aggregated data. The legal, ethical and governance responsibilities must be explored in advance within appropriate timeframes. There may also be implications for consent forms for the trial, allowing further use of data past the initial trial.

### **(5) Data destruction and archiving**

Discuss any regulatory requirements for the archiving period with the data provider, ensuring archiving agreements are compliant with the clinical trials regulations. Discuss any costs associated with holding data for an archiving period, and permissions to retain anonymised data, in original or derived format, beyond the archive period.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalFile1Consultation.pdf](#)
- [AdditionalFile2Consultationresults.docx](#)
- [AdditionalFile3Casestudies.docx](#)