BRIEF REPORT

REVISED Use of routinely collected data in a UK cohort of publicly funded randomised clinical trials [version 2; peer review: 2 approved]

Andrew J. McKay [iD][1], Ashley P. Jones[1], Carrol L. Gamble[1,2], Andrew J. Farmer[3], Paula R. Williamson[2]

[1]Liverpool Clinical Trials Centre, University of Liverpool, a member of Liverpool Health Partners, Liverpool, UK
[2]MRC North West Hub for Trials Methodology Research, Department of Biostatistics, University of Liverpool, Liverpool, UK
[3]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

Abstract
Routinely collected data about health in medical records, registries and hospital activity statistics is now routinely collected in an electronic form. The extent to which such sources of data are now being routinely accessed to deliver efficient clinical trials, is unclear. The aim of this study was to ascertain current practice amongst a United Kingdom (UK) cohort of recently funded and ongoing randomised controlled trials (RCTs) in relation to sources and use of routinely collected outcome data.

Recently funded and ongoing RCTs were identified for inclusion by searching the National Institute for Health Research journals library. Trials that have a protocol available were assessed for inclusion and those that use or plan to use routinely collected health data (RCHD) for at least one outcome were included. RCHD sources and outcome information were extracted.

Of 216 RCTs, 102 (47%) planned to use RCHD. A RCHD source was the sole source of outcome data for at least one outcome in 46 (45%) of those 102 trials. The most frequent sources are Hospital Episode Statistics (HES) and Office for National Statistics (ONS), with the most common outcome data to be extracted being on mortality, hospital admission, and health service resource use.

Our study has found that around half of publicly funded trials in a UK cohort (NIHR HTA funded trials that had a protocol available) plan to collect outcome data from routinely collected data sources. This is much higher than the figure of 8% found in a cohort of 189 RCTs published since 2000, the majority of which were carried out in North America (McCord *et al*., 2019).

Keywords

Electronic Health Records, Data linkage, EHR, NIHR HTA, Randomised Clinical Trial, Randomised Controlled Trial, RCT, Registry, Routinely collected data, Routinely collected health data, RCHD

Corresponding author: Paula R. Williamson (p.r.williamson@liverpool.ac.uk)

Author roles: McKay AJ: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Jones AP**: Conceptualization, Investigation, Methodology, Project Administration, Resources, Visualization, Writing – Review & Editing; **Gamble CL**: Conceptualization, Investigation, Methodology, Project Administration, Resources, Visualization, Writing – Review & Editing; **Farmer AJ**: Conceptualization, Investigation, Methodology, Project Administration, Resources, Visualization, Writing – Review & Editing; **Williamson PR**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

REVISED **Amendments from Version 1**

Clarifications/edits made in response to comments by Reviewer #1. The main changes being (i) we have now made it clear that the "UK cohort" is "NIHR HTA trials with a protocol" ongoing at the stated data extraction date; (ii) we have emphasised that the main result is that of 102 protocols using routinely collected data for an outcome, 46 were using routinely collected data as their sole source for at least one outcome; (iii) we have chosen to use 'routinely collected health data (RCHD)' throughout rather than 'Electronic Health Record (EHR)' for consistency.

Any further responses from the reviewers can be found at the end of the article

Introduction

Routinely collected data about health in medical records, registries and hospital activity statistics is now routinely collected in an electronic form. Progress in achieving connectivity, data linkage and security now offers the possibility of better use of this data for research purposes. For example, recent evidence shows the utility of long-term follow-up of trial patients by linkage to routinely collected health data (RCHD) sources (Fitzpatrick *et al.*, 2018). Innovative data-enabled study designs can answer pressing knowledge gaps in research evidence. However, the extent to which such sources of data are now being routinely employed in research to deliver efficient clinical trials, potentially at a wide scale, is unclear.

The aim of this study was to ascertain current practice amongst a United Kingdom (UK) cohort of recently funded and ongoing randomised controlled trials (RCTs) in relation to sources and use of routinely collected outcome data. We define RCHD to be data collected without specific *a priori* research questions developed prior to using the data for research.

Methods
Inclusion criteria
The following inclusion criteria were used:
1. Ongoing RCT of any type including feasibility or pilot work, funded by the National Institute for Health Research (NIHR) Health Technology Assessment (HTA) programme;

2. availability of a protocol; and

3. use of RCHD for at least one study outcome.

Search methods
A search of the NIHR Journals Library was undertaken to find protocols registered as of 25/10/2019. The search fields and terms used to select were:
1. Search term: 'Random'

2. Research type: 'Primary research'

3. Programme: 'HTA'

4. Status: 'Research in progress'

If the final published report was shown alongside the protocol this was taken to mean that the RCT was not ongoing but the

status had not been updated to 'Published', and the study was excluded.

In the absence of a protocol, the study was excluded. For studies with multiple protocol versions, the most recently available version was used.

Data extraction
One person (AM) extracted the information and categorised each RCHD source, with a second person (PW) checking classifications and explanations. The information extracted was as follows: Lead Investigator surname, year started, ISRCTN, project title, study type, use of RCHD for at least one study outcome, availability of a protocol, any details of data quality assessment of RCHD source prior to use, RCHD source name, reasons for wanting outcome data from RCHD source, specific outcomes and outcome type where clear data to be used will come from named RCHD sources.

Results
Figure 1 shows the study flow diagram. 279 records were identified through database searching and screened for inclusion. 22 were non-RCTs, 1 was a completed RCT, 30 were RCTs but no protocol was available and 10 were unclear. Of the remaining 216 NIHR HTA trials with a protocol available for further study, 102 (47%) planned to use RCHD for at least one outcome.

Table 1 shows the reasons for collecting trial outcome data from routine sources from the 102 eligible trials. The RCHD source was the sole source of outcome data for at least one outcome in 46 (45%) of those 102 trials (categories 3, 4 and 6 in Table 1). In five of these 46 protocols there was reference to prior feasibility work confirming aspects of the quality of the data to be sufficient for the main trial. Of the 102 trials, 14 (categories 7a-7d in Table 1) planned to assess the feasibility of using the RCHD sources during the trial, although details of the assessment were often lacking. Raw data for Figure 1 and Table 1 and Table 2 are available (see *Underlying data*, McKay *et al.* (2020)).

Table 2 shows the RCHD sources of outcome data to be used in these 46 studies. The most frequent RCHD sources are Hospital Episode Statistics (HES) and Office for National Statistics (ONS), with the most common outcome data to be extracted being on mortality, hospital admission, and health service resource use (see *Underlying data*, Data Set 5; McKay *et al.* (2020)). The full list of RCHD sources is given in *Extended data*, Supplementary Table 1 (McKay *et al.*, 2020).

Discussion
Our study has found that around half of publicly funded trials in a UK cohort (NIHR HTA funded trials that had protocol available) plan to collect outcome data from RCHD sources. This is much higher than the figure of 8% found in a cohort of 189 RCTs published since 2000, the majority of which were carried out in North America (McCord *et al.*, 2019).

Very few trial teams described any assessments of data quality from RCHDs in the protocol. Work is ongoing that should
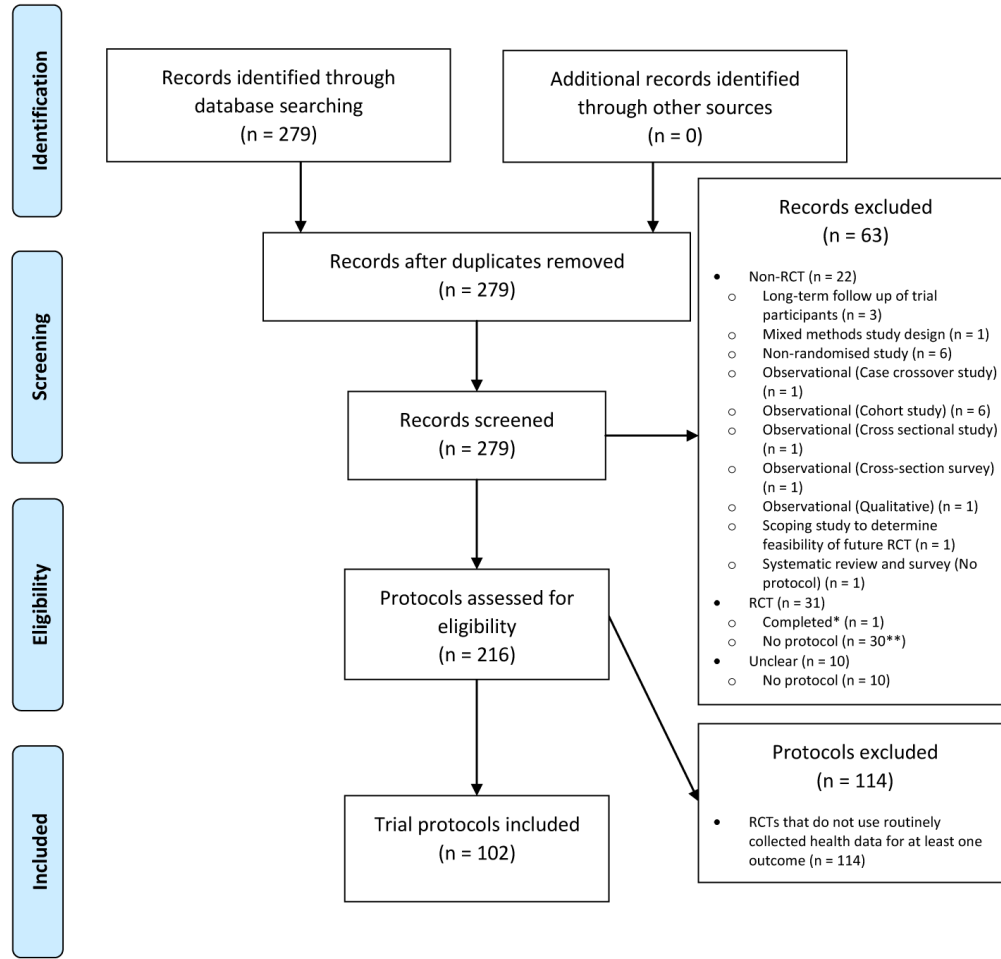
Figure 1. **PRISMA flow diagram.**

Table 1. **Reasons for sourcing outcome data from RCHD sources in 102 studies.** Multiple categories can apply to a single study.

	Categories	Total
(1)	(1a) 'Supplementing data collection for withdrawn patients (consent asked for at time of withdrawal)'	7
	(1b) 'Supplementing data collection for lost-to-follow-up patients'	8
	(1c) 'Supplementing data collection for withdrawn patients (consent NOT ASKED FOR at time of withdrawal)'	2
	(1e) 'Continued data collection for withdrawn patients (consent asked for at time of withdrawal)'	1
(2)	(2) 'Supplementing data collection for unobtainable/missing data'	3
(3)	(3a) 'As the sole source of all outcome data'	0
	(3b) 'As the sole source of all outcome data except for data related to protocol adherence and adverse event reporting being collected using CRFs'	0
(4)	(4) 'As the sole source of some outcome data'	43
(5)	(5a) 'As a source of some outcome data, alongside other sources for the same outcome data (e.g. CRF)'	51
	(5b) 'As a source of some outcome data, but collected by CRF if unable to access data'	3

	Categories	Total
(6)	(6a) 'Registry trial*: As the sole source of outcome data with purpose-built Module to collect remaining outcome data'	1
	(6b) 'Registry trial*: All outcome data collected through multiple RCHD sources except for questionnaire data'	1
	(6c) 'Registry trial*: All outcome data collected through multiple RCHD sources except for some baseline data, questionnaire data and other patient-reported data'	1
(7)	(7a) 'RCHD compared to trial collected data as part of feasibility assessment criteria'	11
	(7b) 'RCHD compared to trial collected data as a main trial secondary outcome'	1
	(7c) 'RCHD compared to trial collected data and then collect long-term follow-up data as part of trial'	1
	(7d) 'RCHD compared to trial collected data and then collect long-term follow-up data after trial has been completed'	1
	(7e) 'Representativeness of randomised patients compared with all eligible patients using RCHD as part of feasibility assessment criteria'	1
(8)	(8a) 'Participants flagged with NHS Digital/other: Check health status of patient prior to contacting in case patient has died'	2
	(8b) 'Participants flagged with NHS Digital/other: Check health status/notification of any deaths, causes'	12
(9)	(9) 'Set up mechanisms for long-term follow-up'	4
(10)	(10) 'Patients asked to provide written consent for continuation in the study once have regained capacity. Those who prefer not to be actively involved in the study follow-up, then asked to provide consent to using their routinely collected NHS data'	1
	Total	**155**

* A registry trial is a RCT conducted using clinical observational registries as the main source of outcome data collection

Table 2. Categories of RCHD sources of outcome data in 46 studies where this was the sole source for at least one outcome.

Source	Number (%)
(i) Primary care data (all regional equivalents)	8 (17%)
(ii) HES (and/or regional equivalents)	27 (59%)
(iii) ONS (and/or regional equivalents)	27 (59%)
(iv) Data collected specifically for patient group or healthcare intervention (to include patient registries, ICNARC, ambulance, etc)	26 (57%)
(v) Other	5 (11%)

determine whether such information should be reported in the trial publication (Kwakkenbos *et al.*, 2018). An extension to the SPIRIT guidelines for trials using RCHD is soon to be initiated, and will determine whether this information should be included in the trial protocol. As a minimum, it is recommended that trialists provide evidence in any funding application about the quality of the data from the RCHD source.

Data availability
Underlying data
Figshare: Use of routinely collected data in a UK cohort of publicly funded randomised clinical trials. https://doi.org/10.6084/m9.figshare.12185193 (McKay *et al.*, 2020).

This project contains the following underlying data:
- Data_Set_1_Details_and_Figure_1_v1.0.csv. (Study identifiers and raw data used for Figure 1.)
- Data_Set_2_Table_1_v1.0.csv. (Raw data used for Table 1.)
- Data_set_3_Supp_Table_1_v1.0.csv. (Raw data used for Supplementary Table 1.)
- Data_set_4_Table_2_v1.0.csv. (Raw data used for Table 2.)
- Data_set_5_Outcomes_using_EHR_data_v1.0.csv. (Raw data showing details of outcomes using data from RCHD sources.)

Extended data
Figshare: Use of routinely collected data in a UK cohort of publicly funded randomised clinical trials. https://doi.org/10.6084/m9.figshare.12185193 (McKay *et al.*, 2020).

This project contains the following extended data:
- Supplementary Table 1 - EHR sources of outcome data v1.0.pdf. (Supplementary Table 1.)

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

References

Fitzpatrick T, Perrier L, Shakik S, *et al.*: **Assessment of Long-term Follow-up of Randomized Trial Participants by Linkage to Routinely Collected Data: A Scoping Review and Analysis.** *JAMA Netw Open.* 2018; **1**(8): e186019.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kwakkenbos L, Juszczak E, Hemkens LG, *et al.*: **Protocol for the development of a CONSORT extension for RCTs using cohorts and routinely collected health data.** *Res Integr Peer Rev.* 2018; **3**: 9.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

McCord KA, Ewald H, Ladanie A, *et al.*: **Current use and costs of electronic health records for clinical trial research: a descriptive study.** *CMAJ Open.* 2019; **7**(1): E23–E32.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

McKay A, Jones A, Gamble C, *et al.*: **Data sets used and Supplementary Table 1.** *figshare.* Dataset. 2020.
http://www.doi.org/10.6084/m9.figshare.12185193.v1

Open Peer Review

Current Peer Review Status: ✓ ✓

✓ **Alison Howie** iD
Western University, London, ON, Canada
Merrick Zwarenstein
Centre for Studies in Family Medicine, Department of Family Medicine, Western University, London, Canada

A sentence could be included to explain your rationale for selecting trials from the Health Technology Assessment Programme at NIHR as against some other source of publicly funded trials. For example your study excludes UKMRC funded trials. The implications of excluding these mechanism-of-action and therapeutic intervention trials could be explained in the discussion.

Your flow diagram indicates that you identified articles through database searching (NIHR library) (n=279) and other sources (n=0). If NIHR HTA trials were of interest, what "other sources" were searched, and why?

Table 2 includes an "other" category for sources of outcome data. It only includes 5 studies, so perhaps you could list these "other" sources, in an appendix if necessary.

You mention that McCord *et al.* found that 8% of trials published since 2000 used outcome data from RCHD sources.
  ○ The McCord paper states, "In most (84%) of the trials in our sample, outcomes were measured with the use of EHRs". It's not entirely clear where the 8% is from?
  ○ I think it's important to note that they searched up to 2017; thus, since your study included trials up to almost the end of 2019, the higher percentage you found may represent a shift over time, especially as the appreciation for pragmatic trials and routinely collected data for these trials has increased in recent years.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Randomized trials, esp pragmatic trials; Epidemiology.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 02 June 2020

https://doi.org/10.5256/f1000research.26780.r64143

✔ **Sharon Love** [iD]
MRC Clinical Trials Unit, University College London, London, UK

Thank you for making the suggested changes. I have no further comments.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Trial conduct, particularly monitoring and the use of routinely collected health data.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

- -

Version 1

Reviewer Report 13 May 2020

?

Sharon Love (iD)

MRC Clinical Trials Unit, University College London, London, UK

This is a review of the protocols of RCTs, currently in progress, funded by NIHR, UK. RCTs were selected from NIHR HTA funding stream list if they claimed to be using routinely collected data for at least one study outcome. The authors found 102 trial protocols matching this criteria and from data extraction that 46 of these were using routinely collected data solely for at least one outcome. The research also found that a handful referenced previous feasibility work confirming the quality of the EHR and also gives a useful table categorising for the 102 trials how they used EHR.

Major Comments
I have only one major comment and it is the reason for both the 'partly' options below. The sample was selected to be using routinely collected data for at least one study outcome. Therefore I think the main result should contain this information. I consider that "in a UK cohort" is not enough of a description of the cohort. The fact that the sample was selected based on using routine data for an outcome is crucial in the interpretation.

The main result is that of 102 protocols using routinely collected data for an outcome, 46 were using routinely collected data as their sole source for at least one outcome. 46/102=45%. Around a half of NIHR HTA funded trials that had an uploaded protocol and used routinely collected health data for at least one study outcome, used solely routinely collected data for at least one trial outcome.

I think this is an important result.

Minor comments
1. Abstract – last part of the last sentence has a word missing "The majority of **which** were carried out in North America".

2. If you have space in the text, it would be useful to add the information that 30 were omitted due to not having a protocol.

3. The flow chart shows you selected the papers by selecting RCT, those that had a protocol and then those using routinely collected data for at least one outcome. I would be tempted to list the inclusion criteria in the paper in the same order.

4. The second inclusion criteria is "use of routinely collected health data". Elsewhere you use the term EHR. I would be tempted to be consistent.

5. Table 1: category 10 description appears incomplete.

6. Table 1: could you add a footnote of the definition of a registry trial?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: With others, I have conducted a review of the use of routinely collected health data by using release lists from registries which has been accepted for publication but is not yet published.

Reviewer Expertise: Trial conduct, particularly monitoring and the use of routinely collected health data.

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 May 2020

Andrew McKay, Liverpool Clinical Trials Centre, University of Liverpool, a member of Liverpool Health Partners, Liverpool, UK

Major comments: Thanks for your important comments. We have made these clearer within the article update to version 2.

Major comments part 1: We have now made it clear that the "UK cohort" is "NIHR HTA trials with a protocol" ongoing at the stated data extraction date.

Major comments part 2: We have now made this clearer.

Minor comments: Thank you for your comments. We have addressed them all within the article update to version 2. In relation to one specific comment, we have chosen to use 'routinely collected health data (RCHD)' throughout rather than 'Electronic Health Record (EHR)' for consistency.

Competing Interests: None.