Identification of COVID-19 clinical studies intending to share individual participant data for secondary use: Protocol for a pilot study

Authors: S. Canham, G. Felder, C. Ohmann, M. Panagiotopoulou (ECRIN)

Date: 9 September 2022

Version: Final version

1. Abstract

A pilot study will be performed to evaluate a simple classification system for data sharing statements of registered clinical studies, characterising the degree of willingness for data sharing, initially and specifically within COVID-19 studies. The evaluation will be performed by 3 experts on a random sample of 200 studies with a data sharing statement in a trial registry, extracted using a search of the metadata repository (MDR, https://crmdr.org/) of ECRIN. The bilateral inter-observer variability between experts will be investigated. In case of disagreement between experts a consensus will be derived to serve as the 'source of truth' (so called "gold standard") for further investigations exploring the use of semi-automatic classification algorithms. Subsequent to the data sharing statement categorisation, the intention is to contact the sponsors and / or principal investigators of those trials that appear likely to provide individual participant data (IPD) and ask whether they would be willing to share IPD in the context of the BY-COVID project.

2. Rationale

Within WP5 of the BY-COVID project (https://by-covid.org/) the intention is to use individual participant data (IPD) from COVID-19 vaccine trials (for example to test existing vaccines against emerging variants). This task will take advantage of the COVID-19 clinical research data repository developed in EOSC-Life WP14 (https://www.eosc-life.eu/). As an instrument for clinical trial data sharing, this repository intends to host the data from the H2020 VACCELERATE phase 2 and phase 3 platform trials on COVID-19 vaccines and to allow data sharing and secondary use in the BY-COVID project. There is a risk, however, that these trials will not be completed for some time and thus the data will only be shared after considerable delay. To be able to cover this risk and to have available IPD within the BY-COVID project timelines, it was decided to search for other COVID-19 trials willing to share individual participant data and try to involve them.

ECRIN has developed a metadata repository (MDR, https://crmdr.org/) for supporting findability of clinical trials. This tool, which can be assessed openly, currently covers around 700.000 trials from around 20 trial registries and more than 1.000.000 digital objects related to these trials (e.g., trial registry and / or results entry, study protocol, case report form, dataset, publication). This tool allows filtering of trials according to different criteria. The databases behind the MDR will be used for the discovery of COVID-19 trials in this study.

For feasibility assessment a preliminary search was performed for completed COVID-19 trials with the support of the MDR. 589 studies were identified with data sharing statements (DSS). From these studies 30 were analysed manually with respect to intention to share IPD from COVID-19 trials. From the analysis a preliminary classification of the DSS was derived and used as input for this study.

3. Objectives

Primary objective is to develop and evaluate a classification system for DSS of registered trials, characterising the degree of willingness for data sharing of COVID-19 trials. Secondary objective is to identify those COVID-19 trials with a high degree of willingness to share IPD.

At a further stage, the sponsors and principal investigators of these trials will be contacted and asked, whether they would be willing to share IPD in the context of the BY-COVID project.

4. Methods

4.1 Eligibility criteria

The following eligibility criteria for clinical trials / clinical studies were used:

- Clinical trials/clinical studies about COVID-19 or SARS-Cov-2
- Data Sharing Statement (DSS) available in the trial registry

Both completed and ongoing studies were considered.

4.2 Information sources

The identification of COVID-19 clinical studies with an intention to share IPD will be performed with the MDR developed by ECRIN. The principal aim of the MDR is to make the data objects generated from clinical research easier to locate, and to describe how each of those data objects can be accessed, providing direct links to them where that is possible. The central idea is to develop systems that can collect the *metadata* about the data objects, including object provenance, location and access details, from a variety of source systems (e.g., trial registries, data repositories, bibliographic systems) and aggregate it into a single repository, the MDR. The MDR provides access to the standardised metadata through a single system, accessed via a web portal (https://crmdr.org/). The metadata schemas, the data structures, the data extraction and the portal are described in the project's wiki (http://ecrin-mdr.online/index.php/Project_Overview).

The MDR covers the following trial registries:

- ClinicalTrials.gov
- EUCTR
- ISRCTN

- WHO ICTRP (providing access to data from a further 15 repositories)

In addition, the following repositories and other data sources are covered:

- Pubmed
- BioLINCC
- YODA

The MDR is updated weekly. As of 20 July 2022 it includes metadata from 715.569 clinical trials / clinical studies and from 1.116.168 related digital objects.

4.3 Search strategy

The search is done directly with the MDR data, using SQL statements against the database rather than the portal. The search strategy is based on the 4 terms:

- **'covid',**
- **'coronavirus',**
- **'sars-2'**
- **'sars2'**

All studies with non-null DSS and including one or more of the 4 terms listed above, either in their titles (all titles are considered) or their related ‘topics’ (associated keywords) were identified and extracted as suitable for DSS categorisation.

N.B. All searches are case-insensitive.

Some of the SQL statements used for the search are presented in the Appendix. Initial checks on the data were performed, using the core MDR dataset on 29 June 2022. These were done to check the consistency of the data between different tables belonging to the MDR.

The ‘study_search’ table in the MDR is used in these statements because it includes fields composed of lexemes (words and word stems) constructed by text indexing processing on the MDR data, on both the sets of titles for each study (title_lexemes field) and the sets of keywords (topic_lexemes field). In most cases the keywords are MESH coded, and both the original topic term and the MESH coded version are included within the lexemes list.

The table is rebuilt during each weekly aggregation as an aid to searching against study attributes in the portal, but in this context it also makes searching against COVID related terms much easier.

In the Appendix code

- 7.1 identifies and extracts the core study data into a new table from those studies that fulfil the COVID-19 linkage criteria (n=**14,308**)
- 7.2 identifies and extracts the core study data into a new table from those studies that fulfil the COVID-19 linkage criteria and which say they have IPD available *now* (n=**4**), while
- 7.3 identifies and extracts the core study data into a new table from those studies that fulfil the COVID-19 linkage criteria and which have a non-null data sharing statement (n=**2,953)**

A full set of the SQL statements used and an explanation of the tables and fields can be obtained by request to the authors.

A subset of the fields extracted by 7.3 (the COVID-19 studies with a DSS) are then extracted into a spreadsheet for easier viewing and analysis by those without direct access to the source database (which is restricted to a few individuals). The fields used are:

- Study Id (of the record in the MDR)
- Source id (usually a registry id)
- Id in source (usually the trial registry id)
- Study display (default) title
- Brief description
- Data sharing statement

If necessary, it will be possible to go back to examine study type and status, and start year/month, to see if there are any differences between them in terms of data sharing. This step is foreseen after the pilot study, when the categorisation system for the DSS has been evaluated.

4.4 Selection process

All trials identified by the code 7.3 as having the COVID-19 linkage criteria, and a non-null data sharing statement (n=**2,953)** will be taken into consideration for the study. At this stage no attempt is made to distinguish negative from positive DSS. From the full list, a random sample of 200 studies will be taken for the pilot categorisation study. The sample will not be stratified according to the length of the DSS – it is a simple random sample from the whole source population.

4.5 Data collection process

The data collection process will consist of three steps:

a) A training session, where the data collection process is introduced (virtual meeting, 1 hour)

b) The process of data collection done independently by each expert

c) Consensus meeting between the experts to derive a consensus in case of disagreement (virtual meeting)

Three experts from ECRIN will independently explore the DSS of the 200 selected studies and will provide an assessment according to the following classification:

Code	**Name**	**Definition**
#	Unclear	DSS not understood, and in particular the plans for sharing IPD are not clear or understood.
N	No IPD sharing	DSS states or implies that there will be no sharing of IPD. Different specific reasons provide different subtypes of this category.
0	No Decision	DSS states that there are currently no plans regarding IPD sharing or a decision has not yet been made.
V	Vaguely positive	DSS states or more normally implies that there will be some degree of IPD sharing, but no details are given or sharing appears limited.
R	By Request	DSS implies that IPD can be obtained by request, though to whom may not be explicitly stated. If stated it is usually the investigator, sometimes the sponsor.

S	Separate Storage	DSS states that IPD will be transferred to a repository – may be general, specific or institutional. More often not named specifically.
C	Complex	DSS states or implies that IPD will be available but under a more complex regime than any of those described above.

This classification has been derived from a manual inspection and analysis of DSS during an initial but non-systematic exploration of the data.

A sample size of 200 has been selected to give a 95%-confidence interval of around 10% for a given proportion. For a predicted agreement rate between two experts of 80%, the 95%-confidence interval will be between 74% and 86%, using asymptotic Wald method.

4.6 Study risk of bias assessment

Not relevant for this pilot study.

4.7 Effect measures

The main outcome of the pilot study is the inter-observer variability between two experts in the assessment of the DSS.

A further outcome is the agreement between the expert(s) and the consensus (“source of truth”/“gold standard”).

4.8 Synthesis methods

The bilateral assessment of the three experts will be tabulated in a cross table. A summary statistic of the agreement/disagreement rate between two experts will be generated.

For measuring interrater reliability between the assessment of two experts the kappa coefficient developed by Fleiss will be applied (1). The statistical analysis will be performed using the statistical software R.

In case of disagreement of the assessment, the three experts from ECRIN will meet to find a consensus (virtual meeting). If there is no agreement, the majority vote will be taken as consensus. The result of this consensus process will be documented (agreed, not agreed) and the assessment of each expert will be compared with the consensus.

Exploratory analysis

It is planned to develop an algorithm for semi-automatic classification of DSS. First feasibility studies have been performed and showed promising results, at least for short DSS. In a series of around 500 studies, it was investigated whether the following items could be derived semi-automatically by text analysis:

- Not understood
- Reason not sharing
- Reason no plan
- Reason no decision

- Vague positive statement
- Simple on request statement
- On request to investigator statement
- On request to sponsor statement
- Mentions 'repository'

The system will be further refined and optimised after the pilot study. To validate a semi-automatic classification system, a gold standard is needed for comparison. For the sample in the pilot study an expert consensus on the classification will be reached, which will be taken as the gold standard for validation of semi-automatic classification algorithms in the future.

4.9 Reporting bias assessment

Not relevant for this study.

4.10 Certainty assessment

95%-confidence intervals will be used to assess the certainty of the agreement between two experts as well as between an expert and the consensus.

5. Other information

5.1 Registration and protocol

The protocol for the pilot study will be registered on ZENODO. The protocol follows the PRISMA 2020 checklist (2).

5.2 Support

The study is performed in the context of the BY-COVID project. BY-COVID is funded by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101046203.

5.3 Competing interest

There are no competing interests of the authors of this protocol

5.4 Availability of data, code and other materials

All individual datasets used in the study will be made publicly available.

6. References

1. Fleiss, Joseph L. (1971) "Measuring nominal scale agreement among many raters." *Psychological Bulletin* , Vol. 76, No. 5 pp. 378–382

2. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. PLoS Med. 2021;18(3):e1003583.

7. Appendix: SQL statements used for data extraction

```
--****************************************************
-- Create tables of relevant records
--****************************************************
```

7.1 Create table of all distinct studies related to covid / sars2 (code)

```
drop table if exists st.cov;
create table st.cov
as
select s.*
from core.study_search ss
inner join core.studies s
on ss.id = s.id
where
(title_lexemes ilike '%covid%'
or title_lexemes ilike '%sars-2%'
or title_lexemes ilike '%coronavirus%'
or title_lexemes ilike '%sars2%'
or topic_lexemes ilike '%covid%'
or topic_lexemes ilike '%sars-2%'
or topic_lexemes ilike '%coronavirus%'
or topic_lexemes ilike '%sars2%');
```

7.2 Create table of source studies reporting that they have IPD now (code)

```
drop table if exists st.cov_ipd;
create table st.cov_ipd
as
select sids.study_id, sids.source_id,
sids.sd_sid, sids.is_preferred,
s.* from core.study_search ss
inner join core.studies s
on ss.id = s.id
inner join nk.study_ids sids
on s.id = sids.study_id
where
```

```
(title_lexemes ilike '%covid%'
or title_lexemes ilike '%sars-2%'
or title_lexemes ilike '%coronavirus%'
or title_lexemes ilike '%sars2%'
or topic_lexemes ilike '%covid%'
or topic_lexemes ilike '%sars-2%'
or topic_lexemes ilike '%coronavirus%'
or topic_lexemes ilike '%sars2%')
and has_ipd = true
```

7.3 Create table of all distinct studies related to covid / sars2 with non-null / non-empty DSS (code)

```
drop table if exists st.cov_dss;
create table st.cov_dss
as
select s.*
from core.study_search ss
inner join core.studies s
on ss.id = s.id
where
(title_lexemes ilike '%covid%'
or title_lexemes ilike '%sars-2%'
or title_lexemes ilike '%coronavirus%'
or title_lexemes ilike '%sars2%'
or topic_lexemes ilike '%covid%'
or topic_lexemes ilike '%sars-2%'
or topic_lexemes ilike '%coronavirus%'
or topic_lexemes ilike '%sars2%')
and s.data_sharing_statement is not null
and trim(s.data_sharing_statement) <> ''
```