Report on DIVERSE Task 1a2

Study Reference: DIVERSE

Protocol title: DIVERSE project: protocol for the scoping review

Protocol version: 1.0

File: Seafile

Status: Draft

Version: 1.1

Written by: Romin Pajouheshnia

Contents

Aim

The aim of Task 1a was to 1) develop a final search strategy for the DIVERSE scoping review (Task 1) and 2) derive a final PubMed search string, through a phase of development and testing.

Details of the overall background and aims of the project are given in the DIVERSE Seafile: DIVERSE_manuscript / Task folders / Objective1 / Task_1_0 / protocol.

Method

Overall strategy

A search strategy was developed in line with the study protocol (Task 1_0, EUPAS39757). The search strategy was developed with the aim to identify, as far as is feasible, a set of documents that fit within the scope of the scoping review, as defined by Task1a1. A summary of the scope is presented in appendix 1, derived from the study protocol and final report from Task1a1. The initial search strategy comprised of 4 sections:

- A set of papers identified by experts within DIVERSE
- PubMed search
- Grey literature search
- A snowball search of included papers

Core reference papers

A set of relevant papers for the review will be identified based on expert knowledge within the group (Task 1a1 - complete).

Grey literature search

For efficiency, the task 1a2 team agreed to reuse the grey literature search designed and conducted in the EMA MINERVA project (Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability: EUPAS39322). An overview of the strategy is reported here: https://www.ema.europa.eu/en/documents/presentation/presentation-session-2-preliminary-list-metadata-romin-pajouheshnia_en.pdf

In short, a list of organizations and consortia with experience in multi-database pharmacoepidemiology were identified. The websites of these organizations/consortia were searched for grey literature documents that report detailed descriptions of multiple data sources or methodology on how to describe data sources. Documents in the public domain that fell within the scope of the scoping review (Task1a1) were identified for inclusion in screening and selection.

Snowball search

The snowball search strategy was discussed in a meeting of Task1a2 members. It was agreed that the snowball search would be a *backwards* snowball search (screening reference lists of included papers) and *only* for the core reference papers, in order to maintain a manageable number of hits. It was agreed that if the snowball search retrieved a large number of unique papers, not identified in other sections of the search strategy (>100 unique papers), a random sample of the core papers would be selected for inclusion in the snowball search (~10 papers)

PubMed search string

A search string in PubMed was developed through the following process:

1. A draft strategy for the search string, based on the strategy from the study protocol (from Hunt et al. 2021, Supplementary Table S1) was shared as a document and presentation to co-authors for approval
2. The initial search string was executed in PubMed and the proportion of core reference papers that were detected by the string (sensitivity) and number of hits were examined.
3. The search string was iteratively improved (see below for the process).

The development process for the search string went as follows:

1. The list of core reference papers identified in Task1a1 were entered into the format of a PubMed extraction csv file, in order to be read by R software and be used a *validation set.* 21/24 of the papers were found to be PubMed indexed. The three papers that were not indexed (2 were white paper documents, 1 was in a journal that is not PubMed indexed) were not included in the validation set.
2. The search string was entered into PubMed by a task member (RP, KS) and the full list of results was extracted into a csv file.
3. The core reference set and the search string results were read into R. An R script was used to calculate the proportion of core papers that were detected by the search string (*sensitivity*), and identify a list of the papers that were not identified.
4. The search string was adapted in an iterative process by adding and removing terms in order to increase the search sensitivity and reduce the total number of hits.
5. To increase the sensitivity of the search, the individual components of the search were tested against the core papers that were not detected using the whole string, in order to identify the reason why the core paper was not detected. Where possible, general terms were added or redundant terms were excluded, in order to change the string so that it could detect the missing core paper without excluding other papers or greatly increasing the number of hits.
6. The optimization process stopped when the search strategy yielded a sensitivity of >80% and the number of hits was ~<= 300 papers (deemed a feasible number of papers to screen in Task 1a5).

Results

A schematic of the overall search strategy can be found in Appendix 2.

Core reference papers, grey literature and snowball

- Initially, 38 papers were identified, or which **24 were included** following screening during task 1a1.
- The grey literature search identified **10 documents.**

The snowball search will be conducted as a part of task 1a3.

Search string

The initial search, derived from Hunt et al 2021, was conducted on 22 April 2021. The search can be broken down into the combination of 5 sections, which are combined with Booleans:

1. (Medicines terms
2. AND Epidemiology and study design terms
3. AND Database terms
4. AND Multi-database studies terms)
5. NOT Clinical trials

The search had poor sensitivity (45%) and specificity (38739 hits). The search was revised by two investigators (KS, RP) as follows:

1. The search was simplified by using only title/abstract search terms for sections 1-4.
2. Redundant terms were removed
3. Terms related to real world evidence were included in section 1
4. Section 4 was expanded using additional terms identified in Task1a1.

The final iteration of the search string comprises of the same 5 sections and can be found in Appendix 2. The search was executed in PubMed on 9th July 2021 and gave the following results, deemed acceptable according to the specifications in the protocol:

- **336 records were identified by the search**
- 17/21 of the core reference papers were identified **(81% sensitivity)**

The four papers in the validation set that were not identified by the final search string are reported in Appendix 3, along with the reason why they were not detected.

Conclusion

A search strategy has been developed in accordance with the study protocol. A PubMed search string was developed based on a published search strategy and adapted such that it met the criteria for *sensitivity* and *specificity* decided in the protocol. The strategy is ready to be implemented in Task 1a3.

Appendix 1: Inclusion/exclusion criteria of the scoping review

i) Inclusion criteria

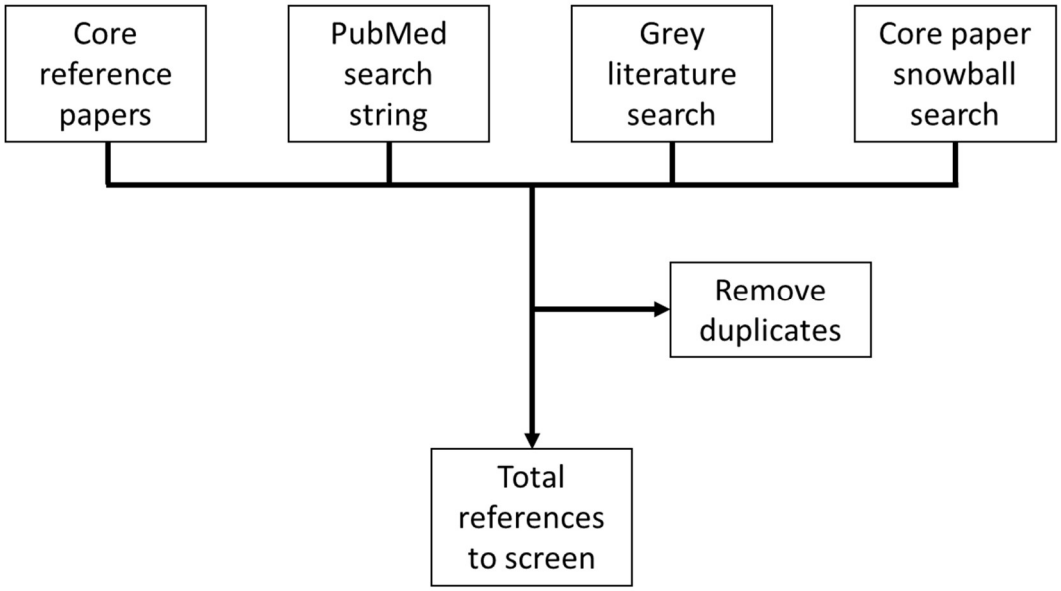
Reviews or methodological documents:
contains recommendations/guidelines for the collection/ reporting of (heterogeneity of) data sources
contains tools to describe data sources (e.g. questionnaires)
created by an organization or a network of organizations conducting MDS to describe contributing data sources
A methodological paper that provides methods/tools to describe data sources
A review or methodological paper that describes data sources
Documents reporting a pharmacoepidemiologic study of specific issue:
a significant description of the data sources involved in the MDS (beyond a description of the contents of the data)
strategies to exploit data source diversity to improve the quality of the generated evidence
strategies to exploit data source diversity to assist interpretation of the generated evidence
Addresses the study questions by reporting/describing:
a tool/ method for collecting data on heterogeneity
a tool/ method for reporting on heterogeneity
a tool/ method for classifying heterogeneity
Heterogeneity leveraged to improve quality of evidence
Heterogeneity leveraged to improve interpretation

ii) Exclusion criteria

- Documents that only describe statistical methods for heterogeneity in results
- Documents that only describe single database studies
- Documents that only describe clinical trials and not observational research

Appendix 2: Full search strategy

* Execution of the strategy is to be completed in Task 1a3



Appendix 3: PubMed Search string, version 1.1 (09 July 2021)

1	("drug*"[Title/Abstract] OR "medicat*"[Title/Abstract] OR "pharmaco*"[Title/Abstract] OR "medical product*"[Title/Abstract] OR "medicinal product*"[Title/Abstract] OR "postmarketing"[Title/Abstract] OR "post-marketing"[Title/Abstract] OR "real-world"[Title/Abstract] OR "real-world"[Title/Abstract])
2	("follow-up studies"[Title/Abstract] OR "prospective studies"[Title/Abstract] OR "cross-sectional studies"[Title/Abstract] OR "pharmacoepidemiol*"[Title/Abstract] OR "epidemiol*"[Title/Abstract] OR "case-control"[Title/Abstract] OR "case-control"[Title/Abstract] OR "cohort"[Title/Abstract] OR "population based"[Title/Abstract] OR "nation wide"[Title/Abstract] OR "nationwide"[Title/Abstract] OR "case crossover"[Title/Abstract] OR "case time control"[Title/Abstract] OR "self controlled case series"[Title/Abstract] OR "surveillance"[Title/Abstract] OR "drug safety monitoring"[Title/Abstract] OR "pharmacovigilance"[Title/Abstract] OR "observational"[Title/Abstract] OR "confounder*"[Title/Abstract] OR "confounding*"[Title/Abstract]) OR "incidence rate"[Title/Abstract] OR "prevalence"[Title/Abstract]
3	("database"[Title/Abstract] OR "data base"[Title/Abstract] OR "data bases"[Title/Abstract] OR "data source"[Title/Abstract] OR "data sources"[Title/Abstract] OR "register"[Title/Abstract] OR "registry"[Title/Abstract] OR "registries"[Title/Abstract] OR "biobank"[Title/Abstract] OR "administrative claims"[Title/Abstract] OR "administrative data"[Title/Abstract] OR "claims data"[Title/Abstract] OR "medical records"[Title/Abstract] OR "patient records"[Title/Abstract] OR "healthcare records"[Title/Abstract] OR "health record data"[Title/Abstract] OR "health records"[Title/Abstract])
4	((("multiple databases"[Title/Abstract] OR "multi-database"[Title/Abstract] OR "multidatabase"[Title/Abstract] OR "multiple data sources"[Title/Abstract] OR "multi data source"[Title/Abstract] OR "multiple centres"[Title/Abstract] OR "multi centre"[Title/Abstract] OR (("multinational"[All Fields] OR "multinationals"[All Fields]) AND "OR"[Title/Abstract])) AND "multi national"[Title/Abstract]) OR "multiple regions"[Title/Abstract] OR "multi country"[Title/Abstract] OR "multiple countries"[Title/Abstract] OR "multi cohort"[Title/Abstract] OR "multi site"[Title/Abstract] OR "multiple sites"[Title/Abstract] OR "distributed data"[Title/Abstract] OR "distributed network*"[Title/Abstract] OR "distributed database*"[Title/Abstract] OR "safety network*"[Title/Abstract] OR "research network*"[Title/Abstract] OR "common data model*"[Title/Abstract] OR "common protocol*"[Title/Abstract] OR "common study protocol*"[Title/Abstract] OR "distributed algorithm*"[Title/Abstract] OR "database heterogeneity"[Title/Abstract] OR "data source heterogeneity"[Title/Abstract] OR "multi-database"[Title/Abstract] OR "multiple health-care databases"[Title/Abstract])
5	("pilot projects"[MeSH Terms] OR "double-blind"[Text Word] OR "placebo-controlled"[Text Word] OR "case reports"[Publication Type] OR "published erratum"[Publication Type] OR "randomized controlled trial"[Publication Type] OR "clinical trial, phase i"[Publication Type] OR "clinical trial, phase ii"[Publication Type] OR "clinical trial, phase iii"[Publication Type] OR "clinical trial, phase iv"[Publication Type] OR "controlled clinical trial"[Publication Type])
6	(#1 AND #2 AND #3 AND #4) NOT #5

Appendix 4: Core reference papers not detected by search string

Article	Reason why it was not captured by search string	Search string section that failed
Burgun A, Bernal-Delgado E, Kuchinke W, et al. Health Data for Public Health: Towards New Ways of Combining Data Sources to Support Research Efforts in Europe. *Yearb Med Inform* 2017;26(1):235-40. doi: 10.15265/IY-2017-034 [published Online First: 2017/09/11]	Very short and general abstract, so hard to identify based on title/abstract screening	Not included by #1, #2, #4
Cave A, Kurz X, Arlett P. Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. *Clinical Pharmacology & Therapeutics* 2019;106(1):36-39. doi: https://doi.org/10.1002/cpt.1426	Includes "clinical trial" in title	Excluded by #5
Dedman D, Cabecinha M, Williams R, et al. Approaches for combining primary care electronic health record data from multiple sources: a systematic review of observational studies. *BMJ open* 2020;10(10):e037405. doi: 10.1136/bmjopen-2020-037405	Doesn't include drug/real-world evidence term	Not included by #1
Su C-C, Chia-Cheng Lai E, Kao Yang Y-H, et al. Incidence, prevalence and prescription patterns of antipsychotic medications use in Asia and US: A cross-nation comparison with common data model. *Journal of psychiatric research* 2020;131:77-84. doi: https://doi.org/10.1016/j.jpsychires.2020.08.025	Does not report any terminology relating to databases/data sources in the abstract	Not included by #3