DIVERSE project: protocol for the scoping review

This document contains the protocol of a scoping review which constitutes Task 1.0 of the DIVERSE project. The DIVERSE Project is an initiative of the Database Special Interest Group of the International Society for Pharmacoepidemiology (ISPE). Funding for this project is provided by ISPE.

This document follows the JBI methodology for scoping reviews highlighted in Chapter 11 of the JBI Manual for Evidence Synthesis (https://wiki.joannabriggs.org/display/MANUAL)

Document history

Version	**Date**	**Author(s)**	**Changes**
0.1	3 Jan 2021	Rosa Gini	First draft for the review of participants to Task 1.0 of the DIVERSE Project, containing comments and questions
0.9	15 Jan 2021	Rosa Gini	Included comments and results from discussion with Gillian Hall, Romin Pajouheshnia, Robert Platt, Soko Setoguchi
1.0	21 Jan 2021	Rosa Gini	Incorporated comments from Gillian Hall and Romin Pajouheshnia

Title of the scoping review

Data source heterogeneity in multi-data base pharmacoepidemiologic studies: a scoping review

Scoping review questions

Primary question: what are the existing tools and recommendations for the collection and reporting of heterogeneity in data sources used in multi-data base pharmacoepidemiologic studies? In particular, what are the existing descriptors of data source heterogeneity?

Secondary question: how is heterogeneity leveraged to improve the quality of the evidence generated in multi-data base pharmacoepidemiologic studies and to assist its interpretation?

Introduction

Multi-data base studies (MDS) are increasingly performed in pharmacoepidemiologic research. A MDS is as a study using at least two healthcare databases, which are not linked with each other at an individual person level, either because they cover and capture information on different individuals, or because, even if populations overlap, local regulations forbid record linkage. In a MDS, analyses are carried out in parallel across each data source applying a common study protocol.[1] Regulatory authorities often require data from multiple data sources to be used in a single study, to enhance the generalizability of results or to obtain sufficient sample size when the exposure and/or outcome is rare.[1-3] MDS pose a number of challenges, including how to manage heterogeneity between the different included data sources.[4-8] Herein we make a distinction between heterogeneity in data sources and heterogeneity in results, where the former may result in homogeneous or heterogeneous results across between data sources in a MDS.

Data sources included in MDS can be different in how they are generated and/or for what purposes. For example, there may be billing/administrative databases generated by healthcare facilities for the purpose of requesting funding from healthcare payers; electronic healthcare records made available by networks of physicians; population drug/disease registries established because of health policies; and research centres with multiple, linked data sources, altogether providing a heterogenous mix of data. These data might also be recorded in different structures, coding systems or languages, which can imply different granularity in clinical content between databases.[9] Variation in information being collected in healthcare databases ultimately arises due to differences between healthcare, administrative and surveillance systems across regions or nations, which also lead to differences in the range of, access to and delivery of healthcare services.

It is extremely challenging to interpret the findings of MDS without a clear understanding of the context and purpose of data collection for each data source. This has been evidenced by the recent retraction of high-impact papers on drug therapies for COVID-19 due to transparency issues when multiple data sources were used.[10,11] Despite calls for the implementation of strategies to improve replicability, increase transparency and reduce bias in MDS,[7,12,13] and despite general recommendations to assess the comparability of data sources in MDS,[14] to our knowledge, there is currently no guidance for how database heterogeneity should be evaluated or even identified and recorded. A preliminary search for existing scoping reviews on this topic was performed in PubMed on January 2nd 2021, using

the search string *(database OR "data source" OR data source) AND heterogeneity AND "scoping review"* and none of 222 results were pertinent.

This scoping review is intended to inform the development of guidelines for the identification, collection and reporting of heterogeneity in MDS, and to identify areas for further research. This activity is the Objective 1 of the DIVERSE project, of the Database Special Interest Group of the International Society for Pharmacoepidemiology (ISPE). The list of the tasks included in Objective 1 of the DIVERSE Project is described in Table 1.

Task #	Task name	Task description	Task leader
1.0	**Protocol & workplan**	i. Draft the scoping review protocol (high-level, including timelines). ii. Complete "OVERVIEW tasks and members document"	Rosa Gini
1a1	**Reference papers**	i. Identify reference set of papers based on expert knowledge. ii. Refine inclusion/ exclusion criteria for literature review based on the chosen papers	Gillian Hall
1a2	**Develop search string**	i. Draft the search strategy and string ii. Validate against references from a1a. iii. Revise to meet criteria from 1a1 and improve sens/spec of search	Romin Pajouheshnia
1a3	**Extract papers**	i. Apply the search string and extract papers into reference managing software, with help from informatics specialist	Romin Pajouheshnia
1a4	**Screening and selection**	i. Conduct screening and selection of articles identified in 1a3, as per protocol. (ii. Optional - test and apply machine-automated tool to support screening)	Romin Pajouheshnia
1a5	**Draft extraction tool**	i. Compile set of variables to extract from selected literature from 1a4 ii. Draft protocol for 1a8, Analyse the extracted review data, and share with group for feedback iii. Draft the extraction tool	Romin Pajouheshnia
1a6	**Pilot the extraction tool**	i. Pilot the extraction tool ii. Revise extraction tool where needed iii. Revise protocl for 1a8, where needed	Rosa Gini
1a7	**Data extraction**	i. Extract data from literature identified in 1a4 using extraction tool from 1a5/6 ii. Conduct necessary quality checks in accordance with protocol	Romin Pajouheshnia
1a8	**Analyse the extracted review data**	i. Analyse the data extracted in 1a7 - create tables and figures	Romin Pajouheshnia
1a9	**Draft report from literature**	i Draft report from literature ii. Send to whole group for feedback	Romin Pajouheshnia

Table 1. Tasks of the DIVERSE Project Objective 1.

Scoping review objective

The objective of this scoping review is to list and summarize existing tools and recommendations for the collection and reporting of heterogeneity in data sources used in MDS, in particular listing and classifying existing descriptors of such heterogeneity. A

secondary objective is to describe how heterogeneity is leveraged to improve the quality of the evidence generated in a MDS and to assist its interpretation.

Glossary

A glossary will be created, in the form of a Google document

https://docs.google.com/document/d/1GXgaR9RWk6pEUg_qGCzbUP6QHd8A5kRbOOBJbfayesw/edit

This will be a living tool that will evolve during the scoping review.

Inclusion criteria, search strategy and source of evidence

Sources included in the scoping review will be published papers and grey literature (e.g. documents or web pages). They will include

- documents or published reviews containing recommendations or guidelines for the collection and reporting of (heterogeneity of) data sources (e.g. RECORD-PE)
- tools to describe data sources (e.g. questionnaires)
- documents created by an organization or a network of organizations conducting MDS to describe the data sources that contribute to its own studies
- published reviews describing tools to describe data sources
- published reviews describing data sources
- in exceptional cases (e.g., when explicitly included in the list of reference sources by one of the experts, see below; this criterion will not contribute to the search string), we will include published papers whose main focus is a specific pharmacoepidemiologic study, but include
    - a significant description of the data sources involved in the MDS (beyond a description of the contents of the data e.g. Table 1), and/or
    - strategies to exploit data source diversity to improve the quality of the evidence, and/or
    - strategies to exploit data source diversity to assist interpretation of the evidence generated by the study

They will *not* include

- Papers that describe statistical methods for heterogeneity in results
- Papers describing single database studies
- Papers describing MDS outside of the field of pharmacoepidemiology

In detail, the process to create the final list of sources of evidence will be as follows.

1. **List of reference sources**: the members of the DIVERSE project will be requested to list sources they consider of interest to respond to the scoping review question (Task 1a1)
2. **Detailed inclusion and exclusion criteria and screening tool**: based on the reference sources, a detailed list of inclusion and exclusion criteria will be created and a screening tool will be developed and tested.[1] When finalised, the list of detailed

[1] The screening tool will consist of a checklist of inclusion and exclusion criteria. The tool will be piloted at each stage (TIAB and full text) on a random sample of 50 records. The pilot extraction will be conducted independently by two reviewers, any uncertainties will be flagged, and interrater agreement will be measured (Cohen's kappa) and disagreements discussed. Based on the experience of the pilot, the screening tool will be revised and the remaining screening for inclusion at each stage will be completed in duplicate. Interrater agreement will again be measured and disagreements discussed. Where an agreement cannot be reached, additional reviewers will be consulted

criteria and the screening tool will be published as an appendix to this protocol (Task 1a2)

3. **List of candidate sources of evidence**: based on the inclusion/exclusion criteria
    - A backwards snow ball strategy will be applied to the reference list (point 1 above) to include additional literature they mention that is deemed potentially relevant (Task 1a2)
    - a search string for PubMed and possibly other databases will be developed and tested, to ensure that it includes at least 80% of the published sources in the reference list (Task 1a2);
    - at the end, a unique list of candidate sources will be obtained with duplicates eliminated (Task 1a3)
4. **Final list of sources of evidence**: the screening tool will be applied to the list of candidate sources (point 3 above); in case of need, use of an automated version of the screening tool will be considered; the result of the screening, along with the initial list of reference sources, will be the source of evidence of the scoping review (Task 1a4).

Data extraction

Information will be extracted from each source of evidence based on a tool that will be developed by participants to the Task 1a5. The information extracted will be finalised in this task and will include the following variables:

- A unique code to be used in the analysis
- Authors
- Title
- Year of publication or finalisation, if not published; for electronic publications: most recent update
- Country of the main affiliation of the first author, or publication organisation where there is no author
- If applicable: DOI
- If applicable: web link
- Type of source of evidence (tool, document, published review, published MDS, etc)
- List of the countries whose population is (possibly partially) included in the data sources mentioned by the source of evidence (if applicable)
- To describe each data source: if the following descriptors are mentioned/recommended
    - Which organization makes the data accessible for research
    - Which organization collects the data and for which purpose
    - What prompts the recording of the data
    - What is the population for which the data is collected, including reasons for being included/excluded from the data collection and for which period of time
    - What is the content of the data
    - What is the data dictionary, including coding systems or free text
- How is heterogeneity in the above dimensions summarised, if applicable
- If heterogeneity is leveraged to improve evidence, how it is leveraged (if applicable)
- If heterogeneity is leveraged to assist interpretation of evidence, how it is leveraged (if applicable)
- Any outstanding research challenges or issues regarding data source heterogeneity described by the authors (e.g. in the discussion section)

A detailed data extraction tool will be developed by participants to the Task 1a5 and will be piloted by participants to Task 1a6. In this phase the data items will be refined, whenever possible a data dictionary will be developed, and a data extraction manual will be completed.

The data extraction tool and manual will be published as appendices to this protocol.

The final data extraction will be conducted by participants to the Task 1a7. Each source of evidence will be analysed independently by two reviewers and conflicts will then be resolved by consensus.

Analysis of the evidence

A detailed analysis plan will be developed by Task 1a8 following the guidelines of section 11.2.8 and 11.2.9 of the JBI Manual for Evidence Synthesis (https://wiki.joannabriggs.org/display/MANUAL) and will be included as an appendix to this protocol. In line with the Arksey and O'Malley framework,[34] a template for implementing an analysis framework, to collate and summarize the results will be developed. Data will be analysed through numerical analysis of the extracted data and through a narrative summary, as described in the section "*Presentation of the results*", below. Numerical data analyses will be preceded by cleaning of the data in the extraction tool and analysed using descriptive statistics in R software.

Presentation of the results

A report will be drafted by Task 1a9, including as supplementary material this protocol, all of its appendixes, the populated data extraction tool and tables/figures. The report will include an introduction, the summary of methods, and a description of the main results. The results will support the discussion in summarising and classifying existing tools, recommendations and descriptors for the collection and reporting of heterogeneity in data sources used in MDS. Moreover, based on the results, the discussion will summarise how heterogeneity is leveraged to improve the quality of the evidence generated in a MDS and to assist its interpretation. Finally, areas where additional research and guidance are needed will be identified.

Consultation
As recommended by the Arksey and O'Malley framework[34], the final report will be reviewed by relevant stakeholders and consumers of the review content. First, the report will be reviewed and feedback provided by the whole DIVERSE working group. Second, the report will be reviewed by the ISPE manuscript committee before publication.

The report will be published in the public domain and submitted for publication in Pharmacoepidemiology and Drug Safety.

Bibliography

1. Gini R, Sturkenboom MC, Sultana J, Cave A, Landi A, Pacurariu A, Roberto G, Schink T, Candore G, Slattery J, Trifirò G. Different Strategies to Execute Multi-Database Studies for Medicines Surveillance in Real-World Setting: A Reflection on the European Model. Clinical Pharmacology & Therapeutics. 2020 Apr 3.

2. Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, Mazzaglia G, Giaquinto C, Corrao G, Pedersen L, van der Lei J. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiology and drug safety. 2011 Jan;20(1):1-1.

3. Pacurariu A, Plueschke K, McGettigan P, Morales DR, Slattery J, Vogl D, Goedecke T, Kurz X, Cave A. Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. BMJ open. 2018 Sep 1;8(9):e023090.

4. Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, Suchard MA, DuMouchel W, Berlin JA. Evaluating the impact of database heterogeneity on observational study results. American journal of epidemiology. 2013 Aug 15;178(4):645-51.

5. Trifirò G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, Bonhoeffer J, Schuemie M, van der Lei J, Sturkenboom M. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how?. Journal of internal medicine. 2014 Jun;275(6):551-61.

6. Lai EC, Stang P, Yang YH, Kubota K, Wong IC, Setoguchi S. International multi-database pharmacoepidemiology: potentials and pitfalls. Current Epidemiology Reports. 2015 Dec 1;2(4):229-38.

7. Klungel OH, Kurz X, De Groot MC, Schlienger RG, Tcherny-Lessenot S, Grimaldi L, Ibáñez L, Groenwold RH, Reynolds RF. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. Pharmacoepidemiology and drug safety. 2016 Mar;25:156-65.

8. Burcu M, Dreyer NA, Franklin JM, Blum MD, Critchlow CW, Perfetto EM, Zhou W. Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. Pharmacoepidemiol Drug Saf. 2020 Mar 11. doi: 10.1002/pds.4975. Epub ahead of print. PMID: 32162381.

9. Avillach P, Coloma PM, Gini R, Schuemie M, Mougin F, Dufour JC, Mazzaglia G, Giaquinto C, Fornari C, Herings R, Molokhia M. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. Journal of the American Medical Informatics Association. 2013 Jan 1;20(1):184-92.

10. Mehra MR, Desai SS, Kuy SR, Henry TD, Patel AN. Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19 (The New England journal of medicine (2020)). The New England journal of medicine. 2020 Jun 18;382(25).

11. Mehra MR, Ruschitzka F, Patel AN. Retraction-Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. Lancet. 2020 Jun 13;395(10240):1820.

12. Bazelier MT, Eriksson I, de Vries F, Schmidt MK, Raitanen J, Haukka J, Starup-Linde J, De Bruin ML, Andersen M. Data management and data analysis techniques in pharmacoepidemiological studies using a pre-planned multi-database approach: a systematic literature review. pharmacoepidemiology and drug safety. 2015 Sep;24(9):897-905.

13. Platt RW, Platt R, Brown JS, Henry DA, Klungel OH, Suissa S. How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias. Pharmacoepidemiology and Drug Safety. 2020 Jan;29:3-7.

14. Hall GC, Sauer B, Bourke A, Brown JS, Reynolds MW, Casale RL. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiology and drug safety. 2012 Jan;21(1):1-0.

15. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, Van Der Lei J. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics. 2015;216:574

16. Lai EC, Ryan P, Zhang Y, Schuemie M, Hardy NC, Kamijima Y, Kimura S, Kubota K, Man KK, Cho SY, Park RW. Applying a common data model to Asian databases for multinational pharmacoepidemiologic studies: opportunities and challenges. Clinical Epidemiology. 2018;10:875.

17. Toh S, Pratt N, Klungel O, Gagne JJ, Platt RW. Distributed networks of databases analyzed using common protocols and/or common data models. Pharmacoepidemiology, Sixth Edition. 2019 Nov 6:617-38.

18. Schneeweiss S, Brown JS, Bate A, Trifirò G, Bartels DB. Choosing among common data models for real-world data analyses fit for making decisions about the effectiveness of medical products. Clinical Pharmacology & Therapeutics. 2020 Apr;107(4):827-33.

19. Lai EC, Man K, Toh D, Platt R, Hallas J, Setoguchi S. Heterogeneity and Validity in National and International Multi-Database Pharmacoepidemiologic Studies (MPES): Lessons Learned in North America, Europe, and Asia. Abstracts Pharmacoepidemiol Drug Saf. 2019; 28(S2): 25-26

20. Gini R, Lanes S, Bollaerts K, Trifiro G, Kirchmayer U, Hall GC. Data diversity in multi-database Pharmacoepidemiologic studies and its role in outcome misclassification: A curse or a blessing?. Abstracts Pharmacoepidemiol Drug Saf. 2019; 28(S2): 241-241.

21. Pajouheshnia R, Gardarsdottir H, Platt R, Toh D, Klungel O. Analysis of data from distributed pharmacoepidemiologic networks. Abstracts Pharmacoepidemiol Drug Saf. 2019; 28(S2): 414-414

22. Lai EC, Shin JY, Kubota K, Man KK, Park BJ, Pratt N, Roughead EE, Wong IC, Kao Yang YH, Setoguchi S. Comparative safety of NSAIDs for gastrointestinal events in Asia-Pacific populations: A multi-database, international cohort study. Pharmacoepidemiology and drug safety. 2018 Nov;27(11):1223-30.

23. Platt RW, Dormuth CR, Chateau D, Filion K. Observational studies of drug safety in multi-database studies: methodological challenges and opportunities. eGEMs. 2016;4(1).

24. Roberto G, Leal I, Sattar N, Loomis AK, Avillach P, Egger P, Van Wijngaarden R, Ansell D, Reisberg S, Tammesoo ML, Alavere H. Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the EMIF project. PloS one. 2016 Aug 31;11(8):e0160648.

25. Gini R, Dodd CN, Bollaerts K, Bartolini C, Roberto G, Huerta-Alvarez C, Martín-Merino E, Duarte-Salles T, Picelli G, Tramontan L, Danieli G. Quantifying outcome misclassification in multi-database studies: the case study of pertussis in the ADVANCE project. Vaccine. 2019 Oct 31.

26. Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, Cazzola W, Coloma P, Berni R, Diallo G, Oliveira JL. Data extraction and management in networks of observational health care databases for scientific research: a comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE strategies. Egems. 2016;4(1).

27. Hall GC, Lanes S, Bollaerts K, Zhou X, Ferreira G, Gini R. Outcome misclassification: Impact, usual practice in pharmacoepidemiology database studies and an online aid to correct biased estimates of risk ratio or cumulative incidence. Pharmacoepidemiology and Drug Safety. 2020 Aug 28.

28. Hunt NB, Gardarsdottir H, Bazelier MT, Klungel OH, Pajouheshnia R. A Systematic Review Of How Missing Data Is Handled And Reported In Multi-database Pharmacoepidemiology. Poster. ICPE ALL ACCESS 2020

29. Peters MD, Godfrey C, McInerney P, Baldini Soares C, Khalil H, Parker DA, Munn Z. Chapter 11: Scoping reviews. JBI Reviewer's Manual [Internet]. Adelaide: JBI, 2017 [cited 2019 July 19].

30. Van de schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, Kramer B, Huijts M, Hoogerwerf M, Ferdinands G, Harkema A. ASReview: Open Source Software for Efficient and Transparent Active Learning for Systematic Reviews. arXiv preprint arXiv:2006.12166. 2020 Jun 22.

31. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MD, Horsley T, Weeks L, Hempel S. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of internal medicine. 2018 Oct 2;169(7):467-73.

32. Equator Network. Enhancing the QUAlity and Transparency Of health Research. https://www.equator-network.org 2020 Sep.

33. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, Klungel O, Petersen I, Sorensen HT, Dixon WG, Guttmann A. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). bmj. 2018 Nov 14;363:k3532.

34. Hilary Arksey & Lisa O'Malley (2005) Scoping studies: towards a methodological framework, International Journal of Social Research Methodology, 8:1, 19-32