





















Describing diversity of real world data sources in pharmacoepidemiologic studies: The DIVERSE scoping review

Rosa Gini¹  | Romin Pajouheshnia^{2,3}  | Helga Gardarsdottir^{2,4,5}  |
 Dimitri Bennett⁶  | Lin Li⁷  | Claudia Gulea⁸  |
 Angelika Wientzek-Fleischmann⁹  | Marloes T. Bazelier²  | Mehmet Burcu¹⁰  |
 Caitlin Dodd¹¹  | Carlos E. Durán⁴  | Sigal Kaplan¹²  | Stephan Lanes¹³  |
 Karine Marinier¹⁴  | Giuseppe Roberto¹  | Kanaka Soman²  |
 Xiaofeng Zhou¹⁵  | Robert Platt¹⁶  | Soko Setoguchi¹⁷  | Gillian C. Hall¹⁸ 

¹ARS Toscana, Florence, Italy

²Division of Pharmacoepidemiology & Clinical Pharmacology, Utrecht University, Utrecht, The Netherlands

³Department of Epidemiology, RTI Health Solutions, Barcelona, Spain

⁴Department of Data Science & Biostatistics, University Medical Center Utrecht, Utrecht, The Netherlands

⁵University of Iceland, Reykjavik, Iceland

⁶Takeda Development Center Americas, Cambridge, Massachusetts, USA

⁷Epidemiology and Benefit Risk, Sanofi, Bridgewater, New Jersey, USA

⁸Center for Observational and Real-World Evidence, MSD, Zürich, Switzerland

⁹Daiichi-Sankyo Europe, Germany

¹⁰Department of Epidemiology, Merck & Co., Inc., Rahway, New Jersey, USA

¹¹Panalgo, USA

¹²Teva Pharmaceutical Industries Ltd., Israel

¹³Carelon Research, Wilmington, Delaware, USA

¹⁴IQVIA, Courbevoie, France

¹⁵Global Medical Epidemiology, Pfizer Inc. New York, USA

¹⁶McGill University, Montreal, Canada

¹⁷Rutgers University, New Jersey, USA

¹⁸Gillian Hall Epidemiology, UK

Correspondence

Rosa Gini, Agenzia regionale di sanità della Toscana, Osservatorio di epidemiologia, Via Pietro Dazzi 1, 50141, Florence, Italy.
 Email: rosa.gini@ars.toscana.it

Funding information

International Society for Pharmacoepidemiology

Abstract

Purpose: Real-world evidence (RWE) is increasingly used for medical regulatory decisions, yet concerns persist regarding its reproducibility and hence validity. This study addresses reproducibility challenges associated with diversity across real-world data sources (RWDS) repurposed for secondary use in pharmacoepidemiologic studies. Our aims were to identify, describe and characterize practices, recommendations and

Prior postings and presentations of the manuscript: (1) Symposium presentation at 13th Asian Conference on Pharmacoepidemiology and 28th Conference of Korean Society for Pharmacoepidemiology and Risk Management (ACPE/KOPERM), October 2021, Seoul, South Korea. (2) Poster presentation and symposium presentation at the 39th International Conference on Pharmacoepidemiology and Therapeutic Risk Management (ICPE), August 24–28, 2022, Copenhagen, Denmark.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Authors. *Pharmacoepidemiology and Drug Safety* published by John Wiley & Sons Ltd.

tools for collecting and reporting diversity across RWDSs, and explore how leveraging diversity could improve the quality of evidence.

Methods: In a preliminary phase, keywords for a literature search and selection tool were designed using a set of documents considered to be key by the coauthors. Next, a systematic search was conducted up to December 2021. The resulting documents were screened based on titles and abstracts, then based on full texts using the selection tool. Selected documents were reviewed to extract information on topics related to collecting and reporting RWDS diversity. A content analysis of the topics identified explicit and latent themes.

Results: Across the 91 selected documents, 12 topics were identified: 9 dimensions used to describe RWDS (organization accessing the data source, data originator, prompt, inclusion of population, content, data dictionary, time span, healthcare system and culture, and data quality), tools to summarize such dimensions, challenges, and opportunities arising from diversity. Thirty-six themes were identified within the dimensions. Opportunities arising from data diversity included multiple imputation and standardization.

Conclusions: The dimensions identified across a large number of publications lay the foundation for formal guidance on reporting diversity of data sources to facilitate interpretation and enhance replicability and validity of RWE.

KEYWORDS

database and multi-database observational studies, diversity in secondary real-world data sources, pharmacoepidemiology, real-world evidence methods, reproducibility, scoping review

Key Points

- This is the first systematic exploration of approaches to describe diversity across data sources generating RWE, and the challenges and opportunities implied by such diversity.
- The review identified 91 documents offering recommendations and examples for characterizing diversity, or leveraging data diversity in pharmacoepidemiologic research.
- Our scoping review provides the basis for a framework to characterize data sources used to generate RWE.
- Further research including formal guidance is needed to enhance replicability and validity of RWE derived from diverse data sources.

Plain Language Summary

Health information collected in electronic databases is commonly used in studies of the safety and effectiveness of medicines. These so-called real world data sources are typically diverse in the kinds of information they contain, as well as their structure and format. Enhancing the understanding of the features of and differences between data sources is essential for researchers to understand how to reproduce studies, and in turn determine whether the results are valid. This is the first systematic exploration of approaches to describe diversity across data sources generating real world evidence (RWE), and of the challenges and opportunities implied by such diversity. The review identified 91 documents offering recommendations and examples for characterizing diversity, or leveraging data diversity in pharmacoepidemiologic research to help researchers interpret or improve the validity of findings. Our scoping review provides the basis for a framework to characterize data sources used to generate RWE.

1 | INTRODUCTION

Real-world evidence (RWE) plays a prominent role in regulatory decision-making throughout the phases of medicines development,¹

especially the post-marketing phase.^{2,3} However, there is still hesitancy to trust the validity of RWE.⁴ One concern is that replicability of RWE remains far from optimal,⁵ yet reproducibility and replicability are necessary to ensure validity of scientific studies.⁶ In some cases,

the variability in results may be authentic, stemming from differences between populations.⁷ Inconsistencies in results, however, are often due to subtle differences in assumptions hidden in differences in study design,^{8–11} or in small differences in implementation details.¹² Strategies to improve reproducibility have been suggested, including standardization of documentation, transparent reporting of design choices,^{13,14,16–20} public sharing of protocols, statistical analysis plans,²¹ and programming code.^{22,23}

Nevertheless, lack of reproducibility continues to be an issue even when study design and implementation are standardized. Conflicting results have been observed between data sources included in multi-database studies (MDS) when analyses are conducted in parallel across more than one data source using a common study protocol and often after conversion to a common data model using the same analytical program.^{12,24–32} When employing an identical study design and implementation on inherently diverse data sources, heterogeneous results across sites may be the consequence of diversity in the RWD and RWD environments rather than authentic differences in populations: variations in how crucial assumptions are met may cause different biases, leading to different results. This may ultimately be a cause of heterogeneity in results observed in MDS and of lack of reproducibility.^{24,29,33–35}

A comprehensive understanding of real-world data sources (RWDS) is therefore essential. However, to our knowledge, there is currently no published guidance on the identification and recording of data source diversity, or on how to leverage this information when interpreting results. In 2020, following three related symposia at its annual conference,^{36–38} this scoping review was funded by ISPE to provide a foundation for developing best practices guidelines.

The primary objective of the scoping review was to identify, describe, and characterize practices, recommendations and tools for collecting and reporting diversity between RWDSs used in pharmacoepidemiologic studies, specifically focusing on different dimensions of data source diversity.

The secondary objective aimed to explore how diversity can be effectively leveraged to enhance the quality of evidence generated in pharmacoepidemiologic studies and facilitate its interpretation.

2 | METHODS

2.1 | Study protocol

This scoping review was conducted following the guidelines of the JBI Manual for Evidence Synthesis³⁹ and reported in accordance with the PRISMA-ScR guideline for scoping reviews (see Supplementary Material S1 for completed checklist). The study protocol received input from experts in various disciplines, including biostatistics, mathematics, pharmacoepidemiology, and medicine. It was registered with the European Union electronic Register of post-authorisation studies, now HMA-EMA Catalogues of real-world data sources and studies, (registration number: EUPAS39757) and is publicly available.⁴⁰

The key methods of the scoping review are summarized below with further details in Supplementary Material S2.

2.2 | Eligibility criteria, search and selection of sources

The authors identified documents that they believed met the inclusion criteria outlined in the protocol. These documents include:

- Documents or published reviews offering recommendations or guidelines for collecting and reporting on the diversity of data sources, tools for describing data sources, or actual descriptions of data sources
- Documents produced by an organization or a collaborative network of organizations engaged in MDS, detailing the data sources contributing to their studies.
- In exceptional circumstances, MDS that provide a substantial description of data sources, or strategies leveraging diversity to enhance evidence quality or interpret results more effectively.

Keywords for the search were established based on the identified documents. Moreover, using the proposed documents as a reference, the exclusion/inclusion criteria were refined, and a selection tool was developed to screen and select documents for inclusion in the review. Each nominated document was reviewed against this tool. Those selected formed the 'core' set of documents for the scoping review, finalizing the selection tool as described below

The keywords from these core documents informed the systematic search conducted in three steps: (1) a snowball search of the core documents' reference lists (2) development of a PubMed search string based on a published strategy⁴¹ and (3) a search for gray literature online. The search process underwent iterative refinement to encompass a minimum of 80% of the core documents covering all publications up to December 2021. Further details on the search are provided in Supplementary Material S2

Each document identified through the literature search underwent a two-step eligibility review using the final selection tool. Initially, titles and, where available, abstracts were screened and excluded based on four criteria: focus on clinical trials; emphasis on statistical methods for heterogeneity of results; lack of methodological focus; or other, to be specified. Subsequently, full texts of the remaining documents were reviewed with the same exclusion criteria, including those that reported relevant information on diversity in data sources (DDS) (e.g., MDS with pertinent DDS descriptions, tools for DDS reporting, and DDS reporting guidelines) or strategies to utilize DDS for evidence enhancement. This process was independently conducted by two reviewers, with any disagreements resolved by a third. The screening tool and results are documented in a Zenodo library (<https://zenodo.org/records/10633913>).

2.3 | Data charting process

Each of the selected documents was then reviewed by one reviewer using a data extraction tool designed to retrieve topics relevant to the two research objectives. The tool included 10 topics associated with the primary objective and 2 topics associated with the secondary

BOX 1 Topics extracted from the documents and analyzed in the scoping review.

Primary objective. Dimensions used or recommended to describe and/or report on diversity across data sources and related methods or tools

Dimensions

Organization. Description of the organization that makes the data accessible for research.

Data originator. Description of the organization that collects the data and for which purpose.

Prompt. Description of the event(s) that prompted the recording of the data.

Inclusion in the population. Description of the event(s) that cause persons to be included in the data source population.

Content. High-level description of the information captured in the data

Data dictionary. Description of the data dictionary, including coding systems or free text.

Time span. Description of the time span when the data source is available.

Healthcare system and culture. Description of the healthcare system and/or the culture of the area where the data source is generated.

Data quality. Description of aspects of data quality of the data source.

Related methods or tools

Summary. Methods or tools to summarize diversity in the above dimensions.

Secondary objective. Data diversity is represented as a challenge and/or opportunity, and how this is addressed/used.

Diversity as a challenge. Is data diversity mentioned as a challenge?

Diversity as an opportunity. Is data diversity leveraged to improve evidence and/or to assist interpretation, and if so, how?

objective (Box 1). Among the topics associated with the primary objectives, nine were dimensions used or recommended to describe diversity, and the remainder referred to methods and tools used or recommended to summarize diversity. Among the nine dimensions, six (Organization, Data originator, Prompt, Inclusion in the population, Content, Data dictionary) were chosen based on the study protocol, and three (Time span, Healthcare system and culture, and Data quality) were added based on reviewers feedback. The topics associated with the secondary objective were Diversity as a challenge and Diversity as an opportunity. Possible responses for all topics comprised: yes, partly, not clear, or no, with supporting text extracted from the document. Besides the topics in Box 1, we extracted the year of publication, affiliation of the first author, and type of document (Review/guideline, Original research article, Other). Finally, the reviewers were invited to describe any other topics found in the documents that could support either study objectives. The extraction tool, including results, is available as Supplementary Material 3 in a Zenodo library (<https://zenodo.org/records/10633913>).

2.4 | Synthesis of results

The extracted documents were described based on their year of publication, categorized as <2013, 2013–2015, 2016–2018, 2019–2021, document type, and affiliation of the first author, categorized as North America, Europe, Asia/Oceania, Africa/South America, and NA.

Furthermore, a content analysis of the 12 topics in Box 1 was conducted. The content analysis is a form of qualitative analysis. It is described as “a method designed to identify and interpret meaning in recorded forms of communication by isolating small pieces of the data that represent salient concepts and then applying or creating a framework to organize the pieces in a way that can be used to describe or explain a phenomenon”.⁴² Text extracted for each topic was independently reviewed by two reviewers. Both reviewers coded from this extracted text recurring explicit or latent concepts that described or provided a deeper understanding of the respective topic. If needed, for example, when extracted text was unclear or insufficiently informative, the original document's full text was evaluated to ensure a thorough comprehension of the authors' rationale and reasoning. Recurring concepts that were identified through review of text extractions from the different articles generated the ‘themes’ as presented in the results. Themes were discussed by each pair of reviewers, and summarized in an overview table, with their description and a point-wise discussion.

3 | RESULTS

3.1 | Search and selection of sources

A total of 91 documents were included in the review: 24 core reference documents and 67 documents found through the literature search. The full list of included documents and selection variables is presented as Supplementary Material 3 in a Zenodo library (<https://zenodo.org/records/10633913>), and the process is summarized in Figure 1.

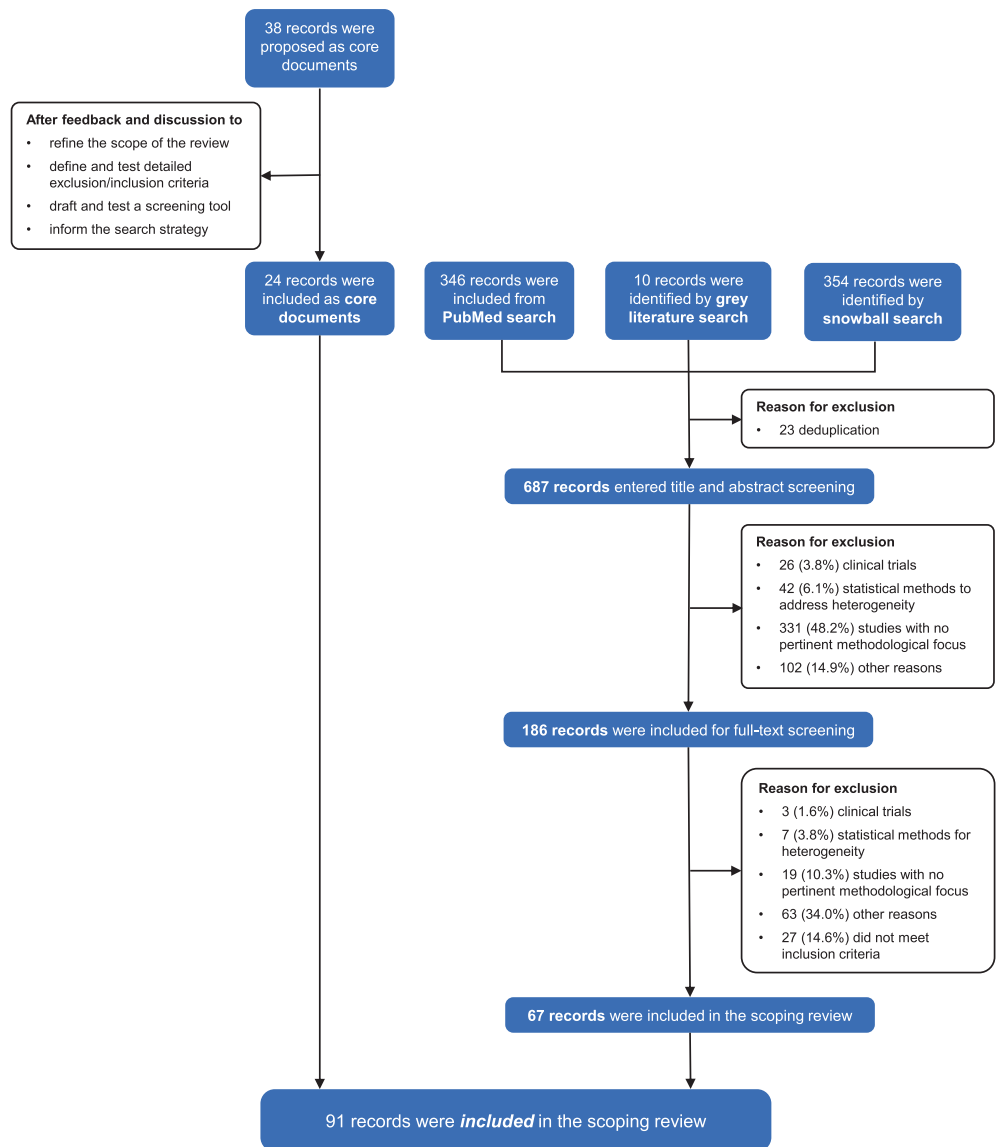
3.2 | Charted data

The number of documents that addressed either study question increased over the search period with 40% published in the last 3 years (see Table 1 and the full dataset in Supplementary Material 4 in the Zenodo library <https://zenodo.org/records/10633913>). Before 2015, most of the documents (<2013, 68.4% and 2013–2015, 64.7%) were authored in North America. Publications from Europe became prevalent in 2016–2018 (61.1%) while those from Asia/Oceania provided a relatively stable contribution (10.5% in the first time period and 10.8% in the last). The first documents identified from Africa/South America were published in 2019–2021. A total of 75 documents addressed the primary research objective, the majority ($N = 39$) were reviews or guidelines. The majority ($N = 30$) of the 59 documents relevant to the secondary research objective were original research articles.

3.3 | Content analysis

The content analysis resulted in identification of three to five themes per dimension, and two to nine themes per each of the other topics

FIGURE 1 Flowchart of the selection of the documents included in the scoping review. On the left column, preliminary selection of the ‘core’ documents. On the right, results from the systematic search and two-steps eligibility review.



(Table 2). Overlapping themes and those noted by reviewers as ‘other’ were all recategorized with the most logical topic. We describe below three examples of how themes were identified and discussed during the content analysis.

An example of a description that was extracted in the topic Prompt reads as follows: “a registry-based study of a drug for rheumatoid arthritis which may focus on EHR records from rheumatology clinics. Some care for this condition, however, may be delivered by primary care doctors whose services might not be tracked in these systems”.⁴³ Several other documents,^{25,34,44–50} highlighted that certain aspects of healthcare may not be adequately captured in certain data sources, leading to missing information. This absence of data might occur within the same data source over time, across different diseases, or among specific population strata, such as individuals with limited health-seeking behavior, for example, those who are unvaccinated. Additionally, data gaps may arise across various data sources due to the specific mechanism that prompts data generation in each source. For instance, data sources relying solely on primary care

diagnoses may overlook acute myocardial infarctions, whereas those with hospital-based diagnoses may miss diabetes diagnoses. In all such instances, it was noted in the literature that the missing data are not randomly distributed, constituting a notable limitation. This set of concepts was identified as a theme. It was labeled as ‘Data are missing not-at-random’ and is described more in detail in Table S1 in Supplementary Material S2.

An example of a description that was extracted in the topic “Diversity as an opportunity” reads as follows: “Multiple imputation adapted to distributed data settings is a feasible method to reduce bias from unmeasured but measurable confounders when at least one database contains the variables of interest.”⁵¹ We found in a second document⁵² a similar idea: that a form of multiple imputation could be enacted across data sources. This concept was identified as a theme. It was labeled as ‘Multiple imputation’ and is described more in detail in Table S1 in Supplementary Material S2.

In the topic ‘Summary’ the analysts chose to explore nine themes, one per dimension. In most dimensions, tools to summarize diversity

TABLE 1 Description of the documents included in the scoping review.

Time period		<2013 (N [%])	2013–2015 (N [%])	2016–2018 (N [%])	2019–2021 (N [%])	Total (N [%])
N		19	17	18	37	91
Type of document	Review/guideline	8 (42.1)	8 (47.1)	7 (38.9)	19 (51.4)	42 (46.2)
	Original research article	7 (36.8)	7 (41.2)	10 (55.6)	13 (35.1)	37 (40.7)
	Other	4 (21.1)	2 (11.8)	1 (5.6)	5 (13.5)	12 (13.2)
Affiliation first author	North America	13 (68.4)	11 (64.7)	3 (16.7)	12 (32.4)	39 (42.9)
	Europe	4 (21.1)	5 (29.4)	11 (61.1)	16 (43.2)	36 (39.6)
	Asia/Oceania	2 (10.5)	1 (5.9)	4 (22.2)	4 (10.8)	11 (12.1)
	Africa/South America				2 (5.4)	2 (2.2)
	NA				3 (8.1)	3 (3.3)
Primary objective. Dimensions used or recommended to describe and/or report on diversity across data sources and related methods or tools						
Number of documents included		17 (89.5)	16 (94.1)	14 (77.8)	28 (75.7)	75 (82.4)
Type of document	Review/guideline	8 (47.1)	7 (43.8)	7 (50.0)	17 (60.7)	39 (52.0)
	Original research article	5 (29.4)	7 (43.8)	6 (42.9)	7 (25.0)	25 (33.3)
	Other	4 (23.5)	2 (12.5)	1 (7.1)	4 (14.3)	11 (14.7)
Secondary objective. Data diversity is represented as a challenge and/or opportunity, and how this is addressed/used						
Number of documents included		12 (63.2)	11 (64.7)	13 (72.2)	23 (62.2)	59 (64.8)
Type of document	Review/guideline	3 (25.0)	6 (54.5)	2 (15.4)	9 (39.1)	20 (33.9)
	Original research article	6 (50.0)	4 (36.4)	10 (76.9)	10 (43.5)	30 (50.8)
	Other	3 (25.0)	1 (9.1)	1 (7.7)	4 (17.4)	9 (15.3)

were not found, and concepts and ontologies to articulate diversity were rarely available. An example where concepts seemed consistent was the topic Time span, and they included: time since data source inception, frequency of updating, time of last update, time spans included in the study, and person time.

The full results from the content analysis of all topics are available in Table S1 in Supplementary Material S2. The tools used in the content analysis are available as Supplementary Material 5 in a Zenodo library (<https://zenodo.org/records/10633913>).

4 | DISCUSSION

To the best of our knowledge, this is the first systematic exploration of approaches to describe diversity across data sources generating RWE, and the challenges and opportunities associated with such diversity. The review identified 91 documents offering recommendations or examples for characterizing diversity or leveraging data diversity in pharmacoepidemiologic research. Prior to 2013, most of these documents were published in North America, but they have gradually expanded to the rest of the world and have become more frequent recently. The study identified nine dimensions to characterize

diversity and investigated tools to summarize them. Challenges and opportunities to leverage data diversity were also identified.

Describing a data source across the nine dimensions clarifies assumptions that can be made when reusing the data for research. For example, describing which events prompt the recording of a diagnosis allows us to clarify whether it can be assumed that chronic diseases are recorded immediately after diagnosis, or a delay should be assumed. The themes identified in each dimension highlight areas where common assumptions may fail to be supported by diverse data. Dimensions and themes provide the basis for a framework to characterize data sources used to generate RWE.

Data sources used to generate RWE are typically embedded in a healthcare system, which is shaped by the policies and culture of the respective country. These data are generated by specific organizations for specific purposes, which dictate who is included in the data source and determine which events prompt data records to be created, as well as how and by whom they are created. Organizations that access the data to conduct studies and produce evidence are (in most cases) different from the data generators, and may have different levels of expertise in research, leading to differences in their ability to identify assumptions, strengths, and limitations for reuse in the context of a specific study.

TABLE 2 Themes identified by the content analysis with in each topic of Box 1.

Primary objective. Dimensions used or recommended to describe and/or report on diversity across data sources and related methods or tools			
Organization. Description of the organization which makes the data accessible for research	Data originator. Description of the organization which collects the data and for which purpose	Prompt. Description of the event(s) that prompt the recording of the data	Inclusion in the population. Description of the event(s) that cause persons to be included in the data source population
Research database partner Governance/accessibility Provenance/funding	(Non-national) Health care provider-specific originator Research specific originator National level originator Regional level originator Private & payer originators	Prompts implicit Types of prompts mentioned Availability of variables entangled with data being non missing Data are missing not-at-random	Qualitative reasons to be included Data source entry and exit dates not mentioned Data sources as dynamic cohorts
Data dictionary. Description of the data dictionary, including coding systems or free text	Time span. Description of the time span when the data source is available	Healthcare system and culture. Description of the healthcare system and/or the culture of the area where the data source is generated	Data quality. Description of aspects of data quality of the data source
International National Regulatory Free text	Frequency of data refresh Data collection start Lag time/date of last collection Duration of follow-up	Health care system general Health care system access Reimbursement and/or regulation Not provided	Completeness/missingness Validity Quality governance Data governance
Secondary objective. Diversity is represented as a challenge and/or opportunity, and how this is addressed/used			
Diversity as a challenge. Is diversity mentioned as a challenge?			
Reasons for diversity General discussion on diversity	Multiple imputation Standardized processes Understanding the impact of diversity in healthcare system and culture on results Plan methods to address data diversity		Content. High-level description of the information captured in the data
Diversity as an opportunity. Is diversity leveraged to improve evidence and/or to assist interpretation, and if so, how?			
Summary. Method or tools to summarize diversity in the above dimensions			
Summarize diversity in organizations that make data accessible, organizations that generated the data, prompts, populations, content, dictionary, time span, culture/healthcare system, and data quality			

The content analysis showed a different level of complexity across the study topics: the results of Content and Data dictionary were lists of themes and none of them had controversial discussion points. This may be because these dimensions are also common in descriptions of prospectively collected data from a study cohort. On the contrary, in the topics Prompt and Inclusion in the population, some themes were more complex. This could be attributed to the fact that the former concept does not exist in the field of primary data collection, and the latter has a different nuance when referring to a data source population rather than a study cohort.

In these topics, we found that a distinction is often not made between the description of data sources, the data source instances extracted for the purpose of a study, and the variables derived from the data source instance. For example, in the topic Inclusion in the population, the theme 'Data source entry and exit dates not mentioned' found cases where dates of entry and exit from the study were mentioned, with no apparent awareness that a study cohort derived from the secondary use of a data source must be nested in the data source in the first place.^{43,47,53,54}

In one paper,⁵⁵ a statement implied that the date of entry in the data source was not retrieved from the data source itself but was instead chosen as the date of the first record prompted in the data source ('population of all patients with at least one record in the database'), and there was an acknowledgement that this may induce important bias. Without transparent reporting on such elements, it becomes challenging to determine their impact on study results, which may ultimately hamper their reproducibility.

At the inception of this scoping review, the guidance available for study reporting offered only rudimentary directions for the characterization of data sources. The RECORD-PE checklist, aimed at study reports, does advocate for the description of the health-care system and the mechanism of data generation, yet its focus is narrowly confined to measuring drug exposure.⁵⁶ A more recent advancement is the HARPER template,¹⁹ which consolidates and updates four earlier templates, including the ENCePP checklist for study protocols.⁵⁷ This template introduces a section specifically dedicated to the description of data sources, comprising both an unstructured entry and a detailed metadata list aimed at facilitating the extraction of the study cohort from the data source. Our proposed framework is designed to bring structure to this unstructured entry, offering a systematic approach to data source characterization.

Furthermore, the structured methodology for identifying fit-for-purpose data, as recently proposed by Gatto and colleagues,^{15,16,17} underscores the necessity of having comprehensive information about potential data sources. This requirement is to ensure they meet the minimum criteria for validly capturing design elements. Our framework not only aligns with this need but also furnishes the essential metadata in a structured format, thereby facilitating a more informed assessment of fitness-for-purpose of data sources for generation of reliable RWE.

In the topic Summary, our main finding was that tools and ontologies for describing data diversity are lacking in most of the nine

dimensions of our framework. Among recent attempts to establish such ontologies, the European Medicines Agency launched an online catalogue of data sources on December 4, 2023, which includes information compatible with our nine dimensions.^{58,59} It relies on an even more comprehensive metadata list developed by the MINERVA project.^{60,61}

The challenges identified in the scoping review were not unexpected and can be attributed to our understanding of the complexity and how we handle its consequences. Dealing with complexity presents an opportunity to enhance the standardization of processes at multiple levels. When examining this opportunity, we highlighted examples of generalization of the multiple imputation methodology to the case of MDS. We also identified awareness of the potential to assess the influence of local culture on study outcomes, which is partially mediated by data diversity. Our scientific community is still poised to develop methods that can fully capitalize on both of these opportunities.

4.1 | Strengths and limitations

The major strengths of this scoping review are a systematic process and transparency, with a pre-specified JBI compliant,³⁹ publicly registered protocol. The involvement of a diverse group of ISPE members with varied backgrounds facilitated contributions to the selection of core documents and enabled content analysis to be completed in pairs. This literature search was restricted to articles within the field of pharmacoepidemiology, aligning with our primary focus. Thus, we may have overlooked pertinent studies from other relevant fields. For practical purposes, a large team of reviewers conducted the data extraction and coding. As a result, the quantitative analysis performed was less extensive than originally planned in the protocol, due to inconsistencies in the extraction of some categorical variables. To prevent this in the content analysis, reviewers frequently referred back to the original articles to locate the necessary information.

5 | CONCLUSION

This systematic and comprehensive study has successfully identified the dimensions that characterize diverse data sources used to generate RWE, facilitating a better understanding and interpretation of the results. However, in most dimensions, tools to summarize diversity of data sources were not found, and ontologies to articulate diversity were rarely available. The findings of this study will establish the foundation for formal guidance on the reporting and conduct of studies utilizing diverse RWDS.

AUTHOR CONTRIBUTIONS

RG, RP, and GCH conceptualized the study. RG, RP, and GCH contributed critical direction and overall project administration. All authors participated in data collection and data verification. RG, RP, GCH, HG,

and DB drafted the initial manuscript draft document. All authors participated in the critical revisions of this article for important intellectual content and approved the final version.

FUNDING INFORMATION

In 2020, following three related symposia at its annual conference, a proposal for funding was approved by ISPE to develop guidance on representing and reporting on data source diversity. This scoping review provides a foundation for such guidance.

CONFLICT OF INTEREST STATEMENT

RG is employed by ARS Toscana, a publicly owned agency that participates in studies funded by pharmaceutical companies and compliant with the ENCePP Code of Conduct. The budget of her unit is partially supported by such studies. CG is an employee of MSD Innovation and Development, Zurich, Switzerland. MSD Innovation and Development did not have any involvement in the study. Lin Li is employed by Sanofi. Mehmet Burcu is an employee of Merck & Co., Inc., Rahway, NJ, United States and owns stocks in Merck & Co., Inc., Rahway, NJ, United States. XZ is an employee of Pfizer Inc. and has received stock from Pfizer Inc.

DATA AVAILABILITY STATEMENT

The dataset used during the current study is available within this article and/or figures, tables, and supplementary material.

ETHICS STATEMENT

The authors state that no ethical approval was needed as this research did not involve human subjects nor was any human subject data captured, extracted, or analyzed.

ORCID

Rosa Gini  <https://orcid.org/0000-0002-6250-877X>
 Romin Pajouheshnia  <https://orcid.org/0000-0002-4208-3583>
 Helga Gardarsdottir  <https://orcid.org/0000-0001-5623-9684>
 Dimitri Bennett  <https://orcid.org/0000-0002-8387-9342>
 Lin Li  <https://orcid.org/0000-0002-3972-735X>
 Claudia Gulea  <https://orcid.org/0000-0001-9607-5901>
 Angelika Wientzek-Fleischmann  <https://orcid.org/0000-0002-5778-3283>
 Marloes T. Bazelier  <https://orcid.org/0000-0002-0795-1381>
 Mehmet Burcu  <https://orcid.org/0000-0003-4572-0987>
 Caitlin Dodd  <https://orcid.org/0000-0002-8784-696X>
 Carlos E. Durán  <https://orcid.org/0000-0003-1853-3641>
 Sigal Kaplan  <https://orcid.org/0000-0002-3352-8480>
 Stephan Lanes  <https://orcid.org/0000-0001-9536-6643>
 Karine Marinier  <https://orcid.org/0000-0002-8715-2220>
 Giuseppe Roberto  <https://orcid.org/0000-0001-6478-6442>
 Kanaka Soman  <https://orcid.org/0009-0006-8966-3944>
 Xiaofeng Zhou  <https://orcid.org/0000-0002-2109-0012>
 Robert Platt  <https://orcid.org/0000-0002-5981-8443>
 Soko Setoguchi  <https://orcid.org/0000-0002-1583-752X>
 Gillian C. Hall  <https://orcid.org/0000-0001-5081-3710>

REFERENCES

- Eskola SM, Leufkens HGM, Bate A, De Bruin ML, Gardarsdottir H. Use of real-world data and evidence in drug development of medicinal products centrally authorized in Europe in 2018–2019. *Clin Pharmacol Ther.* 2022 Jan;111(1):310–320. doi:10.1002/cpt.2462
- Cave A, Kurz X, Arlett P. Real-world data for regulatory decision making: challenges and possible solutions for Europe. *Clin Pharmacol Ther.* 2019 Jul;106(1):36–39.
- Federal Register. Considerations for the Use of Real-World Data and Real-World Evidence To Support Regulatory Decision-Making for Drug and Biological Products; Draft Guidance for Industry. 2021 Accessed March 2024. Available from: <https://www.federalregister.gov/documents/2021/12/09/2021-26640/considerations-for-the-use-of-real-world-data-and-real-world-evidence-to-support-regulatory>
- Orsini LS, Berger M, Crown W, et al. Improving transparency to build Trust in Real-World Secondary Data Studies for hypothesis testing—why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Value Health.* 2020;23(9):1128–1136.
- Madigan D, Ryan PB, Schuemie M, et al. Evaluating the impact of database heterogeneity on observational study results. *Am J Epidemiol.* 2013;178(4):645–651.
- Patil P, Peng RD, Leek JT. A statistical definition for reproducibility and replicability. *bioRxiv.* 2016;66803. doi:10.1101/066803v1 Accessed March 2024.
- Pottegård A, Pedersen SA, Schmidt SAJ, et al. Use of hydrochlorothiazide and risk of skin cancer: a nationwide Taiwanese case-control study. *Br J Cancer.* 2019;121(11):973–978.
- Hernán MA, Robins JM. Using big data to emulate a Target trial when a randomized trial is not available. *Am J Epidemiol.* 2016;183(8):758–764.
- Reynolds RF, Kurz X, de Groot MCH, et al. The IMI PROTECT project: purpose, organizational structure, and procedures. *Pharmacoepidemiol Drug Saf.* 2016;25(Suppl 1):5–10.
- Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, Bartels DB. Transparency and reproducibility of observational cohort studies using large healthcare databases. *Clin Pharmacol Ther.* 2016;99(3):325–332.
- Wang SV, Sreedhara SK, Schneeweiss S. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nat Commun.* 2022;13(1):5126.
- Klungel OH, Kurz X, de Groot MCH, et al. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. *Pharmacoepidemiol Drug Saf.* 2016;25(S1):156–165.
- Wang S, Schneeweiss S, Berger M, et al. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. *Pharmacoepidemiol Drug Saf.* 2017;26(9):1018–1032.
- Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ.* 2021;12(372):m4856.
- Gatto NM, Wang SV, Murk W, et al. Visualizations throughout pharmacoepidemiology study planning, implementation, and reporting. *Pharmacoepidemiol Drug Saf.* 2022;31(11):1140–1152. doi:10.1002/pds.5529
- Gatto NM, Vititoe SE, Rubinstein E, Reynolds RF, Campbell UB. A structured process to identify fit-for-purpose study design and data to generate valid and transparent real-world evidence for regulatory uses. *Clin Pharmacol Ther.* 2023;113(6):1235–1239.
- Gatto NM, Campbell UB, Rubinstein E, Jaksa A, Mattox P, Mo J, et al. The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. *Clinical Pharmacology & Therapeutics.* 2022;111(1):122–134.

18. Hansford HJ, Cashin AG, Jones MD, et al. Development of the TrAnSPARENT ReportinG of observational studies emulating a TARGET trial (TARGET) guideline. *BMJ Open*. 2023;13(9):e074626.
19. Wang SV, Pottegård A, Crown W, et al. HARmonized protocol template to enhance reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: a good practices report of a joint ISPE/ISPOR task force. *Pharmacoepidemiol Drug Saf*. 2023 Jan; 32(1):44-55.
20. Desai RJ, Wang SV, Sreedhara SK, et al. Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA sentinel innovation center. *BMJ*. 2024;12(384):e076460.
21. Gini R, Fournie X, Dolk H, et al. The ENCePP code of conduct: a best practise for scientific independence and transparency in noninterventional postauthorisation studies. *Pharmacoepidemiol Drug Saf*. 2019; 28(4):422-433.
22. Goldacre B, Morton CE, DeVito NJ. Why researchers should share their analytic code. *BMJ*. 2019;367:l6365. doi:10.1136/bmj.l6365
23. Weberpals J, Wang SV. The FAIRification of research in real-world evidence: a practical introduction to reproducible analytic workflows using Git and R. *Pharmacoepidemiol Drug Saf*. 2024;33(1):e5740. doi: 10.1002/pds.5740 Accessed March 2024.
24. Dieleman J, Romio S, Johansen K, et al. Guillain-Barre syndrome and adjuvanted pandemic influenza A (H1N1) 2009 vaccine: multinational case-control study in Europe. *BMJ*. 2011;343:d3908.
25. Platt RW, Dormuth CR, Chateau D, Filion K. Observational studies of drug safety in multi-database studies: methodological challenges and opportunities. *EGEMS (Wash DC)*. 2016;4(1):1221.
26. Filion KB, Chateau D, Targownik LE, et al. Proton pump inhibitors and the risk of hospitalisation for community-acquired pneumonia: replicated cohort studies with meta-analysis. *Gut*. 2014;63(4): 552-558.
27. Requena G, Huerta C, Gardarsdottir H, et al. Hip/femur fractures associated with the use of benzodiazepines (anxiolytics, hypnotics and related drugs): a methodological approach to assess consistencies across databases from the PROTECT-EU project. *Pharmacoepidemiol Drug Saf*. 2016;25(Suppl 1):66-78.
28. Lai ECC, Shin JY, Kubota K, et al. Comparative safety of NSAIDs for gastrointestinal events in Asia-Pacific populations: a multi-database, international cohort study. *Pharmacoepidemiol Drug Saf*. 2018 Nov; 27(11):1223-1230.
29. Willame C, Dodd C, Durán C, et al. Background rates of 41 adverse events of special interest for COVID-19 vaccines in 10 European healthcare databases—an ACCESS cohort study. *Vaccine*. 2023;41(1):251-262. Accessed March 2024. Available from: <https://www.sciencedirect.com/science/article/pii/S0264410X22014293>
30. Bots SH, Riera-Arnau J, Belitser SV, et al. Myocarditis and pericarditis associated with SARS-CoV-2 vaccines: a population-based descriptive cohort and a nested self-controlled risk interval study using electronic health care data from four European countries. *Front Pharmacol*. 2022;13. Accessed March 2024. doi:10.3389/fphar.2022.1038043
31. Gini R, Sturkenboom MCJ, Sultana J, et al. Different strategies to execute multi-database studies for medicines surveillance in real-world setting: a reflection on the European model. *Clin Pharmacol Ther*. 2020;108(2):228-235.
32. Platt R, Platt R, Brown J, Henry D, Klungel O, Suissa S. How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias. *Pharmacoepidemiol Drug Saf*. 2020;29:3-7.
33. Roberto G, Leal I, Sattar N, et al. Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the EMIF project. *PLoS One*. 2016;11(8):e0160648.
34. Gini R, Dodd CN, Bollaerts K, et al. Quantifying outcome misclassification in multi-database studies: the case study of pertussis in the ADVANCE project. *Vaccine*. 2020;38(Suppl 2): B56-B64.
35. Russek M, Quinten C, de Jong VMT, Cohet C, Kurz X. Assessing heterogeneity of electronic health-care databases: a case study of background incidence rates of venous thromboembolism. *Pharmacoepidemiol Drug Saf*. 2023;32(9):1032-1048. doi:10.1002/pds.5631 Accessed March 2024.
36. Gini R, Lanes S, Bollaerts K, Trifirò G, Kirchmayer U, Hall GC. Data diversity in multi-database pharmacoepidemiologic studies and its role in outcome misclassification: a curse or a blessing? In: abstracts of the 35th international conference on pharmacoepidmiology & therapeutic risk management. *Pharmacoepidemiol Drug Saf*. 2017; 26(S2):3-636.
37. Lai EC, Man K, Toh D, Platt R, Hallas J, Setoguchi S. Heterogeneity and validity in national and international multi-database pharmacoepidemiologic studies: lessons learned in North America, Europe, and Asia. In: abstracts of the 35th international conference on pharmacoepidmiology & therapeutic risk Management. *Pharmacoepidemiol Drug Saf*. 2017;26(S2):3-636.
38. Pajouheshnia R, Gardarsdottir H, Platt R, Toh D, Klungel O. Analysis of data from distributed pharmacoepidemiologic networks. In: abstracts of the 35th international conference on pharmacoepidmiology & therapeutic risk management. *Pharmacoepidemiol Drug Saf*. 2017;26(S2):3-636.
39. Peters MD, Godfrey C, Mclnerney P, Baldini Soares C, Khalil H, Parker DA. Chapter 11: scoping reviews. In: Munn Z, ed. *JBI Reviewer's manual*. Joanna Briggs Institute (JBI); 2017.
40. Gini R, Pajouheshnia R, Platt R, Setoguchi S, Hall G. DIVERSE project: protocol for the scoping review. EUPAS39757. Accessed March 2024. https://catalogues.ema.europa.eu/sites/default/files/document_files/DIVERSE_protocol_v1.0.pdf
41. Hunt NB, Gardarsdottir H, Bazelier MT, Klungel OH, Pajouheshnia R. A systematic review of how missing data are handled and reported in multi-database pharmacoepidemiologic studies. *Pharmacoepidemiol Drug Saf*. 2021;30:819-826. doi:10.1002/pds.5245
42. Kolbe RH, Burnett MS. Content-analysis research: an examination of applications with directives for improving research reliability and objectivity. *J Consumer R*. 1991;18(2):243-250.
43. Petruski-Ivleva I. The epidemiology of databases: part I: four principles of working with real-world data. Accessed March 2024. <https://aetion.com/evidence-hub/the-epidemiology-of-databases-part-i-four-principles-of-working-with-real-world-data>
44. Robb MA, Racoosin JA, Worrall C, Chapman S, Coster T, Cunningham FE. Active surveillance of postmarket medical product safety in the Federal Partners' collaboration. *Med Care*. 2012 Nov; 50(11):948-953.
45. Ferrer P, Ballarin E, Sabate M, et al. Sources of European drug consumption data at a country level. *Int J Public Health*. 2014;59(5): 877-887.
46. Panaccio MP, Cummins G, Wentworth C, et al. A common data model to assess cardiovascular hospitalization and mortality in atrial fibrillation patients using administrative claims and medical records. *Clin Epidemiol*. 2015;7:77-90.
47. Sills MR, Kwan BM, Yawn BP, et al. Medical home characteristics and asthma control: a prospective, observational cohort study protocol. *EGEMS (Wash DC)*. 2013;1(3):1032.
48. Thurin NH, Pajouheshnia R, Roberto G, et al. From inception to ConcePTION: genesis of a network to support better monitoring and communication of medication safety during pregnancy and breastfeeding. *Clin Pharmacol Ther*. 2022;111(1):321-331.
49. Schmidt M, Schmidt S, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. *Clin Epidemiol*. 2019;11:563-591.

50. Sturkenboom M, Braeye T, van der Aa L, et al. ADVANCE database characterisation and fit for purpose assessment for multi-country studies on the coverage, benefits and risks of pertussis vaccinations. *Vaccine*. 2020;22(38):B8-B21.
51. Secrest MH, Platt RW, Reynier P, Dormuth CR, Benedetti A, Filion KB. Multiple imputation for systematically missing confounders within a distributed data drug safety network: a simulation study and real-world example. *Pharmacoepidemiol Drug Saf*. 2020;29(Suppl 1): 35-44.
52. Reps JM, Rijnbeek PR, Ryan PB. Supplementing claims data analysis using self-reported data to develop a probabilistic phenotype model for current smoking status. *J Biomed Inform*. 2019;97:103264.
53. Ankrah D, Hallas J, Odei J, Asenso-Boadi F, Dsane-Selby L, Donneyong M. A review of the Ghana National Health Insurance Scheme claims database: possibilities and limits for drug utilization research. *Basic Clin Pharmacol Toxicol*. 2019;124(1): 18-27.
54. Seesaghur A, Petruski-Ivleva N, Banks V, et al. Real-world reproducibility study characterizing patients newly diagnosed with multiple myeloma using clinical practice research datalink, a UK-based electronic health records database. *Pharmacoepidemiol Drug Saf*. 2021; 30(2):248-256.
55. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323-337.
56. Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ*. 2018;14(363): k3532.
57. ENCePP. ENCePP Checklist for study protocols (revision 4). 2018 Accessed March 2024. https://encepp.europa.eu/encepp-toolkit/encepp-checklist-study-protocols_en
58. European Medicines Agency. HMA/EMA Big Data Stakeholder Forum. 2023 Accessed March 2024. <https://www.ema.europa.eu/en/events/hma-ema-big-data-stakeholder-forum-2023>
59. HMA-EMA Catalogues of real-world data sources and studies. Accessed March 2024. catalogues.ema.europa.eu
60. MINERVA. Full metadata list. Accessed March 2024. https://catalogues.ema.europa.eu/sites/default/files/document_files/MINERVA%20Full-metadata-list%2010Jan2022.pdf
61. Pajouheshnia R, Gini R, Gutierrez L. MINERVA study—Supplementary Material 2: Final Metadata List. 2023 Accessed March 2024. <https://zenodo.org/records/10422428>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gini R, Pajouheshnia R, Gardarsdottir H, et al. Describing diversity of real world data sources in pharmacoepidemiologic studies: The DIVERSE scoping review. *Pharmacoepidemiol Drug Saf*. 2024;33(5): e5787. doi:10.1002/pds.5787