Describing Diversity of Real-World Data Sources: A New Framework Adapted From the DIVERSE Initiative Scoping Review

Romin Pajouheshnia,[1,2] Rosa Gini,[3] Helga Gardarsdottir,[2,4,5] Dimitri Bennett,[6] Lin Li,[7] Claudia Gulea,[8] Angelika Wientzek-Fleischmann,[9] Marloes T. Bazelier,[2] Mehmet Burcu,[10] Caitlin Dodd,[11] Carlos E. Durán,[4] Sigal Kaplan,[12] Stephan Lanes,[13] Karine Marinier,[14] Giuseppe Roberto,[3] Kanaka Soman,[2] Xiaofeng Zhou,[15] Robert Platt,[16] Soko Setoguchi,[17] Gillian C. Hall[18]

[1] Department of Epidemiology, RTI Health Solutions, Barcelona, Spain; [2] Utrecht University, Utrecht, The Netherlands; [3] ARS Toscana, Florence, Italy; [4] University Medical Center Utrecht, Utrecht, The Netherlands; [5] University of Iceland, Reykjavik, Iceland; [6] Takeda Development Center Americas, Cambridge, MA, US; [7] Epidemiology and Benefit Risk, Sanofi, Bridgewater, NJ, US; [8] Center for Observational and Real-World Evidence, MSD, Zurich, Switzerland; [9] Daiichi-Sankyo Europe, Munich, Germany; [10] Department of Epidemiology, Merck & Co., Inc., Rahway, NJ, US; [11] Panalgo, Boston, MA, US; [12] Teva Pharmaceutical Industries Ltd., Netanya, Israel; [13] Carelon Research, Wilmington, DE, US; [14] IQVIA, Paris, France; [15] Global Medical Epidemiology, Pfizer Inc., New York, NY, US; [16] McGill University, Montreal, Canada; [17] Rutgers University, New Brunswick, NJ, US; [18] Gillian Hall Epidemiology, London, UK

DISCLOSURES

RPa is an employee of RTI Health Solutions, a unit of RTI International, a nonprofit organization that conducts work for government, public, and private organizations, including pharmaceutical companies. RG is employed by ARS Toscana, a publicly owned agency that participates in studies funded by pharmaceutical companies and is compliant with the ENCePP Code of Conduct. The budget of her unit is partially supported by such studies. DB is an employee of Takeda. LL is an employee of Sanofi and may hold shares and/or stock options in the company. CG is an employee of MSD Innovation and Development, Zurich, Switzerland. MSD Innovation and Development did not have any involvement in the study. MBu is an employee of and owns stocks in Merck & Co., Inc., Rahway, NJ, US. SK is an employee of Teva Pharmaceutical Industries Ltd. KM was an employee of IQVIA at the time of conducting the study. GR is a consultant for ARS Toscana. XZ is an employee of Pfizer Inc. and has received stock from Pfizer Inc. RPI has received fees, outside of the scope of this work, for consulting from Biogen, Merck, and Pfizer.

Funding statement: This work was supported by a manuscript proposal grant to Dr. Pajouheshnia and Dr. Gini from the International Society for Pharmacoepidemiology (ISPE).

Registration: The study protocol is publicly available in the HMA-EMA Catalogue of real-world studies (registration number: EUPAS39757).

BACKGROUND

Variation in results across studies may stem from differences in foundational characteristics between real-world data sources (RWDS). This study addresses the reproducibility challenges associated with data diversity across RWDS in pharmacoepidemiologic studies.

OBJECTIVES

To identify and characterize practices, tools, and recommendations for describing and reporting diversity between RWDS used in pharmacoepidemiologic studies.

METHODS

The full methods and results of this study are available online at Pharmacoepidemiology and Drug Safety.[1]

A scoping review was conducted. First, a 'core' set of documents was identified by co-authors based on their expert subject knowledge, including:

- Documents offering recommendations, guidelines, or tools for collecting/reporting diversity of data sources.
- Documents describing data sources produced by organizations/networks that conduct multi-database studies.
- Multi-database studies with substantial description of data sources or strategies to leverage data source diversity.

A systematic literature search was conducted in 3 steps, and documents were screened and selected using a standardized selection tool. Information was recorded by reviewers using an extraction tool designed to capture information on topics related to collecting and reporting RWDS diversity. Content analysis was conducted to identify common themes within the topics.

DISCUSSION

- The 9 dimensions provide a framework to describe data sources when reporting future studies.
- 36 common themes highlighted common terminologies and understandings, as well as inconsistencies and misconceptions when describing diverse RWDS.
- Dimensions, such as Content and Data dictionary, had the most consistent and simple themes.
- Others, including Prompt and Inclusion in the population, were more complex and were approached less consistently across articles.
- Describing these dimensions will clarify assumptions that are made when interpreting studies using single or multiple routinely collected health data.

Reference

1. Gini R, Pajouheshnia R, Gardarsdottir H, Bennett D, Li L, Gulea C, et al. Describing diversity of real world data sources in pharmacoepidemiologic studies: the DIVERSE scoping review. Pharmacoepidemiol Drug Saf. 2024 May;33(5):e5787. doi:10.1002/pds.5787.

Contact Information

Romin Pajouheshnia, PhD
Research Epidemiologist

RTI Health Solutions
Av. Diagonal, 605, 9-1
08028 Barcelona, Spain

Telephone: +34.933.62.28.18
Email: rpajouheshnia@rti.org

Scan the QR code to access an online version of the poster

RESULTS

Ninety-one articles were included after screening. Mostly review articles and original research articles were included, primarily from North America and Europe (see Figures 1, 2A, 2B).

Nine dimensions to describe RWDS were identified along with 36 associated themes (see Figure 3). No tools to summarize diversity were found for most of the 9 dimensions to describe data source diversity; there were few examples of concepts or ontologies.

Examples from the content analysis:

Example 1: **Prompt**

Several articles reported that information may be missing from data sources when specific aspects of healthcare are not captured completely. Gaps may arise across data sources due to the specific mechanism that prompts data generation in each data source resulting in information that is missing-not-at-random.

Example 2: **Inclusion in the population**

Four groups of conditions for in/exclusion from a data source were found, which could form the basis of an ontology: inhabitance of a country/region, enrollment in a health plan, hospital assistance, and primary care assistance.

The concept of date of entry and exit from a data source was commonly missing. However, several articles did distinguish a study population as being nested within an underlying population with the potential to have a record in the data source.

Figure 1. Flowchart of the Selection of the Documents Included in the Scoping Review

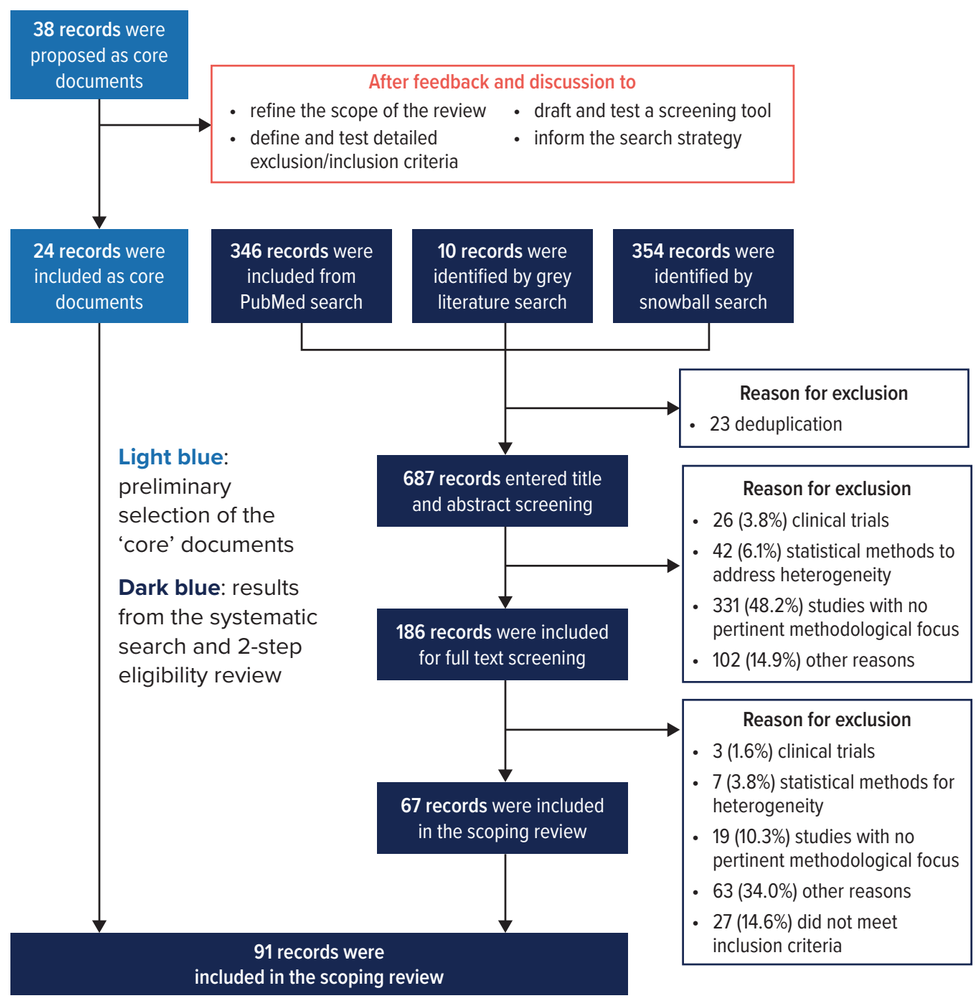

Figure 2. Overview of the Articles Included in the Scoping Review

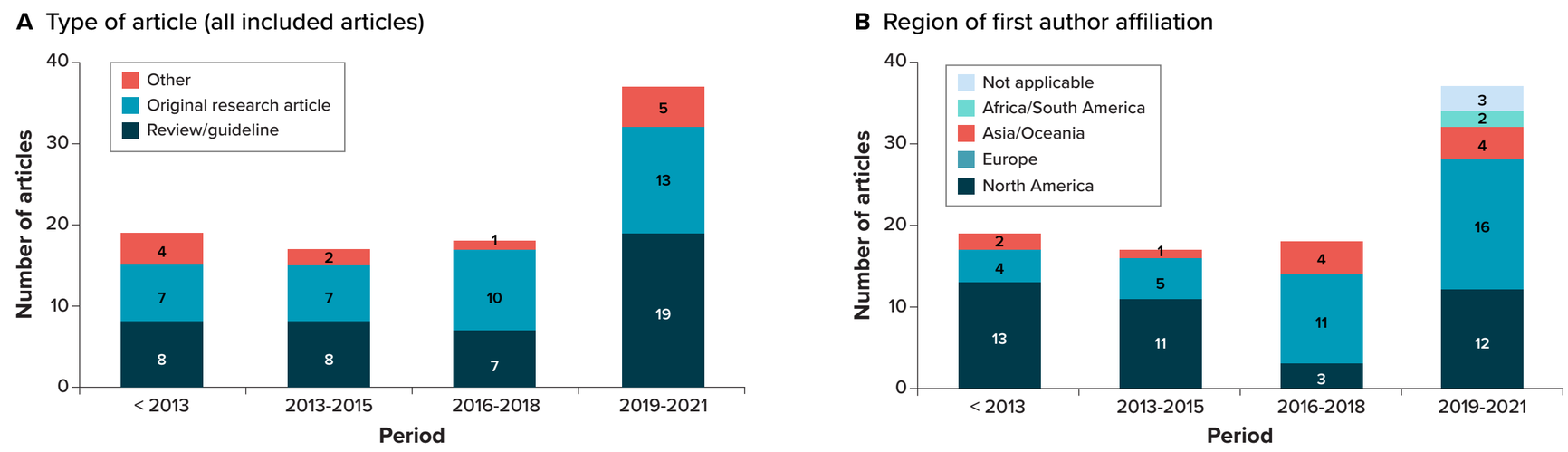

Figure 3. Nine Dimensions to Describe Diverse Data Sources and Related Themes

Dimensions		Themes		
Organization	Description of the organization that makes the data accessible for research	• Research database partner • Governance/accessibility • Provenance/funding		
Data originator	Description of the organization that collects the data and for which purpose	• (Non-national) Healthcare provider–specific originator • Research specific originator	...	• National level originator • Regional level originator • Private & payer originators
Prompt	Description of the event(s) that prompted the recording of the data	• Prompts implicit • Types of prompts mentioned	...	• Availability of variables entangled with data being nonmissing • Data are missing not-at-random
Inclusion in the population	Description of the event(s) that cause persons to be included in the data source population	• Qualitative reasons to be included • Data source entry and exit dates not mentioned	...	• Data sources as dynamic cohorts
Content	High-level description of the information captured in the data	• Socio-demographics • Behavioral • Clinical	...	• Death • Other
Data dictionary	Description of the data dictionary, including coding systems or free text	• International • National • Regulatory • Free text		
Time span	Description of the time span when the data source is available	• Frequency of updating • Time since data source inception	...	• Lag time/date of last update • Timespans within a study (e.g., duration of follow-up)
Healthcare system and culture	Description of the healthcare system and/or the culture of the area where the data source is generated	• Healthcare system general • Healthcare system access	...	• Reimbursement and/or regulation • Not provided
Data quality	Description of aspects of data quality of the data source	• Completeness /missingness • Validity • Quality governance • Data governance		

CONCLUSIONS

- The 9 dimensions provide a framework for describing data sources, thereby further facilitating reproducibility and the interpretation of studies.
- More formal guidance is being developed by the DIVERSE initiative, complementing concurrent initiatives.
- We encourage journals to include the DIVERSE framework in their guidelines for authors for submissions of RWE studies.

Presented at: 2024 ISPE Annual Meeting; 24-28 August 2024; Berlin, Germany