Pharmacoepidemiology and Drug Safety

WILEY

EDITORIAL

Advancing Research Transparency and Reproducibility in Pharmacoepidemiology

Shirley V. Wang[1,2] | Anton Pottegård[3]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Boston, Massachusetts, USA | [2]Department of Msedicine, Harvard Medical School, Boston, Massachusetts, USA | [3]Clinical Pharmacology, Pharmacy and Environmental Medicine, Department of Public Health, University of Southern Denmark, Odense, Denmark

Correspondence: Shirley V. Wang (swang1@bwh.harvard.edu)

Received: 24 December 2024 | **Accepted:** 26 December 2024

Funding: The authors received no specific funding for this work.

While there is universal agreement about the importance of transparent and reproducible science, the building blocks of open science practices have not yet been built into routine workflows for many pharmacoepidemiology and outcomes researchers. In light of the great potential for real-world evidence generated from real-world data to influence public health and policy, this special issue of *Pharmacoepidemiology and Drug Safety* (PDS) called for papers that advance research reproducibility in pharmacoepidemiology.

Several themes emerged in this 19 paper special issue, including (1) understanding data sources and assessing data quality, (2) development of reusable software tools and code sharing, (3) clarity in measurement of key study parameters, (4) demonstration of replicability of results across networks, and (5) infrastructure and processes for conducting reproducible research.

1 | Understanding Data Sources and Assessing Data Quality

A comprehensive understanding of the data is vital for ensuring that a given real-world data (RWD) source is both relevant and reliable for a specific research question. While there are numerous databases that could potentially be used for any given study, with great diversity in data structure and data quality, in practice only a limited set are accessible to the study team. This necessitates careful evaluation to ensure that the database(s) being considered are fit for purpose. The two studies in this theme underscore the importance of clearly describing data sources, their characteristics, and their alignment with the intended research objectives to improve transparency and reproducibility.

Gini et al. [1] present the DIVERSE scoping review, identifying 12 dimensions for describing real-world data sources, including data origin, quality, and content, and highlighting opportunities for improving evidence through better characterization of data heterogeneity. Rivera et al. [2] describe the Oncology QCARD Initiative, which provides a structured framework to conduct a high-level assessment of the fitness of RWD for oncology research as part of early engagement between a sponsor and scientific reviewer, emphasizing data relevance, reliability, and study design transparency in an initial study proposal.

2 | Development of Reusable Software Tools and Code Sharing

Reusable tools play a pivotal role in advancing research efficiency and reproducibility. By providing standardized and accessible frameworks, such tools enable consistent implementation of methods across diverse datasets, fostering transparency and collaboration in pharmacoepidemiology. Proper documentation is critical to facilitate their usability, adaptability, and continual improvement over time. Within this theme, five studies introduce innovative open-source software tools addressing specific research needs and one reviews the prevalence of code sharing in pharmacoepidemiology. Collectively, these papers illustrate the potential of open science practices to democratize access to advanced methodologies and promote robust evidence generation.

https://doi.org/10.1002/pds.70096

Russo and Wang present an open-source R package for Tree-Based Scan Statistics (TBSS) [3], replicating prior results from previous proprietary software, providing users with a framework from which to build new innovations for detecting adverse drug effects through hierarchical outcome screening. Dernie et al. [4] describe a reproducible process for phenotyping in the DARWIN EU network, involving code list generation, review, and diagnostics, a process which facilitates consistency and traceability. Burkard et al. [5] develop standardized formulas to calculate daily drug doses using the OMOP CDM, which were applicable to > 85% of records across diverse databases, with calculated median daily doses that aligned with WHO-defined daily doses. Raventos et al. [6] introduce the IncidencePrevalence R package for estimating disease incidence and prevalence in OMOP-formatted datasets, showing high agreement with published studies and supporting timely epidemiological research. Shen et al. [7] propose a two-step validation framework for comparing causal estimates from observational data and RCTs, accompanied by the RCTrep R package to facilitate implementation. Tazare et al. [8] analyze trends in programming code sharing in pharmacoepidemiology, finding limited adoption, but offering specific recommendations for improving transparency and reproducibility through better documentation and open science practices.

3 | Clarity in Measurement of Key Study Parameters

Validation studies are an important backbone for secondary data studies, strengthening the foundation of pharmacoepidemiology and real-world evidence research. Researchers who undertake these endeavors play a critical role in advancing the reproducibility and replicability of scientific studies. Under this theme, three studies address the validity of algorithms and definitions used in our studies.

Yap et al. [9] highlight variability in algorithm performance by validating rule-based algorithms for detecting major and clinically relevant non-major bleeding events in electronic health records, demonstrating high sensitivity and negative predictive value, which contrasted with lower sensitivity when relying solely on diagnosis codes. Campos et al. [10] tested two validated algorithms for identifying breast cancer incidence across diverse datasets and populations, finding significant variability in performance, and highlighting the importance of assessing algorithms in new contexts to ensure reliability. Ericksen et al. [11] conducted a systematic review to standardize diagnostic codes for sickle cell disease complications, providing a harmonized set of ICD codes to enhance transparency and reduce heterogeneity in real-world evidence studies.

4 | Demonstration of Replicability of Results Across Networks

Collaborative use of federated data networks strengthens the reliability of research findings by enabling replication across diverse data sources. However, effective collaboration requires clear governance, harmonized protocols, and awareness of potential pitfalls such as inconsistent data quality or misaligned analytical practices. The four studies in this theme exemplify efforts to overcome these challenges, providing insights into the benefits and complexities of leveraging federated networks for reproducible research.

Van Baalen et al. [12] discussed how disease-specific federated data networks like HONEUR for multiple myeloma and PHederation for pulmonary hypertension improved transparency and reproducibility in rare disease research by harmonizing data and employing distributed analytics. Gillies et al. [13] conducted a multi-center replication study using Australian dispensing data to assess metformin treatment dynamics, revealing significant variability in results across sites due to differing operational definitions and emphasizing the importance of detailed analytical protocols. Conover et al. [14] evaluated the reproducibility and generalizability of results regarding GLP-1 receptor agonists' effects on chronic lower respiratory disease using the OHDSI network, confirming the robustness of findings while identifying variations within drug classes. Their evaluation provides an example of how standardized tools within distributed data networks can be used to evaluate the replicability of results from RWE studies. Rai et al. [15] described the U.S. FDA's Sentinel System, highlighting how its structured framework, including data harmonization and a standardized querying system, ensures transparency, reproducibility, and replicability of drug safety studies across a national network of data partners, a system that is being used to generate regulatory-grade evidence at scale.

5 | Infrastructure and Process for Conducting Reproducible Research

Transparent and reproducible workflows are foundational to advancing the credibility and reliability of pharmacoepidemiology research. Establishing robust infrastructure and standardized processes promotes clarity in methodologies, ensures accountability, and facilitates collaboration among researchers. The four studies in this theme showcase innovative approaches and tools that address reproducibility challenges and best practices for implementing open and reproducible research.

Abdelaziz et al. [16] demonstrated a step-by-step approach for using R and parquet file formats to manage and analyze large real-world data efficiently, highlighting significant reductions in data size and improved performance compared to traditional SAS workflows. Muntner et al. [17] introduced the concepts of “staging” and “clean rooms” to safeguard the integrity of real-world data analyses by structuring multi-stage analyses with restricted data access and decision-making processes to reduce bias and increase transparency. Nab et al. [18] described the OpenSAFELY platform, designed for secure and reproducible research using electronic health records, offering tools to standardize workflows, enable public code sharing, and provide audit trails for increased transparency. Weberpals and Wang [19] provided a practical tutorial for implementing FAIR (Findable, Accessible, Interoperable, and Reproducible) analytic workflows in real-world evidence studies using Git and R, emphasizing the advantages of version control systems for collaboration and transparency.

6 | Shaping a Transparent Future

This special issue of *Pharmacoepidemiology and Drug Safety* represents an important step forward in the journey toward more transparent and reproducible pharmacoepidemiological research. Across its 19 papers, contributors have explored themes ranging from data quality and reusable tools to the replicability of results across different data sources to leveraging new infrastructure for reproducibility. This body of work offers concrete solutions and frameworks to address the challenges our field faces. As the impact of real-world evidence on healthcare practice and policy continues to expand, the key battle in the coming years will be to embed transparency and reproducibility as a core element of routine research conduct.

The responsibility to lead this charge will increasingly fall to new generations of epidemiologists, whose ingenuity and commitment to open science will shape the future of our discipline. Within the themes discussed in this issue, these researchers have a rich foundation to build from, whether by innovating reusable tools, enhancing the rigor of data evaluation, or fostering collaboration across networks. We are both hopeful and confident that junior researchers will champion a more transparent and reproducible epidemiology, working with key stakeholders to drive progress in methods, practices, and culture. This special issue underscores the many paths forward and serves as both a resource and a call to action for all who aspire to make our science better.

Conflicts of Interest

SVW reports ad hoc consulting for Exponent Inc, Cytel Inc, and MITRE a federally funded research and development center for the Centers for Medicare and Medicaid on unrelated work.

References

1. R. Gini, R. Pajouheshnia, H. Gardarsdottir, et al., "Describing Diversity of Real World Data Sources in Pharmacoepidemiologic Studies: The DIVERSE Scoping Review," *Pharmacoepidemiology and Drug Safety* 33, no. 5 (2024): e5787, https://doi.org/10.1002/pds.5787.

2. D. R. Rivera, J. C. Eckert, C. Rodriguez-Watson, et al., "The Oncology QCARD Initiative: Fostering Efficient Evaluation of Initial Real-World Data Proposals," *Pharmacoepidemiology and Drug Safety* 33, no. 11 (2024): e5818, https://doi.org/10.1002/pds.5818.

3. M. Russo and S. V. Wang, "An Open-Source Implementation of Tree-Based Scan Statistics," *Pharmacoepidemiology and Drug Safety* 33, no. 3 (2024): 5765, https://doi.org/10.1002/pds.5765.

4. F. Dernie, G. Corby, A. Robinson, et al., "Standardised and Reproducible Phenotyping Using Distributed Analytics and Tools in the Data Analysis and Real World Interrogation Network (DARWIN EU)," *Pharmacoepidemiology and Drug Safety* 33, no. 11 (2024): e70042, https://doi.org/10.1002/pds.70042.

5. T. Burkard, K. López-Güell, A. Gorbachev, et al., "Calculating Daily Dose in the Observational Medical Outcomes Partnership Common Data Model," *Pharmacoepidemiology and Drug Safety* 33, no. 6 (2024): e5809, https://doi.org/10.1002/pds.5809.

6. B. Raventós, M. Català, M. Du, et al., "IncidencePrevalence: An R Package to Calculate Population-Level Incidence Rates and Prevalence Using the OMOP Common Data Model," *Pharmacoepidemiology and Drug Safety* 33, no. 1 (2024): e5717, https://doi.org/10.1002/pds.5717.

7. L. Shen, E. Visser, F. van Erning, G. Geleijnse, and M. Kaptein, "A Two-Step Framework for Validating Causal Effect Estimates," *Pharmacoepidemiology and Drug Safety* 33, no. 9 (2024): 5873, https://doi.org/10.1002/pds.5873.

8. J. Tazare, S. V. Wang, R. Gini, et al., "Sharing Is Caring? International Society for Pharmacoepidemiology Review and Recommendations for Sharing Programming Code," *Pharmacoepidemiology and Drug Safety* 33, no. 9 (2024): e5856, https://doi.org/10.1002/pds.5856.

9. A. J. Y. Yap, D. C. H. Teo, P. S. Ang, et al., "Validation of a Major and Clinically Relevant Nonmajor Bleeding Phenotyping Algorithm on Electronic Health Records," *Pharmacoepidemiology and Drug Safety* 33, no. 8 (2024): 5875, https://doi.org/10.1002/pds.5875.

10. A. Campos, R. Ramasubramanian, C. Wong, and A. F. Marcus, "Application of Multiple Validated Algorithms for Identifying Incident Breast Cancer Among Individuals With Atopic Dermatitis," *Pharmacoepidemiology and Drug Safety* 33, no. 5 (2024): 5808, https://doi.org/10.1002/pds.5808.

11. P. N. Ericksen, F. Dabbous, R. Ghosh, et al., "Standardization of Coding Definitions for Sickle Cell Disease Complications: A Systematic Literature Review," *Pharmacoepidemiology and Drug Safety* 33, no. 9 (2024): e5769, https://doi.org/10.1002/pds.5769.

12. V. Van Baalen, E. M. Didden, D. Rosenberg, et al., "Increase Transparency and Reproducibility of Real-World Evidence in Rare Diseases Through Disease-Specific Federated Data Networks," *Pharmacoepidemiology and Drug Safety* 33, no. 4 (2024): e5778, https://doi.org/10.1002/pds.5778.

13. M. B. Gillies, C. Bharat, X. Camacho, et al., "Medicine Utilization Studies in Australian Individual-Level Dispensing Data: A Blinded, Multi-Center Replicated Analysis," *Pharmacoepidemiology and Drug Safety* 33, no. 3 (2024): e5776, https://doi.org/10.1002/pds.5776.

14. M. M. Conover, Y. Albogami, J. Hardin, et al., "Glucagon-Like Peptide 1 Receptor Agonists and Chronic Lower Respiratory Disease Among Type 2 Diabetes Patients: Replication and Reliability Assessment Across a Research Network," *Pharmacoepidemiology and Drug Safety* 34, no. 1 (2025): e70087, https://doi.org/10.1002/pds.70087.

15. A. Rai, J. C. Maro, S. Dutcher, P. Bright, and S. Toh, "Transparency, Reproducibility, and Replicability of Pharmacoepidemiology Studies in a Distributed Network Environment," *Pharmacoepidemiology and Drug Safety* 33, no. 6 (2024): e5820, https://doi.org/10.1002/pds.5820.

16. A. I. Abdelaziz, K. A. Hanson, C. E. Gaber, and T. A. Lee, "Optimizing Large Real-World Data Analysis With Parquet Files in R: A Step-By-Step Tutorial," *Pharmacoepidemiology and Drug Safety* 33, no. 3 (2024): 5728, https://doi.org/10.1002/pds.5728.

17. P. Muntner, R. K. Hernandez, S. T. Kent, et al., "Staging and Clean Room: Constructs Designed to Facilitate Transparency and Reduce Bias in Comparative Analyses of Real-World Data," *Pharmacoepidemiology and Drug Safety* 33, no. 3 (2024): e5770, https://doi.org/10.1002/pds.5770.

18. L. Nab, A. L. Schaffer, W. Hulme, et al., "OpenSAFELY: A Platform for Analysing Electronic Health Records Designed for Reproducible Research," *Pharmacoepidemiology and Drug Safety* 33, no. 6 (2024): e5815, https://doi.org/10.1002/pds.5815.

19. J. Weberpals and S. V. Wang, "The FAIRification of Research in Real-World Evidence: A Practical Introduction to Reproducible Analytic Workflows Using Git and R," *Pharmacoepidemiology and Drug Safety* 33, no. 1 (2024): 5740, https://doi.org/10.1002/pds.5740.