Sharing is caring? A literature review of programming code sharing in pharmacoepidemiological studies

Date:16 May 2023

Version: v3.0

Protocol Authors and Affiliations:

Anna Schultze*, PhD, Assistant Professor at London School of Hygiene and Tropical Medicine, London, UK.
John Tazare*, PhD, Assistant Professor at London School of Hygiene and Tropical Medicine, London, UK.
Shirley V. Wang, PhD, Assistant Professor at Brigham and Women's Hospital, Harvard Medical School. Boston, USA.
Daniel Prieto Alhambra, PhD, Professor of Pharmaco- and Device Epidemiology, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.
Ian Douglas, PhD, Professor at London School of Hygiene and Tropical Medicine, London, UK.
Sebastian Schneeweiss, ScD, Professor at Brigham and Women's Hospital, Harvard Medical School. Boston, USA. Co-Founder and Advisor at Aetion, New York, USA.
Peter Arlett, Head, Data Analytics and Methods Task Force at European Medicines Agency, Netherlands.
Rosa Gini, PhD, Head of the Pharmacoepidemiology Unit at Agenzia Regionale di Sanita della Toscana, Florence, Italy.
Caroline Morton, Epidemiologist/Software Developer, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK.
John Logie, PhD, Director within Real World Data and Analytics Team, Value Evidence and Outcomes, GlaxoSmithKline, London, UK.
Jennifer Popovic, PhD, Senior Director and Therapy Area Lead, Oncology, GlaxoSmithKline, London, UK.
Katherine Donegan, PhD, Head of Epidemiology, UK Medicines and Healthcare Products Regulatory Agency, London, UK.

**joint project leads*

Protocol History

This section will detail any protocol updates. Please note that minor updates - for example to correct errors will result in an increment update only (from v1.0 to v1.1). Major protocol changes, classified as those which will impact the study conduct or analysis will result in a new protocol version (from v1.0 to v2.0).

Protocol Location/Section	***Change***	***Rationale***
V1.0 (v1.0 folder on google drive)	*n/a*	*Original draft shared with the working group*
V2.0 (v2.0 folder on google drive)	*Update extraction fields, add other open science indicators and clarify the phrasing of some fields*	*To incorporate comments from the working group on the original draft.*
V3.0	*Added country of databases used in studies as an extraction field*	*Following conversation with the ENCEPP sharing initiative, to enable an assessment as to whether open science indicators vary according to the regulatory setting*

Introduction

The issue of scientific reproducibility has received increased attention in the past decade, as replication studies of a range of disciplines have estimated that a high proportion of published scientific findings cannot be independently reproduced[1,2]. The REPEAT initiative recently found that reproducibility appears higher in studies using electronic healthcare databases than in other disciplines, but noted that more transparent reporting might aid future replication efforts[3]. Several tools can help promote transparent conduct and reporting of pharmacoepidemiology studies, including the ENCePP code of conduct[4], use of standardized protocol templates[3], study visualizations[5], and reporting checklists and guidance[6,7].

The additional sharing of research materials, which for most pharmacoepidemiology studies involves patient-level data and the programming code, could further help facilitate computational reproduction efforts. Although data sharing in studies using administrative health data is rarely feasible due to data protection legislation, programming code sharing often is. Making programming code publicly available also has other potential benefits, including facilitating the detection of programming errors[8]. Well documented and commented programming code could also help clarify the study implementation, encourage greater re-use of code and promote the uptake of novel analytical methods [9–11]. Another dimension of interest in terms of transparency is whether the code comes with a ‘Sandbox’ environment, including synthetic data, that allows verifying in practice what the program does. Notably, this option may be possible even in case the code is not open itself, and the program is shared with no source code.

Code sharing is increasingly advocated in applied health and methodological research[12], and with journals such as PLOS Medicine and International Journal of Epidemiology mandating the publication of code on acceptance this is an area which is likely to grow rapidly in importance for many pharmacoepidemiologists[13,14]. Despite this, there has been little research on code sharing in either applied or methodological pharmacoepidemiology.

Our aim is to conduct a literature review to quantify the extent of, and trends in, code sharing in pharmacoepidemiology. The review will form part of broader programme of work exploring code sharing in our discipline.

Objectives

Our primary objective is to quantify the number and proportion of papers in a key pharmacoepidemiology journal (Pharmacoepidemiology and Drug Safety [PDS]) which published all or part of their programming code over the period 2017 - 2022.

The secondary objectives are:

- To describe the prevalence of code sharing according to key paper characteristics (article type, funding source, and whether the research topic covered COVID-19)
- Among papers who shared all or part of their code, to describe the method/platform used to share the code
- To describe the prevalence of other open research practices (study pre-registration, protocol sharing, data sharing, code-list sharing, use of checklists in the reporting, preprinting)

Note, for the final secondary objective we will rely on information reported in the paper itself, as cross-referencing other databases (ENCePP, or medRxiv) would be outside the scope of this review.

Methods

Search strategy and study selection

Publication and Timeframe

We will include articles published in Pharmacoepidemiology and Drug Safety (PDS) after the 1 January 2017 and before the 31 December 2022.

Types of studies

We will include articles of the following PDS article types:

- Original research articles
- Review articles
- Brief reports
- Real-World data sources

Commentaries and letters to the editor will be excluded, as will articles where there is no analysis of data (real or simulated). Duplicate publications will be excluded if they occur.

Study identification

We will use the following search string:

'"Pharmacoepidemiology and drug safety"[Journal] AND (("2017/01/01"[Date - Publication] : "2022/12/31"[Date - Publication]))'

We will consider using the R packages (for example, {easyPubMed}[15]) to scrape paper dois and basic information.

Data Collection

Collection and Processing of Potentially Eligible Studies

The list of potentially eligible publication DOIs will be exported as a csv, and stored for record keeping. We will then screen each abstract for eligibility. The relevance of code sharing is likely to vary depending on article type, but it is likely not possible to assess whether code sharing would be applicable to a given paper on the basis of information provided in the abstract alone. Article eligibility will therefore be determined in two steps:

1. **Abstract screen:** Exclude non-eligible article types and articles for which code sharing is clearly not applicable.
    - Article eligibility will be categorised as YES (include), NO (exclude) and MAYBE (screen the full-text)
2. **Full-text screen:** this will assess whether code sharing is applicable by considering whether the article involves analysis of real or simulated data, or a description of an algorithm or method which requires writing programming code.
    - Article eligibility categorised as YES (include) or NO (exclude)

The resulting list of eligible papers will be saved, and the selection process displayed as a flowchart.

The eligibility screen will be done by a single reviewer as there is anticipated to be limited uncertainty regarding the eligibility criteria. Where there is uncertainty, papers will be discussed by the reviewers (AS and JT) and a joint decision taken on whether to include that article. All joint decisions will be documented for transparency.

Data Extraction from Eligible Studies

Basic information, including publication year, author list and doi, will be extracted on each paper. We will use a google form to extract additional information requiring manual screening. A draft of the data extraction elements are provided in appendix 1. The primary outcome is “programming code sharing’, defined as some or all of the code used to process and/or analyse the data being publicly available without requiring further contact with the corresponding author.

Data Analysis

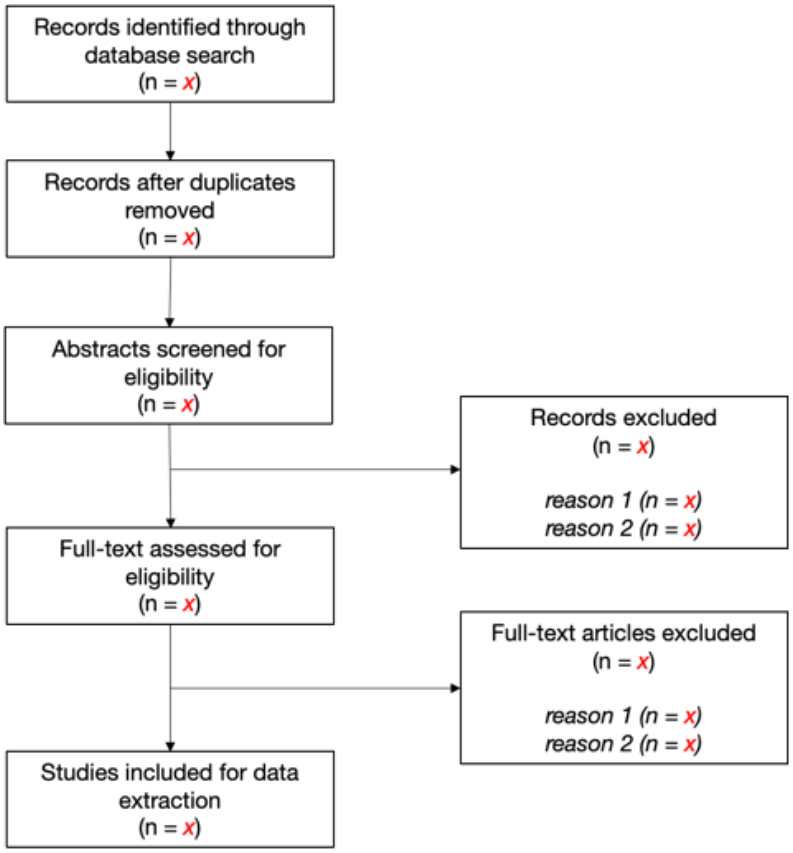
We will describe the prevalence of programming code sharing overall, as well as over calendar time and pipeline stage. Where programming code is shared, we will also describe basic information on the programming code sharing, including where the programming code was shared, and whether any instructions for how to run the programming code are provided. For the secondary objectives, we will describe the prevalence of programming code sharing according to the paper characteristics:

- Type of paper (applied vs. methodological)
- Author affiliation (at least one author with industry affiliation vs not)
- Funding (industry vs not)
- Funding (funder which requires programming code sharing vs not)

Descriptive tables and figures will be generated. The raw data, clean data, and programming code will be shared on github under an MIT license.

Table Shells

[F1] Flowchart of study selection



[T1] Characteristics of Included Papers

		N (%)
Publication year	*2017*	
	2018	
	etc	
Publication type (PDS category)	*Original research*	
	Review	
	etc	
Article type (applied vs. methodological)	*Applied*	

	Methodological	
COVID-19 related research	*Yes*	
Countries covered by the database(s)	*(list)*	
Author affiliation (any author)	*Academic*	
	Industry	
	Regulatory	
	Other	
Industry funding	*Yes*	
Funder which mandates code sharing	*Yes*	
Other Transparency Practices (as mentioned in article)	*At least one*	
	Study preregistration	
	Protocol sharing	
	etc	

[T2] Code sharing characteristics

		N (%)
Did the paper share code?	*Yes*	
	No	
Among those who shared,		
Where was code shared?	*Web Appendix hosted by Journal*	
	OSF	
	etc	
What fileformat was shared?	*.pdf*	
	.R	

	etc	
What was the programming language?	*R*	
	SAS	
	etc	
Was there a description of how to run the code?	*Yes, a README*	
	Yes, comments	
	None	
What type of code was shared?	*All*	
	Partial - data management	
	Partial - analysis	
	Other	

[T3] Code Sharing Over Time

	N eligible papers	**N shared code**	**% shared code**
Year			
2017			
2018			
Etc			

[F2] Point and line chart of code sharing over time

[T4] Code sharing by other characteristics

	N eligible papers	**N shared code**	**% shared code**
Paper type			
Affiliation			
Others..			

[T5] Other transparency practices over time

	N eligible papers	Study pre-registered		Shared protocol		etc
Year		**N**	**%**	**N**	**%**	
2017						
2018						
etc						

[F3a-x]. Point and line charts of other transparency practices over time

Repeat F2 with other outcomes

[T5] Code sharing by other transparency practices

	N eligible papers	N shared code	% shared code
None			
At least one			
Study pre-registered			
Shared study protocol			
etc			

References

1. Errington TM, Mathur M, Soderberg CK, et al. Investigating the replicability of preclinical cancer biology. Pasqualini R, Franco E, eds. *eLife*. 2021;10:e71601. doi:10.7554/eLife.71601
2. Nosek BA, Lakens D. Registered reports: A method to increase the credibility of published results. *Soc Psychol*. 2014;45:137-141. doi:10.1027/1864-9335/a000192
3. Wang S, Pottegård A, Crown W, et al. HARmonized Protocol Template to Enhance Reproducibility (HARPER) of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force. Published online May 6, 2022. Accessed September 27, 2022. https://osf.io/6qxpf/
4. ENCePP Code of Conduct. Published September 27, 2022. Accessed September 27, 2022. https://www.encepp.eu/code_of_conduct/
5. 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
6. 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;372:m4856. doi:10.1136/bmj.m4856
7. 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;363:k3532. doi:10.1136/bmj.k3532
8. Vable AM, Diehl SF, Glymour MM. Code Review as a Simple Trick to Enhance Reproducibility, Accelerate Learning, and Improve the Quality of Your Team's Research. *Am J Epidemiol*. 2021;190(10):2172-2177. doi:10.1093/aje/kwab092
9. Morton C, Devito N, Morley J, et al. Software development skills for health data researchers. *BMJ Health Care Inform*. 2022;29(1):e100488. doi:10.1136/bmjhci-2021-100488
10. Goldacre B, Morton CE, DeVito NJ. Why researchers should share their analytic code. *BMJ*. 2019;367:l6365. doi:10.1136/bmj.l6365
11. Cadarette SM, Ban JK, Consiglio GP, et al. Diffusion of Innovations model helps interpret the comparative uptake of two methodological innovations: co-authorship network analysis and recommendations for the integration of novel methods in practice. *J Clin Epidemiol*. 2017;84:150-160. doi:10.1016/j.jclinepi.2016.12.006
12. Towards open health analytics: our guide to sharing code safely on GitHub – NHS-R Community. Accessed September 27, 2022. https://nhsrcommunity.com/towards-open-health-analytics-our-guide-to-sharing-code-safely-on-github/
13. Materials, Software and Code Sharing | PLOS Medicine. Accessed September 15, 2022. https://journals.plos.org/plosmedicine/s/materials-software-and-code-sharing
14. General Instructions. Oxford Academic. Accessed September 15, 2022. https://academic.oup.com/ije/pages/general_instructions
15. Fantini. easyPubMed: Search and Retrieve Scientific Publication Records from PubMed. Published online 2019. https://CRAN.R-project.org/package=easyPubMed

Appendix I: Extraction Fields

Field	**Format**
Eligibility	
Journal	Text
Publication Date	Date
Publication type	PDS categories (Original Research, Review, Brief Report, Real-World Data, Letter to Editor/Commentary)
Could the article share programming code?*	Yes/No
**defined as the article involving analysis of real or simulated data, or a description of an algorithm or method which requires writing programming code*	
Assess eligibility before additional data extraction	
Article Characteristics	
Article URL	
DOI	
Title	
Authors	
Article type	"Applied research"/"Methodological"/"Both"
Is the analysis based on simulated data?	Yes/No/Partially
Author affiliation (any)	Industry/Academic/Regulatory (Tick yes if more than one author had an affiliation in the category)
What was the funder?	Freetext, more than one possible
Which database(s) was/were used?	Freetext, more than one possible
Which country/ies does the database cover?	Freetext, more than one possible
Is this COVID-19 related research?*	Yes/No
**defined as including 'COVID-19' or 'SARS-CoV2' in the title or abstract*	
Code Sharing	
Published code is linked and/or referenced in the text:	Yes/No/"Available on request"
Published code is directly accessible without further contacting authors	Yes/No
If code is accessible, then:	
Where is code shared?	Freetext
In what format is code shared?	Filetype
What parts of the analysis pipeline are covered by the code?	Freetext
Does the code include instructions for how it should be run?	Yes in a separate file/Yes in comments/No
Does the article contain synthetic data?	Yes/No
If yes, does the code run?	Yes/No
What language or platform was used for the analyses?	1) R 2) Stata 3) SAS 4) Python 5) Aetion 6) TriNetX 7) other, write in (more than one option possible)

ISPE Programming Code Sharing
Study Protocol

Is the programming code open source?	Yes/no/Impossible to assess
If yes, what is the license?	Freetext
Other ORPs	
Was the paper pre-printed?	Yes/No
Are underlying data shared?	Yes - raw data/Yes - analytical files/No, but simulated data is made available/No, but data access procedures are described/"available on request"/No
Are codelists shared?	Yes/No/"Available on request"
Scope of codelists shared	Exposure/Outcome/All Variables
Is the study pre-registered?	Yes/No
Is the study protocol or SAP shared?	Yes/No/"Available on request"
Did the study report adhering to reporing guidelines such as RECORD?	Yes/No/Not applicable to study type
If yes, what guidelines	Freetext