VALUE IN HEALTH 20 (2017) 1009–1022

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Original Report

Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0

Shirley V. Wang, PhD, ScM[1,2,], Sebastian Schneeweiss, MD, ScD[1,2], Marc L. Berger, MD[3], Jeffrey Brown, PhD[4], Frank de Vries, PharmD, PhD[5], Ian Douglas, PhD[6], Joshua J. Gagne, PharmD, ScD[1,2], Rosa Gini, PhD, MSc[7], Olaf Klungel, PhD[8], C. Daniel Mullins, PhD[9], Michael D. Nguyen, MD[10], Jeremy A. Rassen, ScD[11], Liam Smeeth, MSc, PhD[6], Miriam Sturkenboom, PhD, MSc[12], on behalf of the joint ISPE-ISPOR Special Task Force on Real World Evidence in Health Care Decision Making*

[1]*Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, MA, USA;* [2]*Department of Medicine, Harvard Medical School, MA, USA;* [3]*Pfizer, NY, USA;* [4]*Department of Population Medicine, Harvard Medical School, MA, USA;* [5]*Department of Clinical Pharmacy, Maastricht UMC+, The Netherlands;* [6]*London School of Hygiene and Tropical Medicine, England, UK;* [7]*Agenzia regionale di sanità della Toscana, Florence, Italy;* [8]*Division of Pharmacoepidemiology & Clinical Pharmacology, Utrecht University, Utrecht, Netherlands;* [9]*Pharmaceutical Health Services Research Department, University of Maryland School of Pharmacy, MA, USA;* [10]*FDA Center for Drug Evaluation and Research, USA;* [11]*Aetion, Inc., NY, USA;* [12]*Erasmus University Medical Center Rotterdam, Netherlands*

ABSTRACT

Purpose: Defining a study population and creating an analytic dataset from longitudinal healthcare databases involves many decisions. Our objective was to catalogue scientific decisions underpinning study execution that should be reported to facilitate replication and enable assessment of validity of studies conducted in large healthcare databases. **Methods:** We reviewed key investigator decisions required to operate a sample of macros and software tools designed to create and analyze analytic cohorts from longitudinal streams of healthcare data. A panel of academic, regulatory, and industry experts in healthcare database analytics discussed and added to this list. **Conclusion:** Evidence generated from large healthcare encounter and reimbursement databases is increasingly being sought by decision-makers. Varied terminology is used around the world for the same concepts. Agreeing on terminology and which parameters from a large catalogue are the most essential to report for replicable research would improve transparency and facilitate assessment of validity. At a minimum, reporting for a database study should provide clarity regarding operational definitions for key temporal anchors and their relation to each other when creating the analytic dataset, accompanied by an attrition table and a design diagram.

A substantial improvement in reproducibility, rigor and confidence in real world evidence generated from healthcare databases could be achieved with greater transparency about operational study parameters used to create analytic datasets from longitudinal healthcare databases.

Keywords: Transparency, reproducibility, replication, healthcare databases, pharmacoepidemiology, methods, longitudinal data.

Contributors to the joint ISPE-ISPOR Special Task Force on Real World Evidence in Health Care Decision Making paper co-led by Shirley V. Wang and Sebastian Schneeweiss. The writing group contributors are the following: Marc L. Berger, Jeffrey Brown, Frank de Vries, Ian Douglas, Joshua J. Gagne, Rosa Gini, Olaf Klungel, C. Daniel Mullins, Michael D. Nguyen, Jeremy A. Rassen, Liam Smeeth and Miriam Sturkenboom. The contributors who participated in small group discussion and/or provided substantial feedback prior to ISPE/ISPOR membership review are the following: Andrew Bate, Alison Bourke, Suzanne Cadarette, Tobias Gerhard, Robert Glynn, Krista Huybrechts, Kiyoshi Kubota, Amr Makady, Fredrik Nyberg, Mary E Ritchey, Ken Rothman and Sengwee Toh. Additional information is listed in Appendix.

This article is a joint publication by Pharmacoepidemiology and Drug Safety and Value in Health.

* *Address correspondence to*: S. V. Wang, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, United States.

E-mail: swang1@bwh.harvard.edu

1098-3015$36.00 – see front matter © 2017 Published by Elsevier Inc. on behalf of International Society for Pharmacoeconomics and Outcomes Research (ISPOR).

http://dx.doi.org/10.1016/j.jval.2017.08.3018

Introduction

Modern healthcare encounter and reimbursement systems produce an abundance of electronically recorded, patient-level longitudinal data. These data streams contain information on physician visits, hospitalizations, diagnoses made and recorded, procedures performed and billed, medications prescribed and filled, lab tests performed or results recorded, as well as many other date-stamped items. Such temporally ordered data are used to study the effectiveness and safety of medical products, healthcare policies, and medical interventions and have become a key tool for improving the quality and affordability of healthcare [1,2]. The importance and influence of such "real world" evidence is demonstrated by commitment of governments around the world to develop infrastructure and technology to increase the capacity for use of these data in comparative effectiveness and safety research as well as health technology assessments [3–12].

Research conducted using healthcare databases currently suffers from a lack of transparency in reporting of study details [13–16]. This has led to high profile controversies over apparent discrepancies in results and reduced confidence in evidence generated from healthcare databases. However, subtle differences in scientific decisions regarding specific study parameters can have significant impacts on results and interpretation—as was discovered in the controversies over 3rd generation oral contraceptives and risk of venous thromboembolism or statins and the risk of hip fracture [17,18]. Clarity regarding key operational decisions would have facilitated replication, assessment of validity and earlier understanding of the reasons that studies reported different findings.

The intertwined issues of transparency, reproducibility and validity cut across scientific disciplines. There has been an increasing movement towards "open science", an umbrella term that covers study registration, data sharing, public protocols and more detailed, transparent reporting [19–28]. To address these issues in the field of healthcare database research, a Joint Task Force between the International Society for Pharmacoepidemiology (ISPE) and the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) was convened to address transparency in process for database studies (e.g. "what did you plan to do?") and transparency in study execution (e.g. "what did you actually do?). This paper led by ISPE focuses on the latter topic, reporting of the specific steps taken during study implementation to improve reproducibility and assessment of validity.

Transparency and reproducibility in large healthcare databases is dependent on clarity regarding 1) cleaning and other pre-processing of raw source data tables, 2) operational decisions to create an analytic dataset and 3) analytic choices (Figure 1). This paper focuses on reporting of design and implementation decisions to define and create a temporally anchored study population from raw longitudinal source data (Figure 1 Step 2). A temporally anchored study population is identified by a sentinel event—an initial temporal anchor. Characteristics of patients, exposures and/or outcomes are evaluated during time periods defined in relation to the sentinel event.

However understanding how source data tables are cut, cleaned and pre-processed prior to implementation of a research study (Figure 1 Step 1), how information is extracted from unstructured data (e.g. natural language processing of free text from clinical notes), and how the created dataset is analyzed (Figure 1 Step 3) are also important parts of reproducible research. These topics have been covered elsewhere [14,29–36], however we summarize key points for those data provenance steps in the online appendix.

Transparency

Transparency in what researchers initially intended to do protects against data dredging and cherry picking of results. It can be achieved with pre-registration and public posting of protocols before initiation of analysis. This is addressed in detail in a companion paper led by ISPOR [37]. Because the initially planned research and the design and methodology underlying reported results may differ, it is also important to have transparency regarding what researchers *actually did* to obtain the reported results from a healthcare database study. This can be achieved with clear reporting on the detailed operational decisions made by investigators during implementation. These decisions include how to define a study population (whom to study), and how to design and conduct an analysis (what to measure, when and how to measure it).

Reproducibility and replicability

Reproducibility is a characteristic of a study or a finding. A reproducible study is one for which independent investigators implementing the same methods in the same data are able to obtain the same results (direct replication [38]). In contrast, a reproducible finding is a higher order target than a reproducible study, which can be tested by conducting multiple studies that evaluate *the same question and estimand (target of inference)* but use different data and/or apply different methodology or operational decisions (conceptual replication [38]) (Table 1).

Direct replicability is a necessary, but not sufficient, component of high quality research. In other words, a fully transparent and directly replicable research study is not necessarily rigorous nor does it necessarily produce valid findings. However, the transparency that makes direct replication possible means that validity of design and operational decisions can be evaluated, questioned and improved. Higher order issues such as conceptual replication of the finding can and should be evaluated as well, however, without transparency in study implementation, it can be difficult to ascertain whether superficially similar studies address the same conceptual question.

For healthcare database research, direct replication of a study means that if independent investigators applied the *same* design operational choices to the *same* longitudinal source data, they should be able to obtain the *same* results (or at least a near exact reproduction). In contrast, conceptual replication and robustness of a finding can be assessed by applying the same methods to *different* source data (or different years from the same source). Here, lack of replicability would not necessarily mean that one result is more "correct" than another, or refutes the results of the original. Instead, it would highlight a need for deeper inquiry to find the drivers of the differences, including differences in data definitions and quality, temporal changes or true differences in treatment effect for different populations. Conceptual replications can be further evaluated through application of *different* plausible methodologic and operational decisions to the same or *different* source data to evaluate how much the finding is influenced by the specific parameter combinations originally selected. This would encompass evaluation of how much reported findings vary with plausible alternative parameter choices, implementation in comparable data sources or after flawed design or operational decision is corrected. However, the scientific community cannot evaluate the validity and rigor of research methods if implementation decisions necessary for replication are not transparently reporte.

The importance of achieving consistently reproducible research is recognized in many reporting guidelines (e.g. STROBE [34], RECORD [39], PCORI Methodology Report [40], EnCePP [33]) and is one impetus for developing infrastructure and tools to scale up capacity for generating evidence from large healthcare database research [3,41–45]. Other guidelines, such as the ISPE Guidelines for Good Pharmacoepidemiology Practice (GPP) broadly cover many aspects of pharmacoepidemiology from protocol development, to responsibilities of research personnel