Diabetes Care
American Diabetes Association

A Road Map for Peer Review of Real-World Evidence Studies on Safety and Effectiveness of Treatments

Almut G. Winterstein, Vera Ehrenstein, Jeffrey S. Brown, Til Stürmer, and Meredith Y. Smith

Diabetes Care 2023;46(8):1448–1454 | *https://doi.org/10.2337/dc22-2037*

Suggested best practices for peer review of real-world evidence studies that make causal inferences on drug effects
1. Checklists & guidance
Provide reference to good pharmacoepidemiologic study and reporting practices.
Require (not recommend) use of RECORD-PE checklist.
2. A priori protocols
Encourage prior registration of study protocol and analysis plan and justification for protocol deviations during study conduct.
3. Reviewer expertise
Ensure inclusion of epidemiologist with demonstrated experience with the real-world data source on peer review team.
4. Data provenance, characterization, & custodianship
Ensure availability of detail about the underlying source, author knowledge of the data source, and how data were transformed and curated into the research database.
Ensure availability of population-based and cohort-based metrics that enable reviewers to assess appropriateness and generalizability of the data source.

ARTICLE HIGHLIGHTS

- The marked expansion of real-world evidence (RWE) studies is paralleled by increasing concern about study validity, violating causal inference principles by the inappropriate use of data and methods.
- Recent study retractions highlight opportunities to improve RWE research conduct and strengthen the peer review and editorial processes.
- Four practices to strengthen the peer review process include the requirement of checklists, availability of predetermined study protocols and analysis plans, inclusion of appropriate pharmacoepidemiologic expertise among peer reviewers, and provision of detail on data provenance and characterization to support assessment of appropriateness, validity, and generalizability of the data source.

A Road Map for Peer Review of Real-World Evidence Studies on Safety and Effectiveness of Treatments

Almut G. Winterstein,[1,2,3] *Vera Ehrenstein,*[2,3,4] *Jeffrey S. Brown,*[2,3,5,6] *Til Stürmer,*[2,3,7] *and Meredith Y. Smith*[2,3,8,9]

Diabetes Care 2023;46:1448–1454 | https://doi.org/10.2337/dc22-2037

The growing acceptance of real-world evidence (RWE) in clinical and regulatory decision-making, coupled with increasing availability of health care data and advances in automated analytic approaches, has contributed to a marked expansion of RWE studies of diabetes and other diseases. However, a recent spate of high-profile retractions highlights the need for improvements in the conduct of RWE research as well as in the associated peer review and editorial processes. We review best pharmacoepidemiologic practices and common pitfalls regarding design, measurement, analysis, data validity, appropriateness, and generalizability of RWE studies. To enhance RWE study assessments, we propose that journal editors require *1*) study authors to complete RECORD-PE, a reporting guideline for pharmacoepidemiological studies on routinely collected data, *2*) availability of predetermined study protocols and analysis plans, *3*) inclusion of pharmacoepidemiologists on the peer review team, and *4*) provision of detail on data provenance, characterization, and custodianship to facilitate assessment of the data source. We recognize that none of these steps guarantees a high-quality research study. Collectively, however, they permit an informed assessment of whether the study was adequately designed and conducted and whether the data source used was fit for purpose.

Peer-reviewed articles are rarely retracted (1). Thus, when a retraction does occur, it provides an opportunity to reflect on how to improve not only the scientific study but also the peer review and editorial processes.

Two retractions of real-world evidence (RWE) studies concerning coronavirus disease 2019 (COVID-19) treatments, published in leading medical journals, are a prominent case in point (2,3). Both studies used the same data source. This source was previously unknown to the research community, but by virtue of referencing established data standards, it appeared to be of acceptable quality. Even if the data source had been valid, and there is now ample evidence to suggest that it was not, both reports omitted crucial information about study design and measurement, thus limiting a meaningful assessment of bias and adequate peer review. One article evaluated adverse effects of renin-angiotensin inhibitors on COVID-19 mortality, a drug class commonly involved in RWE studies in patients with diabetes, while the other evaluated hydroxychloroquine as a potential treatment for COVID-19.

Concerns about the retracted studies' validity quickly emerged, including doubts about the quality of the underlying data. The database in question had been assembled by Surgisphere, a U.S.-based health care analytics company, which vanished in

[1]*Department of Pharmaceutical Outcomes and Policy, Department of Epidemiology, and Center for Drug Evaluation and Safety, University of Florida, Gainesville, FL*
[2]*International Network for Epidemiology in Policy, American College of Epidemiology, Washington Avenue Extension, Albany, NY*
[3]*International Society for Pharmacoepidemiology, Bethesda, MD*
[4]*Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark*
[5]*Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Cambridge, MA*
[6]*TriNetX, LLC, Cambridge, MA*
[7]*Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC*
[8]*Evidera, Inc., PPD, Boston, MA*
[9]*School of Pharmacy, University of Southern California, Los Angeles, CA*

Corresponding author: Almut G. Winterstein, almut@ufl.edu

Received 18 October 2022 and accepted 5 May 2023

the wake of the retractions (4). Other flaws identified provenance, and clinically implausible hydroxychloroquine doses.

These two retractions are examples of a bigger problem involving publication of flawed RWE studies, perhaps exacerbated through the COVID-19 pandemic, when the urgent demand for data elucidating potential treatment strategies may have created pressures to relax standards regarding the level and rigor of scientific review, thus contributing to what has been dubbed "pandemic-related research waste." For example, estimates from observational studies of the effects of hydroxychloroquine on COVID-19 mortality have ranged from more than a twofold reduction to a twofold increase, with some findings likely attributable to immortal time and selection bias (5,6,7,8,9).

As illustrated in the foregoing example, deficiencies in study design, reporting, and peer review collectively contributed to propagating misinformation, thereby damaging public confidence in research and the policies built upon it (10). What conclusions can be drawn from this experience? The marked inadequacies of these retracted studies highlight important shortcomings in the peer review and editorial processes and their ability to prioritize studies that rest on the two pillars that support valid causal inference: *1*) use of real-world data that are fit for purpose and *2*) the application of appropriate observational research methods. On behalf of the International Network for Epidemiology in Policy (INEP) and the International Society of Pharmacoepidemiology (ISPE), we provide an overview of common shortcomings in RWE studies and related best practices to guide peer reviewers and *Diabetes Care* readers who may seek to engage in RWE-based research. We then propose four practices to strengthen the peer review and editorial process and, ultimately, to maximize the value of RWE for informing health practice and policy.

HOPE AND HYPE OF RWE

RWE studies are particularly valuable if they can complement available clinical trial evidence, e.g., by offering longer follow-up to evaluate long-term effects, providing drug-to-drug comparisons, or capitalizing on large sample sizes to evaluate the heterogeneity of treatment effects or manifestation of rare outcomes. *Diabetes Care* publishes RWE studies in virtually every issue (11,12,13). Indeed, the growing acceptance of RWE in clinical and regulatory decision-making (14) as well as the increasing availability of health care data coupled with automated analytic approaches have contributed to a marked growth in RWE studies. Commensurate with this growth, there has been increasing concern about inappropriate conduct of RWE studies that may violate causal inference principles via inappropriate use of data and methods. Of particular concern is the application of data-driven (focused on model fit) rather than design-driven (focused on isolating cause and effect) approaches. A major concern with relying on data-driven approaches is that they fail to consider the processes that generated the data and are therefore prone to bias. RWE originates from routinely collected data, i.e., real-world data (RWD) that are typically a "by-product" of clinical or health care administrative processes. As a result, appropriate design, analysis, and interpretation of studies based on such data require specific expertise. Such expertise extends beyond the clinical subject matter and falls squarely within the field of pharmacoepidemiology, a discipline dedicated to the evaluation of medical interventions using epidemiologic methods (15). Below we discuss the key methodological issues pertinent to RWE studies that should be understood and considered by reviewers, along with relevant examples pertaining to diabetes care. We also recommend a recently published collection of 30 must-read articles for newcomers to pharmacoepidemiology for those interested in learning more about these methods (16).

KEY METHODOLOGICAL ISSUES IN RWE STUDIES

Study Design

A paramount consideration in RWE research is the selection of the study design. The choice of what analytic (i.e., statistical or other computational) methods to use is a secondary concern, as even complex analytic techniques fall short of removing biases that could have been addressed (or introduced) by the study design itself.

The two main observational designs for RWE intervention studies using automated health care data are cohort and self-controlled (e.g., case-crossover) designs. Cohorts are usually defined by treatment initiation (a hypothetical intervention) and followed over time to compare the incidence of outcomes. Self-controlled studies examine treatments and outcomes within individuals rather than across individuals by analyzing different treatment periods within the same person, assuming intermittent treatments and their transient effects on outcomes. Both design types, cohort studies and self-controlled studies, have experimental counterparts that share similar prerequisites. Cohort studies require a comparator exposure that is a plausible alternative to treatment (an active comparator or, in some cases, no treatment). Providers' and/or patients' treatment selection is, however, rarely random, thus confounders must be accounted for. In contrast, self-controlled studies can control time-fixed confounders because comparisons are made within and not across patients. As in cross-over randomized controlled trials (RCTs), however, measured outcomes must be fully reversible and not associated with the probability of exposure in the future. We refer the reader to published guidance on self-controlled studies (17,18) and focus here on the most common observational study design, i.e., the cohort design.

The initiation of a treatment is usually based on indication and contraindications, i.e., expected benefit and harm in patients based on such factors as disease severity, patient comorbidities, comedications, degree of frailty, and socioeconomic status. In most RWD studies, many of these factors are not measured with sufficient detail to allow analytic control for differences across cohorts. The best way to minimize confounding by indication is therefore by study design, i.e., to compare the initiation of the treatment of interest with the initiation of an alternative treatment for the same indication (19). A placebo-controlled RCT design addresses the question of whether to treat or not to treat. In contrast, the Active Comparator, New User (ACNU) study design addresses the question of which treatment (of two or more possible treatments) should be given to a patient after the decision to treat has been made (20). While both questions are clinically relevant, answering the first with RWD will suffer from strong (and mostly unmeasured) confounding by indication, whereas the second can often be answered robustly, given an adequate active comparator.

In an ACNU design, an active comparator that presents an appropriate alternative

to the treatment in question, selected based on guidelines and clinical input, can balance differences between comparison groups. Such differences, considering all relevant risk factors for the study outcome, should be presented in a baseline characteristics table (before any further analytic adjustments are made). For example, to examine the impact of a long-acting insulin treatment regimen for patients with type 2 diabetes, initiators of insulin glargine could be compared with initiators of NPH insulin, thus allowing researchers to balance BMI by design, removing confounding by the arguably most important indication for long-acting insulin in patients with type 2 diabetes (high BMI). This allows the researcher to evaluate cancer outcomes even in the absence of BMI data (21). Deviations from the ACNU design are possible but require a strong rationale for doing so, and ensuing limitations must be carefully evaluated (22). Challenging examples of diabetes-related studies in this regard include evaluations of agents that are indicated for subpopulations with specific comorbidities that are added to an existing regimen or of newer agents with higher copays. Appropriate comparisons for the former would, ideally, consider patients with similar treatment history and comorbidities; comparisons for the latter would consider patients who are initiated on another brand of active comparator, although such patients may not always be available in the data set. The strikingly similar baseline characteristics of initiators of rosiglitazone and pioglitazone in the evaluation of rosiglitazone cardiovascular safety, both approved at the same time as second-line agents, exemplify the strengths of an ACNU design (23).

Other examples of bias-reducing techniques that should be applied at the design stage for RWE studies include alignment of follow-up start with the treatment initiation (considering induction periods until drug effects can manifest and latent periods until disease manifestation can be recognized and diagnosed), establishment of appropriate censoring criteria, and exclusion of groups with high probability to be only in one exposure group to enhance covariate balance (24). One important consideration when addressing latency periods or requiring a certain treatment duration is the avoidance of immortal time bias, where follow-up includes a time period in which the treatment outcome cannot occur (25,26).

Measurement

While the concept of confounding is widely appreciated as a major threat to RWE studies, measurement bias is less recognized. Unlike in protocol-based baseline and outcome assessments in clinical trials, RWD generation depends on other factors, whether it be a patient's decision to seek health care, a health care provider's decision regarding when and how to evaluate the patient, or a patient's decision to use the prescribed treatment and seek follow-up care. In addition, outcome capture, e.g., via use of International Classification of Disease (ICD) codes, relies on conventions that are not established for research purposes. Thus, measurement in RWE studies, i.e., the process of converting the available data into key study variables, is as important as the choice of an appropriate study design.

Similar to weighing the importance of sensitivity and specificity of a diagnostic test, the impact of measurement validity on effect estimates must be considered when evaluating RWE studies (27). As a general principle, measure specificity is critical for the assessment of outcomes, as noise (false positives) dilutes the effect attributable to the drug exposure and might lead to null findings. To assess the potential for such measurement bias, peer reviewers should ensure that validation studies of outcome measures in RWE studies are provided. Typically, outcomes that result inevitably in health care utilization because of their severity are less prone to misclassification errors than outcomes that require patient and clinician decisions for evaluation.

In measuring exposure, outcomes and even confounders, nondifferential misclassification, i.e., measurement error that affects comparison groups to the same extent, and differential misclassification can occur. For example, patients who use a drug that is known to cause diabetes (e.g., antipsychotics) or who have comorbidities associated with diabetes (e.g., hypertension) will be more frequently screened for diabetes than a potential comparison group of individuals who do not use antipsychotics or antihypertensives (28). If it is impossible to find comparison groups with similar screening frequencies, testing frequency must be considered if the design is to arrive at unbiased estimates (29). If omitted, diabetogenic risk of these medications will be overestimated.

Exposure measurement problems are also prominent in RWD because of patient nonadherence issues and missing end dates in utilization. Other challenges include scenarios where the outcome follows soon after drug discontinuation or a switch from one study drug to the other. Assignment of "active" exposure times informed by pharmacological properties is critical to avoid misclassification and biased exposure effect estimates. A particular challenge is protopathic bias, where drug exposure ends because of early signs of outcome manifestation, which, if not considered in the design and measurement, will inevitably result in an underestimate of the drug effect. Thus, follow-up is commonly extended for a short time period after the estimated exposure time, such as in the rosiglitazone safety study (23). Of note, the core rationale for intention-to-treat designs in superiority trials to retain randomization does not apply to RWE studies, thus as-treated exposure assignments should be used if possible, as they avoid the dilution of drug effects via misclassification. An example of this approach is available in the RWD-based replication of the Cardiovascular Outcome Study of Linagliptin Versus Glimepiride in Patients With Type 2 Diabetes (CAROLINA) trial, which evaluates the cardiovascular safety of linagliptin versus glimepiride (30).

Finally, while confounding and measurement bias are two separate mechanisms and require different approaches for mitigation, bias can occur in the measurement of confounders as well, and adjustment for a misclassified confounder can exaggerate rather than mitigate bias. Unmeasured and poorly measured confounders should be formally evaluated in well-conducted RWE studies rather than addressed via boilerplate disclaimers in the study limitation section. This can be accomplished via simple or probabilistic bias analyses or, if possible, linkage of subgroups to external data sets that allows propensity score calibration or multiple imputations in sensitivity analyses, as exemplified in a study on glucagon-like peptide 1 receptor agonists on chronic lower respiratory disease exacerbations (13,31).

Analytical Methods

Even with an ACNU study design, there will be (hopefully small) differences in patient characteristics between treatment

cohorts. Such differences could be exacerbated if the available data set does not capture disease severity. For example, since heart failure severity measures are usually not available in RWD, excluding patients with heart failure might be preferable over analytic control. Measurable differences can be removed using propensity scores (PS), which are increasingly used in RWE research as an alternative to multivariable outcome models to control for measured confounding (32). Of note, these methods still assume no unmeasured confounding (after balancing the measured covariates) to arrive at unbiased effect estimates. PS balancing methods (e.g., PS matching, weighting, or [fine] stratification) will all result in the same treatment effect estimate if the treatment effect is uniform for all patients. Estimates will be different, however, if some patient subgroups experience more benefit or harm, in which case a decision will need to be made about the most appropriate population of interest (or estimand) (33). The PS also allow us to identify patient subgroups that only receive one of the treatments and for whom we cannot analyze treatment effects (so-called nonpositivity) and patients treated contrary to prediction (per the PS model). Trimming the tails of the (overlapping) PS distribution can be used to potentially reduce unmeasured confounding by frailty and bias due to measurement error (34). Relevant to treatment modalities in diabetes, changes in treatment due to switching or add-on therapy need to be considered, oftentimes resulting in time-varying confounding where changes in the original confounders at cohort entry may have occurred (35,36). It is beyond the scope of this brief summary to cover these situations, but it is important to highlight that, like in an RCT, patients should never be excluded from the cohort based on events that occur during follow-up.

Generalizability

While RWE studies are usually appreciated for their superior generalizability compared with clinical trials, it should be noted that patients who are available in an RWD database may be systematically different from the patients we assume they represent. For example, patients in an employer-sponsored health insurance claims database may not represent the general population of patients with a specific condition. This notwithstanding, the sample size that is commonly accomplished with RWD might support assessments of treatment effect heterogeneity across a broad spectrum of subgroups that may not be achievable with RCT data.

In summary, the multitude of considerations in design, measurement, and analysis, and the magnitude of their impact on the validity of RWE studies, poses challenges for successful peer review. Here, we propose four best practices that may aid peer reviewers and the review process of RWE studies (Table 1).

PROPOSED BEST PRACTICES IN THE PEER AND EDITORIAL REVIEW OF RWE STUDIES

Use of Checklists and Guidance Documents

Best practices for the conduct of pharmacoepidemiologic studies (37) have been incorporated into checklists to ensure adequate reporting of study design, measurement, and analysis and to facilitate assessment of bias (38,39). Recent efforts have generated additional guidance for transparent reporting aided by graphic representation of study designs and use of a standard terminology to define design and analytic concepts (40,41), including specific guidance for research in diabetes (42).

While the use of checklists does not guarantee a high-quality study, report of comprehensive detail on research methods is a prerequisite to assessment of study quality. Postpublication assessment of the two Surgisphere articles showed that while their proportion of adequately reported STROBE checklist items was 71–85%, the proportion of reported items from the checklists specific for pharmacoepidemiologic studies on routinely collected data (i.e., RECORD and RECORD-PE) (39) was comparatively lower at 15–63% (43). The chosen statistical analysis method (logistic regression) was not suitable for data with variable follow-up. Not least, the insufficient description of the database population and data flow and linkage alone should have raised red flags during peer review.

The Surgisphere examples are symptomatic of a larger issue. A recent review of the submission requirements of 257 journals that publish observational studies reported that only 5% required completion of the STROBE checklist or its extensions, 9% suggested use, 5% recommended a "relevant guideline," 28% had indirect references in editorial policies or International Committee of Medical Journal Editors recommendations, and 54% did not make any reference to the STROBE checklist or its extensions (44). We suggest that journals require the submission of RECORD-PE to ensure appropriate use and reporting and thereby adequate evaluation of RWE studies.

Availability of the Study Protocol

Another good practice for RWE studies includes the registration of a study protocol,

Table 1—Suggested best practices for peer review of RWE studies that make causal inferences on drug effects

1. Checklists and guidance documents Provide reference to good pharmacoepidemiologic study and reporting practices. Require (not recommend) use of RECORD-PE checklist.
2. A priori protocols Encourage prior registration of study protocol and analysis plan and justification for protocol deviations during study conduct.
3. Reviewer expertise Ensure inclusion of epidemiologist with demonstrated experience with the RWD source on peer review team.
4. Data provenance, characterization, and custodianship Ensure availability of detail about the underlying source, author knowledge of the data source, and how data were transformed and curated into the research database. Ensure availability of population-based and cohort-based metrics that enable reviewers to assess appropriateness and generalizability of the data source.

as widely adopted for RCTs and recently advocated by relevant professional societies, including ISPE (45,46). The European Medicines Agency also requires registration of both protocols and final reports of mandated postauthorization safety studies in its publicly available EU PAS Register. At minimum, RWE investigators should have a predetermined and publicly available, detailed study protocol and analysis plan prior to study initiation, and they should be prepared to justify protocol deviations so as to avoid design decisions that are solely driven by the data encountered during analysis. We noted in our own review of journal submission requirements that some encourage protocol preregistration (e.g., *The Lancet*) or submission of the original institutional review board–approved study protocol (e.g., *Annals of Internal Medicine*), and we recommend that such practices are broadly adopted to enhance the peer review process.

Reviewer Expertise in Epidemiology

Scientists who are qualified to evaluate design and measurement decisions need to have both training in epidemiologic reasoning to project the impact of such decisions on study validity and a solid understanding of the RWD source, including the underlying health care delivery process that generated these data, to anticipate biases. While editorial boards may maintain a cadre of statisticians to perform reviews of statistical analyses, we are not aware of a similar standard practice regarding epidemiologists with expertise in RWD. We recommend that biomedical journals include an epidemiologist with demonstrated experience with the utilized RWD source on the peer review team. Although such a requirement is neither a prerequisite for nor a guarantee of a high-quality research article, it represents a necessary (if not sufficient) step in that direction and is consistent with the position taken by prominent grant funders such as the National Institutes of Health (NIH). With the notable exception of *Diabetes Care*, many high-ranking biomedical journals rely heavily on clinical trialists, who typically possess little to no expertise in RWE studies, to serve as reviewers. Additionally, we argue that the assessment could include the expertise of the authoring team (e.g., documented history of research conducted with the RWD source). Assessment of study team qualifications is standard practice in the peer review of funding applications, and something akin to this could be adopted for use in the peer review of articles.

Data Provenance, Characterization, and Custodianship

RWD sources include product and disease registries, administrative claims data, electronic health records (EHR), patient-generated data, and various combinations of these sources. These data sources collect information for purposes unrelated to any specific study; rather, the data are collected for purposes ranging from documentation of care (e.g., EHR data) to reimbursement (e.g., claims data) to public health surveillance (e.g., vaccine registry) (47,48). Each source's strengths and limitations need to be understood by both the study investigators and peer reviewers to ensure the data are fit for purpose and suitable for the study design.

Administrative claims data that are generated through health care billing transactions between providers and public and private insurers are often used for RWE studies, because the insurance enrollment period allows for calculation of defined person-time. This is a period during which all medically attended events are expected to be observed, because the medical care will be submitted for reimbursement by the provider. During an enrollment period, the absence of care is meaningful. EHR data sources that are used to document care during a medical encounter are overseen by individual health systems or academic medical centers. They are appreciated for the clinical details that are often missing from claims data (e.g., laboratory results, BMI, and blood pressure), but these data lack the concept of person-time, meaning that the absence of care indicates either that care was not received or that care was provided in a different health system.

Further, research involving data from multiple sources requires understanding of the degree to which differences in health care delivery, data privacy regulations, documentation, and other site-, system-, or region-specific issues affect data quality and interpretability. In the Surgisphere example, the authors claimed to have had access to standardized EHRs of hundreds of hospitals across the globe, a claim that would have been recognized as implausible by a reviewer with expertise in RWE research (49).

Peer reviewers unfamiliar with a data source should identify the type of data source used (e.g., EHR, claims, registry, etc.) and the institution responsible for collecting and managing the data, and they should search for other publications that have used the data to ascertain the extent to which the data have been previously used for research. Reviewers also should examine baseline cohort characteristics as a form of plausibility assessment of the measures derived from the data source. An adequate evaluation of an RWE study also depends on the degree to which the authors have shared documentation (e.g., algorithms and codes) regarding how the RWD were used to define key concepts like person-time, patient entry into and exit from a study cohort, study eligibility criteria, medical encounters, outcomes, exposures, and other important variables. In addition, the documentation should specify how data flowed from routine collection to assembly into a database and describe the associated data curation, transformation, and linkage steps (50). The latter includes providing details on the level of completeness and reasons for missingness by data domain.

The newly created U.S. National COVID Cohort Collaborative (N3C) database of data from COVID-19–positive patients and test-negative comparators is an example of good practice in this regard (51). The database was rapidly established by the NIH to respond to the COVID-19 pandemic and now includes more than 80 data contributors. Data are standardized into a common data model that is updated frequently and curated by a database oversight team. The N3C intended to enable high-throughput research from many investigators, with more than 200 projects underway. *Diabetes Care* has recently published several articles using N3C data (52,53). However, even though N3C follows good processes for data set creation, the complexities in using EHR data from multiple disparate health systems for evaluation of treatment effects (54), and the novelty of the data source and processes that created and curated it, pose significant challenges for determining whether it is fit for purpose in any specific use case and method. Peer review with experts knowledgeable in RWD studies is needed to help ensure findings from databases such as N3C are robust and appropriately reported (55).

In summary, the use of RWE in clinical, policy, and regulatory decision-making is rapidly expanding. As a result, the peer review process must evolve. In this analysis, we have outlined the key methodological issues pertinent to RWE studies that should be considered by reviewers. We have proposed four key approaches that can enhance the RWE publishing process and help optimize the value of RWE studies. These include *1*) the requirement for comprehensive reporting of study methods via checklists; *2*) the availability of an a priori–developed study protocol and analysis plan; *3*) the inclusion of adequate expertise in observational research methods and RWD sources among the peer review team; and *4*) the availability of sufficient detail on data provenance and characterization that allows assessment of whether the data are fit for purpose and valid. We recognize that none of these steps guarantees a high-quality research study. Collectively, however, they permit an informed assessment of whether the study was adequately designed and conducted and whether the data source used was fit for purpose.

CONTRIBUTORS AND SOURCES

This article was developed on behalf of the International Network for Epidemiology in Policy (INEP) to address a growing concern in the epidemiologic community regarding RWE studies that do not follow good epidemiologic design practices and violate causal inference principles and to address the peer review of related study publications. The authors, recruited from leadership within the International Society of Pharmacoepidemiology (ISPE), have academic, industry, and regulatory backgrounds in North America and Europe, and all have RWE research experience. Drs. Winterstein and Stürmer are ISPE past presidents and chairs of large pharmacoepidemiology programs in academia. Dr. Winterstein has served as chair of the U.S. Food and Drug Administration (FDA) Drug Safety and Risk Management Advisory Committee, and Dr. Ehrenstein has served on the steering committee of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP), coordinated by the European Medicines Agency. She also coauthored the RECORD-PE guidelines. With a research focus on data quality and the development of distributed data networks, Dr. Brown has leadership roles in FDA Sentinel, PCORnet, and the Innovation in Medical Evidence and Development Surveillance (IMEDS) program. Dr. Smith serves on the INEP board and cochairs the ISPE Public Policy Committee. The authors developed this article collaboratively.

Acknowledgments. This article was submitted on behalf of the INEP, an international consortium of epidemiology societies and associations that strives to develop ethical and transparent policies informed by translation of high-quality epidemiologic evidence to protect the health of communities and people. The manuscript was also endorsed by the ISPE.

Funding and Duality of Interest. A.G.W. has received research funding from the NIH, Agency for Healthcare Research and Quality, Patient-Centered Outcomes Research Institute, Centers for Disease Control and Prevention, FDA, the State of Florida, the Bill and Melinda Gates Foundation, and Merck Sharpe & Dohme. She has received consulting honoraria from Arbor Pharmaceuticals, Bayer, Ipsen, and Genentech, Inc. V.E. is a salaried employee of Aarhus University, which receives institutional research funding from public and private entities, including regulators, pharmaceutical companies, and contract research organizations. J.S.B. was an employee of the Harvard Pilgrim Health Care Institute (HPHCI) during the project. He is now Chief Scientific Officer at TriNetX, LLC. He has received research funding via HPHCI from the NIH, Patient-Centered Outcomes Research Institute, FDA, the Biologics and Biosimilars Collective Intelligence Consortium, the Reagan-Udall Foundation, GSK, Janssen, and Pfizer. He has received consulting fees from Forian, Inc., Intercept Pharmaceuticals, Jazz Pharmaceuticals, Mathematica Policy Research, and IBM. He is on an external advisory board of PicnicHealth. T.S. receives investigator-initiated research funding and support as a principal investigator (R01 AG056479) from the National Institute on Aging and as a coinvestigator (R01 HL118255 and R01MD011680) from the NIH. He also receives salary support as director of comparative effectiveness research (CER), NC TraCS Institute, from a UNC clinical and translational science award (UL1TR002489), from the Center for Pharmacoepidemiology (current members GSK, UCB BioSciences, Takeda, AbbVie, and Boehringer Ingelheim), from a pharmaceutical company (Novo Nordisk), and by a generous contribution from Dr. Nancy A. Dreyer to the Department of Epidemiology, University of North Carolina at Chapel Hill. T.S. does not accept personal compensation of any kind from any pharmaceutical company. He owns stock in Novartis, Roche, and Novo Nordisk. M.Y.S. is a full-time employee of Evidera, Inc., PPD, a part of Thermo Fisher Scientific. When this article was prepared, she was a full-time employee of AstraZeneca, PLC, and she is a shareholder in the company.

Prior Presentation. Content similar to this paper was presented and discussed in a Hot Topic Session at the 38th International Conference on Pharmacoepidemiology, Copenhagen, Denmark, 24–28 August 2022.

References

1. Brainard J. Rethinking retractions. Science 2018;362:390–393
2. Mehra MR, Desai SS, Ruschitzka F, Patel AN. Retracted: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. Lancet 2020;395:1820
3. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Retraction: Cardiovascular disease, drug therapy, and mortality in Covid-19. N Engl J Med 2020;382:2582
4. Hopkins J, Gold R. The big-data mystery behind retracted Covid-19 studies of hydroxychloroquine, other drugs. Published 11 June 2020. Accessed 20 May 2021. Available from https://www.wsj.com/articles/the-big-data-mystery-behind-retracted-covid-19-studies-of-hydroxychloroquine-other-drugs-11591867981
5. Renoux C, Azoulay L, Suissa S. Biases in evaluating the safety and effectiveness of drugs for Covid-19: designing real-world evidence studies. Am J Epidemiol 2021;190:1452–1456
6. Magagnoli J, Narendran S, Pereira F, et al. Outcomes of hydroxychloroquine usage in United States veterans hospitalized with COVID-19. Med 2020;1:114–127.e3
7. Geleris J, Sun Y, Platt J, et al. Observational study of hydroxychloroquine in hospitalized patients with Covid-19. N Engl J Med 2020;382:2411–2418
8. Membrillo de Novales F, Ramírez-Olivencia G, Estébanez M, et al. Early hydroxychloroquine is associated with an increase of survival in COVID-19 patients: an observational study. Accessed 11 June 2021. Available from https://www.preprints.org/manuscript/202005.0057/v1
9. Axfors C, Schmitt AM, Janiaud P, et al. Mortality outcomes with hydroxychloroquine and chloroquine in COVID-19 from an international collaborative meta-analysis of randomized trials. Nat Commun 2021;12:2349
10. Goodman JL, Borio L. Finding effective treatments for COVID-19: scientific integrity and public confidence in a time of crisis. JAMA 2020;323:1899–1900
11. Patorno E, Pawar A, Bessette LG, et al. Comparative effectiveness and safety of sodium-glucose cotransporter 2 inhibitors versus glucagon-like peptide 1 receptor agonists in older adults. Diabetes Care 2021;44:826–835
12. Wang T, Yang JY, Buse JB, et al. Dipeptidyl peptidase 4 inhibitors and risk of inflammatory bowel disease: real-world evidence in U.S. adults. Diabetes Care 2019;42:2065–2074
13. Albogami Y, Cusi K, Daniels MJ, Wei YJ, Winterstein AG. Glucagon-like peptide 1 receptor agonists and chronic lower respiratory disease exacerbations among patients with type 2 diabetes. Diabetes Care 2021;44:1344–1352
14. Breckenridge AM, Breckenridge RA, Peck CC. Report on the current status of the use of real-world data (RWD) and real-world evidence (RWE) in drug development and regulation. Br J Clin Pharmacol 2019;85:1874–1877
15. Strom BL. Pharmacoepidemiology. 3rd ed. New York, Wiley, 2000
16. Pottegård A, Morin L, Hallas J, et al. Where to begin? Thirty must-read papers for newcomers

to pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2022;31:257–259
17. Cadarette SM, Maclure M, Delaney JAC, et al. Control yourself: ISPE-endorsed guidance in the application of self-controlled study designs in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2021;30:671–684
18. Hallas J, Pottegård A. Use of self-controlled designs in pharmacoepidemiology. J Intern Med 2014;275:581–589
19. Sendor R, Stürmer T. Core concepts in pharmacoepidemiology: confounding by indication and the role of active comparators. Pharmacoepidemiol Drug Saf 2022;31:261–269
20. Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. Curr Epidemiol Rep 2015;2:221–228
21. Stürmer T, Marquis MA, Zhou H, et al. Cancer incidence among those initiating insulin therapy with glargine versus human NPH insulin. Diabetes Care 2013;36:3517–3525
22. Johnson ES, Bartman BA, Briesacher BA, et al. The incident user design in comparative effectiveness research. Pharmacoepidemiol Drug Saf 2013;22:1–6
23. Graham DJ, Ouellet-Hellstrom R, MaCurdy TE, et al. Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone. JAMA 2010;304:411–418
24. Westreich D, Cole SR. Invited commentary: positivity in practice. Am J Epidemiol 2010;171:674–681
25. Suissa S. Immortal time bias in observational studies of drug effects. Pharmacoepidemiol Drug Saf 2007;16:241–249
26. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 2010;340:b5087
27. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 2012;65:343–349.e2
28. Winterstein AG, Kubilis P, Bird S, Cooper-DeHoff RM, Nichols GA, Delaney JA. Misclassification in assessment of diabetogenic risk using electronic health records. Pharmacoepidemiol Drug Saf 2014;23:875–881
29. Cooper-DeHoff RM, Bird ST, Nichols GA, Delaney JA, Winterstein AG. Antihypertensive drug class interactions and risk for incident diabetes: a nested case-control study. J Am Heart Assoc 2013;2:e000125
30. Patorno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM. Using real-world data to predict findings of an ongoing phase IV cardiovascular outcome trial: cardiovascular safety of linagliptin versus glimepiride. Diabetes Care 2019;42:2204–2210
31. Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. Curr Epidemiol Rep 2014;1:175–185
32. Webster-Clark M, Stürmer T, Wang T, et al. Using propensity scores to estimate effects of treatment initiation decisions: state of the science. Stat Med 2021;40:1718–1735
33. Stürmer T, Wyss R, Glynn RJ, Brookhart MA. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. J Intern Med 2014;275:570–580
34. Stürmer T, Webster-Clark M, Lund JL, et al. Propensity score weighting and trimming strategies for reducing variance and bias of treatment effect estimates: a simulation study. Am J Epidemiol 2021;190:1659–1670
35. Patorno E, Garry EM, Patrick AR, et al. Addressing limitations in observational studies of the association between glucose-lowering medications and all-cause mortality: a review. Drug Saf 2015;38:295–310
36. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ 2017;359:j4587
37. Public Policy Committee, International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practice (GPP). Pharmacoepidemiol Drug Saf 2016;25:2–10
38. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61:344–349
39. 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
40. Schneeweiss S, Rassen JA, Brown JS, et al. Graphical depiction of longitudinal study designs in health care databases. Ann Intern Med 2019;170:398–406
41. 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
42. Patorno E, Schneeweiss S, Wang SV. Transparency in real-world evidence (RWE) studies to build confidence for decision-making: reporting RWE research in diabetes. Diabetes Obes Metab 2020;22(Suppl. 3):45–59
43. Benchimol EI, Moher D, Ehrenstein V, Langan SM. Retraction of COVID-19 pharmacoepidemiology research could have been avoided by effective use of reporting guidelines. Clin Epidemiol 2020;12:1403–1420
44. Sharp MK, Tokalić R, Gómez G, Wager E, Altman DG, Hren D. A cross-sectional bibliometric study showed suboptimal journal endorsement rates of STROBE and its extensions. J Clin Epidemiol 2019;107:42–50
45. International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practices (GPP). Published June 2015. Accessed 18 August 2021. Available from https://www.pharmacoepi.org/resources/policies/guidelines-08027/#4
46. 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;23:1128–1136
47. Lin KJ, Schneeweiss S. Considerations for the analysis of longitudinal electronic health records linked to claims data to study the effectiveness and safety of drugs. Clin Pharmacol Ther 2016;100:147–159
48. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323–337
49. Watson J. An open letter to Mehra et al and The Lancet. Published 28 May 2020. Accessed 20 May 2021. Available from https://zenodo.org/record/3871094#.YKqD2S9h1pQ
50. Schmidt M, Schmidt SAJ, 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
51. National Center for Advancing Translational Sciences. National Covid Cohort Collaborative (N3C). Accessed 13 July 2021. Available from https://ncats.nih.gov/n3c
52. Kahkoska AR, Abrahamsen TJ, Alexander GC, et al. Association between glucagon-like peptide 1 receptor agonist and sodium-glucose cotransporter 2 inhibitor use and COVID-19 outcomes. Diabetes Care 2021;44:1564–1572
53. Wong R, Vaddavalli R, Hall MA, et al.; N3C Consortium. Effect of SARS-CoV-2 infection and infection severity on longer-term glycemic control and weight in people with type 2 diabetes. Diabetes Care 2022;45:2709–2717
54. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51(Suppl. 3):S30–S37
55. Brown JS, Bastarache L, Weiner MG. Aggregating electronic health record data for COVID-19 research–caveat emptor. JAMA Netw Open 2021;4:e2117175