# Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making

Jessica M. Franklin[1], Robert J. Glynn[1], David Martin[2] and Sebastian Schneeweiss[1,*]

**The analysis of longitudinal healthcare data outside of highly controlled parallel-group randomized trials, termed real-world evidence (RWE), has received increasing attention in the medical literature. In this paper, we discuss the potential role of RWE in drug regulation with a focus on the analysis of healthcare databases. We present several cases in which RWE is already used and cases in which RWE could potentially support regulatory decision making. We summarize key issues that investigators and regulators should consider when designing or evaluating such studies, and we propose a structured process for implementing analyses that facilitates regulatory review. We evaluate the empirical evidence base supporting the validity, transparency, and reproducibility of RWE from analysis of healthcare databases and discuss the work that still needs to be done to ensure that such analyses can provide decision-ready evidence on the effectiveness and safety of treatments.**

The analysis of longitudinal healthcare data outside of highly controlled traditional randomized trials, termed real-world evidence (RWE), has received increasing attention in the medical literature.[1–4] In the United States, the 21st Century Cures Act and the sixth reauthorization of the Prescription Drug User Fee Act both underscore the potential utility of data generated during the routine operation of the healthcare system for the regulation of therapeutics and treatment decision making by patients. Both also require that the US Food and Drug Administration (FDA) publish guidance addressing when and how RWE may be used to support the assessment of safety and effectiveness in regulatory submissions.[5,6]

The majority of RWE studies published to date rely on healthcare databases that aggregate routinely collected health information for large numbers of patients, such as health insurance claims databases, electronic health record databases, or some disease registries.[7] These databases may be derived from source data that are structured or unstructured and subsequently curated. Retrospective studies resulting from these databases are uniformly lacking randomization because, by definition, the treatments provided to patients enrolled in these databases are simply observed and are not randomly assigned. Patients receive treatments as part of routine clinical care selected in consultation with their treating provider, not by study investigators. Therefore, when comparing patients selectively receiving alternative treatments, confounding bias is a major concern that can diminish the validity of such studies.[8]

In this paper, we discuss the potential roles of RWE in drug regulation with a focus on the analysis of healthcare databases, as the role of pragmatic randomized trials has been discussed previously.[9] We present several cases in which RWE is already used as well as cases in which RWE could potentially support regulatory decision making. We also summarize key issues that investigators and regulators should consider when designing or evaluating such studies, and we propose a structured process for implementing analyses that facilitates regulatory review. We further evaluate the empirical evidence base supporting the validity, transparency, and reproducibility of RWE from healthcare database analyses, and we discuss the work that still needs to be done to ensure that such analyses can provide decision-ready evidence on the effectiveness and safety of treatments.

## POTENTIAL USE CASES FOR RWE

In **Figure 1**, we summarize four key use cases in which RWE could support regulatory decision making. The first example, case 1, uses RWE to support primary approval of a new medication. Because medications are not yet available outside of trials at the time of approval, data on patients receiving the new medication would need to be collected from trial data. However, data on control patients could be derived from real-world data (RWD) sources to form "external controls."[10] This strategy may be particularly desirable when the disease under study is rare, making the conduct of a well-powered randomized controlled trial (RCT) more difficult, as in rare diseases or highly targeted therapies with low-frequency molecular subsets.[11] Such approaches are more likely to succeed if the condition in question has a predictable and measurable decline of health without the new treatment, and the treatment is expected to have a large effect.[12] Control patients could be identified as patients with indications for the new medication who are receiving an older alternative treatment or are receiving no treatment in cases where

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA; [2]Office of Medical Policy, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA. *Correspondence: Sebastian Schneeweiss (schneeweiss@post.harvard.edu)

Received 15 October, 2019; accepted 25 November, 2018. doi:10.1002/cpt.1351

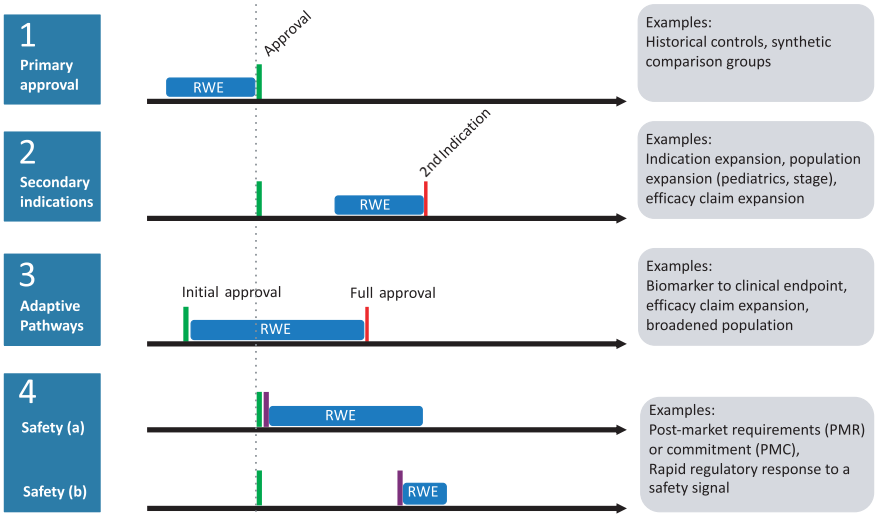


**Figure 1** Real-world evidence (RWE) in regulatory decision making: key use cases.

there is no currently approved therapy. In either case, the goal of using RWD is to quantify the typical disease trajectory in the absence of the new medication to provide a counterfactual to the disease trajectory observed on the experimental treatment. Regulatory agencies have used external controls derived from RWD and other sources, and guidance has been published in this area.[10] Considerations for how to define the control group and conduct the analysis are discussed in later sections.

In case 2, RWE is used to support a supplemental indication. For example, many drugs receive a primary approval based on intermediate outcomes, such as lowering blood pressure for antihypertensives or lowering Hemoglobin $A1_c$ for antidiabetic medications. Once the drug is marketed, RWE can be used to evaluate drug effects on clinical outcomes, such as myocardial infarction or stroke, thereby supporting an expanded indication for prevention of the clinical outcomes. Alternatively, RWE could be used, along with other evidence, to support broadening the population, for example, to other disease stages or to pediatrics if initially approved in adults. RWE may be particularly useful in supporting pediatric approvals, because drugs are sometimes approved for pediatric patients based on demonstrated effectiveness in adults along with data supporting extrapolation to younger populations without new RCTs to evaluate efficacy conducted specifically in pediatric patients.

Note that RWE submitted to support supplemental applications may sometimes be based on off-label use. Patients using the drug off-label may be quite different from the larger pool of patients who would use the drug if it were approved in the new therapeutic area because they may have already exhausted approved treatment regimens or they may be generally more aggressive in their treatment strategy. However, there are sometimes challenges in reliably ascertaining why a n particular drug is being used. Therefore, the design of an RWE study to evaluate effectiveness in a new population should consider how off-label use may differ from use of an approved comparator therapy and how evidence may need to be updated once the expanded approval is attained.

Case 3 depicts the use of RWE to support adaptive or accelerated approval. For example, the European Medicines Agency conducted a pilot project to explore the use of adaptive pathways with medicines under development. In the adaptive pathway, an initial conditional marketing authorization was made for a limited population with high unmet medical need on the basis of biomarker data or small clinical trials. The sponsor was then required to collect additional data to substantiate therapeutic effectiveness and safety to receive full marketing authorization.[13,14] One way to provide that additional data is to record the outcomes of patients receiving the drug during the conditional market authorization period in a registry or other RWD source. Similar to the supplemental approval case described above, the full approval could further broaden the indication to clinical outcomes or to additional populations not included in the initial marketing authorization.

Although the FDA does not have an adaptive pathways program or a conditional approval mechanism, RWE could be used in an analogous fashion to support further assessments of effectiveness and safety after accelerated approval. Accelerated approval allows for earlier approval of drugs for serious diseases based on a surrogate end point that is reasonably likely to predict clinical benefit or on an intermediate clinical end point that is reasonably likely to predict an effect on irreversible morbidity or mortality or other clinical benefit, taking into account the availability of alternative treatments.[15] Drug companies are still required to conduct phase IV confirmatory studies to substantiate the expected clinical benefit. In the majority of accelerated approvals, the confirmatory study is a clinical trial. If the trial shows that the drug actually provides a clinical benefit, then the FDA grants traditional approval for the drug. Among the subset of accelerated approvals in which confirmatory studies are not randomized trials, RWE might be applicable.[16]

Although the first three scenarios focus on using RWE to provide evidence on medication effectiveness to support approval, the fourth scenario includes the assessment of drug safety. The FDA and other regulatory agencies have a long history of using RWE to assess the safety of regulated medications. For example, RWE played a role in raising the alarm of increased cardiovascular risk on rofecoxib[17] and in alleviating concerns about bleeding risk on

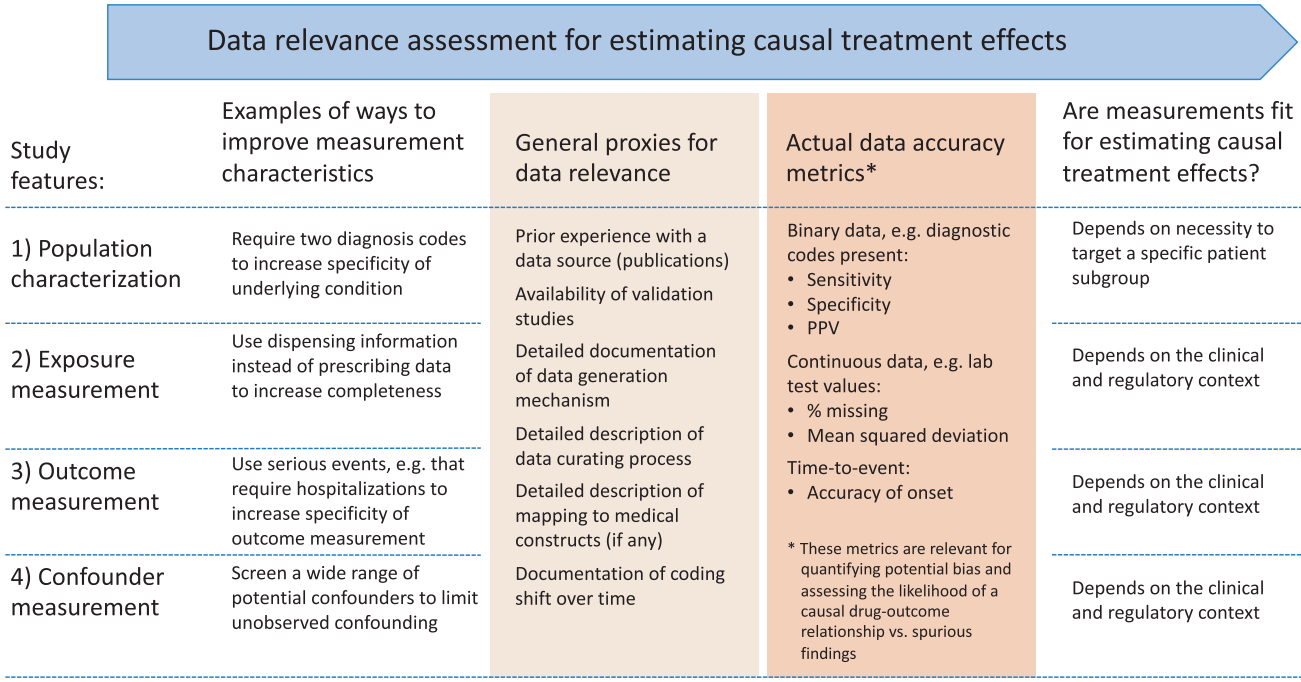


**Figure 2** A structured approach to real-world data relevance. PPV, positive predictive value.

dabigatran.[18,19] The Sentinel Initiative, launched in 2008, created a national system of health insurance claims databases that can be used for rapid safety assessment.[20] The first safety labeling change in response to a Sentinel study was prompted by an assessment of the risk of intussusception after rotavirus vaccine.[21] In general, RWE safety studies, such as postmarketing requirements or imposed postauthorization studies, can arise as a result of preapproval or premarketing authorization pharmacovigilance and risk management planning, as in case 4(a), or as part of a rapid regulatory response to a new safety signal that arises at any time after marketing, as in case 4(b). However, many postmarketing studies of safety continue to rely on randomized trials. The mandate for studies of cardiovascular safety for all newly approved antidiabetic drugs has resulted in at least 19 large clinical end point trials that are either completed or ongoing, although many of them have been initiated with the secondary aim of demonstrating efficacy in preventing cardiovascular events in order to receive supplemental approval for that indication.[22,23]

## DATA RELEVANCY CONSIDERATIONS

In order to study a causal treatment effect in a healthcare database, the following four features need to be measured: (i) the population inclusion and exclusion criteria to characterize the target population and understand its generalizability, (ii) the exposure status, (iii) the outcome, and (iv) the confounding factors that influence both the treatment decision and the outcome. The measurement of each of those features comes with specific measurement characteristics that are quantified by metrics like sensitivity and specificity for binary variables or mean squared difference and proportion missing for continuous variables (**Figure 2** middle column). For example, accuracy of measurement of diabetes based on the presence of a relevant diagnosis code is characterized with sensitivity, specificity, and positive predictive value. Accuracy of measurement of a continuous measurement, such as Hemoglobin $A1_c$ is characterized by the proportion of patients who have the measurement of interest in the time period relevant for the study, for example, during the 3 months prior to treatment initiation, and by the error in the measurement.

In a specific study, most of these measurement characteristics are unknown and approximated from prior experience in the same or similar data. Investigators often apply general principles of constructing measurements in such data to increase the likelihood that the measurement characteristics will improve in the desired direction (**Figure 2** left column). However, even if the exact measurement characteristics of the four study features are perfectly known, it is not obvious how good is good enough for unbiased treatment effect estimation (**Figure 2** right column).[14] It is well known that in randomized trials with primary data collection not all measurements are perfect[24]; nevertheless, we see them usually as the best evidence available. Often, the values of the measurement characteristics that are required depend on the setting. For illustration, if one wishes to estimate a risk ratio, it is advisable to have the highest possible specificity of the outcome measurements, whereas if the estimate of interest is the risk difference, a high sensitivity is preferred to reduce bias.[7] Data relevance is also informed by additional issues, such as the potential representativeness of the population, drug uptake, and the person-time under observation.[25]

## ASSESSING THE VALIDITY OF INDIVIDUAL RWD ANALYSES

Building on the considerations for data relevancy, both sponsors and regulators need to be able to identify the features of a high-quality study that are likely to result in actionable evidence. This evaluation must incorporate assessment of when an RWE study is appropriate and how to design the study in a way that promotes

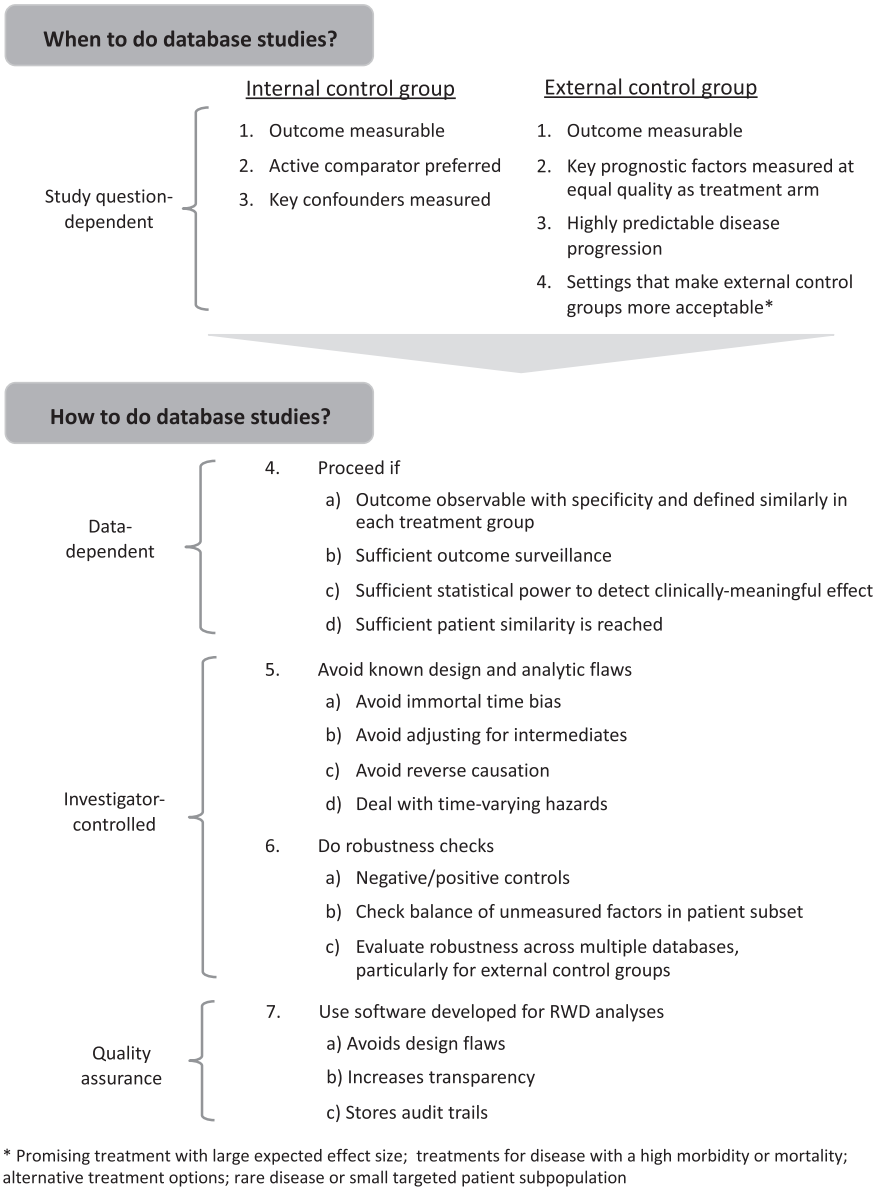


**Figure 3** A summary of general recommendations on when and how database studies of drug effects are likely to produce actionable causal insights similar to randomized controlled trials using an internal control group (a comparative study where both treatment groups are derived from real-world data (RWD)) or an external control group (where only patients in the control group are derived from RWD).

validity. We have previously discussed when and how to conduct RWD analyses in the context of a comparative study with an internal control group.[26] In **Figure 3**, we summarize the considerations for comparative studies, and we additionally note the considerations for single-arm RWD analyses that are used as an external control to a single-arm trial.

A comparative database study is appropriate only when both the outcome and exposure are measurable in the available data and when key confounders that are not balanced through study design are measured as well. Although nonuser comparator groups are possible in RWE studies, active comparators are preferred because patients taking an alternative therapy with similar indications as the therapy under study are likely to be more similar to the treated patients, thereby reducing concerns about unmeasured confounding. Indeed, nonuser comparator groups have been shown in several examples to be very different from patients receiving active treatments.[27,28] Therefore, the key confounders that must be measured are those that are unlikely to be balanced through selection of the comparator group or other features of the study design.

The outcome must be measurable also when using RWD to form an external control group (**Figure 3**). In addition, the disease under study should have a highly predictable progression with low variability across patients, and key prognostic factors that predict disease progression should be measured in the database at approximately equal quality to that in the treatment arm from the clinical trial. As noted in earlier sections, other factors that make an external control group more acceptable include scenarios with a promising treatment with a large expected effect size or a treatment for

a highly fatal or debilitating disease and few alternative treatments, where it would be unethical to randomize patients. In rare diseases or in scenarios with targeted therapies for a small subpopulation, such as cancers with a rare genetic subtype, it may be difficult to identify enough patients to conduct a well-powered clinical trial; therefore, identifying external control patients from a healthcare database may be preferable.

Although the considerations for when to conduct an RWE study are quite different between internal and external control group studies, the key considerations for how to conduct the analysis are largely overlapping (**Figure 3**). Data-dependent issues include whether the outcome is observable in the RWD database with high specificity and defined similarly in the two treatment groups, allowing for unbiased estimates of relative treatment effect[29] and with sufficient sensitivity to yield meaningful conclusions. Investigators should also check, prior to conducting comparative analyses, that there is a sufficient number of outcomes recorded to result in good statistical power in the population under study.

Confounding adjustment methods, such as propensity score methods, have not often been used in the context of external control groups. However, a single-arm trial with an external control group is a type of observational study. Indeed, it is known that external control groups tend to have worse outcomes than a similar control group in an RCT.[30,31] This discrepancy may be due to the fact that control groups in a randomized trial need to meet inclusion and exclusion criteria that are generally more stringent and identify a less-sick population than is typical of external control groups. In the setting of a contemporaneous external control group identified in RWD, inclusion and exclusion criteria can be applied, secular differences can be avoided, and propensity score–based methods may be expected to reduce differences between the treatment group and the external control group, leading to more valid comparisons and less biased estimates of treatment effect.

For RWE studies with either internal or external control groups, investigators should confirm that the strategy for confounding adjustment is able to achieve balance in the covariates between treatment groups.[32] If balance has not been achieved, investigators can consider modifying the confounding adjustment strategy prior to evaluating comparative results. However, if balance can only be achieved with a strategy that prunes a large proportion of the study sample, then the treatment groups may simply be too different to be compared in the available data.

Common avoidable design and analytic flaws have been discussed primarily in the context of comparative RWE studies with an internal control group, but they are also potential concerns in an RWE study using an external control group in which one or more existing treatments are available. For example, problems with immortal time can result if a nonuser external control group is identified by selecting patients with recorded indications for treatment who are not treated throughout follow-up.[33,34] In this situation, patients who survived long enough to receive treatment during follow-up are systematically excluded from the control group, potentially making the control group seem worse. As in RWE studies with internal control groups, the identification of external control groups should also ensure that covariates are measured prior to the start of exposure to an existing treatment to avoid adjusting for causal intermediaries.[35] Finally, design of the external control group should acknowledge the impact of time-varying hazards and, ideally, identify patients at a similar point in their disease course as the patients in the treated arm in order to avoid depletion of susceptibles.[36]

Once a design and analysis plan is selected, robustness checks can help regulators and sponsors gauge the likely reliability of the findings (**Figure 3**). Replicating negative or positive control comparisons, where the true effect or lack of a true effect is known from prior studies, can increase confidence in the database and the study design.[37–39] If some confounders are measured only for a subset of the study population, and, therefore, not used in confounding adjustment procedures, then confirming that balance is achieved on those confounders by proxy in the subset can mitigate concerns about unmeasured confounding.[40] If differences remain between treatment groups, the amount of imbalance can inform formal sensitivity analyses that evaluate the potential impact of unmeasured confounders on study conclusions.[41] Evaluating the variability of findings across multiple databases can also be helpful for establishing reliability, and it is especially important in the context of an external control group, because a well-chosen control group that is a good match for the treated patients should vary little across databases or should vary only in ways that are predictable given the measured prognostic factors.[12]

Relying on RWE for regulatory decisions also requires measures to ensure that analyses are completed and reported accurately. There have been several recent initiatives aimed at improving transparency in reporting of RWE database studies by encouraging standardized and comprehensive reporting of design and analytic choices.[42,43] RWD analyses can further be enhanced by using a software developed specifically for RWD.[44–46] Such software can reduce human error that can occur when programming code is written *de novo* for specific analyses, encourage good design by prohibiting designs that are known to lead to bias, increase transparency by clearly documenting design and analysis choices in human-readable language rather than statistical programming code alone, and store audit trails of all analyses done, ensuring that reported analyses have not been cherry-picked based on results that agree with the sponsor's goals.

## A STRUCTURED PROCESS FOR USING RWD TO SUPPORT DRUG REGULATION

RCTs are a trusted source of evidence in drug regulation not only because randomization allays concerns about confounding but also because there is a strong and well-understood process in place at the FDA, the European Medicines Agency, and other regulatory agencies for how RCTs can be designed and implemented in a way that facilitates regulatory decision making, accompanied by several decades' experience in translating the results from these studies into decisions. Although postmarket RWD has been used to evaluate medication safety, the use of healthcare database analyses to potentially support product effectiveness and approval is a relatively new concept. In addition, as noted in the previous section, there are many clinical scenarios that are unlikely to result in a successful RWD analysis. Therefore, study sponsors will require guidance on deciding

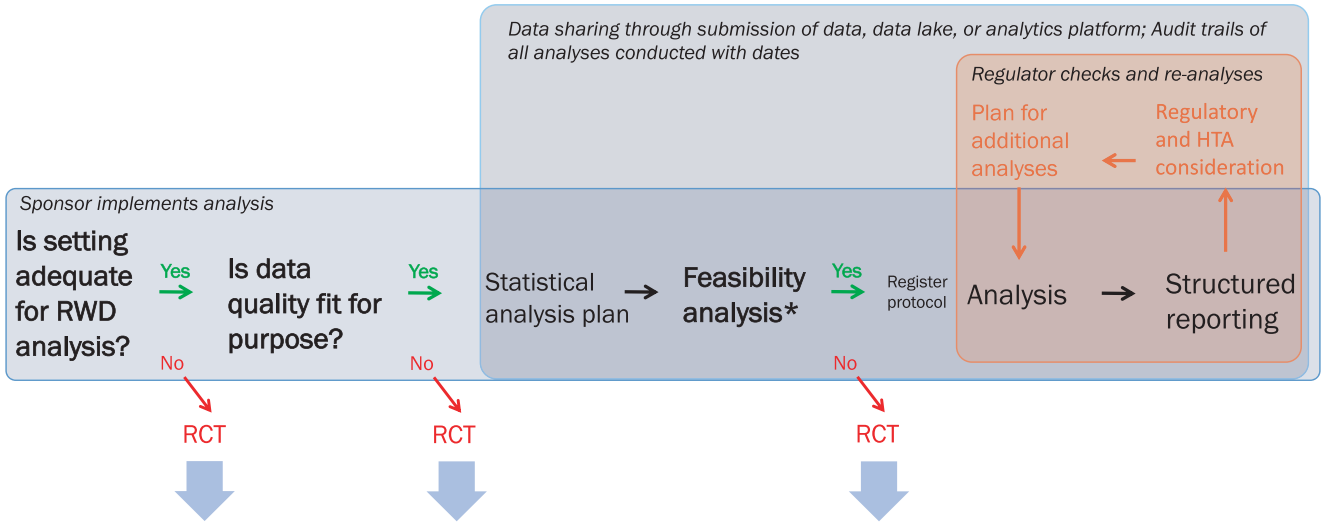


**Figure 4** A structured process for implementation of real-world data (RWD) analyses for regulatory decision-making. *Feasibility analysis can include checking covariate balance after applying the chosen confounding adjustment strategy, checking statistical power, evaluating positive or negative control outcomes, and other analyses without evaluating the study outcomes in the two treatment groups. HTA, Health Technology Assessment; RCT, randomized controlled trial.

when to conduct an RWD analysis vs. when to make the investment in an RCT and how to structure design and analysis activities. We propose one template that could be considered for this process, summarized in **Figure 4**.

The proposed process incorporates the suggestions for when and how to do RWD analyses from **Figure 2** into a structured process of elimination (blue box in **Figure 4**) with the default being an RCT with primary data collection. For example, if the outcomes or exposures are not measured in RWD, if there are no alternative treatments to use as a comparator so that placebo control is required, or if strong confounding is expected that will be unmeasured in RWD, then the study setting is inappropriate for RWD, and an RCT is required. If the study setting is believed to be amenable to an RWD analysis, then a specific RWD source must be selected and evaluated to determine whether the population and measures are sufficient to answer the question with high validity.

Assuming an appropriate database is available, the investigators can proceed with developing the protocol and statistical analysis plan, detailing selection of the study population, methods for confounding adjustment, and specification of primary vs. secondary analyses. After initial cohorts are extracted from the RWD, investigators are encouraged to evaluate the statistical power expected in a comparative analysis given the observed number of events, combining data from the two treatment groups, to ensure comparative results do not influence the decision to proceed. In addition, investigators should evaluate the level of covariate balance that can be achieved between treatment arms in the RWD without proceeding to the analysis of outcome data. The check of covariate balance could provide a final decision point for regulators and investigators to determine whether the selected RWD source is capable of answering the question of interest. At this point, if all parties agree that the RWD analysis is likely to result in actionable evidence, then the study protocol is registered and comparative analyses can subsequently be completed.

The study process further incorporates the need for registration of RWD analyses, transparent implementation of a study, complete reporting, and audit trails of every analysis conducted (gray box in **Figure 4**). The audit trail allows for regulators to verify that the recommended process sequence was respected and that comparative results were not evaluated before the protocol was finalized and registered. This system would enable regulators to repeat the exact same study and change assumptions or definitions in the design and statistical analysis either through submission of data or by providing access to the data. When data belong to third parties, licensing agreements that enable sharing through data lakes and analytics platforms may need to be drafted (orange box in **Figure 4**).

## EMPIRICAL EVIDENCE ON THE REPRODUCIBILITY OF RWD ANALYSES

If the results of nonrandomized studies from healthcare databases are to be trusted, then they must be reported with enough transparency that an independent group could replicate the results. Direct replicability the ability to exactly reproduce an analysis in a given dataset based on reporting of the methods without accompanying programming code(47), is distinct from reproducibility, which assumes programming code is available, as well as conceptual replicability, which is focused on whether findings replicate in a new experiment or new dataset.[47,48] Direct replicability of RWD analyses is important, as it not only allows other investigators to independently verify the findings, but also ensures that details of the RWD analysis are reported completely such that other groups can appropriately judge the rigor and scientific merit of the design and analysis. In order to quantify the level of replicability of the published RWD literature, past work has attempted to replicate 40 published studies using one healthcare claims database.[44] Although both treatment effects and baseline characteristics were typically similar between the original findings and the replications, there were some large differences (**Figure 5**). In addition,

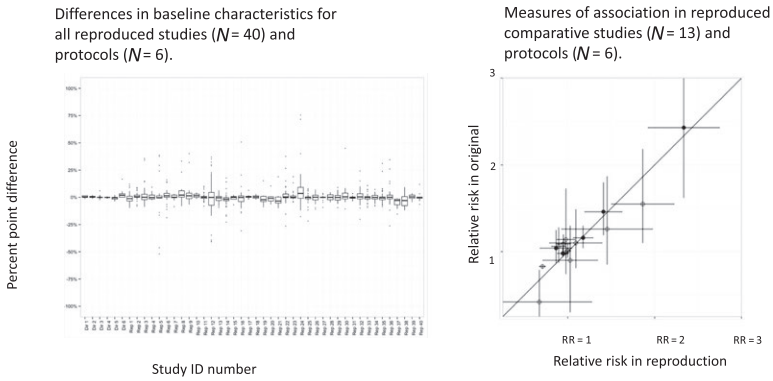


**Figure 5** The ability to replicate real-world data (RWD) analyses is a foundational requirement for regulatory decision making. Reprinted with permission from Wang *et al.*[44]

several replications required some trial and error to find the correct specification that led to similar findings when the original publication did not report the specification sufficiently clearly. Replicability was generally better when the replication investigators had the original protocol available, as details that are often left out of a publication were clearly specified in the protocol, highlighting the need for additional reporting on implementation outside of published papers, which often have space constraints that prevent full reporting.

Although this study provided some data on the level of replicability that can be expected in the current literature, it included a small and highly selected set of publications. The REPEAT initiative (www.repeatinitiative.org) seeks to extend these data by evaluating replicability more broadly across published RWD analyses of medication effectiveness and safety. This initiative plans to assess 250 publications based on analyses of healthcare claims for completeness in reporting, of which 150 will be replicated. Moving forward, confidence in the findings of RWD analyses requires that at least some RWD analyses, particularly those with high-impact or important regulatory consequences, continue to be replicated. However, there is currently little incentive for replication activities, either by the original investigator or by other teams. Therefore, new strategies will be needed for encouraging replication of RWD analyses, which may include funding streams for replications, an independent global consortium of database analysts who work together to sample and replicate findings, or replication groups within regulatory agencies.

## EMPIRICAL EVIDENCE ON THE VALIDITY OF RWD ANALYSES

### Past assessments of RWD

Despite procedures to ensure careful design and implementation of high quality RWD analyses, many in the clinical research community remain unpersuaded that a nonrandomized study, particularly one based on a healthcare database, can routinely match the results of an RCT asking a similar question. The history of nonrandomized RWD analyses is full of high-profile failures, where findings from nonrandomized studies were overturned in subsequent RCTs. Although *post hoc* analyses sometimes reveal which specific design choice caused the misleading findings,[36] many researchers and practicing physicians have come to believe that nonrandomized studies of treatment effects are categorically not valid and, therefore, not a source of information for treatment decision making. These failures are drowning out a more nuanced message that appropriately designed RWE studies have often aligned with the evidence from RCTs. The field has strengthened considerably in the last 2 decades, as epidemiologists have consolidated knowledge about when and how valid healthcare database studies can be conducted.[26] Many of the biases found to plague earlier studies can now be avoided by adhering to principled design and analytic choices for RWD analysis (**Figure 3**).

Several systematic reviews have attempted to answer the question as to whether RWD analyses can systematically match the results of well-controlled RCTs by comparing the results of published nonrandomized and randomized studies of the same clinical question.[49] Nearly all of these reviews have found at least some large differences in estimated treatment effects between the two study types, but none of them attempted to understand why a specific study implementation failed or to remove those studies that had obvious known design flaws, such as immortal time bias. Therefore, these systematic reviews do not provide an accurate representation of the agreement that could be achieved between RCTs and RWD analyses using current state-of-the-art methods.

### New evidence on the validity of RWD analyses

Recent healthcare database studies that follow modern epidemiologic principles could provide a new evidence base on the agreement between RWD analyses and RCTs (**Figure 6**). In the assessment of potentially harmful drug side effects, healthcare database studies identified an increased risk of death following cardiac surgery for patients treated with aprotinin,[50] which was later confirmed in the Blood Conservation Using Antifibrinolytics in a Randomized Trial (BART) trial.[51] A database study of tocilizumab in patients with rheumatoid arthritis supported the cardiovascular safety of this drug relative to alternative treatments,[52] and another database study supported no difference in the combined cardiovascular safety of allopurinol vs. febuxostat for gout.[53] Both findings were presented at scientific meetings prior to the presentation of trial results, which confirmed those findings.[54,55] Studies

| | | Real World Data Analysis | Randomized Controlled Trial |
|---|---|---|---|
| **Studies of harmful effects (Safety)** | (1) | Schneeweiss S, et al. **Aprotinin during coronary-artery bypass grafting and risk of death**. NEJM 2008; 358:771–83* | Fergusson DA, et al. **A comparison of aprotinin and lysine analogues in high-risk cardiac surgery**. NEJM 2008; 358:2319–31 |
| | (2) | Kim SC, et al. **Cardiovascular safety of tocilizumab versus tumor necrosis factor inhibitors in patients with rheumatoid arthritis** - a multi-database cohort study. Arthritis Rheumatol 2017; 69(6):1154–64* | Giles JT, et al. **Comparative cardiovascular safety of tocilizumab vs etanercept in rheumatoid arthritis**: Results of a randomized, parallel-group, multicenter, non-inferiority, phase 4 clinical trial [abstract]. Arthritis Rheumatol 2016; 68 (suppl 10) |
| | (3) | Zhang MA, et al. **Risk of cardiovascular events in older patients with gout initiating febuxostat vs allopurinol**: A population-based cohort study. Circulation 2018; 138:1116–26* | White WB, et al. **Cardivascular safety of febuxostat or allopurinal in patients with gout**. NEJM 2018; 378:1200–10 |
| **Studies of beneficial effects (Effectiveness)** | (4) | Patorno E, et al. **Cardiovascular safety of canagliflozin versus other non-gliflozin antidiabetic agents**: A population-based cohort study . BMJ 2018; 360:k119* | Neal B, et al. **Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes**. NEJM 2017; 377:644–57 |
| | (5) | Gökbudget N, et al. **Blinatumomab vs historical standard therapy of adult relapsed/ refractory acute lymphoblastic leukemia**. Blood Cancer J 2016; 6:e473 | Kantarjian H, et al. **Blinatumomab versus chemotherapy for advanced acute lymphoblastic leukemia**. NEJM 2017; 376:836-47 |
| | (6) | Seeger JD, et al. **Safety and effectiveness of dabigatran and warfarin in routine care of patients with atrial fibrillation**. Thromb Haemostasis 2015; 114:1277–89 | Connolly SJ, et al. **Dabigatran versus warfarin in patients with atrial fibrillation**. NEJM 2009; 361(12):1139–51. |
| | (7) | Fralick M, et al. **Using healthcare claims databases to identify supplemental indications of approved medications**. JAMA Internal Medicine 2018; 178:55–63 | ONTARGET Investigators, Yusuf S, et al. **Telmisartan, ramipril, or both in patients at high risk for vascular events.** NEJM 2008; 358:1547–59 |

* We provide citation to the published manuscript here. Initial results were presented at FDA or scientific conferences before the corresponding RCT findings were known.

**Figure 6** Seven examples of real-world data studies that followed recommendations in **Figure 3** and matched the results of a randomized controlled trial (RCT). For each example, the study that was released first is shaded.

of treatment safety are sometimes considered to be more predictably valid if there is little patient channeling into treatment groups based on concerns about a yet unknown adverse event.

In contrast, nonrandomized phase IV studies of beneficial treatment effects that may support secondary approvals are considered more challenging, as physicians will make treatment decisions on the basis of disease severity and prognosis, leading to strong confounding by indication in many situations. However, even in the setting of effectiveness studies, there are recent encouraging examples of alignment (**Figure 6**). The findings of the Canagliflozin Cardiovascular Assessment Study (CANVAS) trial evaluating the effectiveness of canagliflozin for prevention of cardiovascular events were predicted by an RWE study presented months before the trial was completed, confirming the reduction in heart failure hospitalizations.[56,57] The effectiveness of blinatumomab for the treatment of acute lymphoblastic leukemia was documented in a study that utilized data from a single-arm trial of blinatumomab vs. an external control group derived from RWD sources.[58] Analyses controlled for prognostic factors that may have differed between the blinatumomab arm and the external control group using weighting and propensity score methods. Study findings were later confirmed by the Blinatumomab Versus Standard of Care Chemotherapy in Patients With Relapsed or Refractory Acute Lymphoblastic Leukemia (ALL) (TOWER) trial.[59]

Using RWD to mimic the design of randomized trials and predict trial results prior to their release, as in the previous examples, provides important evidence that RWD can sometimes be used to support regulatory and treatment decision making. However, producing such examples requires identifying ongoing high-quality trials that are likely to be replicable in RWD. Therefore, there remain only a small number of these examples completed. To supplement this evidence base, some investigators have also begun to replicate published RCTs to identify whether RWD analyses can *post hoc* match the results of the RCT.

A retrospective example, mirroring the data availability in 2009 when telmisartan received a secondary approval for the reduction of cardiovascular risk, showed that the Ongoing Telmisartan Alone and in Combination with Ramipril Global Endpoint Trial (ONTARGET) trial findings of noninferiority could have been generated from RWD at that time (**Figure 6**).[37,60] As noninferiority findings in database studies can be the consequence of measurement error biasing results toward the null, the study also included an analysis confirming ramipril's increased risk of angioedema to demonstrate "assay sensitivity" of the analysis.[61] In addition, a comparison of dabigatran and warfarin that was designed to match the Randomized Evaluation of Long-Term Anticoagulation Therapy (RE-LY) trial replicated the effects sizes observed in the RCT.[19,62] All of these studies used new-user, active-comparator cohort designs with extensive covariate adjustment via propensity score methods, following the guidance summarized in **Figure 3**.

## Further building the evidence base for RWD

Given the potential uses of RWE for augmenting the information from RCTs in drug regulation and the growing list of RWD analyses that have either predicted or *post hoc* matched the results from RCTs, how much confidence should physicians and patients have

in the results of such studies? How can the FDA translate these findings into standards for regulatory decision making? Just as a few anecdotes of failed database studies do not provide sufficient evidence to condemn all such studies, the examples of successful studies provided here cannot ensure the validity of all future RWD studies. Two key issues need to be resolved through additional experience with RWE studies: How will we know from the outset with high confidence that an RWE study, after its completion, will provide us with a level of evidence that is acceptable for decision making? Among completed studies, how will we reliably identify high-accuracy RWEs and separate out low-quality RWEs in a time-efficient way?

A starting point is the consideration on when and how to use RWD analyses summarized in **Figure 3**. However, to better understand how those criteria can guide decision makers, we now need a systematic evaluation of the ability of different RWD study designs to consistently match the results of RCTs across a number of selected clinical questions. Comparing published RWD and RCT findings, accounting for methodological quality in the RWD implementation, provides one avenue of research. However, the published RWD studies may be subject to publication bias, where results may have gone unpublished, depending on the findings. In addition, published RWD studies often were not designed specifically to match the design of RCTs and may instead intentionally include populations not included in the RCTs to evaluate the generalizability of the RCT findings.

An alternative approach is to prospectively design new RWD studies to match the design of a large sample of published RCTs. This approach avoids publication bias by prespecifying all RCTs that are being replicated, and it allows for studies to be designed to match the RCT as closely as possible. Creating new RWE studies also allows for extensive sensitivity analyses to identify whether alternative designs or analyses could have improved agreement between the RWE finding and the RCT finding. This type of study could, therefore, provide an empirical evidence basis capable of quantifying the level of confidence that can be assumed for high-quality RWD analyses using state-of-the-art methods across several clinical areas. The FDA has recently funded an initiative to do just that, focused on replicating at least 30 published RCTs in healthcare claims data (www.rctduplicate.org). This project will also inform future work focused on replicating ongoing RCTs with results not yet released to evaluate whether RWD analyses can *a priori* predict the results of RCTs. Based on this empirical evidence, the agency still has the unenviable task to produce guidance. What rate of replication in a specific therapeutic area is acceptable, and are there other measurable features related to the clinical and regulatory question or the data source that partly determine RWE replication success?

## CONCLUSION

High-quality nonrandomized studies from healthcare databases may improve treatment decision making by patients and their providers by supplementing the evidence available from RCTs and other forms of RWE. If only a small proportion of postmarketing trials were replaced with nonrandomized studies from real-world healthcare databases, this change would translate into faster availability of relevant information to patients and providers using substantially fewer resources in terms of patients randomized, who could instead contribute to other research that requires randomization. Over time, enhanced curation of unstructured secondary healthcare data as well as prospective data capture from registries and mobile technologies could also potentially be leveraged to broaden the utility of existing secondary RWD. In this paper, we attempted to summarize major considerations for building an evaluation framework that can reliably interpret and utilize evidence on therapeutics generated from nonrandomized RWE studies. These considerations include the key use cases, recommendations on when and how to conduct database studies, a process for implementation, and empirical evidence on validity. Refining the science on these issues is crucial to meet the high standards demanded for regulatory approval.

Nonrandomized studies from RWE databases have many advantages for use in regulatory decision making. The large size of many databases can often yield powerful analyses for even rare diseases or rare safety events. In addition, the patient populations derived from RWD reflect the wide spectrum of patients actually using the treatment in routine clinical care, including many patients who are unlikely to ever be studied in a clinical trial. However, these advantages are accompanied by several potential disadvantages, including lack of controlled and standardized measurements and lack of baseline randomization to control confounding resulting from treatment selection. Developing transparent and standardized approaches for RWD analyses and an empirical evidence base that supports those approaches will allow regulators to balance these advantages and disadvantages in decision making.

## FUNDING

This work was funded partially by the Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and by the US Food and Drug Administration (FDA) (HHSF223201710186C). This article reflects the views of the authors and should not be construed to represent the FDA's views or policies.

## CONFLICT OF INTEREST

S.S. is a consultant to WHISCON and to Aetion of which he also owns equity. He is the principal investigator of grants to the Brigham and Women's Hospital from Bayer, Vertex, and Boehringer Ingelheim unrelated to the topic of this paper. R.J.G. has received support from grants to the Brigham and Women's Hospital from AstraZeneca, Kowa, Novartis, and Pfizer for research unrelated to the topic of this paper. J.M.F. and D.M. report no conflict of interest.



1. Schneeweiss, S. *et al.* Real world data in adaptive biomedical innovation: a framework for generating evidence fit for decision-making. *Clin. Pharmacol. Ther.* **100**, 633–646 (2016).
2. Jarow, J.P., LaVange, L. & Woodcock, J. Multidimensional evidence generation and FDA regulatory decision making: defining and using "real-world" data. *JAMA* **318**, 703–704 (2017).
3. Sherman, R.E. *et al.* Real-world evidence—what is it and what can it tell us. *N. Engl. J. Med.* **375**, 2293–2297 (2016).
4. McDonald, L., Lambrelli, D., Wasiak, R. & Ramagopalan, S.V. Real-world data in the United Kingdom: opportunities and challenges. *BMC Med.* **14**, 97 (2016).
5. Bonamici, S. H.R. 34 – 114th Congress (2015-2016): 21st Century Cures Act [Internet]. <https://www.congress.gov/

bill/114th-congress/house-bill/34> (2016). Accessed June 15, 2017.
6. US Food and Drug Administration. Prescription Drug User Fee Act (PDUFA) – PDUFA VI: fiscal years 2018–2022 [Internet]. <https://www.fda.gov/forindustry/userfees/prescriptiondruguserfee/ucm446608.htm>. Accessed June 15, 2017.
7. Schneeweiss, S. & Avorn, J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J. Clin. Epidemiol.* **58**, 323–337 (2005).
8. Schneeweiss, S. Confounding. In *Pharmacoepidemiology and Therapeutic Risk Management* (eds. Hartzema, A.G., Chan, A. & Tilson, H.H.) 345–354 (Harvey Whitney, Cincinnati, OH, 2008).
9. Sugarman, J. & Califf, R.M. Ethics and regulatory complexities for pragmatic clinical trials. *JAMA* **311**, 2381–2382 (2014).
10. US Department of Health and Human Services (DHHS), US Food and Drug Administration (FDA). Guidance for Industry: E10 Choice of Control Group and Related Issues in Clinical Trials. <https://www.fda.gov/downloads/Guidances/UCM073139.pdf> (2001).
11. Sasinowski, F.J.. Quantum of effectiveness evidence in FDA's approval of orphan drugs. Drug Inf J **46**, 238–263 (2012).
12. Simon, R. *et al.* The role of nonrandomized trials in the evaluation of oncology drugs. *Clin. Pharmacol. Ther.* **97**, 502–507 (2015).
13. Eichler, H.-G. *et al.* Adaptive licensing: taking the next step in the evolution of drug approval. *Clin. Pharmacol. Ther.* **91**, 426–437 (2012).
14. Eichler, H.-G. *et al.* From adaptive licensing to adaptive pathways: delivering a flexible life-span approach to bring new drugs to patients. *Clin. Pharmacol. Ther.* **97**, 234–246 (2015).
15. US Food and Drug Administration. Guidance for industry: expedited programs for serious conditions – drugs and biologics [Internet]. <https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM358301.pdf> (2014). Accessed October 2, 2018.
16. Beaver, J.A. *et al.* A 25-year experience of US Food and Drug Administration accelerated approval of malignant hematology and oncology drugs and biologics: a review. *JAMA Oncol.* **4**, 849–856 (2018).
17. Ray, W.A., Stein, C.M., Daugherty, J.R., Hall, K., Arbogast, P.G. & Griffin, M.R. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *Lancet* **360**, 1071–1073 (2002).
18. Southworth, M.R., Reichman, M.E. & Unger, E.F. Dabigatran and postmarketing reports of bleeding. *N. Engl. J. Med.* **368**, 1272–1274 (2013).
19. Seeger, J.D., Bykov, K., Bartels, D.B., Huybrechts, K., Zint, K. & Schneeweiss, S. Safety and effectiveness of dabigatran and warfarin in routine care of patients with atrial fibrillation. *Thromb. Haemost.* **114**, 1277–1289 (2015).
20. US Food and Drug Administration. The Sentinel Initiative: A National Strategy for Monitoring Medical Product Safety. <https://www.fda.gov/downloads/Safety/FDAsSentinelInitiative/UCM124701.pdf> (2008).
21. Yih, W.K. *et al.* Intussusception risk after rotavirus vaccination in U.S. infants. *N. Engl. J. Med.* **370**, 503–512 (2014).
22. Turner, J.R., Karnad, D.R. & Kothari, S. In: *Cardiovascular Safety in Drug Development and Therapeutic Use*. 239–250 (Adis, Cham, Switzerland, 2017).
23. Mannucci, E., Mosenzon, O. & Avogaro, A. Analyses of results from cardiovascular safety trials with DPP-4 inhibitors: cardiovascular outcomes, predefined safety outcomes, and pooled analysis and meta-analysis. *Diabetes Care* **39** (suppl. 2), S196–S204 (2016).
24. Ioannidis, J.P. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
25. Daniel, G., Silcox, C., Bryan, J., McClellan, M., Romine, M. & Frank, K. Characterizing RWD quality and relevancy for regulatory purposes. <https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf> (2018).
26. Franklin, J.M. & Schneeweiss, S. When and how can real world data analyses substitute for randomized controlled trials? Real world evidence and RCTs. *Clin. Pharmacol. Ther.* **102**, 924–933 (2017).
27. Glynn, R., Schneeweiss, S., Wang, P., Levin, R. & Avorn, J. Selective prescribing led to overestimation of the benefits of lipid-lowering drugs. *J. Clin. Epidemiol.* **59**, 819–828 (2006).
28. Glynn, R.J., Knight, E.L., Levin, R. & Avorn, J. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology* **12**, 682–689 (2001).
29. Kelsey, J.L., Whittemore, A.S., Evans, A.S. & Thompson, W.D. *Methods in Observational Epidemiology*, 2nd edn (Oxford University Press, New York, NY, 1996).
30. Sacks, H., Chalmers, T.C. & Smith, H. Randomized versus historical controls for clinical trials. *Am. J. Med.* **72**, 233–240 (1982).
31. Schulz, K.F., Chalmers, I., Hayes, R.J. & Altman, D.G. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* **273**, 408–412 (1995).
32. Franklin, J.M., Rassen, J.A., Bartels, D.B. & Schneeweiss, S. Prospective cohort studies of newly marketed medications: using covariate data to inform the design of large-scale studies. *Epidemiology* **25**, 126–133 (2014).
33. Suissa, S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol. Drug Saf.* **16**, 241–249 (2007).
34. Suissa, S. Immortal time bias in pharmacoepidemiology. *Am. J. Epidemiol.* **167**, 492–499 (2008).
35. Schisterman, E.F., Cole, S.R. & Platt, R.W. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* **20**, 488–495 (2009).
36. Hernán, M.A. *et al.* Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* **19**, 766 (2008).
37. Fralick, M., Kesselheim, A.S., Avorn, J. & Schneeweiss, S. Use of health care databases to support supplemental indications of approved medications. *JAMA Intern. Med.* **178**, 55–63 (2017).
38. Lipsitch, M., Tchetgen, E.T. & Cohen, T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* **21**, 383 (2010).
39. Lipsitch, M., Tchetgen, E.T. & Cohen, T. Negative control exposures in epidemiologic studies. *Epidemiology* **23**, 351–352 (2012).
40. Patorno, E. *et al.* Claims-based studies of oral glucose-lowering medications can achieve balance in critical clinical variables only observed in electronic health records. *Diabetes Obes. Metab.* **20**, 974–984 (2018).
41. Schneeweiss, S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol. Drug Saf.* **15**, 291–303 (2006).
42. von Elm, E. *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med.* **4**, e296 (2007).
43. Wang, S.V. & Schneeweiss, S.. Joint ISPE-ISPOR Special Task Force on RWE in healthcare decision-making. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies v2.0 [Internet]. <https://www.ispor.org/RWD/2017-06-12-Manuscript-Draft-JTF-ISPE-for-membership-review.pdf> (2017). Accessed June 22, 2017.
44. Wang, S.V., Verpillat, P., Rassen, J.A., Patrick, A., Garry, E.M. & Bartels, D.B. Transparency and reproducibility of observational cohort studies using large healthcare databases. *Clin. Pharmacol. Ther.* **99**, 325–332 (2016).
45. Gagne, J.J., Houstoun, M., Reichman, M.E., Hampp, C., Marshall, J.H. & Toh, S. Safety assessment of niacin in the US Food and Drug Administration's mini-sentinel system. *Pharmacoepidemiol. Drug Saf.* **27**, 30–37 (2018).
46. Innovation in everything we do – Sanofi [Internet]. <https://www.sanofi.com/en/science-and-innovation/innovation-in-our-dna/innovation-in-everything-we-do>. Accessed October 15, 2018.
47. Peng, R.D., Dominici, F. & Zeger, S.L. Reproducible epidemiologic research. *Am. J. Epidemiol.* **163**, 783–789 (2006).
48. Goodman, S.N., Fanelli, D. & Ioannidis, J.P.A. What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341 ps12 (2016).

49. Anglemyer, A., Horvath, H.T. & Bero, L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials [Internet]. In: *Cochrane Database of Systematic Reviews* (ed. The Cochrane Collaboration) (John Wiley & Sons, Chichester, UK, 2014). <https://doi.org/doi.wiley.com/10.1002/14651858.mr000034.pub2>. Accessed May 12, 2016.
50. Schneeweiss, S., Seeger, J.D., Landon, J. & Walker, A.M. Aprotinin during coronary-artery bypass grafting and risk of death. *N. Engl. J. Med.* **358**, 771–783 (2008).
51. Fergusson, D.A. *et al.* A comparison of aprotinin and lysine analogues in high-risk cardiac surgery. *N. Engl. J. Med.* **358**, 2319–2331 (2008).
52. Kim, S.C. *et al.* Cardiovascular safety of tocilizumab versus tumor necrosis factor inhibitors in patients with rheumatoid arthritis – a multi-database cohort study: cardiovascular safety of tocilizumab compared with TNF inhibitors. *Arthritis Rheumatol.* **69**, 1154–1164 (2017).
53. Zhang, M. *et al.* Assessment of cardiovascular risk in older patients with gout initiating febuxostat versus allopurinol: a population-based cohort study. *Circulation* **138**, 1116–1126 (2018).
54. Giles, J. *et al.* Comparative cardiovascular safety of tocilizumab vs etanercept in rheumatoid arthritis: results of a randomized, parallel-group, multicenter, noninferiority, phase 4 clinical trial [abstract]. *Arthritis Rheumatol.* **68**(suppl. 10). <https://acrabstracts.org/abstract/comparative-cardiovascular-safety-of-tocilizumab-vs-etanercept-in-rheumatoid-arthritis-results-of-a-randomized-parallel-group-multicenter-noninferiority-phase-4-clinical-trial/> (2016).
55. White, W.B. *et al.* Cardiovascular safety of febuxostat or allopurinol in patients with gout. *N. Engl. J. Med.* **378**, 1200–1210 (2018).
56. Patorno, E. *et al.* Cardiovascular outcomes associated with canagliflozin versus other non-gliflozin antidiabetic drugs: population based cohort study. *BMJ* **360**, k119 (2018).
57. Neal, B. *et al.* Canagliflozin and cardiovascular and renal events in type 2 diabetes. *N. Engl. J. Med.* **377**, 644–657 (2017).
58. Gökbuget, N. *et al.* Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood Cancer J.* **6**, e473 (2016).
59. Kantarjian, H. *et al.* Blinatumomab versus chemotherapy for advanced acute lymphoblastic leukemia. *N. Engl. J. Med.* **376**, 836–847 (2017).
60. Mann, J.F. *et al.* Dual inhibition of the renin–angiotensin system in high-risk diabetes and risk for stroke and other outcomes: results of the ONTARGET trial. *J. Hypertens.* **31**, 414–421 (2013).
61. Temple, R. & Ellenberg, S.S. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Ann. Intern. Med.* **133**, 455–463 (2000).
62. Connolly, S.J. *et al.* Dabigatran versus warfarin in patients with atrial fibrillation. *N. Engl. J. Med.* **361**, 1139–1151 (2009).