When Can Nonrandomized Studies Support Valid Inference Regarding Effectiveness or Safety of New Medical Treatments?

Jessica M. Franklin[1,2,*], Richard Platt[3], Nancy A. Dreyer[4], Alex John London[5], Gregory E. Simon[6], Jonathan H. Watanabe[7], Michael Horberg[8], Adrian Hernandez[9] and Robert M. Califf[10]

The randomized controlled trial (RCT) is the gold standard for evaluating the causal effects of medications. Limitations of RCTs have led to increasing interest in using real-world evidence (RWE) to augment RCT evidence and inform decision making on medications. Although RWE can be either randomized or nonrandomized, nonrandomized RWE can capitalize on the recent proliferation of large healthcare databases and can often answer questions that cannot be answered in randomized studies due to resource constraints. However, the results of nonrandomized studies are much more likely to be impacted by confounding bias, and the existence of unmeasured confounders can never be completely ruled out. Furthermore, nonrandomized studies require more complex design considerations which can sometimes result in design-related biases. We discuss questions that can help investigators or evidence consumers evaluate the potential impact of confounding or other biases on their findings: Does the design emulate a hypothetical randomized trial design? Is the comparator or control condition appropriate? Does the primary analysis adjust for measured confounders? Do sensitivity analyses quantify the potential impact of residual confounding? Are methods open to inspection and (if possible) replication? Designing a high-quality nonrandomized study of medications remains challenging and requires broad expertise across a range of disciplines, including relevant clinical areas, epidemiology, and biostatistics. The questions posed in this paper provide a guiding framework for assessing the credibility of nonrandomized RWE and could be applied across many clinical questions.

The randomized controlled trial (RCT) has been the gold standard for evaluating the effectiveness and safety of medications for more than 50 years.[1] Despite the many advantages of traditional RCTs, there are concerns that the narrowly defined patient population and tightly controlled treatments and settings required in many RCTs for drugs may not reflect treatment effects or outcomes in usual care. In addition, the high costs of both implementation and long-term follow-up in a traditional RCT often constrain the focus to outcomes that can be measured in the shorter term with smaller sample sizes, including intermediate outcomes, biomarkers, or surrogates. For these reasons, real-world evidence (RWE) has been proposed as a complementary source of evidence that can better capture treatments as used in routine care and the subsequent outcomes that are most meaningful to patients.[2,3] RWE has been defined as any evidence regarding the risks and benefits of medications derived from data sources other than traditional RCTs, i.e., real-world data (RWD).[4,5] Under this definition, RWE can be either randomized or nonrandomized.

Randomly allocating treatment to study patients ensures that, on average, treatment groups will be similar with respect to all patient characteristics that may impact risk for the outcome.[6] The resulting balance in patient characteristics enables one to infer that differences in outcomes between treatment groups can be attributed to differences in the treatments under study, rather than other factors. Nonrandomized or observational studies, in contrast, do not use random treatment allocation. As patients and their providers make treatment decisions on the basis of individual patient characteristics and circumstances, patients receiving alternative therapies may differ on many important factors affecting outcomes. While confounders known to influence both treatment assignment and outcomes can be measured and adjusted for in the design or analysis of nonrandomized studies, it can never be guaranteed that all such confounding factors have been controlled. There is always a possibility that there were additional factors unknown to the investigators that may be confounding the observed relationships between treatments and outcomes, leading to inaccurate estimates of treatment effects.

Given the strong control of both known and unknown confounding factors that is a benefit of randomization, why pursue the use of nonrandomized RWE for informing treatment decision

[1]Optum Epidemiology, Boston, Massachusetts, USA; [2]Division of Pharmacoepidemiology & Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA; [3]Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, Massachusetts, USA; [4]IQVIA Real World Solutions, Cambridge, Massachusetts, USA; [5]Philosophy Department & Center for Ethics and Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; [6]Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA; [7]School of Pharmacy and Pharmaceutical Sciences, University of California Irvine, Irvine, California, USA; [8]Kaiser Permanente Mid-Atlantic Permanente Research Institute and Mid-Atlantic Permanente Medical Group, Bethesda, Maryland, USA; [9]Duke Clinical Research Institute, Durham, North Carolina, USA; [10]Verily Life Sciences and Google Health, Cambridge, Massachusetts, USA. *Correspondence: Jessica M. Franklin (Jessica.franklin@optum.com)

Received February 9, 2021; accepted March 25, 2021. doi:10.1002/cpt.2255

making and drug regulation, rather than limiting focus to randomized RWE? There will always be more open clinical questions than there are resources to answer them all with traditional RCTs. For example, large nonrandomized RWE studies can provide evidence on how an average treatment effect, estimated from a traditional RCT, varies across trial subpopulations or across patient populations not included in the trial. In rare diseases, it may be impossible to recruit a sufficiently large number of patients into multiple randomized trials exploring varying clinical questions. Traditional RCTs are often too small to investigate rare adverse medication events, and increasingly globalized drug development means that there is little incentive to conduct postmarketing RCTs in order to access new geographic markets. In addition, the recent proliferation of research-ready longitudinal healthcare databases, including health insurance claims, electronic health records, and patient registries, provide abundant opportunities for nonrandomized research to quickly and efficiently answer questions on medications that are widely used.

There are many well-established uses of such data for providing evidence on medications, including evaluations of drug prescribing, utilization patterns, or adherence.[7] Due to the issues of confounding, use of RWD to infer the causal effects of a treatment is more difficult. The US Food and Drug Administration (FDA) relied on RWD to create the Sentinel System, which uses claims data from multiple databases to quickly investigate medication safety concerns as they arise.[8,9] Even before the Sentinel System, the FDA, the European Medicines Agency (EMA), and other drug regulators have long accepted nonrandomized data to inform regulatory decisions on medication safety.[7,10] Strong confounding is generally less likely in assessment of adverse effects that are unexpected or unrelated to a treatment's therapeutic "target."[11] The use of nonrandomized RWE from healthcare databases intended to support a claim of either drug safety or effectiveness deserves careful consideration, as such studies can more easily lead to erroneous conclusions because of bias, including a greater risk of manipulation to meet desired outcomes.

In this paper, we discuss what is required for valid assessment of medication safety and effectiveness from nonrandomized studies, the topic of the third component of the National Academies of Science, Engineering and Medicine (NASEM) Forum on Drug Discovery, Development, and Translation Workshop Series on RWE, sponsored by the FDA.[12] Although the focus on nonrandomized research is spurred in part by the availability of longitudinal healthcare databases, we discuss principles applicable to nonrandomized research more broadly, including nonrandomized studies using primary data collection and single-arm trials using external nonrandomized control groups. Accurate measurement of study variables remains of fundamental importance to the validity of nonrandomized studies, and measurement needs are further complicated in nonrandomized studies by the need for accurate assessment of confounders in addition to treatments and outcomes. However, we do not directly address data integrity and relevancy concerns here, as they are already discussed in an accompanying paper.

STRATEGIES TO MINIMIZE BIAS

As described above, the principal challenge of nonrandomized research on medications or other medical treatments is controlling for confounding, driven by differing characteristics of patients receiving alternative treatment strategies in real-world care. The alternative treatment strategies being compared may include different medications, different doses or formulations, or treatment vs. no treatment, commonly referred to as a nonuser comparator group. Although *confounding by indication*, caused by intentional choices of clinicians or patients to select different treatments for patients with differing characteristics, is typically thought of as the most pernicious type of confounding,[11] confounding can arise from several other sources. Differences in costs between compared treatments can result in treatment groups with differing socioeconomic status and therefore differing burden of disease and access to high-quality care.[13] Patients who take preventive medications are often more likely to practice other healthy behaviors such as diet and exercise and are less likely to be suffering from major chronic conditions that consume the focus of medical care such as cancer or end-stage renal disease.[14,15]

Does the design emulate a hypothetical randomized trial design?

Control of confounding in nonrandomized research depends strongly on study design, as no analytic method can rescue a seriously flawed design. Although the details of the study design for nonrandomized RWE must be tuned to the specific clinical question, consensus has emerged in the last decade that the design of a nonrandomized study should fit within the "target trial framework" by emulating the design of a hypothetical randomized trial.[16–18] A hypothetical target trial does not need to be a real RCT; it doesn't even need to be a trial that could feasibly be conducted. It serves only as a guiding framework for design of the corresponding nonrandomized study, and there is now strong evidence that using a hypothetical target trial as a design guide can eliminate many of the most egregious design mistakes in nonrandomized studies of medications. It can also help focus thinking on how randomization can be emulated through control of confounding factors. Broad knowledge across relevant clinical, epidemiological, and biostatistical domains is needed. If using an existing database, deep knowledge of the data provenance is also required, as discussed in the accompanying paper on data quality.

Designing a nonrandomized study through emulation of an RCT typically favors cohort studies, where cohorts of patients receiving different treatment strategies are followed over time. However, case-control studies nested within a cohort, self-controlled designs, and some other designs can also be thought of as emulations of trials. Emulation of an RCT also favors new user designs, where patients are followed from the beginning of treatment initiation, vs. prevalent user studies that include patients at varying points along the treatment pathway; alternative designs can focus on patients who do or do not switch from one medication to another, or on patients who do or do not continue treatment past a given time period.[19–21]

The important feature in all of these designs is that the design clearly identifies the inception point of the study, which serves both to anchor the study to the timing of the treatment decision and to anchor all other study measurements, comparable to the time of randomization in an RCT. Specifically, "baseline variables"

are focused on characterizing the study participants at the inception point and should be checked for balance to evaluate comparability of treatment groups, just as is commonly done in RCTs to evaluate the success of randomization. Follow-up begins immediately or shortly after the inception point, and thus, shortly after the treatment decision is made, minimizing the chance that patients with early adverse events are excluded from study follow-up. A clear inception point tied to the timing of the treatment decision also eliminates many errors related to measurement of inclusion/exclusion criteria, treatment status, and baseline covariates during the follow-up period, which can lead to large biases in effect estimates, such as immortal time bias.[17,22]

Thus, if the planned nonrandomized design cannot be envisioned as a corresponding target trial, then the validity of the design and resulting findings are questionable. With this perspective, it becomes clear that many of the published nonrandomized studies of treatment effects using RWD, including a large proportion of studies evaluating treatments for coronavirus disease 2019 (COVID-19), have used less valid designs. For example, several studies of COVID-19 treatments assessed inclusion/exclusion criteria or treatment assignment during study follow-up, a design that clearly would not be possible in a randomized trial.[23] These design issues may account for many of the conflicting findings of such studies.

Is the comparator or control condition appropriate?

Another aspect of the design that is known to have a large impact on confounding is the type of comparator selected. Although the choice of comparator is largely driven by the research question, focusing on active treatment comparator(s) with similar indications and similar treatment modality as the treatment of interest can greatly mitigate the risk of unmeasured confounding.[15,24,25] Nonuser comparator groups are highly suspect, as patients who are receiving treatment are often very different with respect to their disease severity and risk of adverse events compared with patients who are not receiving treatment for the same disease. Even if researchers are interested in demonstrating effectiveness or safety for a given medication, rather than comparing the effectiveness of alternative medications, evaluating medication outcomes against an active alternative treatment may be sufficient to answer the question with lower risk of unmeasured confounding.[26] For example, if research interest is focused on evaluating the cardiovascular effects of a sodium-glucose cotransporter 2 inhibitor for type 2 diabetes mellitus, comparing against another antihyperglycemic drug, such as a dipeptidyl peptidase 4 inhibitor, whose cardiovascular effects are already well known, may be preferred. Use of an active comparator is also considered to yield findings more relevant to real-world decision making in diseases with at least one indicated treatment with known efficacy. Thus, if a nonuser comparator group is used, there should be a strong justification for its necessity and extreme care taken to measure and balance all potential confounders. In addition to similar indications and treatment modality, active comparators with similar formulary access and good availability or market share in the geographic areas of the study are preferred.

Does the primary analysis adjust for measured confounders?

The goal of adjustment is to compare outcomes between the treatment groups, only among patients who are similar with respect to the confounders, thereby eliminating the impact of confounder differences on the estimation of treatment effect. One approach is ordinary multivariable regression, including linear, logistic, proportional hazards or other forms of regression.[27] While multivariable regression is simple and has been in use for estimating the effects of explanatory variables for many decades, it can lead to bias when there are regions of nonpositivity, i.e., patients who are outliers with respect to one or more of the confounders and who have no similar patients in the alternative treatment group against whom they can be compared.[28] Diagnosing areas of nonpositivity and checking balance of confounders in general is difficult in a regression model, as the balancing of confounders is part of the regression procedure itself.

Alternatively, propensity score methods can create balance in confounders and easily diagnose regions of nonpositivity. The propensity score is the probability of treatment assignment, given the confounders.[29] In a randomized trial, the probability of receiving each treatment is known for all patients, as it is defined by the randomization scheme (0.5 for each arm in an RCT with equal distribution between two arms). In a nonrandomized study, the propensity to treatment is typically unknown and must be estimated using observed data, often with a logistic regression model. It has been shown that creating balance in the distribution of the propensity score between treatment groups will on average balance the variables that went into estimation of the propensity score model.[29,30] Therefore, use of propensity score methods can emulate how randomization balances baseline factors in an RCT, except that propensity score methods balance only those factors measured and included in the propensity score model, while randomization theoretically balances all factors, both measured and unmeasured.

There are now many variations on how the propensity score can be utilized to create balanced treatment groups, but two of the simplest and most common approaches in RWE on medications are matching and weighting on the propensity score.[31–33] An important advantage of these methods is the fact that balance on confounders can be directly evaluated in the matched or weighted patient sample, similar to the evaluation of balance in an RCT.[34] If acceptable balance has not been achieved, the propensity score approach must be modified and reimplemented until acceptable balance is reached, potentially including removal of patients from the study sample if there are no comparable patients in the alternative treatment group. Evaluation of balance should consider the study question and the importance of individual confounders, as tighter balance may be required for risk factors with stronger relationships with the outcome.

In all of the adjustment approaches discussed above, confounders must be selected prior to implementing adjustment. Confounders should include at minimum all variables that impact both treatment assignment and outcomes, but adjusting for all variables that impact the outcome, regardless of whether they impact treatment assignment, has been shown to lead to the most precise treatment effect estimates.[35,36] Instrumental variables

(IVs) that impact treatment assignment but not outcome except through treatment should not be adjusted for as they can increase bias from unmeasured confounding as well as decrease estimate precision.[37,38] Therefore, when selecting confounders for adjustment for an outcome with a previously developed risk score, all factors incorporated into the risk score should be considered. Other factors not typically found in risk scores that can nonetheless impact the outcome, such as socioeconomic status and access to care, should also be considered. As noted previously, good measurement of confounders is critical to the success of confounding adjustment, and increasing misclassification or mismeasurement in confounders leads to increasing bias in effect estimates.[39]

Although the selection of variables for confounding adjustment is clear in theory, in any given nonrandomized RWE study, it can never be known with certainty which variables are predictive of outcome, of treatment, or neither. In nonrandomized research in existing healthcare databases, selection of confounders from the thousands of measured variables can be especially difficult. There are now many approaches to automated variable selection, which are capable of sifting through a large number of measured variables to identify those most likely to contribute to confounding by evaluating variable associations with outcome and treatment. For example, the freely accessible high-dimensional propensity score algorithm, created for use with health insurance claims data, automatically creates binary variables describing the frequency of unique diagnoses, procedures, and medication dispensations in the claims data.[40] It then calculates an approximation of the expected bias in the treatment effect estimate due to each variable based on the variable's prevalence and univariate associations with treatment and outcome. The bias calculation is used for prioritizing variables for inclusion in the propensity score model. Alternatively, several other machine learning–based approaches focus on modeling the propensity score and/or the outcome while simultaneously selecting variables for adjustment.[41–44] Many of these have been shown to have very good theoretical properties, but may need additional work to scale to large healthcare databases common in RWE.[45,46]

Debate is still ongoing regarding whether investigator selection of covariates or automated confounder selection is preferred in studies based on existing healthcare databases. However, there is increasing consensus that automated approaches, guided by knowledgeable investigators, may provide the best of both worlds.[33,47,48] At minimum, if investigator selection is used for primary analyses, automated procedures can provide a useful sensitivity analysis to identify whether there are any important variables that were missed by investigators. If automated procedures are used for primary analyses, then investigators should review the list of confounders selected by the automated approaches to evaluate whether additional variables should be included. Allowing automated procedures to determine adjustment for primary analyses without reporting and assessment of the adjustment variables and achieved balance is not recommended. Often, it is worthwhile to seek collaboration or consultation from developers of the primary data source to ensure appropriate employment of data in development of covariates.

DO SENSITIVITY ANALYSES QUANTIFY THE POTENTIAL IMPACT OF RESIDUAL CONFOUNDING?

Even if the study design and adjustment strategy has attempted to thoroughly account for potential differences between treatment groups that could lead to confounding of the treatment effect estimate, residual confounding cannot be ruled out. Therefore, nonrandomized studies that will be used for regulatory and treatment decision making must consider the potential impacts of residual confounding in sensitivity analyses. Sensitivity analyses could directly evaluate unmeasured confounders by, for example, evaluating the balance of confounders that were unavailable for adjustment for the full study population but are measured in a subset of the study population through data linkage.[49] Another possibility is assessing the effects of treatments on control outcomes that share a similar confounding mechanism as the outcome of interest but have a known relationship with treatment.[50,51] Replicating the known effect on the control outcome provides some assurance that confounding has been well controlled. Even when there is no suitable control outcome available and no information on confounders available through data linkage, quantitative bias analysis can be performed to identify the magnitude of unmeasured confounding that would be required to invalidate the conclusions from the study.[52,53] This information can then be compared with the plausible range of associations of likely confounders with treatment assignment and outcome to identify whether unmeasured confounding remains a significant concern in the study.

IV methods are an alternative confounding adjustment approach that, unlike the propensity score methods described above, can adjust for unmeasured confounders. Utilizing an IV adjustment approach requires the availability of a valid IV that is predictive of treatment but affects outcome only through treatment.[54,55] Randomization in an RCT can be thought of as an IV, as randomization does not impact outcomes except through its influence on patient treatment. Occasionally, formulary or policy changes that lead to significant modifications in treatment choice can provide a "natural experiment" that can be used for constructing an IV analysis.[56] Other suggestions for IVs in nonrandomized RWE studies of medications include physician prescribing preference, hospital or regional preferences, or specialist access.[57–60] However, all of these potential IVs have been criticized as likely correlated with outcome and therefore confounded.[61] IV analyses also typically have lower statistical power than traditional adjustment approaches. Thus, these approaches may be inappropriate for primary analyses or for studying rare medication safety events but can provide supplemental assessment of unmeasured confounding in studies of intended medication effects, where the risk of residual confounding is higher.

ARE METHODS OPEN TO INSPECTION AND (IF POSSIBLE) REPLICATION?

Protocol registration

Despite the progress made over the last few decades in understanding the study designs and methods that allow for valid inferences from nonrandomized studies, one major hurdle that remains is agreement on study processes that can ensure transparency and

integrity of analyses. For example, all randomized trials that are either published in a major medical journal or used for a regulatory decision are required to register a detailed protocol at ClinicalTrials.gov prior to enrolling the first study patient. Registration of the trial protocol ensures that design and analytic choices were made prior to evaluation of study outcomes and were therefore not influenced by the results. In the context of a nonrandomized study, protocol registration can enhance the likelihood that the investigators or other parties do not modify the design and analysis in order to produce a desired result, as changes to the protocol occurring during the conduct of the study and their rationale would be documented.

Comparable to registration in randomized trials, registration of nonrandomized studies is often completed prior to study start, i.e., prior to conducting any analytic work on the study question of interest (but likely after an appropriate data source with a sufficient number of patients of interest has been identified). This approach is often used, for example, for studies registered on the European Union postauthorization studies register prior to data accrual (www.encepp.eu). However, some investigators may be resistant to this recommendation, as nonrandomized database studies are often designed adaptively with initial learnings on confounding mechanisms observed in the database contributing to decisions on study design. Alternatively, one could simply blind herself to outcome information, but use other data, such as balance on confounders, to contribute to study design.[62] This approach guards against data dredging but does not allow information on observed outcome risks to contribute to the study design. While a strong need for evidence on important safety questions sometimes means that even underpowered analyses should proceed, assessments of effectiveness are often postponed if there is not sufficient power to provide meaningful results. Thus, a more flexible option is to allow use of outcome information overall (not separated by treatment groups) in order to estimate likely study power given the number of observed events.[63] Registration of the protocol could then take place after these initial analyses of feasibility and validity have been conducted and the design and analysis are finalized.

A related question concerns to what extent outcome information can be used to aid investigators in the selection of confounders to be used for adjustment. As noted in prior sections, many machine learning–based approaches now utilize outcome modeling in order to simultaneously identify confounders and estimate treatment effect, making prespecification of the final set of covariates impossible. However, the approach to automated confounder selection should be prespecified in the protocol, thereby maintaining the benefits of preregistration and allowing for a diverse set of confounder adjustment methods.

The challenge in all of these approaches is guaranteeing that investigators did not evaluate treatment effect estimates prior to selecting the design and analysis. Even registration of the protocol does not eliminate this concern, as there is no way to know whether investigators conducting retrospective RWE previously identified favorable designs and analyses prior to registering the protocol. There are now some analytic platforms for healthcare databases that provide an audit trail of all analyses that have been conducted that could then be shared with regulators to verify that comparative analyses of outcomes were not conducted until after the protocol was registered.[64] However, if data were available to investigators outside of the platform, it may still be possible to evaluate treatment effect estimates outside of the platform prior to protocol registration.

Sharing results

Rather than relying on the integrity of investigators or on analytic platforms to guarantee that study design and analysis were prespecified, an alternative approach is to instead allow regulators or other evidence consumers to replicate or reproduce findings independently of the sponsor. This approach could include supporting replication of study findings in a separate real-world database or setting, using the same methodology as the original study. Alternatively, sponsors could provide data supporting primary analyses so that regulators could reproduce primary analyses and examine whether results are robust to changes in design or analysis. Submission of study data is required for FDA submissions involving RCTs and is typically provided for nonrandomized studies based on primary data collection, but is not uniformly required across regulatory agencies. Given the increased concerns about the quality of both data and design in nonrandomized RWE based on healthcare databases, one might expect that the submission of study data for regulatory submission of such studies is more imperative. However, sharing data is more difficult in this context. Databases derived from patient records may be controlled by healthcare systems or payers who license use of the data to investigators under agreements that prohibit the further sharing of patient-level data. While they may make exceptions for sharing data with regulatory agencies, which data should be shared? Extensive modifications to the design or analysis (using different comparators, different washout periods for defining new use of a drug, etc.) could require access to the underlying healthcare database rather than simply sharing of a final analytic data set.

Sharing of individual-level data should mitigate risks of reidentification. Several privacy-preserving options for data sharing for RWE based on healthcare databases have been described previously in the literature.[65] For example, sharing healthcare databases with regulators may be done via an "archive," where regulators or other research consumers can access data directly via an analytics platform or data access portal, or via an "enclave," where regulators can submit queries and receive aggregate results. One example of the enclave model is the Centers for Medicare and Medicaid Services (CMS) Virtual Research Data Center (VRDC), which allows remote access to CMS data and transfer of aggregate results through the submission of Statistical Analysis System (SAS) queries.[66] Similar systems could be setup for other large research databases but would require substantial additional investment from database aggregators. Other techniques to mitigate reidentification risks include using statistical methodologies for deidentification, requiring contractual commitments not to reidentify, and limiting data access only for legitimate research purposes.[67] Given that the need for transparency, reproducibility, and rigor must be weighed against the imperative to protect patient and health system interests, there likely will not be a single solution that is uniformly applicable to all studies.

Sharing programming code used for creating all analytic results, as well as code for cohort creation in the context of healthcare database

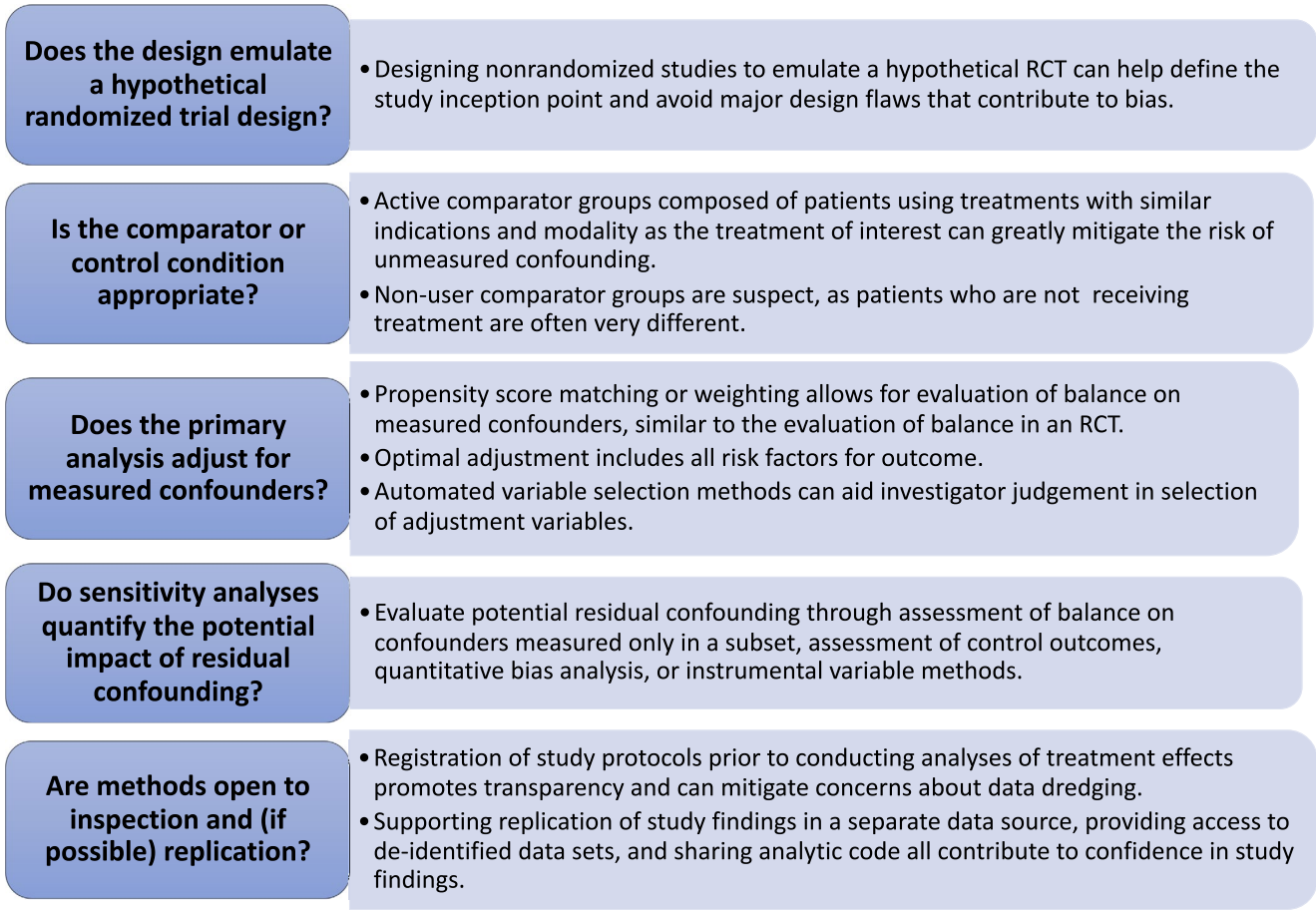

Figure 1 Valid inference on medication effects from nonrandomized studies. RCT, randomized controlled trial.

studies, should be required in regulatory submissions and highly encouraged in published literature. For example, the FDA's Sentinel program posts code for planned analyses prior to running the code to produce results, thereby serving as both preregistration of planned analyses and sharing of code.[68] Similarly, websites developed for registration of nonrandomized study protocols could allow for sharing of study code, before or after completion of analyses. However, implementing or understanding publicly available code may be difficult, even when it is developed for use with a common data model. Thus, along with the programming code, protocols registered prior to the study should be updated and shared to ensure that the final implementation of the study is described accurately in the protocol and changes to the protocol are clearly documented with rationale. Several resources have become available recently to promote thorough and transparent reporting of nonrandomized studies,[69–72] and standardization of reporting may improve the ability of regulators and other interested parties to comprehend and synthesize study findings and would allow other investigators to replicate findings.

CONCLUSION

Understanding of when and how nonrandomized studies can lead to valid quantification of the benefits and risks of medications has greatly improved over the last few decades.[73,74] In particular, the emulation of RCT designs in nonrandomized studies has helped to clarify design thinking in both primary data collection and analysis of existing healthcare databases. The recent proliferation of structured healthcare databases, such as health insurance claims, electronic health records, and registries has further stimulated interest in RWE on drugs based on nonrandomized designs. However, designing a high-quality nonrandomized study of medications remains challenging. Addressing confounding requires that a study can build on existing causal knowledge, for example, the relationships of potential confounders to exposure and outcome. Thus, evaluation of confounding and other potential biases requires broad expertise across a range of disciplines, including relevant clinical areas, epidemiology, and biostatistics. There is unlikely to ever be a simple checklist that can differentiate a high-quality nonrandomized study from a low-quality study as these judgements will always require subject matter expertise, but the guiding questions and recommendations provided in this paper detail strategies that could be applied across many clinical questions (**Figure 1**). Nonrandomized studies that cannot respond adequately to these questions lack credibility needed for decision making. Standardization of study processes, including protocol registration, sharing of data and analytic code, and reporting of results will further improve the reliability of nonrandomized research, just as it has for RCTs.

FUNDING
J.M.F.: NHLBI Grant R01HL141505. G.E.S.: NIMH Cooperative Agreement U19MH092201.

CONFLICT OF INTEREST
J.M.F. is an employee of Optum Epidemiology. G.E.S. is an employee of Kaiser Permanente Washington. N.A.D. is an employee of IQVIA Real World Solutions. M.H. is an employee of Kaiser Permanente Mid-Atlantic. R.M.C. is an employee of Verily and Google Health, and a board member for Cytokinetics. All other authors declared no competing interests for this work.

1. Concato, J., Shah, N. & Horwitz, R.I. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* **342**, 1887–1892 (2000).
2. 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).
3. Franklin, J.M. & Schneeweiss, S. When and how can real world data analyses substitute for randomized controlled trials? *Clin. Pharmacol. Ther.* **102**, 924–933 (2017).
4. 21st Century Cures Act, HR 34, 114th Congress (2015–2016). <https://www.congress.gov/bill/114th-congress/house-bill/34> (2016). Accessed June 15, 2017.
5. US Food and Drug Administration. Prescription Drug User Fee Act (PDUFA) - PDUFA VI: Fiscal Years 2018 - 2022. <https://www.fda.gov/industry/prescription-drug-user-fee-amendments/pdufa-vi-fiscal-years-2018-2022>. Accessed June 15, 2017.
6. Suresh, K. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *J. Hum. Reprod. Sci.* **4**, 8–11 (2011).
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. Food and Drug Administration. The sentinel initiative: a national strategy for monitoring medical product safety <https://www.fda.gov/downloads/Safety/FDAsSentinelInitiative/UCM124701.pdf> (May 2008).
9. Behrman, R.E., Benner, J.S., Brown, J.S., McClellan, M., Woodcock, J. & Platt, R. Developing the Sentinel System—a national resource for evidence development. *N. Engl. J. Med.* **364**, 498–499 (2011).
10. Plueschke, K., McGettigan, P., Pacurariu, A., Kurz, X. & Cave, A. EU-funded initiatives for real world evidence: descriptive analysis of their characteristics and relevance for regulatory decision-making. *BMJ Open.* **8**, e021864 (2018).
11. Walker, A.M. Confounding by indication. *Epidemiology* **7**, 335–336 (1996).
12. National Academies of Sciences Engineering and Medicine. Examining the impact of real-world evidence on medical product development: proceedings of a workshop series <https://www.nationalacademies.org/our-work/examining-the-impact-of-real-world-evidence-on-medical-product-development-a-workshop-series> (2019).
13. Gopalakrishnan, C. *et al*. Evaluation of socioeconomic status indicators for confounding adjustment in observational studies of medication use. *Clin. Pharmacol. Ther.* **105**, 1513–1521 (2019).
14. Glynn, R., Schneeweiss, S., Wang, P.S., Levin, R. & Avorn, J. Selective prescribing led to overestimation of the benefits of lipid-lowering drugs. *J. Clin. Epidemiol.* **59**, 819–828 (2006).
15. 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).
16. Hernán, M.A. *et al*. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* **19**, 766–779 (2008).
17. Hernán, M.A., Sauer, B.C., Hernández-Díaz, S., Platt, R. & Shrier, I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J. Clin. Epidemiol.* **79**, 70–75 (2016).
18. Hernán, M.A. & Robins, J.M. Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183**, 758–764 (2016).
19. Ray, W.A. Evaluating medication effects outside of clinical trials: new-user designs. *Am. J. Epidemiol.* **158**, 915–920 (2003).
20. Brookhart, M.A. Counterpoint: the treatment decision design. *Am. J. Epidemiol.* **182**, 840–845 (2015).
21. Suissa, S., Moodie, E.E.M. & Dell'Aniello, S. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores: Prevalent New-user Designs. *Pharmacoepidemiol. Drug Saf.* **26**, 459–468 (2017).
22. Suissa, S. Immortal time bias in pharmacoepidemiology. *Am. J. Epidemiol.* **167**, 492–499 (2008).
23. Califf, R.M., Hernandez, A.F. & Landray, M. Weighing the benefits and risks of proliferating observational treatment assessments: observational cacophony, randomized harmony. *JAMA* **324**, 625–626 (2020).
24. Matthews, K.A., Kuller, L.H., Wing, R.R., Meilahn, E.N. & Plantinga, P. Prior to use of estrogen replacement therapy, are users healthier than nonusers? *Am. J. Epidemiol.* **143**, 971–978 (1996).
25. Setoguchi, S. *et al*. Influence of healthy candidate bias in assessing clinical effectiveness for implantable cardioverter-defibrillators: cohort study of older patients with heart failure. *BMJ* **348**, g2866 (2014).
26. Huitfeldt, A., Hernan, M.A., Kalager, M. & Robins, J.M. Comparative effectiveness research using observational data: active comparators to emulate target trials with inactive comparators. *EGEMs* **4**, 1234 (2016).
27. McCullagh, P. *Generalized Linear Models* (Routledge, London, 2019).
28. 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).
29. Rosenbaum, P.R. & Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
30. Rosenbaum, P.R. & Rubin, D.B. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* **79**, 516–524 (1984).
31. Dehejia, R.H. & Wahba, S. Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Stat.* **84**, 151–161 (2002).
32. Hirano, K. & Imbens, G.W. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv. Outcomes Res. Methodol.* **2**, 259–278 (2001).
33. Franklin, J.M., Eddings, W., Austin, P.C., Stuart, E.A. & Schneeweiss, S. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Stat. Med.* **36**, 1946–1963 (2017).
34. Franklin, J.M., Rassen, J.A., Ackermann, D., Bartels, D.B. & Schneeweiss, S. Metrics for covariate balance in cohort studies of causal effects. *Stat. Med.* **33**, 1685–1699 (2014).
35. Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J. & Stürmer, T. Variable selection for propensity score models. *Am. J. Epidemiol.* **163**, 1149–1156 (2006).
36. Austin, P.C., Grootendorst, P. & Anderson, G.M. A comparison of the ability of different propensity score models to balance

measured variables between treated and untreated subjects: a Monte Carlo study. *Stat. Med.* **26**, 734–753 (2007).
37. Myers, J.A. *et al*. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am. J. Epidemiol.* **174**, 1213–1222 (2011).
38. Pearl, J. *On a Class of Bias-Amplifying Variables that Endanger Effect Estimates*. Uncertain. Artif. Intell. (2010).
39. Hartzema, A.G. & Schneeweiss, S. Addressing misclassification in pharmacoepidemiologic studies. In: *Pharmacoepidemiology and Therapeutic Risk Management* (Hartzema, A.G., Tilson, H.H. & Chan, K.A., eds.). (Harvey Whitney, Cincinnati, OH, 2008).
40. Schneeweiss, S., Rassen, J.A., Glynn, R.J., Avorn, J., Mogun, H. & Brookhart, M.A. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20**, 512–522 (2009).
41. van der Laan, M.J. & Gruber, S. Collaborative double robust targeted maximum likelihood estimation. *Int. J. Biostat.* **6**, Article 17 (2010).
42. Shortreed, S.M. & Ertefaie, A. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics* **73**, 1111–1122 (2017).
43. Koch, B., Vock, D.M. & Wolfson, J. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics* **74**, 8–17 (2018).
44. Ju, C., Wyss, R., Franklin, J.M., Schneeweiss, S., Häggström, J. & van der Laan, M.J. Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data. *Stat. Methods Med. Res.* **28**, 1044–1063 (2019).
45. Wyss, R., Schneeweiss, S., van der Laan, M., Lendle, S.D., Ju, C. & Franklin, J.M. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology* **29**, 96–106 (2018).
46. Ju, C. *et al*. Scalable collaborative targeted learning for high-dimensional data. *Stat. Methods Med. Res.* **28**, 532–554 (2019).
47. Dorie, V., Hill, J., Shalit, U., Scott, M. & Cervone, D. Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Stat. Sci.* **34**, 43–68 (2019).
48. Toh, S., García Rodríguez, L.A. & Hernán, M.A. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiol. Drug Saf.* **20**, 849–857 (2011).
49. 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).
50. Lipsitch, M., Tchetgen, E.T. & Cohen, T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* **21**, 383–388 (2010).
51. Tchetgen, T.E. The control outcome calibration approach for causal inference with unobserved confounding. *Am. J. Epidemiol.* **179**, 633–640 (2014).
52. Schneeweiss, S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol. Drug Saf.* **15**, 291–303 (2006).
53. Lash, T.L., Fox, M.P. & Fink, A.K. *Applying Quantitative Bias Analysis to Epidemiologic Data* (Springer, New York, NY, 2009).
54. Angrist, J.D., Imbens, G.W. & Rubin, D.B. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* **91**, 444–455 (1996).
55. Hernán, M.A. & Robins, J.M. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* **17**, 360–372 (2006).
56. Brookhart, M.A., Rassen, J.A. & Schneeweiss, S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol. Drug Saf.* **19**, 537–554 (2010).
57. Brookhart, M.A., Wang, P.S., Solomon, D.H. & Schneeweiss, S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* **17**, 268–275 (2006).
58. Brookhart, M.A. & Schneeweiss, S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *Int. J. Biostat.* **3**, Article 14 (2007).
59. Huybrechts, K.F., Gerhard, T., Franklin, J.M., Levin, R., Crystal, S. & Schneeweiss, S. Instrumental variable applications using nursing home prescribing preferences in comparative effectiveness research. *Pharmacoepidemiol. Drug Saf.* **23**, 830–838 (2014).
60. Desai, R.J. *et al*. Association of osteoporosis medication use after hip fracture with prevention of subsequent nonvertebral fractures: an instrumental variable analysis. *JAMA Netw. Open* **1**, e180826 (2018).
61. Garabedian, L.F., Chu, P., Toh, S., Zaslavsky, A.M. & Soumerai, S.B. Potential bias of instrumental variable analyses for observational comparative effectiveness research. *Ann. Intern. Med.* **161**, 131–138 (2014).
62. Rubin, D.B. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* **26**, 20–36 (2007).
63. Franklin, J.M., Glynn, R.J., Martin, D. & Schneeweiss, S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin. Pharmacol. Therap.* **105**, 867–877 (2019).
64. 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).
65. Simon, G.E. *et al*. Data sharing and embedded research. *Ann. Intern. Med.* **167**, 668–670 (2017).
66. CMS Research Data Assistance Center. CMS Virtual Research Data Center (VRDC) FAQs <https://www.resdac.org/cms-virtual-research-data-center-vrdc-faqs>. Accessed November 19, 2019.
67. Lo, B. Sharing clinical trial data: maximizing benefits, minimizing risk. *JAMA* **313**, 793–794 (2015).
68. FDA Sentinel Initiative. Sentinel analytic packages <https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse>. Accessed November 20, 2019
69. von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C. & Vandenbroucke, J.P. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* **370**, 1453–1457 (2007).
70. Benchimol, E.I. *et al*. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* **12**, e1001885 (2015).
71. Wang, S.V. *et al*. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. *Pharmacoepidemiol. Drug Saf.* **26**, 1018–1032 (2017).
72. Gatto, N.M., Reynolds, R.F. & Campbell, U.B. A structured preapproval and postapproval comparative study design framework to generate valid and transparent real-world evidence for regulatory decisions. *Clin. Pharmacol. Ther.* **106**, 103–115 (2019)
73. Velentgas, P., Dreyer, N.A., Nourjah, P., Smith, S.R. & Torchia, M.M. *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide* (Agency for Healthcare Research and Quality (US), Rockville, MD, 2013).
74. Berger, M.L. *et al*. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol. Drug Saf.* **26**, 1033–1039 (2017).