



Replication of randomized clinical trial results using real-world data: paving the way for effectiveness decisions

Journal of **Comparative Effectiveness Research**

Kristin M Sheffield^{*,1} , Nancy A Dreyer² , James F Murray¹, Douglas E Faries³  & Megan N Klopchin¹

¹Global Patient Outcomes & Real World Evidence, Eli Lilly & Company, Indianapolis, IN 46225, USA

²IQVIA Real-World & Analytic Solutions, Cambridge, MA 02139, USA

³Real World Analytics, Eli Lilly & Company, Indianapolis, IN 46225, USA

*Author for correspondence: Tel.: +1 317 503 8821; Sheffield.kristin_m@lilly.com

The FDA is preparing guidance about using real-world evidence (RWE) to support decisions about product effectiveness. Several ongoing efforts aim to replicate randomized clinical trial (RCT) results using RWE with the intent of identifying circumstances and methods that provide valid evidence of drug effects. Lack of agreement may not be due to faulty methods but rather to the challenges with emulating RCTs, differences in healthcare settings and patient populations, differences in effect measures and data analysis, bias, and/or the efficacy–effectiveness gap. In fact, for some decisions, RWE may lead to better understanding of how treatments work in usual care settings than a more constrained view from RCTs. Efforts to reconcile the role and opportunities for generating complementary evidence from RWE and RCTs will advance regulatory science.

First draft submitted: 30 July 2020; Accepted for publication: 26 August 2020; Published online: 11 September 2020

Keywords: causal inference • real-world evidence • regulatory decision-making • replication

Background

Regulators and other decision makers have primarily relied on evidence from randomized clinical trials (RCTs) to support drug effectiveness determinations. RCTs provide a straightforward and well-known approach to minimize bias between treatment groups, as well as tightly controlled measurements to maximize data quality. These features provide *a priori* confidence that the results from well-designed and executed RCTs will be causally interpretable [1]. In contrast, evidence generated from observational studies of real-world data (RWD) is often considered inferior because nonrandom treatment assignment and less rigorous data collection may compromise internal validity, making causal interpretation of results challenging [1]. Moreover, the results of observational studies evaluating a treatment effect do not always agree with RCTs and these inconsistencies engender uncertainty and skepticism about observational research. While we may always need RCT results, a possibility exists that for some decisions, real-world evidence (RWE) can provide substantial evidence of treatment effect and may lead to a better understanding of how a treatment works in usual care settings versus a more constrained view from RCTs.

The use of RWE to inform drug effectiveness decisions has been limited to the support of approvals in rare diseases and oncology and the comparative effectiveness of preventative vaccines [2]. The FDA RWE Program plans to evaluate observational study designs and issue guidance regarding whether and how they may generate evidence to support decisions related to product effectiveness [2]. Evidence used to support product effectiveness determinations needs to establish a causal relationship between product use and healthcare outcomes [2]. While there is a considerable statistical literature on causal frameworks [3–6], standard and emerging methods for causal inference [7–9] and methods for assessing assumptions for causal inference [10], significant challenges remain. The Framework for US FDA's RWE Program suggests that replications of RCTs using RWD may provide insights into the opportunities and limitations of observational studies. Assessing the comparability of results between RCTs and rigorously designed observational studies may shed light on the clinical scenarios, study designs, end points and

statistical methods that lead to comparable results, increasing confidence that these types of studies can provide reliable evidence of drug effect. In this paper we explore the issues of RCTs versus RWE replication, including clinical and statistical reasons for potential differences in results and recommendations for future efforts.

Current efforts to reconcile the role of RWE as a complement to RCT

There are several ongoing efforts that aim to replicate the results of RCTs using rigorously designed and analyzed observational studies. The results of these efforts are likely to receive considerable attention from both advocates and critics of RWE and to raise additional questions. The most prominent effort is RCT DUPLICATE, a collaboration between FDA, Brigham and Women's Hospital and Harvard Medical School Division of Pharmacoepidemiology, to replicate 30 completed Phase III or IV trials and to predict the results of seven ongoing Phase IV trials using Medicare and commercial claims data [11]. The Multi-Regional Clinical Trials Center and OptumLabs are leading another effort called Observational Patient Evidence for Regulatory Approval and Understanding Disease (OPERAND) [12]. They have funded Brown University and Harvard Pilgrim Health Care Institute to replicate the ROCKET-AF trial for atrial fibrillation and the LEAD-2 trial for Type 2 diabetes control using claims from commercial and Medicare Advantage plans and electronic medical record data from OptumLabs Data Warehouse. Finally, FDA has funded the Yale University-Mayo Clinic Center of Excellence in Regulatory Science and Innovation to predict the results of three to four ongoing safety trials using OptumLabs claims data [13].

These replication efforts will evaluate comparability of RCT and observational study results based on prespecified measures of agreement – regulatory agreement and estimate agreement [14]. Regulatory agreement indicates that the observational study replicates the direction and statistical significance of the RCT and the conclusions drawn (including decisions made) would be same [14]. Estimate agreement means that the effect estimate from the observational study lies within the 95% CI for the treatment effect from the RCT [14]. The selection of the metrics used to assess agreement matters when judging the success of these replication exercises and determining implications regarding confidence in future observational studies. Regulatory agreement seems most relevant when considering the application of observational study results to support future regulatory decisions. RCT DUPLICATE investigators expect the probability of regulatory agreement in the absence of bias to be in the 80–90% range for trials showing significant effects. For trials that failed to find significant effects, the expected probability of regulatory agreement is 95% for truly negative trials and <20% for false negative trials [14]. However, investigators have not prespecified an *a priori* expectation of regulatory agreement for the overall set of 30 studies. Such a quantification of the expected value of agreement statistic may be helpful when evaluating the results of these exercises and could be set by considering how well RCTs replicate each other according to these agreement measures.

The purpose of these replication exercises is to identify the clinical scenarios (e.g., indications and outcomes), study designs and analytic approaches that lend themselves to valid study implementation with RWD. The premise is that developing such an empirical evidence base will enable regulators to appraise with greater confidence when observational studies may be used to support regulatory decision making [14,15]. When considering these replication efforts, several questions arise for relevant stakeholders and the scientific community. For example, why do some observational studies fail to replicate the effect from the RCT while others succeed? [14] When will there be sufficient empirical evidence to predict with high certainty the validity of an observational study of treatment effectiveness?

Interpreting replication results & recommendations

Even with rigorously designed observational studies intended to mimic the target RCTs, some variation in results should be expected. There are valid reasons for why RCTs and observational studies may not agree and it is important to consider possible reasons in the context of these replication exercises. While agreement between RCTs and observational studies may be interpreted as strengthening support for the causal relationship, the absence of concurrence does not mean one approach or the other is wrong. Previous comparisons of observational studies and RCTs have primarily attributed discrepancies in results to bias and confounding in the treatment effect estimates from the observational studies [16]. However, other factors also may be responsible, such as challenges with emulating the target RCT, differences in healthcare settings, inclusion of more vulnerable or diverse patients, differences in effect measures and data analysis and the efficacy-effectiveness gap. It is worth noting that even highly cited RCTs have been later contradicted, [17] and the reasons for failure to replicate can shed insight on the design and validity of RCTs themselves.

Differences in patient populations, end point measurement & calendar time

Despite efforts to approximate and apply the eligibility criteria from the RCT, important differences may remain between patient populations in the target RCT and the observational study. First, it can be challenging to apply trial eligibility criteria to real-world data sources and criteria that involve clinical data or physiological measurements will not be measurable in administrative claims data, or even in electronic medical records [18]. Even after applying similar eligibility criteria, there may be differences between the RCT and real-world populations in the underlying distributions of baseline patient characteristics, such as age or prevalence of comorbid conditions. RCT DUPLICATE researchers made the decision not to reweight the real-world populations to match the reference RCT population in primary analyses; however, this approach may be explored in sensitivity analyses. It is possible that the patient populations from which the real-world study samples are drawn (e.g., Medicare and commercial claims databases) may have different baseline risks for the primary outcome compared with the RCT patients or even to each other, as observed in a recent database study predicting the results of the CAROLINA trial [19]. The distributions of treatment effect modifiers also might differ between the RCT and real-world population, or there may be heterogeneity of treatment effect based on population variables that may be unknown or unaccounted for in each dataset. Finally, geographic differences need to be considered. Most pivotal, registrational RCTs are global studies and patients from the USA represent a subset of the overall study population. Replication efforts based on US data only should conduct sensitivity analyses to compare results to the treatment effect estimates for US trial patient subgroups where possible.

Differences in end point measurement also may impact agreement. The outcome definitions and severity of events that are captured in RCTs may differ from those used in observational studies based on secondary data, meaning that the end points could represent somewhat different concepts. Claims-based studies require rules or algorithms to identify cohorts, exposures and outcomes events. This speaks to the desirability of using validated algorithms with known operating characteristics in RWE studies. The capture of outcome events may only occur for events that required medical care or were queried about during a medical encounter. The surveillance periods, duration of follow-up and timing of end point measurement may differ as well. Trials evaluate end points at protocol-specified time intervals, while in clinical practice patients are monitored according to usual care, which varies across practices, health systems and regions and it is not possible to determine the intent of observed tests and procedures (e.g., routine surveillance vs suspicion of disease).

Finally, there may be calendar time differences. In some cases, the observational study will require a later time period than the corresponding pivotal Phase III RCT because the drug was either not yet on the market or only used off-label at the time the RCT was conducted. Improvements in clinical practice and changes in the treatment landscape or patient population may occur over time, which can impact treatment effect estimates.

Differences in data analysis & effect measures

When comparing treatment effect estimates from the RCTs and observational studies, it is important to consider differences in data analysis and the causal effect measures under study. RCTs typically conduct intent-to-treat analysis based on how patients were randomly assigned into treatment groups, whereas observational studies group patients into treatment groups based on what was observed (as used) rather than what was intended (intent to treat) and conduct the observational analog of a 'per protocol' analysis [20]. Direct comparisons of effect estimates between RCTs and observational studies may require reanalysis of the RCT data to estimate the per-protocol effect and make adjustments for adherence and loss to follow-up [20]. Patterns of intercurrent events such as treatment adherence, dose changes, and treatment augmentation or switching are expected to vary between RCTs and RWD. In an RCT, the study protocol defines rules for treatment changes and adherence is closely monitored, while in clinical practice there is less oversight of patients and much greater variability in treatment changes and the influence of other factors, such as cost or insurance coverage and social determinants of health. RCT DUPLICATE investigators plan to censor patients when they discontinue the study treatment or switch treatments [14]. It is not clear what other approaches may be used to address ITT versus per-protocol analysis or account for treatment changes during the follow-up period in the observational studies, but this issue highlights the potential complementarity of RCTs and RWD.

Bias & confounding

It is well known that observational studies may be affected by bias and unmeasured confounding. Real-world data sources are often missing key sociodemographic and clinical variables that may confound the treatment-outcome

relationship. There may be channeling bias related to patient, physician and healthcare system factors influencing prescribing and use of a newly approved drug [21] or other biases known to harm the internal validity of observational studies, such as confounding by indication, confounding by frailty and the healthy user effect [22]. There also may be misclassification or measurement error in the exposure and baseline predictors or covariates that can impact the estimation of treatment effects and alter their interpretation [23].

None of these issues have a simple or guaranteed solution that can offer the reassurance provided by randomization. Comparative analyses based on observational data are challenging. Standard bias control methods rely on the assumption of no unmeasured confounding, which may be untenable in many observational studies of treatment effectiveness. However, one is not without some recourse. The literature is rich with methods and evaluations of methods for adjusting for confounders, such as the use of matching, stratification, weighting and regression-based methods often through use of the propensity score [4,7,8,24–27]. Emerging methods such as doubly robust approaches, double matching (propensity and prognostic scores) and even machine learning approaches such as model averaging may provide improved estimators of treatment effects [28–31]. Recent review papers are emphasizing the use of an expanding toolkit of sensitivity analysis methods to assess the potential impact of unmeasured confounding [32–35]. It is expected that competent researchers will avoid common study design flaws [1] and apply accepted methods to adjust for measured confounding, as well as more novel methods to address unmeasured confounding. It is critical to conduct extensive sensitivity analyses where researchers diligently seek and include all factors that can address any underlying bias or confounding.

The impact of known but unmeasured confounding and measurement error could be minimized even further by the selection of real-world data sources that are ‘fit for purpose’ for replicating the treatment effect from the RCT. The decision to use claims data for the RCT DUPLICATE project was logistical rather than scientific. It is likely that for some RCTs, administrative claims data alone will be insufficient to support replication since it may not be able to match the inclusion or exclusion criteria or have the necessary variables to adequately find and correct the bias or confounding. Linkage with laboratory data or electronic health records data may be necessary to provide richer information. If linking is not possible, then replication in different types of real-world data sources may be helpful.

Efficacy–effectiveness gap

The efficacy–effectiveness gap refers to the longstanding observation that products perform differently in clinical practice than in RCTs. There is considerable literature on this topic [36], but it has not been directly addressed in recent discussions on replications of RCTs using RWD. Imperfections of the healthcare system and delivery of care may contribute to differences in outcomes between RCTs and observational studies [36]. In routine practice, there are barriers to accessing healthcare resources, variability in testing, diagnosis, and treatment, and physician behavior and patient adherence are not optimal [36]. True differences in outcomes may be expected between highly protocolized care of RCTs and usual care in real-world settings. The efficacy–effectiveness gap may also be the result of complex interactions between the drug’s biologic effects and patient, provider and healthcare-related factors [36]. While the methodological considerations discussed earlier in this paper may partially explain the efficacy–effectiveness gap, it is also possible that observational study results represent part of a continuum of truth about a treatment effect.

Recommendations

Disagreement in results of these RCT replication efforts could be evidence of challenges with emulating the target trials using RWD, differences in the data analysis and effect measures, bias or confounding in the observational study, or the efficacy–effectiveness gap. It is likely that all these factors will contribute in some way to any observed discrepancies in effect estimates. Arguably, the most important work will involve disentangling the reasons for differences to understand the results of the exercises and the implications for confidence in the internal validity of observational studies. A recent paper by the lead investigators of RCT DUPLICATE considers the potential challenges in emulating a target trial with RWD and provides a list of measures of emulation differences related to study populations, treatment strategies and outcomes (e.g., proportion of patients with labs available and the length of follow-up in RWE vs RCT) that are observable [37]. These measures, along with targeted sensitivity analyses, may help investigators understand the potential impact of emulation differences between the RCT and the RWE replication. Lodi and colleagues [20] describe a systematic approach to improve the comparison of effect estimates from RCTs and observational studies based on RWD, based on harmonization of study protocols and

data analysis to target the same causal effect and the same estimand and sensitivity analyses to investigate remaining discrepancies.

We encourage other researchers to build upon the foundation of current RCT replication efforts and conduct and publish similar replication exercises in additional therapeutic classes, with careful consideration of data and methods in the context of fitness for purpose, with attention to describing agreement and when results are discrepant, to identifying potential explanations for differences. We recommend that future efforts are expanded to include not only health insurance claims but also electronic health records data, registries, linked data sources and other clinically rich data sources. It would be most useful to focus on clinical and regulatory contexts where observational studies could demonstrate substantial evidence of treatment effect to support product effectiveness determinations and labeling changes, especially since this is an important goal identified in the 21st Century Cures Act. While FDA has accepted observational studies to support effectiveness determinations in limited instances in the past, there are a growing number of therapeutic areas where there appears to be an acceptable level of risk, particularly in oncology, in which observational studies may provide primary or supportive evidence sufficient for a label change. These may include long-term effectiveness within an already approved indication, additional claims or end points within an approved indication, changes in the combination therapy, or changes in the indicated patient population. Finally, it will be important to consider other approaches, in addition to RCT replication efforts, to address FDA concerns regarding establishing causality with observational studies, for example, the application of methods to address unmeasured confounding and work to advance methodological and statistical approaches to support causal conclusions. There is a growing literature on causal frameworks and causal estimation in observational studies; however, there is more work to be done to develop and evaluate novel statistical approaches and to advance regulatory understanding and confidence in observational studies of treatment effectiveness. Just as the requirements for an acceptable RCT have become recognized and codified, it will be necessary to develop consensus on the characteristics of high-quality observational research that could meet the standard of an ‘adequate and well-controlled investigation’ in order for these studies to rise to the level of substantial evidence [38,39].

Conclusion

RCTs and RWE are complementary and each contribute valuable information about patient outcomes. The gains from the use of observational studies to support regulatory decisions could be considerable. Efforts to replicate RCTs can support the credibility of observational studies for estimating treatment effectiveness by demonstrating that they can support the same regulatory decision or causal conclusion as RCTs when the clinical setting is carefully chosen, appropriate data are selected and best practices for design and analysis are followed. However, careful review and interpretation of the results will be critical, particularly if there are discrepancies between the target RCT and the observational study. The conclusion should not be that the observational study was flawed or confounded [40]. The learning achieved through investigating the reasons for differences may improve our understanding of when trustworthy causal inferences can be made from observational data but may also provide insights on how better to design RCTs to improve their generalizability and usefulness in decision-making. Efforts to reconcile the role and opportunities for generating complementary evidence from RWE and RCTs will not only advance regulatory science but also progress the learning healthcare system. The replication of RCTs is not an end goal but rather an intermediate step on the way to making statements or hypotheses about the efficacy–effectiveness gap, optimal use of medical products in real-world settings, heterogeneity of treatment effect in various subpopulations, and long-term outcomes. These are complementary insights RWE is uniquely positioned to provide.

Future perspective

FDA guidance regarding whether and how observational study designs may generate evidence to support decisions related to product effectiveness is expected by the end of 2021. Rigorously designed and conducted observational studies may offer valuable information that complements evidence from clinical trials. RCT replication efforts, advances in statistical approaches to causal estimation of treatments effects and experience with using RWE may increase regulator understanding and confidence in observational studies of treatment effectiveness. In 5–10 years from now, we expect that observational study designs will be used often to provide evidence of product effectiveness to support label changes in appropriate regulatory and clinical circumstances.

Executive summary

Background

- The use of observational studies to inform drug effectiveness decisions has been limited in the past, primarily due to challenges with causal inference.
- Replications of randomized clinical trials (RCTs) using real-world data may provide insights into the opportunities and limitations of observational studies for regulatory decision making.

Current efforts to reconcile the role of real-world evidence as a complement to RCT

- Several ongoing efforts aim to replicate the results of RCTs using rigorously designed and analyzed observational studies.
- Comparability will be evaluated based on prespecified measures of agreement: regulatory agreement and estimate agreement.
- The purpose of these efforts is to identify clinical scenarios, data sources, study designs and analytic approaches that provide reliable evidence of drug effect.

Interpreting replication results & recommendations

- Discrepancies in results between RCTs and observational studies should be expected and it is valuable to carefully consider possible reasons.
- Factors that may contribute to disagreement in results include: challenges with emulating the target RCT, differences in healthcare settings, differences in study populations and end point measurement, differences in effect measures and statistical analysis, the efficacy–effectiveness gap, and bias or confounding.
- The most challenging and informative task of these replication efforts will be to disentangle these potential reasons in order to understand the implications for confidence in observational studies.
- Future work should include clinically rich data sources and focus on clinical and regulatory contexts where observational studies could support labeling changes.

Conclusion

- RCTs and observational studies provide valuable, complementary information about patient outcomes. The gains from the use of observational studies to support regulatory decisions could be considerable.
- Careful review of the results of these replication efforts will be critical. The learning achieved may improve understanding of when trustworthy causal inferences can be made from observational studies.

Financial & competing interests disclosure

KM Sheffield, JF Murray, DE Faries and MN Klopchin are employees of Eli Lilly and Company and own stock in Eli Lilly and Company. NA Dreyer is an employee of IQVIA and accepts no personal consulting or speaking fees. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

References

Papers of special note have been highlighted as: • of interest

1. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clin. Pharmacol. Ther.* 102(6), 924–933 (2017).
2. U.S. Food and Drug Administration. Framework for FDA's real-world evidence program (2018). www.fda.gov/media/120060/download
- **Outlines the framework for US FDA's real-world evidence program, including commentary on observational studies and expected guidance documents.**
3. Holland PW. Statistics and causal inference. *J. Am. Stat. Assoc.* 81(396), 945–960 (1986).
4. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social and Biomedical Sciences*. Cambridge University Press, NY, USA (2015).
5. Halpern JY, Pearl J. Causes and explanations: a structural-model approach – Part I: causes. *Brit. J. Phil. Sci.* 56, 843–887 (2005).
6. Halpern JY, Pearl J. Causes and explanations: a structural-model approach – Part II: explanations. *Brit. J. Phil. Sci.* 56, 889–911 (2005).
7. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55 (1983).
8. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 25(1), 1–21 (2010).

9. Zagar A, Kadziola Z, Lipkovich I, Faries D, Madigan D. Evaluating bias control strategies in observational studies using frequentist model averaging. *J. Biopharm. Stat.* (under review).
10. Zhang X, Faries DE, Li H, Stamey JD, Imbens GW. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiol. Drug Saf.* 27(4), 373–382 (2018).
- **Provides a review of methods for addressing unmeasured confounding and includes a flowchart with best practice recommendations.**
11. RCT Duplicate. RCT duplicate: randomized controlled trials duplicated using prospective longitudinal insurance claims: applying techniques of epidemiology (2020). www.rctduplicate.org
12. OptumLabs. Using RWD in regulatory decision-making (2020). www.optumlabs.com/work/data-regulatory-decision.html
13. Yale University Center for Outcomes Research & Evaluation. Understanding the use of existing real-world data for medical product evaluation (2020). <https://medicine.yale.edu/core/current-projects/cersi/research/>
14. Franklin JM, Pawar A, Martin D *et al.* Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. *Clin. Pharmacol. Ther.* 107(4), 817–826 (2020).
- **Describes randomized clinical trial (RCT) DUPLICATE initiative, including the study design implementation process.**
15. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision-making. *Clin. Pharmacol. Ther.* 105(4), 867–877 (2019).
16. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ* 352, i493 (2016).
17. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 294(2), 218–228 (2005).
18. Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. Feasibility of using real-world data to replicate clinical trial evidence. *JAMA Netw. Open* 2(10), e1912869 (2019).
19. Paterno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM. Using real-world data to predict findings of an ongoing phase IV cardiovascular outcome trial – cardiovascular safety of linagliptin vs. glimepiride. *Diabetes Care* 42(12), 2204–2210 (2019).
- **Uses administrative claims data to predict results from an ongoing phase IV cardiovascular outcome trial in patients with Type 2 diabetes.**
20. Lodi S, Phillips A, Lundgren J *et al.* Effect estimates in randomized trials and observational studies: comparing apples with apples. *Am. J. Epidemiol.* 188(8), 1569–1577 (2019).
- **Describe a systematic approach to improve the comparison of effect estimates from RCTs and observational studies based on real-world data.**
21. Schneeweiss S, Gagne JJ, Glynn RJ, Ruhl M, Rassen JA. Assessing the comparative effectiveness of newly marketed medications: methodological challenges and implications for drug development. *Clin. Pharmacol. Ther.* 90(6), 777–790 (2011).
22. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med. Care* 48(Suppl. 6), S114–S120 (2010).
23. Rodriguez de Gil P, Bellara AP, Lanehart RE, Lee RS, Kim ES, Kromrey JD. How do propensity score methods measure up in the presence of measurement error? A Monte Carlo study. *Multivariate Behav. Res.* 50, 520–532 (2015).
24. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* 34, 3661–3679 (2015).
25. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat. Med.* 33, 1057–1069 (2014).
26. Bodory H, Camponovo L, Huber M, Lechner M. The finite sample performance of inference methods for propensity score matching and weighted estimators. *J. Business Econ. Stat.* 38(1), 183–200 (2018).
27. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J. Am. Stat. Assoc.* 113(521), 390–400 (2018).
28. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* 23, 2937–2960 (2004).
29. Faries D, Zhang X, Kadziola Z, *et al.* *Real World Health Care Data Analysis: Causal Methods and Implementation Using SAS*. SAS Institute Inc, NC, USA (2020).
30. Leacy FP, Stuart EA. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Stat. Med.* 33(20), 3488–3508 (2020).
31. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat. Sci.* 22(4), 523–539 (2007).
32. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol. Drug Saf.* 15(5), 291–303 (2006).
33. Uddin MJ, Groenwold RHH, Ali MS *et al.* Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *Int. J. Clin. Pharm.* 38(4), 1–10 (2016).

34. Vanderweele TJ, Ding P. Sensitivity analysis in observational studies: introducing the e-value. *Ann. Intern. Med.* 167(4), 268–274 (2017).
35. Zhang X, Faries DE, Li H, Stamey JD, Imbens GW. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiol. Drug Saf.* 27, 373–382 (2018).
36. Nordon C, Karcher H, Groenwold RHH *et al.* The “efficacy-effectiveness gap”: historical background and current conceptualization. *Value Health* 19(1), 75–81 (2016).
37. Franklin JM, Glynn RJ, Suissa S, Schneeweiss S. Emulation differences vs. biases when calibrating real-world evidence findings against randomized controlled trials. *Clin. Pharmacol. Ther.* 107(4), 735–737 (2020).
- **Provides example measures that may be used to assess emulation differences between RCTs and real-world evidence.**
38. Mahendraratnam N, Eckert J, Mercon K *et al.* Understanding the need for non-interventional studies using secondary data to generate real-world evidence for regulatory decision making and demonstrating their credibility (2020). <https://healthpolicy.duke.edu/sites/default/files/2020-08/Non-Interventional%20Study%20Credibility.pdf>
39. Kesselheim AS, Avorn J. The Food and Drug Administration has the legal basis to restrict promotion of flawed comparative effectiveness research. *Health Aff.* 31(10), 2200–2205 (2012).
40. Dreyer NA, Hall M, Christian JB. Modernizing regulatory evidence with trials and real-world studies. *Ther. Innov. Regul. Sci.* 54, 1112–1115 (2020).