PERSPECTIVE

Emulation Differences vs. Biases When Calibrating Real-World Evidence Findings Against Randomized Controlled Trials

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

Actionable real-world evidence (RWE) requires accurate estimation of causal treatment effects. Calibration of RWE against randomized controlled trials (RCTs) is sometimes done to demonstrate that RWE can support the same causal conclusion as RCTs. Disagreements can occur when studies in each pair asked different questions in different populations or due to the presence of residual bias. Distinguishing among reasons for differences will impact the level of confidence in RWE.

Several projects, such as the RCT DUPLICATE initiative funded by the US Food and Drug Administration (FDA), National Institutes of Health (NIH), and others, have launched in the last few years with the aim of assessing whether nonrandomized database studies can in some circumstances produce conclusions on the effectiveness of medications that are similar to those provided by RCTs.[1,2] While comparison of randomized and nonrandomized findings is not new, heightened current interest is in part spurred by new initiatives at several regulatory agencies focused on assessing the role RWE can play in regulatory decision making.[3,4] The FDA defines RWE as evidence on the benefits and risks of medications derived from routinely collected healthcare data, although other data sources such as patient registries may also be used.

A key challenge for any project attempting to calibrate RWE findings against RCT findings is that differences between treatment effect estimates from the two study types can be driven by bias due to lack of randomization in RWE and/or by other differences in the design, such as inclusion/exclusion criteria, outcome measurement, or motivations for patients to adhere to study medications. Even if RWE studies are designed to match the corresponding RCT as closely as possible, emulation of all study components is typically impossible.

EMULATING TARGET TRIALS

In clinical epidemiology, it has been recommended for several decades to contemplate how a randomized trial would be designed to answer a specific question before designing the nonrandomized counterpart to study the same question. Hernán and Robins call the former a hypothetical "target trial," which would then be emulated by a nonrandomized study.[5] This process has proven very useful in clarifying design choices for nonrandomized research on medications. It is also a highly flexible process, as specification of the target trial is often iterative. Realities of data collection, patient access, and other practical considerations impose constraints to the nonrandomized emulation; the hypothetical target trial can thus be adjusted so that the design that is feasible in a given healthcare database can accord with the design of the target trial.

In calibrating RWE studies against existing RCTs, the process of emulating a target RCT in nonrandomized data is similar, except that in this case the target trial is already underway or completed. Thus, adjusting the design of the target RCT to improve the feasibility of the design in existing data is not possible, making exact emulation more difficult. Instead, investigators must adapt the design elements that they can from the trial, given the constraints of the database, and note the trial specifications that cannot be completely emulated. This process will highlight unavoidable emulation differences between a completed or in-progress RCT and the RWE replication.

When assembling a series of RWE replications of RCTs, as in the RCT DUPLICATE initiative, there will be some trials where RWE specifications can closely emulate the RCT and others that cannot be emulated as closely. Examples of the latter

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA; [2]Departments of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada. *Correspondence: Jessica M. Franklin (jmfranklin@bwh.harvard.edu)

Received October 10, 2019; accepted January 10, 2020. doi:10.1002/cpt.1793

are run-in periods during which nonadherent or drug intolerant patients are excluded before randomization, or, to homogenize patients, run-in periods that place all patients on a common medication before randomization. Therefore, enumerating—and quantifying to the extent possible—such emulation differences can provide insight into whether and to what extent differences in treatment effect estimates between RWE and corresponding RCTs are due to bias related to a lack of randomization vs. other differences in design between the two study types. Specifically, attempting to correlate measures of emulation difference, such as those suggested in **Table 1**, with the magnitude of differences in treatment effect estimates may provide understanding of which emulation differences are most important in contributing to the "efficacy–effectiveness gap" between RWE and RCT findings, as long as bias due to confounding is not also correlated with the emulation difference of interest. Once better understood, such metrics could possibly be transformed into a simple three-point emulation scale.

In addition to emulation differences, potential bias can be partially explored by evaluating alternative methodological approaches to control confounding in the treatment-outcome relationship of interest. If adjustment for additional covariates or use of an adjustment approach that better balances covariates can result in RWE effect estimates that are closer to the effect estimate from the RCT, then this exploration can provide evidence that initial estimates may have been biased due to confounding. Sensitivity analyses exploring unmeasured confounding, such as evaluation of control outcomes or evaluation of confounders measured in a subset of the RWE study population can also substantiate some concerns about bias.[6] Similarly, sensitivity analyses that consider a somewhat more or less restrictive implementation of eligibility/exclusion criteria of the RWE study compared with the trial will illuminate the emulation success. Either way, it is important to differentiate between the differences in treatment effect estimates explained mostly by bias and those explained by emulation differences.

MEASURING EMULATION DIFFERENCES

Table 1 provides a list of potential differences between an RCT and a corresponding RWE emulation, as well as potential measures of the difference that are observable. Note that any of these potential emulation differences can occur and challenge the interpretation of findings when calibrating RWE against RCTs, even when the RWE study has been designed specifically to match the RCT, given the constraints of the real-world data source.

Matching trial exclusion criteria can be challenging, due to lack of complete patient history in some databases, biomarkers or specific imaging results that may be unavailable for some patients, and poor recording of patient symptoms. Many trials also include vague exclusion criteria that are "in the judgement of the investigator," such as "unlikely to survive at least 5 years." Such criteria make replication of the trial population difficult even in the context of a new RCT and are difficult to translate into a measurable criterion in RWE.

RWE may also have difficulty exactly mimicking the RCT treatment strategies, including the adherence to medications and the dosing and treatment augmentation schedule.[7] Depending on the adherence metrics reported in the trial, differential adherence can be described. Evaluating the treatment changes made during the course of follow-up, along with the ordering of labs relevant to treatment changes in the RWE, can indicate how well treatment strategies are aligned between the two study types.

Placebo control in the RCT will always be difficult to emulate in RWE. Investigators may wish to emulate placebo control by identifying patients receiving standard of care therapy who either do or do not add the treatment of interest. However, defining standard of care in the real-world data may be difficult, and the specific treatments used may differ considerably between the RCT and RWE. Better control of bias may be achieved through comparison of initiators with the treatment of interest to initiators of another agent that is neutral with respect to the study outcome.

Despite attempts to harmonize outcome definitions, differences may persist in how the outcome is defined and identified. The CAROLINA trial comparing linagliptin to glimepiride evaluated the risk of moderate or severe hypoglycemia.[8] An RWE prediction of this trial based on healthcare claims data could not directly mimic the definition from the trial, which captured many less severe events that did not require contact with the healthcare system.[9] Instead, hypoglycemia was defined as a primary inpatient or emergency department diagnosis

Table 1 Challenges in calibrating RWE against RCTs and measures of differences between the study types

Emulation differences	Measures of emulation difference
Differences in study populations due to: • Unrecorded physician decision making in RCT • Lack of complete patient histories in RWE • Inconsistent recording of laboratory values in RWE • Poor recording of patient symptoms in RWE	• Proportion of patients excluded due to each exclusion criterion in RCT vs. RWE • Average available look-back in RWE • Proportion of patients with labs available in RWE
Differences in treatment strategies due to: • Placebo control in RCT • Tight control of medication adherence in RCT • Tight control of dosing schedule and treatment augmentation in RCT	• Proportion of patients discontinuing before the end of follow-up in RCT vs. RWE • Proportion of patients reaching target dose during follow-up in RCT vs. RWE • Proportion of patients on other therapies during follow-up in RCT vs. RWE • Ordering of labs in RWE for treatment titration
Differences in outcome measurement due to: • Differing outcome definitions • Differential surveillance for outcomes • Differing lengths of follow-up	• Overall outcome risk in RCT vs. RWE • Distributions of combined outcome events attributable to each cause in RCT vs. RWE • Length of follow-up in RCT vs. RWE • Evaluation of proportional hazards • Side-by-side Kaplan-Meier plots

RCTs, randomized controlled trials; RWE, real-world evidence.

of hypoglycemia. This could partly explain the difference in the estimated hazard ratios (HR) between the RCT and the RWE study for moderate/severe hypoglycemia (0.15 [0.08–0.29] vs. 0.42 [0.32–0.56], respectively). The findings for the cardiovascular outcome were highly overlapping: HR = 0.98 (0.84–1.14) in the RCT vs. HR = 0.91 (0.79–1.05) in the RWE study. Even when outcomes are defined similarly between the RCT and RWE, surveillance for the outcome may be different.

Finally, when emulating an RCT with long-term follow-up, many patients in the RWE study may fail to reach the end of follow-up due to patient churn in and out of healthcare databases or due to medication nonadherence in on-treatment analyses. If the true underlying treatment effect is constant throughout follow-up, then estimates generated from a shorter follow-up would be unbiased. Side-by-side Kaplan-Meier plots from the two studies can help investigators understand how treatment effect estimates compare throughout the course of follow-up when supported by relevant statistical information.

CONCLUSION

Given the difficulty in disentangling the influence of emulation differences vs. confounding bias in RCT-RWE calibration exercises, great care is needed when interpreting the findings. Indeed, finding differences between RCTs and RWE results when they try to answer the same question can be discomforting: Is RWE accurately targeting an effect that is different than in the RCT or is it inaccurately targeting the same effect as the RCT?

Investigators undertaking replication or prediction of an RCT must carefully investigate sources of emulation differences in order to identify whether observed differences in treatment effect estimates are likely to be driven primarily by emulation differences. As published RCT replications continue to accumulate, future modeling exercises could attempt to understand how each emulation difference measure correlates with the magnitude of the difference in findings between RCTs and RWE. However, it should be expected that the impact of each emulation difference will vary from trial to trial, making context expertise important for interpretation of findings. Research should evaluate whether summary scores for the emulation quality can be developed and correlated with the magnitude of observed differences in findings, possibly in conjunction with scores regarding risk of bias in observational studies.[10]

Emulation of a single trial requires hundreds of subjective decisions regarding both how to emulate the RCT design and how to control confounding. While targeted sensitivity analyses can provide some information on how these decisions impact the closeness of the trial emulation and potential for bias in the study, it is not feasible to investigate all reasonable specifications, and other choices could lead to better or worse replication of RCT findings. Investigators conducting calibration exercises should transparently report their methods so that others can modify protocols and investigate additional specifications, careful to publish all findings regardless of whether or not the emulation is successful, thereby adding to the community's learnings on what works and what doesn't work regarding both emulation of RCT features and reduction of bias in nonrandomized RWE studies.

Nevertheless, the fact that RWE is not emulating RCT designs could be considered a specific strength of RWE, as it may be more reflective of actual clinical care. Indeed, RWE is often used to complement evidence from RCTs that may be seen as too restrictive in the profile of the studied population vs. the one that will use the medications in real life.

FUNDING

This study was partially funded by the FDA under contract HHSF223201710186C and partially by the Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital. Dr. Franklin was funded by the National Heart, Lung, and Blood Institute under award number R01HL141505. The opinions expressed in this article are entirely those of the authors.

CONFLICT OF INTEREST

S.Sc. is the principal investigator of investigator-initiated grants to the Brigham and Women's Hospital from Bayer, Vertex, and Boehringer Ingelheim unrelated to the topic of this study. He is a consultant to Aetion, a software manufacturer of which he owns equity. His interests were declared, reviewed, and approved by the Brigham and Women's Hospital and Partners HealthCare System in accordance with their institutional compliance policies. R.J.G. received grant support for the statistical design, monitoring, and analysis of RCTs from Amarin, AstraZeneca, Kowa, Novartis, and Pfizer. His interests were declared, reviewed, and approved by the Brigham and Women's Hospital and Partners HealthCare System in accordance with their institutional compliance policies. S. Su. has received research grants from Bayer, Boehringer-Ingelheim, Bristol-Myers-Squibb, and Novartis and has participated in advisory board meetings or as speaker for AstraZeneca, Boehringer-Ingelheim, and Novartis. J.M.F. declared no competing interests for this work.

1. Franklin, J.M. *et al*. Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. *Clin. Pharmacol. Ther.* **107**, 817–826 (2020).
2. Carrigan, G. *et al*. Using electronic health records to derive control arms for early phase single-arm lung cancer trials: proof-of-concept in randomized controlled trials. *Clin. Pharmacol. Ther.* **107**, 369–377 (2020).
3. Eichler, H.G. *et al*. Are novel, nonrandomised analytic methods fit for decision making? The need for prospective, controlled and transparent validation. *Clin. Pharmacol. Ther.* **107**, 773–779 (2020).
4. US Food and Drug Administration. *Framework for FDA's Real-World Evidence Program* (US Food and Drug Administration, Washington, DC, 2018).
5. 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).
6. 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).
7. van Onzenoort, H.A.W. *et al*. Participation in a clinical trial enhances adherence and persistence to treatment: a retrospective cohort study. *Hypertension* **58**, 573–578 (2011).
8. Rosenstock, J. *et al*. Effect of linagliptin vs glimepiride on major adverse cardiovascular outcomes in patients with type 2 diabetes: The CAROLINA randomized clinical trial. *JAMA* **322**, 1155 (2019).
9. Patorno, E., Schneeweiss, S., Gopalakrishnan, C., Martin, D. & Franklin, J.M. Using real-world data to predict findings of an ongoing phase IV cardiovascular outcome trial: Cardiovascular safety of linagliptin versus glimepiride. *Diabetes Care* **42**, 2204–2210 (2019).
10. Sterne, J.A.C. *et al*. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* **355**, i4919 (2016).