

When Can We Rely on Real-World Evidence to Evaluate New Medical Treatments?

Gregory E. Simon^{1,*}, Richard Platt², Jonathan H. Watanabe³, Andrew B. Bindman⁴, Alex John London⁵, Michael Horberg⁶, Adrian Hernandez⁷ and Robert M. Califf⁸

Concerns regarding both the limited generalizability and the slow pace of traditional randomized trials have led to calls for greater use of real-world evidence (RWE) in the evaluation of new treatments or products. The RWE label has been used to refer to a variety of departures from the methods of traditional randomized controlled trials. Recognizing this complexity and potential confusion, the National Academies of Science, Engineering, and Medicine convened a series of workshops to clarify and address questions regarding the use of RWE to evaluate new medical treatments. Those workshops identified three specific dimensions in which RWE studies might differ from traditional clinical trials: use of real-world data (data extracted from health system records or data captured by mobile devices), delivery of real-world treatment (open-label treatments delivered in community settings by community practitioners), and real-world treatment assignment (including nonrandomized comparisons and variations on random assignment such as before-after or stepped-wedge designs). For any RWE study, decisions regarding each of these dimensions depends on the specific research question, characteristics of the potential study settings, and characteristics of the settings where study results would be applied.

Traditional randomized clinical trials (TRCTs) often fail to provide the timely and relevant evidence necessary for regulatory, clinical, and coverage decisions regarding use of novel medical treatments or new uses for existing treatments.^{1,2} Participants in TRCTs often differ markedly from those treated in community practice in terms of sociodemographic characteristics, prognostic characteristics, co-occurring conditions, and motivation or likelihood of treatment adherence. The tightly controlled (and typically blinded) treatments in TRCTs may yield outcomes quite different from outcomes of more typical treatment by real-world providers. Relatively small sample sizes sometimes limit TRCTs to assessment of intermediate or surrogate outcomes (e.g., tumor shrinkage or reduction in suicidal ideation), rather than the outcomes of greatest interest to patients, clinicians, and purchasers (e.g., cancer-free survival or prevention of suicidal behavior). Concerns regarding the relevance and generalizability of findings from TRCTs have prompted demands for evidence derived from real-world settings.^{2,3}

Efficiency of evidence generation is also a growing concern. Tightly controlled treatment delivery and research-specific data collection contribute significantly to the increasing costs of TRCTs, with the median cost of a phase III clinical trial exceeding \$100 million⁴ and the mean development cost of bring a new medication to market exceeding \$1.3 billion.⁵ Those rising costs threaten to slow or restrict development of innovative treatments, novel usages of established treatments, and efficient comparative effectiveness trials.^{3,6} Narrow eligibility criteria and demanding

assessment protocols slow recruitment, delaying or completely preventing trial completion because of inadequate recruitment.^{7,8}

The COVID-19 pandemic has dramatically revealed the need for more efficient and timely evidence generation. The urgent need for evidence regarding new and re-purposed therapeutics for COVID-19 has prompted both refreshing innovation to improve speed and efficiency^{9,10} and frustration regarding the fragmentation and slow pace of traditional clinical trials initiated during the pandemic.^{11,12}

Mindful of the need for more relevant evidence and more efficient evidence generation, the US Food and Drug Administration (FDA) sponsored a three-part workshop series, organized and hosted by the National Academies Forum on Drug Discovery, Development, and Translation on “Examining the Impact of Real-World Evidence on Medical Product Development.” As described in recently published proceedings,¹³ those workshops focused on aligning incentives to promote appropriate innovation and developing practical approaches to improve the efficiency and relevance of evidence generation. As participants in that workshop series, we now propose specific, practical guidance based on our discussions during and following those workshops.

Dimensions of RWE

As others have pointed out,^{2,14} real-world evidence (RWE) may be best defined by what it is not. By that definition, RWE could include a range of evidence not generated by TRCTs.¹⁴ Recent discussion by the FDA defines RWE as evidence derived from real-world data (RWD) rather than data collected by and for research. Studies to

¹Kaiser Permanente Washington Health Research Institute; ²Harvard Pilgrim Health Care Institute, Harvard Medical School; ³School of Pharmacy and Pharmaceutical Sciences, University of California Irvine; ⁴Kaiser Foundation Health Plan and Hospitals; ⁵Philosophy Department & Center for Ethics and Policy, Carnegie Mellon University; ⁶Kaiser Permanente Mid-Atlantic Permanente Research Institute, Mid-Atlantic Permanente Medical Group; ⁷Duke Clinical Research Institute; ⁸Verily Life Sciences and Google Health. *Correspondence: Gregory E. Simon (gregory.e.simon@kp.org)

Received February 8, 2021; accepted March 24, 2021. doi:10.1002/cpt.2253

Table 1 Variation in design of three real-world evidence studies

	Data sources	Treatment delivery	Treatment assignment
Salford Lung Study	Electronic health records and community pharmacy records	Treatment by community providers following study protocol	Patient-level random assignment
Ofatumumab Trial	Data collected and recorded by research clinicians in research database	Treatment by research clinicians following study protocol	Case series compared with historical control group
IMPACT-Afib Trial	Insurance claims generated by community providers	Treatment by community providers with no guidance from a study protocol	Patient-level random assignment

generate RWE could differ from TRCTs in any of several dimensions or characteristics.^{15,16} The National Academies of Science, Engineering, and Medicine (NASEM) workshop series focused on three of those dimensions. First, generation of RWE usually relies on RWD, such as records created by routine healthcare operations or data recorded outside of healthcare settings by mobile devices and connected consumer devices (fitness monitors, glucometers, and home blood pressure monitors).^{17,18} Second, studies to generate RWE may involve treatments delivered in real-world healthcare settings—or outside of healthcare settings altogether. In those real-world settings, practice may depart from research protocols, patients may vary widely in adherence, and blinding or masking of treatment may not be feasible.^{19,20} Third, while RWE can be derived using individual patient randomization, generation of RWE may involve research designs or mechanisms of treatment assignment other than patient-level randomization. Those alternatives could include cluster-level randomization, stepped-wedge designs, nonrandomized comparisons with contemporaneous controls,^{21,22} or single-arm comparisons with historical controls. Each step on this gradient from patient-level randomization to completely naturalistic nonrandomized comparison may increase the risk that inference will be biased by unmeasured confounding.²³ Any RWE study may depart from traditional clinical trial methods on any or all of these three dimensions: real-world data, real-world treatment, or real-world treatment assignment.

Calls for increasing the use of RWE have focused on two proposed advantages. First, evidence generated from real-world practice settings using data regarding real-world patients could be more relevant to real-world decisions.^{1,2} Second, reliance on existing care-delivery and informatics infrastructure could significantly reduce the time and expense required to generate useful evidence.²⁴ For any specific aspect or dimension of RWE, these two motivations may or may not be aligned. For example, implementation of clinical trials in community practice settings may lead to more relevant evidence, but the resulting variability in treatment delivery could increase the sample size (and cost) required to accurately detect treatment effects. Similarly, a clinical trial focused on relevant clinical outcomes rather than biomarkers or intermediate outcomes may require a larger sample and longer duration of follow-up. The relative importance of relevance, generalizability, cost, and timeliness may lead to different decisions regarding methods for RWE studies.

Matching Questions with RWE Design Decisions

Three examples illustrate the range of RWE studies evaluating effectiveness and/or safety of medical products.

- The Salford Lung Study²⁵ was a community-based randomized trial demonstrating the effectiveness of combination fluticasone/vilanterol for management of obstructive lung disease. Investigators hoped to improve generalizability through enrollment of patients treated in community practice and increase efficiency by using existing health informatics infrastructure. The trial involved traditional patient-level random assignment of treatment, and treatments were delivered according to a specific study protocol. But study data regarding eligibility, baseline prognostic characteristics, treatment adherence, adverse events, and effectiveness outcomes were all extracted from real-world health system and pharmacy records.
- Approval of ofatumumab for treatment of chronic lymphocytic leukemia²⁶ depended on comparison of outcomes in a single-arm trial to those in a historical control group. Choice of that design was motivated by desire to accelerate access to a potential breakthrough treatment for a life-threatening condition. The single-arm trial, however, still included several features of TRCTs: collection of research-specific data regarding patients treated by research clinicians in research settings.
- The IMPACT-Afib trial²⁷ used a patient-level randomized design to examine the effect of encouraging open-label use of oral anticoagulants on treatment initiation (primary outcome) and risk of cerebrovascular events (secondary outcomes) in people with atrial fibrillation with indications for anticoagulation. Design decisions were motivated both by desire to increase generalizability and (given the sample size of ~ 80,000 patients) contain study costs. All treatments were determined and delivered by clinicians in community outpatient practices, with no study-specific training, treatment protocol, or monitoring of adherence. Study data were extracted from linked insurance claims databases organized by the FDA Sentinel program.

Table 1 illustrates how each of these exemplar RWE studies does or does not differ from a traditional clinical trial in each of these three dimensions (data sources, treatment delivery, and treatment assignment). Some discussions of RWE have blurred or conflated these dimensions. For example, the term “observational studies” has been used to refer both to use of existing records data sources (such as EHRs or insurance claims) and to comparisons based on nonrandomized or naturalistic treatment assignment. While these two aspects of study design may often be associated in practice, they are conceptually distinct and have specific advantages and disadvantages compared with TRCTs.

The 21st Century Cures Act instructed the FDA to evaluate greater use of RWE in regulatory decisions regarding medical products. That legislation specifically called for increased use of RWE in decisions regarding new indications for already-approved products and in postmarketing surveillance. It also explicitly permitted use of RWE in other regulatory decisions, such as initial approval of new products. Different departures from traditional clinical trial methods may be more or less appropriate (or practicable) to inform different regulatory decisions. For example, decisions regarding approval of new products can—and have—relied on nonrandomized comparisons²⁸ and use of data from “real world” sources, such as data extracted from EHRs and or pharmacy records.²⁵ But research prior to regulatory approval or licensure of a new product is unlikely to involve loosely controlled treatment delivery by a wide range of community practitioners. Records data from community practice are generally not available prior to initial approval of a novel treatment. In contrast to the initial approval or licensure of new products, decisions regarding new indications for or adverse effects of approved products may be best informed by data from diverse patient populations treated in community practice settings by a wide range of practitioners.²⁹ Real-world research to inform regulatory decisions will often differ from TRCTs, but the specifics of those methodologic differences will depend on the regulatory question and the research setting. This overview paper and three more detailed accompanying papers intend to inform an ongoing discussion regarding the design and interpretation of RWE studies to inform a range of regulatory decisions.

None of the above-described departures from traditional clinical trial methods are new inventions. Initial descriptions of pragmatic or real-world clinical trials are more than 50 years old.³⁰ Early calls for large simple trials pointed out the advantages in efficiency and scale of conducting trials in community settings and relying on data collected and recorded for routine clinical care.^{31,32} Early descriptions of real-world or pragmatic randomized trials described potential generalizability advantages of open-label treatment and more real-world variation in treatment intensity or adherence.³³ Comparative effectiveness research often involves both real-world treatment delivery and nonrandomized comparisons based on treatments “as used.” The RWE label could be considered an umbrella term for a range of research designs and methods already in common use. But the formal call for increased use of RWE in regulatory decisions should have important practical implications for various stakeholders.¹⁴ Developers of new medical products will need to consider the potential advantages and disadvantages of RWE studies in all stages of product development and regulatory submission (initial approval, approval of new indications, and postmarketing surveillance). Regulators will need to consider the trustworthiness of evidence from RWE studies for regulatory decisions regarding a range of medical products and clinical scenarios (potential breakthrough or first-in-class treatments, subsequent products in an existing class, biosimilars, etc.). The ultimate evidence consumers (guideline developers, clinicians, payers, patients, and families) will need to consider is how new forms of evidence can guide clinical and coverage or payment decisions regarding new treatments or products.

Questions for Evidence Generators and Evidence Consumers

Given the distinct dimensions of RWE and the diversity of potential uses, global judgments regarding the validity or utility of RWE are not possible or useful. For any specific regulatory, clinical, or policy question, it is necessary to consider potential departures from TRCT methods along each of the three dimensions described above. For each departure, evidence generators should examine both potential advantages, in either generalizability or in relevance or efficiency, and potential threats to valid inference regarding effectiveness or safety. To support that examination, the NASEM Workshop Series focused on three specific topics or questions regarding use of RWE to assess the effectiveness or safety of medical products:

- Real-world data: When are data from nontraditional sources accurate and reliable enough to support valid inference?^{17,18}
- Real-world treatment delivery: When can treatment delivered outside of traditional research settings (in community practice or outside of healthcare settings altogether) generate valid and relevant evidence while assuring participant safety?³⁴
- Real-world treatment assignment: When can naturalistic or nonrandomized assignment of treatments support unbiased comparisons?^{21,22}

Each of these questions is intentionally framed in conditional terms (i.e., When can nonrandomized comparisons support unbiased comparisons?) rather than absolute terms (i.e., Can nonrandomized designs support unbiased comparisons?). We presume, for example, that nonrandomized comparisons can support valid inference in some cases and not in others. For each of these three questions, the appropriateness of RWE methods to generate valid, reliable, and actionable evidence will depend on the specific regulatory or clinical decision at hand and specific characteristics of the treatments of interest, possible study settings, outcomes of interest, and potential sources of bias. Addressing these questions will often require balancing scientific considerations against practical constraints.

Challenges in delivering timely and robust evaluation of COVID-19 therapeutics illustrate the relevance of each of these questions. Uncertainty regarding the provenance and trustworthiness of RWD from health records was central to controversies regarding both potential adverse effects of cardiovascular drugs³⁵ and therapeutic effects of hydroxychloroquine.³⁶ While many COVID-19 trials have seen slow or failed enrollment,^{11,12} trials allowing more real-world treatment delivery and data collection have yielded rapid and actionable results.¹⁰ Some observational comparisons have yielded encouraging and highly publicized results,^{37,38} accompanied by concern regarding residual confounding by indication—and sometimes followed by discordant results from subsequent randomized comparisons.⁹

These three dimensions of RWD may be linked for practical reasons. For example, retrospective nonrandomized comparisons are often practically limited to use of existing healthcare records regarding treatments delivered under real-world conditions. Nevertheless, these three dimensions are conceptually distinct.

Design and implementation of any RWE study require separate decisions regarding appropriateness of data sources, control or blinding of treatment delivery, and validity of nonrandomized comparisons. Those separate decisions each require careful consideration regarding how specific departures from traditional clinical trial methods will promote efficiency and generalizability while maintaining scientific validity.

Three companion papers consider each of these three questions regarding appropriate methods for developing RWE. For each question, we attempt to identify specific criteria or requirements that should be considered in the design of any RWE study. These criteria are intended to serve as decision aids for investigators or treatment developers designing RWE studies and for the various stakeholders (regulators, health systems, payers, clinicians, and patients) who hope to evaluate the validity and relevance of RWE to a specific decision.

FUNDING

G.E.S.: NIMH Cooperative Agreement U19MH092201.

CONFLICT OF INTEREST

G.E.S. is an employee of Kaiser Permanente Washington. A.B.B. is an employee of Kaiser Foundation Health Plan and Hospitals. M.H. is an employee of Kaiser Permanente Mid-Atlantic. R.M.C. is an employee of Verily and Google Health, Board member for Cytokinetics. All other authors declared no competing interests for this work.

© 2021 The Authors. *Clinical Pharmacology & Therapeutics* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

- Corrigan-Curay, J., Sacks, L. & Woodcock, J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA* **320**, 867–868 (2018).
- 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).
- Eapen, Z.J., Lauer, M.S. & Temple, R.J. The imperative of overcoming barriers to the conduct of large, simple trials. *JAMA* **311**, 1397–1398 (2014).
- Martin, L., Hutchens, M., Hawkins, C. & Radnov, A. How much do clinical trials cost? *Nat. Rev. Drug Discov.* **16**, 381–382 (2017).
- Wouters, O.J., McKee, M. & Luyten, J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA* **323**, 844–853 (2020).
- Berndt, E.R. & Cockburn, I.M. *Price Indexes for Clinical Trial Research: a Feasibility Study* (National Bureau of Economic Research, Cambridge, MA, 2013).
- Carlisle, B., Kimmelman, J., Ramsay, T. & MacKinnon, N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. *Clin. Trials*. **12**, 77–83 (2015).
- Huang, G.D. et al. Clinical trials recruitment planning: A proposed framework from the Clinical Trials Transformation Initiative. *Contemp. Clin. Trials*. **66**, 74–79 (2018).
- 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).
- Recovery Collaborative Group et al. Dexamethasone in hospitalized patients with Covid-19. *N. Engl. J. Med.* **384**, 693–704 (2021).
- Coronavirus drugs trials must get bigger and more collaborative. *Nature* **581**, 120 (2020).
- Glasziou, P.P., Sanders, S. & Hoffmann, T. Waste in covid-19 research. *BMJ* **369**, m1847 (2020).
- National Academies of Sciences Engineering and Medicine. *Examining the Impact of Real-World Evidence on Medical Product Development: Proceedings of a Workshop Series* (2019). (National Academies Press, Washington, DC, 2019).
- US Food and Drug Administration. *Framework for FDA's Real-World Evidence Program* (US Food and Drug Administration, Silver Spring, MD, 2018).
- Fanaroff, A.C., Steffel, J., Alexander, J.Y.H., Lip, G.Y.H., Califf, R.M. & Lopes, R.D. Stroke prevention in atrial fibrillation: re-defining 'real-world data' within the broader data universe. *Eur. Heart J.* **39**, 2932–2941 (2018).
- Schneeweiss, S. Real-world evidence of treatment effects: the useful and the misleading. *Clin. Pharmacol. Ther.* **106**, 43–44 (2019).
- Miksad, R.A. & Abernethy, A.P. Harnessing the Power of Real-World Evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin. Pharmacol. Ther.* **103**, 202–205 (2018).
- Richesson, R.L. et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J. Am. Med. Inform. Assoc.* **20**, e226–231 (2013).
- Loudon, K., Treweek, S., Sullivan, F., Donnan, P., Thorpe, K.E. & Zwarenstein, M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ* **350**, h2147 (2015).
- Thorpe, K.E. et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *CMAJ* **180**, E47–E57 (2009).
- 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. Ther.* **105**, 867–877 (2019).
- 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).
- Califf, R.M. Comparison of observational data and the ONTARGET results for Telmisartan treatment of hypertension: bull's-eye or painting the target around the arrow? *JAMA Intern. Med.* **178**, 63–65 (2018).
- Lauer, M.S., Gordon, D., Wei, G. & Pearson, G. Efficient design of clinical trials and epidemiological research: is it possible? *Nat. Rev. Cardiol.* **14**, 493–501 (2017).
- Woodcock, A. et al. Effectiveness of fluticasone furoate plus vilanterol on asthma control in clinical practice: an open-label, parallel group, randomised controlled trial. *Lancet* **390**, 2247–2255 (2017).
- Lemery, S.J. et al. U.S. Food and Drug Administration approval: ofatumumab for the treatment of patients with chronic lymphocytic leukemia refractory to fludarabine and alemtuzumab. *Clin. Cancer Res.* **16**, 4331–4338 (2010).
- Cocoros, N.M. et al. FDA-Catalyst-Using FDA's Sentinel Initiative for large-scale pragmatic randomized trials: Approach and lessons learned during the planning phase of the first trial. *Clin. Trials*. **16**, 90–97 (2019).
- Jen, E.Y. et al. FDA approval: blinatumomab for patients with B-cell precursor acute lymphoblastic leukemia in morphologic remission with minimal residual disease. *Clin. Cancer Res.* **25**, 473–477 (2019).
- Fralick, M., Kesselheim, A.S., Avorn, J. & Schneeweiss, S. Use of health care databases to support supplemental indications of approved medications. *JAMA Intern. Med.* **178**, 55–63 (2018).
- Schwartz, D. & Lellouch, J. Explanatory and pragmatic attitudes in therapeutical trials. *J. Chronic Dis.* **20**(8), 637–648 (1967).
- Yusuf, S., Collins, R. & Peto, R. Why do we need some large, simple randomized trials? *Stat. Med.* **3**(4), 409–422 (1984).
- Peto, R., Collins, R. & Gray, R. Large-scale randomized evidence: large, simple trials and overviews of trials. *J. Clin. Epidemiol.* **48**, 23–40 (1995).

33. Simon, G., Wagner, E. & VonKorff, M. Cost-effectiveness comparisons using "real world" randomized trials: the case of new antidepressant drugs. *J. Clin. Epidemiol.* **48**, 363–373 (1995).
34. Ford, I. & Norrie, J. Pragmatic trials. *N. Engl. J. Med.* **375**, 454–463 (2016).
35. Mehra, M.R., Desai, S.S., Kuy, S., Henry, T.D. & Retracted, P.A.N. Cardiovascular disease, drug therapy, and mortality in Covid-19. *N. Engl. J. Med.* **382**, 2582 (2020).
36. Mehra, M.R., Desai, S.S., Ruschitzka, F. & Patel, A.N. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* (2020). [https://doi.org/10.1016/S0140-6736\(20\)31180-6](https://doi.org/10.1016/S0140-6736(20)31180-6).
37. Joyner, M.J. et al. Effect of convalescent plasma on mortality among hospitalized patients with COVID-19: initial three-month experience. *medRxiv.* (2020). <https://doi.org/https://doi.org/10.1101/2020.08.12.20169359>.
38. Arshad, S. et al. Treatment with hydroxychloroquine, azithromycin, and combination in patients hospitalized with COVID-19. *Int. J. Infect Dis.* **97**, 396–403 (2020).