Published in final edited form as:
Clin Trials. 2023 December ; 20(6): 689–698. doi:10.1177/17407745231193137.

Assessing the Use of Observational Methods and Real-World Data to Emulate Ongoing Randomized Controlled Trials

Joshua D Wallach[1], **Yihong Deng**[2], **Eric C Polley**[3], **Sanket S Dhruva**[4,5], **Jeph Herrin**[6], **Kenneth Quinto**[7], **Charu Gandotra**[8], **William Crown**[9], **Peter Noseworthy**[2,10], **Xiaoxi Yao**[2,10], **Molly Moore Jeffery**[11], **Timothy D Lyon**[12], **Joseph S Ross**[13,14,15], **Rozalina G McCoy**[2,16,17]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA

[2]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, USA

[3]Department of Public Health Sciences, University of Chicago, Chicago, IL, USA

[4]Section of Cardiology, Department of Medicine, San Francisco Veterans Affairs Health Care System, San Francisco, CA, USA

[5]University of California, San Francisco School of Medicine, San Francisco, CA, USA

[6]Section of Cardiovascular Medicine, Yale School of Medicine, New Haven, CT, USA

[7]Office of Medical Policy, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Springs, MD, USA

[8]Office of New Drugs, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Springs, MD, USA

[9]Florence Heller Graduate School, Brandeis University, Waltham, MA, USA

[10]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA

[11]Division of Health Care Delivery Research and Department of Emergency Medicine, Mayo Clinic, Rochester, MN, USA

[12]Department of Urology, Mayo Clinic, Jacksonville, FL, USA

Corresponding author: Rozalina G. McCoy, MD MS. Division of Endocrinology, Diabetes, and Nutrition, Department of Medicine, University of Maryland School of Medicine. 670 West Baltimore Street, Health Sciences Facility III, Room 4050, Baltimore, MD 21201. Phone: 410-706-7167. rozalina.mccoy@som.umaryland.edu.

Author Contributions

Dr. Wallach, Dr. Deng, and Dr. McCoy had full access to all data in the study and take responsibility for data integrity and accuracy of data analysis.

Concept and design: All authors.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Wallach, Lyon, Jeffery, Ross, McCoy
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Wallach, Deng, Herrin, Polley, McCoy
Obtained funding: Ross and Jeffery.
Administrative, technical, or material support: Ross.
Supervision: McCoy and Ross.

Disclaimer: The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the US government.

[13]Center for Outcomes Research and Evaluation, Yale-New Haven Health, New Haven, CT, USA

[14]Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

[15]Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA

[16]Division of Community Internal Medicine, Geriatrics, and Palliative Care, Department of Medicine, Mayo Clinic, Rochester, MN, USA

[17]OptumLabs, Eden Prairie, MN, USA

Abstract

Background/Aims: There has been growing interest in better understanding the potential of observational research methods in medical product evaluation and regulatory decision-making. Previously, we used linked claims and electronic health record data to emulate two ongoing randomized controlled trials (RCTs), characterizing the populations and results of each RCT prior to publication of its results. Here, our objective was to compare the populations and results from the emulated trials with those of the now published RCTs.

Methods: This study compared participants' demographic and clinical characteristics and study results between the emulated trials, which used structured data from OptumLabs Data Warehouse, and the published PRONOUNCE and GRADE trials. First, we examined the feasibility of implementing the baseline participant characteristics included in the published PRONOUNCE and GRADE trials' using real-world data, and classified each variable as ascertainable, partially ascertainable, or not ascertainable. Second, we compared the emulated trials and published RCTs for baseline patient characteristics (concordance determined using standardized mean differences <0.20) and results of the primary and secondary endpoints (concordance determined by direction of effect estimates and statistical significance).

Results: The PRONOUNCE trial enrolled 544 participants, and the emulated trial included 2226 propensity scorematched participants. In the PRONOUNCE trial publication, one of the 32 baseline participant characteristics was listed as an exclusion criterion on ClinicalTrials.gov but was ultimately not used. Among the remaining 31 characteristics, 9 (29.0%) were ascertainable, 11 (35.5%) were partially ascertainable, and 10 (32.2%) were not ascertainable using structured data from OptumLabs. For one additional variable, the PRONOUNCE trial did not provide sufficient detail to allow its ascertainment. Of the nine variables that were ascertainable, values in the emulated trial and published randomized controlled trial were discordant for 6 (66.7%). The primary endpoint of time from randomization to the first major adverse cardiovascular event and secondary endpoints of nonfatal myocardial infarction and stroke were concordant between the emulated trial and published randomized controlled trial. The GRADE trial enrolled 5047 participants, and the emulated trial included 7540 participants. In the GRADE trial publication, 8 of 34 (23.5%) baseline participant characteristics were ascertainable, 14 (41.2%) were partially ascertainable, and 11 (32.4%) were not ascertainable using structured data from OptumLabs. For one variable, the GRADE trial did not provide sufficient detail to allow for ascertainment. Of the eight variables that were ascertainable, values in the emulated trial and published randomized controlled trial were discordant for 4 (50.0%). The primary endpoint of time to hemoglobin ≥7.0% was mostly concordant between the emulated trial and the published randomized controlled trial.

Conclusion: Despite challenges, observational methods and real-world data can be leveraged in certain important situations for more timely evaluation of drug effectiveness and safety in more diverse and representative patient populations.

Keywords

Target trial emulation; real-world data; observational methods

Introduction

Rigorous double-blind, randomized controlled trials (RCTs) are considered one of the gold standards for comparing the safety and efficacy of medical interventions.[1] Although randomization and blinding minimize confounding and selection bias, RCTs often have important limitations, including strict inclusion and exclusion criteria; enrichment factors; recruitment and retention difficulties; modest sample sizes; under-representation of women, older adults, people from lower socioeconomic strata, and racial and ethnic minorities; high costs; long timeframes; and short follow-up durations, all of which can limit generalizability of their findings to real-world clinical practice.[2] Accordingly, there has been growing interest in better understanding the potential application of observational research methods to medical product evaluation and regulatory decision-making for medical interventions already in clinical use, including the necessary data standards, policies, and methods to ensure that evaluations using real-world data are rigorous and reliable.[3]

Several previous studies have replicated the results of completed RCTs using observational methods applied to real-world data,[4-9] and some researchers have suggested that observational data can be relied upon as evidence when RCTs are not possible or available.[10] However, less is known about emulation of ongoing RCTs. Emulating RCTs prior to, rather than following, their publication could help avoid the potential biases introduced by trying to replicate the results, rather than emulate the methods of RCTs that have already been completed and disseminated among the scientific community. Additionally, real-world emulation can contextualize RCT findings with respect to their generalizability to diverse settings and populations. In order to determine the feasibility of emulating ongoing RCTs, we conducted two trial emulations using linked claims and electronic health record data, anticipating the populations and results of each trial prior to publication of its results.[11, 12] Specifically, we emulated the PRONOUNCE trial,[13] a phase 3b RCT comparing cardiovascular safety of degarelix and leuprolide among patients with prostate cancer and cardiovascular disease,[11] and the GRADE trial,[14] a pragmatic, randomized, parallel-arm trial comparing four second-line glucose-lowering drugs among adults with moderately uncontrolled type 2 diabetes on metformin monotherapy.[12] The primary publications for these emulation trials demonstrated the feasibility of this approach to generate timely data on comparative effectiveness and safety prior to the publication of the actual RCTs.[11, 12] With the recent publication of the primary results of both the PRONOUNCE and GRADE trials,[13, 14] comparison of our findings with those RCT populations and results can enhance our understanding of the appropriate role of using real-world data to emulate clinical trials of medical products.

Our objectives for this study are therefore threefold: 1) to compare the RCT participants to patients identified in real-world data after application of the pre-specified trial eligibility criteria; 2) to compare the final primary and secondary endpoint results from the published RCTs with the emulated trials; and 3) to use these findings to provide insight into the advantages and challenges of conducting emulations of ongoing trials.

Methods

Initial identification of ongoing trials

Using ClinicalTrials.gov, we previously identified two comparative effectiveness and safety trials that were ongoing and unpublished at the time of our search. We limited our sample to trials that: 1) tested a drug that had already received U.S. Food and Drug Administration approval and been available for use for at least 3 years; 2) examined the use of that product for cardiovascular disease or diabetes management; 3) planned to enroll 500 patients or more; 4) recruited patients from sites in the U.S.; and 5) focused on outcomes that could be reasonably ascertained from insurance claims or electronic health record data. Based on our search, the PRONOUNCE (A Trial Comparing Cardiovascular Safety of Degarelix Versus Leuprolide in Patients With Advanced Prostate Cancer and Cardiovascular Disease)[15] and GRADE (A Comparative Effectiveness Study of Major Glycemia-lowering Medications for Treatment of Type 2 Diabetes)[14] trials were selected.

In the emulation of the PRONOUNCE trial, the primary endpoint was the time to first occurrence of a major adverse cardiovascular event, a composite endpoint defined as all-cause death, nonfatal myocardial infarction, or nonfatal stroke. The PRONOUNCE trial major adverse cardiovascular event outcome was modified to include all-cause rather than cardiovascular mortality, as cause of death is not available in our data. In the emulation of the GRADE trial, the primary endpoint was time to primary metabolic failure, calculated as days to hemoglobin A1c ≥7% while receiving the assigned medication, with eligibility for outcome ascertainment starting at month 3 after the index date (analogous to the first quarterly hemoglobin A1c assessment in GRADE). The hemoglobin A1c outcomes were modified from the GRADE trial, where confirmation of an elevated result with a repeat hemoglobin A1c was required, as this is not done in routine care. Secondary outcomes are listed in Supplementary Table 1.

Trial emulation

Two separate retrospective cohort studies were conducted using deidentified administrative claims data from OptumLabs® Data Warehouse, which includes electronic health record, medical and pharmacy claims, laboratory results, and enrollment records for commercial and Medicare Advantage enrollees, representing a diverse range of racial and ethnic groups, ages, and geographic regions across the U.S. The findings of both retrospective cohort studies have been published[11, 12] and the Supplementary Materials includes additional information.

Statistical analysis

Comparison of trials and emulated trials.—First, we compared the total sample sizes and baseline characteristics of participants included in the emulated trials and published RCTs. Variables included in the published RCTs were assessed for feasibility of implementing them in real-world data. Variables were classified as *ascertainable* if it was possible to develop a computable phenotype using available structured data, including diagnosis codes, procedure codes, medications, and laboratory test results. Variables were classified as *partially ascertainable* if it was possible to develop: (a) an approximate, but not exact, phenotype using available structured data, (b) an approximate phenotype for variables where the RCT was not detailed enough in the specification of their eligibility criteria (e.g., only mentioning 'lipid-modifying agents' instead of the specific medications), or (c) an exact phenotype using available structured data, but the data were not available for all participants (i.e., laboratory measures available for a subset of the cohorts as a result of agreements between OptumLabs and commercial laboratories but not the complete sample). Finally, variables were classified as *not ascertainable* if it was not feasible to develop a computable phenotype using available structured data.

We then recorded and summarized the baseline characteristics of participants of all ascertainable or partially ascertainable variables from the RCTs, including those that were not included in the initial emulations but were part of the published versions of the RCTs. Pairwise comparisons were conducted between the total populations for the emulated trial and published RCT variables, with standardized mean differences <0.20 considered concordant.[16]

To compare the primary and secondary endpoints between the emulated trials and published RCTs, we characterized the effect estimates for each outcome based on its statistical significance (e.g., *P-value* <0.05 vs ≥0.05) and direction (i.e., "increased" for relative effect estimates greater than 1 or mean differences greater than 0, and "decreased" for relative effect estimates less than 1 or mean differences less than 0). As in previous evaluations,[5] effect estimates from matched pairs were considered to be concordant if 1) there was concordant direction of the emulated trial and published RCT effect estimates and both effect estimates were statistically significant or 2) the emulated trial and published RCT effect estimates were both non-statistically significant. Effect estimates from matched pairs that did not fulfill criteria (1) or (2) were classified as discordant. As secondary measures of concordance, we determined how often the emulated trial and published RCT effect estimates had overlapping 95% confidence intervals.

The Mayo Clinic institutional review board exempted this study from review and the requirement for informed consent because it used preexisting, deidentified data.

Results

PRONOUNCE trial emulation

Data ascertainment for participant characteristics.—The PRONOUNCE trial stopped enrollment at 544 (275 in the degarelix arm and 269 in the leuprolide arm) of the 900 planned patients because of difficulties with recruitment and fewer than expected primary endpoint events. The emulated trial included 2226 propensity score-matched participants (1113 in the degarelix arm and 1113 in the leuprolide arm).

Among the 32 baseline variables included in Table 1 of the PRONOUNCE trial publication, one variable, prior hormonal therapy (unless terminated at least 12 months prior to trial), was listed as an exclusion criterion on ClinicalTrials.gov but was ultimately not used as an exclusion criterion in the published RCT. Therefore, while investigators of the RCT included patients who had received prior hormonal therapy (8.1% of their cohort), our emulated trial cohort excluded them. Of the remaining 31 baseline variables (Table 1, Supplementary Table 2), 9 (29.0%) were ascertainable, 11 (35.5%) were partially ascertainable, and 10 (32.3%) variables that were not ascertainable using structured data from OptumLabs. Additionally, the published RCT did not provide sufficient detail to ascertain the ‘other’ prostate cancer therapy variable, using structured data from OptumLabs (i.e., did not specify what is included in ‘other’).

Comparison of emulated trial vs. published RCT participants.—Of the nine participant characteristics that were ascertainable using structured data from OptumLabs, 6 (66.7%) were discordant between the emulated trial and published RCT. The men included in the emulated trial were slightly older than the men included in the published RCT (Supplementary Table 2). Men in the emulated trial were also less likely to have received prior radiotherapy and radical prostatectomy than those in the published RCT. Men in the emulated trial were less likely to have a history of myocardial infarction and hypertension, but more likely to have a history of coronary, carotid, or iliofemoral revascularization and dyslipidemia.

Of the 10 variables that were partially ascertainable, 7 (70.0%) were discordant. Men in the emulated trial were more likely to be Black or African American and less likely to be White than those in the published RCT (Supplementary Table 2). Mean total serum cholesterol was higher in the emulated trial than the published RCT. Men in the emulated trial with available prostate specific antigen data (38.0% of the cohort) had higher mean values than those in the published RCT. Men in the emulated trial were less likely to have received cardiovascular disease medications, agents acting on the renin-angiotensin system, and β-blockers than men in the published RCT.

Comparison of emulated trial vs. published RCT results.—The primary endpoint of time from randomization to first major adverse cardiovascular event and secondary endpoints of non-fatal myocardial infarction and stroke were concordant between the emulated trial and the published PRONOUNCE trial (Table 3). Overall, the hazard ratios from the emulated trial were in the same direction (major adverse cardiovascular event: hazard ratio 1.18 [95% confidence interval, 0.86-1.61], myocardial infarction: hazard ratio 1.16 [0.60-2.25]; and stroke: hazard ratio 0.92 [0.45-1.85] as those observed in the published RCT (major adverse cardiovascular event: hazard ratio 1.28 [0.59-2.79]; myocardial infarction: hazard ratio 1.59 [0.38-6.67], and stroke: hazard ratio 0.90 [0.18-4.46]). Importantly, none of the results reached the threshold for statistical significance for either the emulated trial or the published RCT. The hazard ratios for the secondary endpoint of all-cause death were in the opposite direction, with a statistically significant hazard ratio observed in the emulated trial (hazard ratio 1.48 [1.01-2.18]) and a non-statistically significant hazard ratio observed in the published RCT (hazard ratio 0.84 [0.32-2.18]).

GRADE Trial emulation

Data ascertainment for participant characteristics.—The GRADE trial enrolled 5047 participants (approximately 1260 per each of the 4 study arms), while the emulated trial included 7540 patients (weighted) (4168 in the glimepiride arm, 572 in the liraglutide arm, and 2800 in the sitagliptin arm). The glargine arm had to be excluded from the emulation trial analyses due to inability to achieve balance on baseline covariates.

Of the 34 baseline variables described in the GRADE trial publication (Table 2, Supplementary Table 3), 8 (23.5%) were ascertainable, 14 (41.2%) were partially ascertainable, and 11 (32.4%) were not ascertainable using structured data from OptumLabs,

Comparison of emulated trial vs. published RCT participants.—Of the 8 baseline variables that were ascertainable using structured data from OptumLabs and were listed in Table 1 of the GRADE trial baseline characteristics publication, 4 (50.0%) were discordant. Participants in the emulated trial were older and less likely to be male than participants included in the published RCT (Supplementary Table 3). Overall, participants in the emulated trial were less likely to have neuropathy but more likely to have retinopathy than participants in published RCT.

Of the 14 participant characteristics that were partially ascertainable, 9 (64.3%) were discordant. Participants in the emulated trial were more likely to have completed some college and to have received blood pressure and lipid-lowering (medications than those included in published RCT (Supplementary Table 3). Participants in the emulated trial were less likely to be identified or classified as current smokers and have urine albumin-creatinine ratio <30 mg/g. Participants in the emulated trial had higher hemoglobin A1c, total cholesterol, triglycerides, and serum creatinine levels, with correspondingly lower estimated glomerular filtration rate levels than participants in the published RCT.

Comparison of emulated trial vs. published RCT results.—The primary endpoint of time to hemoglobin A1c ≥7.0% was concordant between the emulated trial and the published RCT for some, but not all, pairwise comparisons (Table 4). Overall, the hazard ratios from the emulated trial were concordant in direction and significance (liraglutide vs. glimepiride for primary metabolic failure: hazard ratio 0.57 [95% confidence interval, 0.43-0.75] and liraglutide vs. sitagliptin for primary metabolic failure: hazard ratio 0.55 [0.41-0.73]) with those observed in the published RCT (liraglutide vs. glimepiride: hazard ratio 0.87 [0.79-0.96] and liraglutide vs. sitagliptin: hazard ratio 0.39 [0.63-0.76]). However, the emulated trial found no significant difference between sitagliptin and glimepiride with respect to achieving primary metabolic failure (hazard ratio 1.03 [0.94-1.13]), while the published GRADE trial found a statistically significant difference between them (hazard ratio 1.27 [1.14-1.39]). Similar patterns were seen for the outcome of secondary metabolic failure. Results for the other secondary outcomes were concordant, with neither the emulated trial nor the published RCT finding significant differences between the drugs in achieving tertiary metabolic failure or any of the examined microvascular, cardiovascular, and other secondary endpoints.

Discussion

In our two trial emulations using linked electronic health record and claims data from patients managed in routine practice settings across the U.S., we found that despite differences in baseline characteristics of accrued participants, the results were mostly concordant between our emulation of the PROUNOUNCE and GRADE trials and the published RCTs. Our experience, while demonstrating the feasibility of emulating trials prior to their publication, also identified important challenges and opportunities for the use of emulated trials as complement to RCTs evaluating efficacy and safety of approved drug products. Most notable challenges related to the ability to reasonably ascertain eligibility criteria and certain baseline participant characteristics from insurance claims or electronic health record data and the reliance on data elements with variable levels of missingness. At the same time, our emulated trials suggest that observational methods and real-world data may serve as a complement to RCTs, providing more timely, less expensive, or larger evaluations of drug effectiveness and safety in potentially more diverse and representative patient populations in certain situations.

Our experience highlights that target trial emulations are most likely to be feasible when they are evaluations of the comparative efficacy and/or safety of two or more prescribed therapeutics among relatively broadly-defined populations. Certain categories of target trials are difficult or impossible to emulate using real-world data, including placebo-controlled trials, trials using behavioral interventions as controls, and trials of unapproved therapeutics or those available over-the-counter. Target trials to be emulated should also have indications, key eligibility criteria, and endpoints for which it is possible to develop computable phenotypes using structured data, including diagnosis codes, procedure codes, medications, and laboratory test results. Target trials with characteristics or endpoints based on clinical measures, changes in symptom scoring, patient reported outcomes, or other unstructured data are less feasible to emulate as these variables cannot be routinely ascertained using claims data, even when linked to electronic health record data. While common data models that provide standardized data organization for the detailed patient information available within real-world data may expand the types of target trials that are feasible to emulate, these models have not been widely adopted by health systems.

If these characteristics can be met, there are important advantages of applying observational methods to real-world data. First, trial emulations can be completed at a fraction of the time and cost of a traditional RCT. While the PRONOUNCE and GRADE trial emulations were designed and conducted within 18 months, the RCTs took 5 and 10 years to complete, respectively. When RCTs take a long time to complete, it can undermine their relevance and ability to inform clinical practice. For instance, by the time the GRADE trial was completed and published,[14] more contemporary treatments, such as the sodium-glucose cotransporter-2 inhibitor class of glucose lowering drugs, became available and widely used.[17, 18]

Second, trial emulations do not face the same recruitment challenges as RCTs, which can allow for larger and more generalizable real-word populations. For example, the PRONOUNCE trial was terminated early due to slower than anticipated enrollment, which meant the emulation cohort was nearly four times larger than the underpowered RCT.[13] Our trial emulations also captured diverse populations, including older adults, racial and ethnic minoritized populations, and patients with multiple comorbidities, who are often underrepresented in clinical trials.[19] It is essential that the evidence used to assess treatments is based on a population similar to the people who will ultimately use the treatment. Third, real-world data sources can provide a wide range of potential characteristics and outcomes. Lastly, by following the target trial framework,[10] which was developed to inform the methodological considerations for conducting causal comparative effectiveness analyses using real-world data, emulations can attempt to avoid design flaws that often undermine the evidence generated by nonrandomized studies.

Despite the advantages of conducting trial emulations, we encountered several challenges that limited our ability to fully compare participant characteristics and final results from the RCTs with emulation trials. Although we implemented the eligibility criteria of the RCTs as closely as possible,[11, 12] we found that across both trial emulations, fewer than 30% of the baseline patient characteristics reported in the trial publications were fully ascertainable using structured data in OptumLabs. Nearly all laboratory measures were classified as either partially or not ascertainable because these variables were only available for a subset of the emulation trial cohort because of agreements between OptumLabs and commercial laboratories (and most other claims data do not have any laboratory results available). Notably, while use of a health system’s electronic health record rather than claims data would allow for more complete capture of laboratory and other clinical parameters, this would likely limit overall sample size and generalizability, and require significant effort for extraction of variables from unstructured data. Overall, the emulated trial and published RCT cohorts were significantly different across most fully and partially ascertainable patient characteristics, which highlights the fact that even when closely emulating the eligibility criteria of RCTs, RCT and real-world populations are inherently different under the current RCT paradigm.[20] These barriers to RCT external validity are due to the fact that many RCTs capture a highly selected population because of the kinds of health systems that participate in RCTs (i.e., primarily academic), and the burden of RCT participation on patients (e.g., multiple lengthy in-person visits). While pragmatic trials seek to improve generalizability and external validity of their findings through less stringent eligibility criteria and study procedures, they, too, are often limited by where they can be conducted and the types of patients able to participate.[21] Therefore, trial emulations using real-world data sources are an important source of evidence of treatment effectiveness and safety because they use a population that more closely resembles the general population of patients with the disease of interest, spanning a wide range of disease complexity and comorbidity, and as such may be more representative of the settings and patients using the therapeutic.

Our experience also highlights that trial emulations are unable to account for all the differences between how medications are used in clinical practice versus evaluated in an RCT. Although we were able to identify an adequate number of patients receiving both interventions evaluated in the PRONOUNCE trial emulation, we found that that relatively few patients started on degarelix without switching over to leuprolide in clinical practice and that exposure to degarelix was relatively short.[11] Outside of an RCT, patients with prostate cancer may switch to leuprolide due to the less frequent dosing interval or due to progressive disease. In the GRADE trial emulation, one of the treatment arms (glargine) was excluded because we were not able to achieve balance on baseline patient characteristics even after weighting.[12] This was not surprising, as treatment with basal insulin in the clinical context examined by the GRADE trial (i.e., as a second-line agent after metformin in patients with HbA1c less than 8.5%) is outside the standard of care and mainstream clinical practice.[12] Overall, these findings suggest that trial emulations should carefully consider differences between how medications are studied in trials and whether and how they are used in routine practice. An additional challenge we encountered was that the statistical analysis plan for an RCT can change over time and often without being updated in ClinicalTrials.gov (Supplementary Materials).

Limitations

This study has several limitations. First, as our manuscript focuses on the experiences encountered conducting emulations of two specific RCTs, the findings may not be generalizable to other RCT designs. However, we selected the PRONOUNCE and GRADE trials because they captured different indications, populations, interventions, and outcome types, providing a broad overview of the types of challenges and considerations that could be encountered across different fields. Second, since the study cohorts comprised people with private and Medicare Advantage health plans, the results may not generalize to all people with or without insurance coverage. However, in the absence of an all-payer claims database (and which, regardless, would not have electronic health record data) that captures the entire U.S. population, there are no perfect data sources to conduct trial emulations. Third, we did not consider the concordance between the RCTs and emulated trials across different subgroup analyses, as subgroup results were not consistently reported in the published RCTs. Furthermore, we could only emulate the North American portion of the PRONOUNCE Trial, as half of the population was recruited from centers in Europe. Difference between these populations across race and ethnicity are likely. Lastly, the variables that can and cannot be ascertained depends on the real-world data source being used for the emulation. We used OptumLabs Data Warehouse, which is uniquely robust for trial emulation because it captures a national, large, and diverse population and includes laboratory test results and electronic health record data for a subset of the population.

Conclusion

In our emulations of the PRONOUNCE and GRADE trials using linked electronic health record and claims data, we found that although there were some differences between the characteristics of patients comprising the cohorts of the emulated trial and the published RCTs after application of the pre-specified trial eligibility criteria, the results from primary and secondary endpoints were largely concordant in terms of direction and statistical significance. Despite notable challenges, our emulated trials suggest that observational methods and real-world data may serve as a complement to RCTs, providing more timely, less expensive, or larger evaluations of drug effectiveness and safety in potentially more diverse and representative patient populations in certain situations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Dr Wallach reported receiving research support through the Collaboration for Regulatory Rigor, Integrity, and Transparency (CRRIT) at Yale University from Arnold Ventures and through the Yale–Mayo Clinic Center for Excellence in Regulatory Science and Innovation (CERSI) program (U01FD005938) and being supported by the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under award K01AA028258. Dr McCoy reported receiving support from the National Institute of Health National Institute of Diabetes and Digestive and Kidney Diseases (K23DK114497, R03DK127010), the National Institute on Aging (R01AG079113), and the Patient-Centered Outcomes Research Institute (DB-2020C2-20306). She also serves as a consultant o Emmi(R) (Wolters Kluwer) on developing patient education materials related to diabetes. Dr Dhruva reported being funded by the Department of Veterans Affairs, the National Evaluation System for Health Technology (NEST), and Arnold Ventures. Dr Herrin reported working under contract to the Centers for Medicare & Medicaid Services on the development and evaluations of measures of health care quality. Dr Noseworthy reported receiving personal fees from Optum during the conduct of the study. Dr Shah reported receiving support through the Mayo Clinic from the Food and Drug Administration to establish the Yale–Mayo Clinic CERSI program (U01FD005938); the Centers for Medicare & Medicaid Innovation under the Transforming Clinical Practice Initiative; the Agency for Healthcare Research and Quality (R01HS025164, R01HS025402, R03HS025517, and K12HS026379); the National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health (R56HL130496, R01HL131535, and R01HL151662); the National Science Foundation; and the Patient-Centered Outcomes Research Institute to develop a clinical data research network (LHSNet). Dr. Ross currently receives research support through Yale University from Johnson and Johnson to develop methods of clinical trial data sharing, from the Medical Device Innovation Consortium as part of the National Evaluation System for Health Technology (NEST), from the Food and Drug Administration for the Yale-Mayo Clinic Center for Excellence in Regulatory Science and Innovation (CERSI) program (U01FD005938), from the Agency for Healthcare Research and Quality (R01HS022882), from the National Heart, Lung and Blood Institute of the National Institutes of Health (NIH) (R01HS025164, R01HL144644), and from Arnold Ventures; in addition, Dr. Ross was an expert witness at the request of Relator’s attorneys, the Greene Law Firm, in a qui tam suit alleging violations of the False Claims Act and Anti-Kickback Statute against Biogen Inc. that was settled September 2022.

Funding/Support

This publication is supported by the US Food and Drug Administration (FDA) of the US Department of Health and Human Services (HHS) as part of a financial assistance award (U01FD005938) totaling $250,000 with 100% funded by FDA/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.

Role of the Funder/Sponsor

The FDA was involved in the design and conduct of the study, collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data sharing

This study was conducting using de-identified data from OptumLabs Data Warehouse. Raw data are not publicly available.

References

1. Jones DS and Podolsky SH. The history and fate of the gold standard. Lancet 2015; 385: 1502–1503. [PubMed: 25933270]
2. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016; 375: 2293–2297. [PubMed: 27959688]
3. Wallach JD, Zhang AD, Skydel JJ, et al. Feasibility of Using Real-world Data to Emulate Postapproval Confirmatory Clinical Trials of Therapeutic Agents Granted US Food and Drug Administration Accelerated Approval. JAMA Netw Open 2021; 4: e2133667. [PubMed: 34751763]

4. Melloni C, Washam JB, Jones WS, et al. Conflicting results between randomized trials and observational studies on the impact of proton pump inhibitors on cardiovascular events when coadministered with dual antiplatelet therapy: systematic review. Circ Cardiovasc Qual Outcomes 2015; 8: 47–55. [PubMed: 25587094]

5. Franklin JM, Patorno E, Desai RJ, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. Circulation 2021; 143: 1002–1013. [PubMed: 33327727]

6. Fralick M, Bartsch E, Darrow JJ, et al. Understanding when real world data can be used to replicate a clinical trial: A cross-sectional study of medications approved in 2011. Pharmacoepidemiol Drug Saf 2020; 29: 1273–1278. [PubMed: 32798299]

7. Lonjon G, Boutron I, Trinquart L, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. Ann Surg 2014; 259: 18–25. [PubMed: 24096758]

8. Tannen RL, Weiner MG and Xie D. Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. Pharmacoepidemiol Drug Saf 2008; 17: 671–685. [PubMed: 18327852]

9. Tannen RL, Weiner MG, Xie D, et al. A simulation using data from a primary care practice database closely replicated the women's health initiative trial. J Clin Epidemiol 2007; 60: 686–695. [PubMed: 17573984]

10. Hernán MA and Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016; 183: 758–764. [PubMed: 26994063]

11. Wallach JD, Deng Y, McCoy RG, et al. Real-world Cardiovascular Outcomes Associated With Degarelix vs Leuprolide for Prostate Cancer Treatment. JAMA Netw Open 2021; 4: e2130587. [PubMed: 34677594]

12. Deng Y, Polley EC, Wallach JD, et al. Emulating the GRADE trial using real world data: retrospective comparative effectiveness study. BMJ 2022; 379: e070717. [PubMed: 36191949]

13. Lopes RD, Higano CS, Slovin SF, et al. Cardiovascular Safety of Degarelix Versus Leuprolide in Patients With Prostate Cancer: The Primary Results of the PRONOUNCE Randomized Trial. Circulation 2021; 144: 1295–1307. [PubMed: 34459214]

14. Nathan DM, Lachin JM, Balasubramanyam A, et al. Glycemia Reduction in Type 2 Diabetes - Glycemic Outcomes. N Engl J Med 2022; 387: 1063–1074. [PubMed: 36129996]

15. Melloni C, Solvin SF, Blemings A, et al. Cardiovascular Safety of Degarelix Versus Leuprolide for Advanced Prostate Cancer: The PRONOUNCE Trial Study Design. JACC CardioOncol 2020; 2(1):70–81. [PubMed: 34396210]

16. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009; 28: 3083–3107. [PubMed: 19757444]

17. Davies MJ, Aroda VR, Collins BS, et al. Management of Hyperglycemia in Type 2 Diabetes, 2022. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care 2022; 45: 2753–2786. [PubMed: 36148880]

18. ElSayed NA, Aleppo G, Aroda VR, et al. 9. Pharmacologic Approaches to Glycemic Treatment: Standards of Care in Diabetes-2023. Diabetes Care 2023; 46: S140–S157. [PubMed: 36507650]

19. Franklin JM and Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther 2017; 102: 924–933. [PubMed: 28836267]

20. Groenwold RHH. Trial Emulation and Real-World Evidence. JAMA Netw Open 2021; 4: e213845. [PubMed: 33783521]

21. Loudon K, Treweek S, Sullivan F, et al. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ 2015; 350: h2147. [PubMed: 25956159]

Table 1.

Comparison of the PROUNOUNCE trial and emulated trial cohorts

Characteristics included in the Table 1 of the PRONOUNCE trial	**Ability to ascertain characteristic using OptumLabs data**	**Trial and emulated trial characteristics concordant**[a]
Age	Ascertainable	No
Race	Partially ascertainable	No
Ethnicity	Partially ascertainable	Yes
Weight	Not ascertainable	NA
Body mass index	Not ascertainable	NA
Smoking status	Partially ascertainable	Yes
Baseline blood pressure	Not ascertainable	NA
Total serum cholesterol	Partially ascertainable	No
Type 2 diabetes	Ascertainable	Yes
N-Terminal prohormone B-type natriuretic peptide	Not ascertainable	NA
High-sensitivity C-reactive protein	Partially ascertainable	Yes
Troponin T	Not ascertainable	NA
Radiotherapy	Ascertainable	No
Radical prostatectomy	Ascertainable	No
Hormonal therapy	NA: This was an exclusion criteria	NA
Other	NA: No information provided in PRONOUNCE Trial	NA
Gleason score	Not ascertainable	NA
Stage of prostate cancer	Not ascertainable	NA
Testosterone	Partially ascertainable	Yes
Prostate specific antigen	Partially ascertainable	No
Myocardial infarction	Ascertainable	No
Coronary carotid, or iliofemoral revascularization	Ascertainable	No
Coronary, carotid, or iliofemoral stenosis >50% by angiography,	Not ascertainable	NA
Carotid stenosis >50% by ultrasound	Not ascertainable	NA
Ankle-brachial index <0.9	Not ascertainable	NA
Atrial fibrillation	Ascertainable	Yes
Dyslipidemia	Ascertainable	No
Hypertension	Ascertainable	Yes
Cardiovascular medications	Partially ascertainable	No
Lipid-modifying agents	Partially ascertainable	Yes
Agents acting on the renin-angiotensin system	Partially ascertainable	No
β-Blockers	Partially ascertainable	No

[a] Concordant defined as a standardized mean differences between trial and emulated cohort variables <0.20. NA, not applicable

Table 2.

Comparison of GRADE trial and emulated trial cohorts

Characteristics included in the Table 1 of the GRADE trial	Ability to ascertain characteristic using OptumLabs data	Trial and emulated trial characteristics concordant[a]
Age	Ascertainable	No
Sex	Ascertainable	No
Race	Partially ascertainable	Yes
Ethnicity	Partially ascertainable	Yes
Education completed	Partially ascertainable	No
Duration of diabetes	Not ascertainable	NA
Screening metformin dose	Not ascertainable	NA
Baseline metformin dose	Not ascertainable	NA
Family history of any first-degree relatives with diabetes	Not ascertainable	NA
Heart attack/stroke	Ascertainable	Yes
Retinopathy	Ascertainable	No
Neuropathy	Ascertainable	No
Hypertension	Ascertainable	Yes
Elevated blood lipids	Ascertainable	Yes
Blood pressure medications	Partially ascertainable	No
Lipid-lowering medications	Partially ascertainable	No
Statin	Ascertainable	Yes
Aspirin	Not ascertainable:	NA
Depression/anxiety medication(s)	Not ascertainable	NA
Smoking status	Partially ascertainable	No
Weight	Not ascertainable	NA
Body mass index	Not ascertainable	NA
Blood pressure	Not ascertainable	NA
Hemoglobin A1c	Partially ascertainable	Yes
Cholesterol	Partially ascertainable	No
Triglycerides	Partially ascertainable:	No
High-density lipoprotein	Partially ascertainable	Yes
Low-density lipoprotein	Partially ascertainable	Yes
Urine albumin-creatinine ratio	Partially ascertainable	No
Fasting glucose	Not ascertainable	NA
Estimated glomerular filtration rate	Partially ascertainable	No
Serum creatinine	Partially ascertainable	No
Fasting C-peptide	Not ascertainable	NA
Fasting insulin	Not ascertainable	NA

[a] Concordant defined as a standardized mean differences between trial and emulated cohort variable <0.20. NA, not applicable

Table 3.

Comparison of PRONOUNCE Trial and Emulation Trial Primary and Secondary Endpoint Results

	PRONOUNCE Trial		PRONOUNCE Emulation		Concordance
Endpoints	**Hazard ratio (95% confidence interval)**	**P-value**	**Hazard ratio (95% confidence interval)**	**P-value**	**Same direction; same level of significance; overlapping 95% confidence intervals**
Primary					
Time from randomization to first adjudicated major adverse cardiovascular event	1.28 (0.59-2.79)	0.53	1.18 (0.86-1.61)	0.30	Yes; Yes; Yes
Secondary					
Myocardial infarction	1.59 (0.38-6.67)	0.52	1.16 (0.60-2.25)	0.66	Yes; Yes; Yes
Stroke	0.90 (0.18-4.46)	0.90	0.92 (0.45-1.85)	0.81	Yes; Yes; Yes
All-cause death	0.84 (0.32-2.18)	0.72	1.48 (1.01-2.18)	0.046	No; No; Yes
Angina	Not reported	Not reported	1.36 (0.43-4.27)	0.60	NA; NA; NA

Table 4.

Comparison of GRADE Trial and Emulation Trial Primary and Secondary Endpoint Results

	GRADE Trial		**GRADE Emulation**		**Concordance**
Endpoint	**Hazard ratio (95% confidence interval)**	**P-value**	**Hazard ratio (95% confidence interval))**	**P-value**	**Same direction; same level of significance; overlapping 95% confidence interval**
Primary					
Primary Metabolic Failure (time to Hemaglobin A_{1c} ≥7.0%)					
Liraglutide *v* glimepiride	0.87 (0.79 to 0.96)	≤0.01	0.57 (0.43 to 0.75)	<0.001	Yes; Yes; No
Sitagliptin *v* glimepiride	1.27 (1.14 to 1.39)	≤0.001	1.03 (0.94 to 1.13)	0.48	Yes; No; No
Liraglutide *v* sitagliptin	0.69 (0.63 to 0.76)	≤0.001	0.55 (0.41 to 0.73)	<0.001	Yes; Yes; Yes
Secondary					
Secondary Metabolic Failure (time to Hemaglobin A_{1c} >7.5%)					
Liraglutide *v* glimepiride	0.88 (0.79 to 0.99)	none provided	0.61 (0.43 to 0.87)	0.01	Yes; Yes; Yes
Sitagliptin *v* glimepiride	1.19 (1.08 to 1.33)	none provided	1.04 (0.91 to 1.18)	0.6	Yes; No; Yes
Liraglutide *v* sitagliptin	0.74 (0.66 to 0.83)	none provided	0.59 (0.41 to 0.85)	0.01	Yes; Yes; Yes