JAMA | Original Investigation

Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses
Results of 32 Clinical Trials

Shirley V. Wang, PhD, ScM; Sebastian Schneeweiss, MD, ScD; and the RCT-DUPLICATE Initiative

Editorial page 1352

Supplemental content

IMPORTANCE Nonrandomized studies using insurance claims databases can be analyzed to produce real-world evidence on the effectiveness of medical products. Given the lack of baseline randomization and measurement issues, concerns exist about whether such studies produce unbiased treatment effect estimates.

OBJECTIVE To emulate the design of 30 completed and 2 ongoing randomized clinical trials (RCTs) of medications with database studies using observational analogues of the RCT design parameters (population, intervention, comparator, outcome, time [PICOT]) and to quantify agreement in RCT-database study pairs.

DESIGN, SETTING, AND PARTICIPANTS New-user cohort studies with propensity score matching using 3 US claims databases (Optum Clinformatics, MarketScan, and Medicare). Inclusion-exclusion criteria for each database study were prespecified to emulate the corresponding RCT. RCTs were explicitly selected based on feasibility, including power, key confounders, and end points more likely to be emulated with real-world data. All 32 protocols were registered on ClinicalTrials.gov before conducting analyses. Emulations were conducted from 2017 through 2022.

EXPOSURES Therapies for multiple clinical conditions were included.

MAIN OUTCOMES AND MEASURES Database study emulations focused on the primary outcome of the corresponding RCT. Findings of database studies were compared with RCTs using predefined metrics, including Pearson correlation coefficients and binary metrics based on statistical significance agreement, estimate agreement, and standardized difference.

RESULTS In these highly selected RCTs, the overall observed agreement between the RCT and the database emulation results was a Pearson correlation of 0.82 (95% CI, 0.64-0.91), with 75% meeting statistical significance, 66% estimate agreement, and 75% standardized difference agreement. In a post hoc analysis limited to 16 RCTs with closer emulation of trial design and measurements, concordance was higher (Pearson *r*, 0.93; 95% CI, 0.79-0.97; 94% meeting statistical significance, 88% estimate agreement, 88% standardized difference agreement). Weaker concordance occurred among 16 RCTs for which close emulation of certain design elements that define the research question (PICOT) with data from insurance claims was not possible (Pearson *r*, 0.53; 95% CI, 0.00-0.83; 56% meeting statistical significance, 50% estimate agreement, 69% standardized difference agreement).

CONCLUSIONS AND RELEVANCE Real-world evidence studies can reach similar conclusions as RCTs when design and measurements can be closely emulated, but this may be difficult to achieve. Concordance in results varied depending on the agreement metric. Emulation differences, chance, and residual confounding can contribute to divergence in results and are difficult to disentangle.

JAMA. 2023;329(16):1376-1385. doi:10.1001/jama.2023.4221

Author Affiliations: Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts.

Group Information: The RCT-DUPLICATE Initiative authors appear at the end of the article.

Corresponding Author: Shirley Wang, PhD, Harvard Medical School, 1620 Tremont St, Ste 303, Boston, MA 02120 (swang1@bwh.harvard.edu).

Randomized clinical trials (RCTs) are the accepted standard to understand the efficacy of medical products.[1] Unfortunately, trials cannot be conducted to evaluate every aspect of a drug's effect in all population segments relevant to clinical practice. Decision-makers are interested in using real-world evidence to understand the effectiveness of medical products in clinical practice.[2,3] Real-world evidence is derived from studies conducted with nonrandomized data, including data routinely collected by the health care system, such as longitudinal insurance claims and electronic health records.[2] Although the potential for real-world evidence to inform clinical practice is recognized, the proliferation of studies with variable methodological rigor[4-6] has made it challenging to confidently determine whether real-world evidence studies can yield actionable insights by providing causal conclusions on treatment effects. To understand the validity of real-world evidence studies, a natural comparison is between the results of a real-world evidence study and the accepted standard for valid causal inference, a well-conducted RCT.

Multiple studies that have compared findings between published RCTs and nonrandomized real-world evidence studies have had mixed conclusions.[7-10] However, these comparisons used real-world evidence studies that were not designed to mimic the RCTs. The varying degree of mismatch between the design and the target question makes it difficult to assess agreement in results.

RCT-DUPLICATE is an initiative to better understand to what extent real-world evidence studies conducted using health care databases can provide valid causal inference. The premise is to use RCT results as a reference standard for valid causal inference and to learn whether similar clinical conclusions would have been drawn if the RCT protocol had been translated and implemented as a database study protocol. We aimed to emulate RCT designs under the best possible circumstances by identifying and implementing observational analogues of RCT design parameters that define the research question (population, intervention, comparator, outcome, time-frame [PICOT][11]), apply confounding adjustment methods, and then compare the results of RCT-database study pairs (eMaterials 1 in Supplement 1).[12-14]

We present the results from emulating 30 completed RCTs investigating medication treatment effects. These RCTs were selected because their design and measurements seemed amenable to emulation with health care claims data; we also present results from predicting the findings of 2 trials that were ongoing at the time of the emulation.

Key Points

Question Are database studies that are explicitly designed to emulate past and ongoing randomized clinical trials (RCTs) of medications able to generate similar causal conclusions?

Findings In this highly selected, nonrepresentative sample, real-world evidence studies generally reached similar conclusions as RCTs (Pearson correlation *r*, 0.82; 75% statistical significance agreement, 66% estimate agreement, 75% standardized difference agreement). In a post hoc, exploratory stratified analysis, agreement was higher in RCT-database pairs classified as having closer emulation of the RCT design.

Meaning Selected database studies can complement RCT evidence to enhance understanding of how medications work in clinical practice. Emulation differences, chance, and residual confounding can contribute to divergence in results and are difficult to disentangle.

Methods

Trial Selection

The nonrepresentative trial selection process was described in a previous article.[14] Trials were selected based on (1) the observability of key study parameters in health care claims data and (2) feasibility checks. In other words, the treatment, comparator, outcome, and key trial inclusion-exclusion criteria had to be measurable within the data sources we were using; key confounding variables had to be measured and balanced after propensity score matching; and the patient counts had to be sufficient for the database study to have power at least equal to the RCT. Details on trial selection appear in eMaterials 1 and eTable 1 in Supplement 1.

The selected trials were designed for regulatory submissions and aimed to support superiority or noninferiority claims (**Table 1** and eTable 1 in Supplement 1). They included 8 trials evaluating cardiovascular outcomes of antidiabetic medications; 1 trial of the influence of antidiabetic drugs on hemoglobin A_{1c}, 3 trials of the effectiveness of antiplatelet agents on cardiovascular outcomes, 3 trials of direct oral anticoagulants for atrial fibrillation, 5 trials of direct oral anticoagulants for venous thromboembolism; 2 trials of antihypertensive drugs; 2 trials of osteoporosis therapies; 1 trial of therapy for chronic kidney disease, 1 trial of therapy for heart failure; 2 trials of asthma treatments; 3 trials of treatments for chronic obstructive pulmonary disease (COPD); and 1 trial of cardiovascular outcomes for prostate cancer therapies. The ClinicalTrials.gov NCT registration numbers of these RCTs are listed in eTable 2 in Supplement 1.

Data Sources

We used 3 US health care claims data sources for emulation of RCTs: Optum's deidentified Clinformatics Data Mart Database (2004-2019), IBM MarketScan (2003-2017), and subsets of Medicare Parts A, B, and D (2011-2017 including all patients with a diabetes or heart failure diagnosis, 2009-2017 including all patients with a dispensation for an oral anticoagulant). Each data source contained deidentified information for covered health care encounters of patients enrolled in participating health insurance plans. The data included demographics (age, sex); enrollment start and end dates; dispensed medications with dates, dose, and days of supply; procedures; and medical diagnoses with place of service and associated service dates. Death was captured with high completeness in Medicare data from the master beneficiary summary file or the vital status file. In the 2 commercial databases, out-of-hospital death was captured less completely. Because cause of death was not recorded, we substituted cardiovascular death

Table 1. Effect Estimates and Agreement Metrics

		Effect estimates (95% CI)				Agreement		
			Database study[a]					
Study No.	Trial name	RCT	Adjusted[b]	Crude[b]	Standardized difference[c]	Statistical significance	Estimate	Standardized difference
1	LEADER	0.87 (0.78 to 0.97)	0.82 (0.76 to 0.87)	0.57 (0.54 to 0.61)	0.90	SA	EA	SD
2	DECLARE-TIMI58	0.83 (0.73 to 0.95)	0.69 (0.59 to 0.81)	0.47 (0.41 to 0.53)	1.76	SA		SD
3	EMPA-REG	0.86 (0.74 to 0.99)	0.83 (0.73 to 0.95)	0.63 (0.57 to 0.70)	0.35	SA	EA	SD
4	CANVAS	0.86 (0.75 to 0.97)	0.77 (0.70 to 0.85)	0.58 (0.54 to 0.62)	1.34	SA	EA	SD
5	CARMELINA	1.02 (0.89 to 1.17)	0.90 (0.84 to 0.96)	0.90 (0.86 to 0.95)	1.61	SA	EA	SD
6	TECOS	0.98 (0.88 to 1.09)	0.89 (0.86 to 0.91)	0.81 (0.79 to 0.84)	1.71	SA	EA	SD
7	SAVOR-TIMI	1.00 (0.89 to 1.12)	0.81 (0.76 to 0.86)	0.65 (0.62 to 0.69)	3.16	SA		
8	LEAD-2	0 (−0.20 to 0.20)	0.05 (−0.11 to 0.22)	0.01 (−0.11 to 0.13)	−0.37	SA	EA	SD
9	TRITON-TIMI	0.81 (0.73 to 0.90)	0.88 (0.79 to 0.97)	0.70 (0.65 to 0.76)	−1.11	SA	EA	SD
10	PLATO	0.84 (0.77 to 0.92)	0.92 (0.83 to 1.02)	0.84 (0.78 to 0.91)	−1.31		EA	SD
11	ISAR-REACT 5	1.36 (1.09 to 1.70)	NA[d]	NA[d]	NA[d]	NA[d]		
12	ARISTOTLE	0.79 (0.66 to 0.95)	0.68 (0.61 to 0.76)	0.66 (0.62 to 0.71)	1.36	SA	EA	SD
13	RE-LY	0.66 (0.53 to 0.82)	0.73 (0.60 to 0.90)	0.67 (0.58 to 0.78)	−0.66	SA	EA	SD
14	ROCKET AF	0.79 (0.66 to 0.96)	0.70 (0.62 to 0.80)	0.76 (0.69 to 0.84)	1.00	SA	EA	SD
15	EINSTEIN DVT	0.68 (0.44 to 1.04)	0.75 (0.62 to 0.90)	0.85 (0.76 to 0.95)	−0.42	SAP	EA	SD
16	EINSTEIN PE	1.12 (0.75 to 1.68)	0.67 (0.55 to 0.80)	0.73 (0.64 to 0.83)	2.28	SAP		
17	RE-COVER II	1.08 (0.64 to 1.80)	1.15 (0.74 to 1.78)	1.48 (1.09 to 2.00)	−0.18	SA	EA	SD
18	AMPLIFY	0.84 (0.60 to 1.18)	0.81 (0.54 to 1.23)	0.64 (0.50 to 0.82)	0.13	SA	EA	SD
19	RECORD1	0.25 (0.14 to 0.47)	0.17 (0.10 to 0.29)	0.25 (0.18 to 0.34)	0.63	SA	EA	SD
20	TRANSCEND	0.92 (0.81 to 1.05)	0.88 (0.81 to 0.96)	0.80 (0.74 to 0.85)	0.55		EA	SD
21	ONTARGET	1.01 (0.94 to 1.09)	0.83 (0.77 to 0.90)	0.68 (0.64 to 0.72)	3.46	SAP		
22	HORIZON-PFT	0.59 (0.42 to 0.83)	0.72 (0.55 to 0.94)	1.08 (0.86 to 1.35)	−0.90	SA	EA	SD
23	VERO	0.44 (0.29 to 0.68)	NA[d]	NA[d]	NA[d]	NA[d]	NA[d]	NA[d]
24	DAPA-CKD	0.61 (0.51 to 0.72)	0.80 (0.52 to 1.26)	0.41 (0.29 to 0.58)	−1.10	SD		SD
25	PARADIGM-HF	0.80 (0.73 to 0.87)	1.02 (0.91 to 1.14)	0.95 (0.90 to 1.02)	−3.42			
26	P04334[e,f]	0.56 (0.44 to 0.72)	0.78 (0.62 to 0.97)	0.87 (0.76 to 0.99)	−1.95	SA	SD	SD
27	D5896	1.07 (0.70 to 1.65)	1.38 (0.90 to 2.13)	1.41 (1.00 to 1.98)	−0.81	SA	EA	SD
28	IMPACT[e,g]	0.85 (0.80 to 0.90)	1.13 (1.04 to 1.23)	1.22 (1.15 to 1.30)	−5.46			
29	POET-COPD	0.83 (0.77 to 0.90)	1.02 (0.93 to 1.12)	1.05 (0.99 to 1.12)	−3.27			
30	INSPIRE[h]	0.97 (0.84 to 1.12)	0.93 (0.90 to 0.96)	0.83 (0.81 to 0.85)	0.56	SA	EA	SD
31	CAROLINA[i]	0.98 (0.84 to 1.14)	0.91 (0.79 to 1.05)	0.92 (0.83-1.01)	0.70	SA	EA	SD
32	PRONOUNCE[i]	1.28 (0.59 to 2.79)	1.35 (0.94 to 1.93)	1.70 (1.30 to 2.21)	−0.12	SA	EA	SD

Abbreviations: EA, estimate agreement, adjusted database study point estimates falling within the 95% CI of the corresponding randomized clinical trial (RCT) result; NA, not applicable; SA, full statistical significance agreement, adjusted database study and RCT estimates and CIs on the same side of null; SAP, partial significance agreement, meets the prespecified noninferiority criteria even though the database study may have indicated superiority; SD, standardized difference agreement, SDs $|z|$ less than 1.96.

[a] Pooled across databases.

[b] Crude estimates for the database study cohorts designed to emulate trial designs do not adjust for confounding except through design. Adjusted estimates additionally adjust for confounding through propensity score matching on prespecified risk factors for the outcome that are associated with exposure.

[c] The SD calculations are available in the Methods section. These quantify the difference in effect size between the RCT and database study relative to the pooled standard deviation. Therefore, an SD of 1.00 indicates that the effect estimate from the RCT and the database study are 1 standard deviation apart. Assuming an α level of .05 and assuming that both the database and RCT results are based on large samples, the null hypothesis of no difference would be rejected whenever $|z|>1.96$.

[d] χ^2 test indicated that results were heterogeneous across databases. See database-specific results in Supplement 1.

[e] Trial had coprimary comparisons. The first listed was the primary comparison in the database study emulation protocol.

[f] Because an effect estimate was not reported, we calculated the risk ratio based on results reported for P4334 (PMID: 20678306).

[g] Because of challenges with measurement of recurrent outcomes, the estimated hazard ratio from secondary analyses of time to first occurrence of the primary outcome was used as the comparison for the database study emulation of the IMPACT trial instead of the rate ratio.

[h] Because of challenges with measurement of recurrent outcomes and low recurrence rate, the estimated rate ratio from the INSPIRE trial was approximated with a hazard ratio from the database emulation of the RCT.

[i] PRONOUNCE and CAROLINA were the 2 trials for which the trial design was emulated, protocol registered, and database study results generated before the results of the trials were made public.

in trial outcomes with all-cause death and proceeded on the assumption that after implementation of trial-specified exclusion of patients with cancer and other comorbidities, the majority of deaths would be cardiovascular related.

Emulation Process

As previously outlined,[14,15] we developed a structured process to emulate trials in a transparent and reproducible way. A common protocol template was used for each trial. Each emulation protocol was deposited on ClinicalTrials.gov after feasibility analyses but before analysis of exposure-outcome relationships (links to the NCT database emulations of RCTs are in eTable 2 in Supplement 1).

In brief, for each RCT, we took the following steps:

1. Extracted key study parameters for specifying the RCT design (PICOT).
2. Created measures using administrative claims data that were analogues to the PICOT parameters of the RCT.
3. Conducted feasibility analysis including an assessment of statistical power (requiring power at least equal to the RCT after matching) and evaluation of measurement and balance on key confounders (standardized differences <0.1) after propensity score matching. Feasibility counts were generated unstratified by exposure. Based on the feasibility analyses, we determined whether to continue with the emulation. No treatment-stratified outcome counts or inferential analyses were conducted until after the protocol was registered.
4. Documented the primary as well as secondary analyses, including analyses of control outcomes. Control outcomes with well-described null, positive, or negative associations were used as a proxy for the expected net bias due to residual confounding, measurement issues, or varying follow-up criteria.[16-18]
5. Registered at ClinicalTrials.gov, with a full protocol deposited.
6. Implemented all prespecified designs and ran all analyses.
7. Recorded and compared results.

Additional details including the ClinicalTrials.gov database studies are discussed in eMaterials 1 and eTable 2 in Supplement 1.

Design and Analysis

The selected trials were emulated with drug initiator cohort designs, and balance was sought on more than 100 preexposure characteristics using propensity score matching with a caliper of 1%. Outcome models were primarily Cox proportional hazards models with on-treatment analysis due to recognition that persistence of drug use is shorter in clinical practice whereas adherence is typically higher in RCTs. Analyses of deidentified patient-level data were conducted in each database separately and results were pooled. Additional details are included in eMaterials 1 in Supplement 1. Specific choices for design, analysis methods, control, and sensitivity analyses for each trial emulation are linked in eTable 2 in Supplement 1, including covariate and risk factor distributions before and after propensity score matching.

Database emulations of RCTs were conducted using the Aetion Evidence Platform with supplemental programming using SAS version 9.4 (SAS Institute Inc) and Cran R version 4.1.1.[19] Studies were approved by the Brigham and Women's Hospital institutional review board. Informed consent was waived because we were making secondary use of existing, deidentified data, and the study was considered minimal risk. Although US Health Insurance Portability and Accountability Act regulations do not allow sharing patient-level data, research requests to reproduce findings in our data-analytics environment will be considered.

RCT Emulation Quality

There is no perfect emulation of an actual RCT's design with secondary clinical data.[20] Challenges to emulating aspects of trial design will occur when close observational analogues to trial design elements cannot be identified. Although at a conceptual level, the RCT and the database study pairs were designed to address similar PICOT-defined research questions, in some cases, difficulty identifying observational analogues to trial design parameters may have led the 2 studies to address different operational research questions. We encountered many such elements of RCT design that were difficult to emulate with clinical practice data, including imperfect alignment of outcome measurements, shorter persistence in clinical practice, and lack of placebo in clinical practice.

Predefined Binary Agreement Metrics Between RCT and Database Study Findings

To evaluate whether the 32 database emulations of RCTs would support the same regulatory conclusions as the original RCTs, we computed 3 predefined binary metrics[14] for this research activity in addition to Pearson and intraclass correlation coefficients, calibration, and Bland-Altman plots[21]: (1) *full statistical significance agreement*, defined by estimates and CIs on the same side of the null; (2) *estimate agreement*, defined by whether estimates for the trial emulation fell within the 95% CI for the trial results; (3) *standardized difference agreement* between treatment effect estimates from trials and emulations, defined by standardized differences

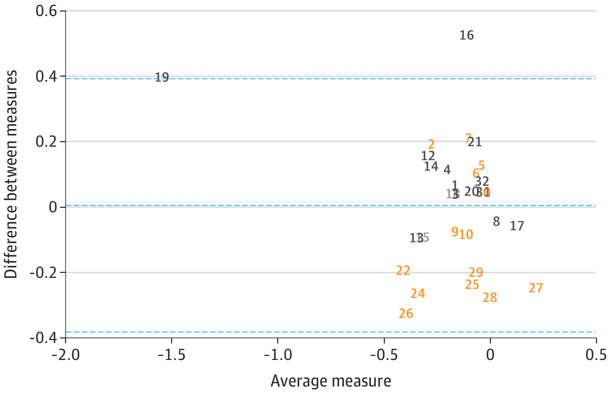
$$|z|<1.96\ (z = \hat{\Theta}_{RCT}\,{}_{-}\hat{\Theta}_{RWE}/\sqrt{\hat{\sigma}^2_{\mathrm{RCT}} + \hat{\sigma}^2_{RWE}})$$

where $\hat{\Theta}$ are the treatment effect estimates (usually log hazard ratios), the $\hat{\sigma}^2$ are associated variances. (*RWE* indicates real-world evidence.) In addition, *partial significance agreement* was defined as meeting the prespecified noninferiority criteria even though the database study may have indicated superiority.

Exploratory and Post Hoc Descriptive Analyses

The specific objective was to have a best-case comparison of results for RCT-database pairs with analogous study designs. Recognizing the range of design emulation differences encountered during the conduct of this study, we developed a post hoc composite binary indicator for how closely the trial design and measurements were emulated and use this measure in a descriptive exploration of agreement metrics (eMaterials 2 in Supplement 1). We additionally computed the Cohen κ for chance corrected agreement (eMaterials 1 in Supplement 1).

Figure 1. Bland-Altman Plot of Agreement in Randomized Clinical Trial–Database Pairs



The difference between the randomized clinical trial (RCT) and database study model coefficients for the effect estimates (eg, log hazard ratio) are plotted against the averaged value for each pair. The 3 blue dashed lines reflect the mean and 95% CIs for the difference in effect estimates for each pair. Each number represents the RCT-database pair listed in Table 1, Table 2, and Figure 2. Black indicates close emulation of the RCT design in exploratory analyses defined in Figure 2; orange, RCT-database pairs with more design emulation differences and not considered close emulations. Some numbers are colored gray for readability. ClinicalTrials.gov NCT registration numbers for RCTs and database studies are provided in eTable 2 of Supplement 1.

Results

Thirteen of 32 RCTs (41%) were superiority trials; the remainder targeted noninferiority. A variety of clinical outcomes were investigated as part of the set of trial emulations (eTable 3 in Supplement 1).

Overall, the Pearson coefficient (*r*, 0.82; 95% CI, 0.64-0.91) and intraclass correlation coefficient (0.81; 95% CI, 0.64-0.91) indicated a positive correlation between the results of RCT-database pairs results (pooled across databases). The mean difference between the coefficients for the effect estimates (eg, log hazard ratio) for the RCT-database study pairs was 0.01, with 95% CIs ranging from −0.38 to 0.39, and there were no clear trends in the Bland-Altman plot (**Figure 1**). Several points deviated from the diagonal line in a calibration plot (eFigure 1 in Supplement 1). Post hoc exploration suggested that the removal of 1 or 2 points that could be outliers in either the Bland-Altman or the calibration plot produced correlation coefficients between 0.44 and 0.86. A minority of RCT-database study pairs missed on 1 or more binary agreement metrics. Specifically, 75% met statistical significance agreement (56% full, 19% partial), 66% met estimate agreement, and 75% met standardized difference agreement (Table 1). Of 8 RCT-database study pairs that did not meet statistical significance agreement, 4 RCTs had statistically significant results at α = .05, whereas the database studies' 95% CIs included the null, 2 of the database studies had results that could not be pooled due to lack of homogeneity in results across data sources, 1 database study had statistically significant results, whereas the RCT did not (upper bound of the RCT CI = 1.05), and 1 pair had statistically significant results on the opposite sides of null. Chance-corrected agreement as measured by the Cohen κ was 0.57 (95% CI, 0.34-0.81).

In post hoc exploration of stratification by whether there was close emulation of design parameters (n = 16) or not (n = 16) for the subset of trials for which PICOT design parameters could be more closely emulated, smaller differences in effect size and higher correlation in results were observed between RCT-database study pairs (Figure 1, Pearson *r*, 0.93; 95% CI, 0.79-0.97; κ, 0.89; 95% CI, 0.69-1.00). The close exploratory subgroup had a higher proportion of results that met the binary agreement metrics than the subgroup with more design emulation differences (full or partial significance agreement, 94% vs 56%; full significance agreement, 75% vs 38%; estimate agreement, 88% vs 50%; and standardized difference agreement, 88% vs 69%). Lower correlation and agreement were observed for trials in which close observational analogues for 1 or more RCT design parameters could not be identified (Pearson *r* = 0.53; 95% CI, 0.00-0.83; κ 0.31; 95% CI, 0.04-0.59; 2 of 16 were excluded due to heterogeneity and inability to combine across databases).

A comparison of RCT and pooled database study results with associated agreement metrics for all 32 trials are shown in Table 1; database-specific and intention-to-treat results are shown in eTable 4 in Supplement 1. The 2 database studies that predicted results of ongoing trials were classified as having close emulation of design using the exploratory indicator. After the trial results were made public, the results indicated agreement between the RCT-database study pairs on all 3 prespecified binary agreement metrics.[22,23]

The 35 control outcomes were evaluated, confirming the expected result in 83% (**Table 2**). Of the 6 control outcomes that did not have expected results, 2 were for the outcome of major bleeding, for which there may be effect modification by age[24] or other characteristics[25]; for another 4 control outcomes, confounding or other biases may explain the results. Five of the 6 corresponding trial emulations nevertheless showed strong agreement.

Emulation Differences

Emulation differences and bias are summarized in **Figure 2** and eMaterials 2 in Supplement 1.

Age and Sex Distribution

We applied the same inclusion-exclusion criteria as the actual trials. This strategy mimicked the actual trial design but like 2 independently conducted trials did not guarantee identical distributions of enrolled participants. Important patient characteristics, such as age, sex, comorbidities, and preexposure medication use, often differed (eFigures 2, 3, and 4 in Supplement 1).

Comparator and Outcome Emulation Quality

Comparator emulation was ranked as good for 21 trials (66%) with active comparators and moderate for 8 trials (25%). Comparator emulation was poor for 3 trials (studies 5, 6, 7 in Table 1, Table 2, and Figure 2), for which the comparator therapy used as placebo proxy a much less costly class of established medications, and we expected residual confounding attributable to socioeconomic factors that are not captured well in claims data.

Table 2. Negative and Positive Control Outcomes to Assess the Potential for Bias[a]

Study No.	Trial name	Outcome[b]	Expected hazard ratio	Database study	Met expectation
1	LEADER	Severe hypoglycemia	<1	0.73 (0.65-0.81)	Yes
2	DECLARE	Diabetic ketoacidosis	>1	1.36 (0.78-2.37)	Yes
3	EMPA-REG	HF hospitalization	<1	0.35 (0.27-0.46)	Yes
		Diabetic ketoacidosis	>1	1.25 (0.89-1.76)	Yes
4	CANVAS	HF hospitalization	<1	0.36 (0.30-0.44)	Yes
		Diabetic ketoacidosis	>1	1.70 (1.29-2.25)	Yes
5	CARMELINA	End-stage kidney disease	≅ 1	1.04 (0.81-1.33)	Yes
6	TECOS	Severe hypoglycemia	<1	0.40 (0.38-0.43)	Yes
7	SAVOR-TIMI	Severe hypoglycemia	<1	0.37 (0.33-0.41)	Yes
8	LEAD-2				
9	TRITON-TIMI 38	Major bleeding	>1	1.17 (1.01-1.34)	Yes
		Pneumonia hospitalization	≅ 1	0.83 (0.73-0.95)	No
10	PLATO	Major bleeding	≅ 1	1.16 (0.98-1.39)	Yes
		Pneumonia hospitalization	≅ 1	1.01 (0.84-1.22)	Yes
11	ISAR-REACT 5	Major bleeding	≅ 1	1.01 (0.75-1.35)	Yes
		Pneumonia hospitalization	≅ 1	0.88 (0.60-1.27)	Yes
12	ARISTOTLE	Major bleeding	<1	0.64 (0.60-0.68)	Yes
13	RE-LY	Major bleeding	≅ 1	0.91 (0.84-0.98)	No
14	ROCKET AF	Major bleeding	≅ 1	1.17 (1.09-1.25)	No
15	EINSTEIN DVT	Major bleeding	≅ 1	1.00 (0.87-1.16)	Yes
16	EINSTEIN PE	Major or clinically relevant nonmajor bleeding	≅ 1	1.12 (0.98-1.28)	Yes
17	RE-COVER II	Major bleeding	≅ 1	1.07 (0.73-1.55)	Yes
18	AMPLIFY	Major bleeding	<1	0.75 (0.53-1.08)	Yes
19	RECORD1	Major bleeding	≅ 1	0.68 (0.40-1.17)	Yes
20	TRANSCEND				
21	ONTARGET	Angioedema	<1	0.89 (0.28-2.82)	No
22	HORIZON-PFT				
23	VERO				
24	DAPA-CKD	Genital infections	>1	2.63 (2.04-3.39)	Yes
25	PARADIGM-HF	Major bleeding	≅ 1	1.08 (0.78-1.50)	Yes
26	P04334	Pneumonia	>1	0.89 (0.50-1.59)	No
27	D5896	Pneumonia	≅ 1	1.25 (0.85-1.83)	Yes
28	IMPACT	Pneumonia	≅ 1	1.09 (0.90-1.32)	Yes
29	POET-COPD	Pneumonia	≅ 1	1.02 (0.84-1.23)	Yes
30	INSPIRE	Pneumonia	>1	1.18 (1.10-1.26)	Yes
31	CAROLINA	Severe hypoglycemia	<1	0.42 (0.32-0.56)	Yes
		End-stage kidney disease	≅ 1	1.08 (0.66-1.79)	Yes
32	PRONOUNCE				

[a] Blank cells indicate that no control outcomes were evaluated.

[b] Control outcomes with well-described null, positive, or negative associations with the compared therapies were identified from peer-reviewed literature or from secondary analyses in the evaluated randomized clinical trial. The database result for the control outcome was considered to have met expectation if the point estimate and CIs were generally consistent with the expected result (hazard ratio, <1, >1, ≅ 1).

Outcome emulation was ranked as good for 19 trials (59%) and moderate for 13 (41%) for which the outcome algorithms had lower specificity or had substantial missing data. Operational definitions of comparator and outcome emulation quality are available in eMaterials 2 in Supplement 1. Event rates in the database studies were mostly lower than in the emulated RCTs (eTable 5 in Supplement 1).

Placebo Control

Ten of the trials (31%) involved a placebo comparator. We emulated placebo groups with new use of an active comparator that was strongly expected to have no effect on the outcome of interest.

Initiate Therapy in the Hospital

Three trials (9%) involved initiation of therapy while patients were in the hospital (studies 9, 10, 11 listed in Table 1, Table 2, and Figure 2), involving head-to-head comparisons of antiplatelet agents after acute coronary syndrome. Given that claims data do not record medication use in hospitals, the index date and follow-up were defined by the day of first drug dispensation after discharge. The separation in cumulative incidence plots for the RCTs indicated that the effect of therapy likely began early, while many participants were still hospitalized.[26-28] In the database studies emulating RCTs, the effect size was closer to the null because it only included patients who survived the index hospitalization.

Figure 2. Emulation Challenges

No.	Trial name	Comparator emulation[a]	Outcome emulation[b]	Age distribution, mean difference, y	Sex distribution, difference in % female	Run-in window[c]	Placebo control	In-hospital start of medication	Dose titration during follow-up	Discontinuation of maintenance therapy at randomization	Delayed effect[d]	Close emulation[e]
1	LEADER	Moderate	Good	-3.4	-17.8	Yes, placebo	Yes	No	Yes	No	No	Yes
2	DECLARE	Moderate	Moderate	1.4	-4.9	Yes, placebo	Yes	No	No	No	No	No
3	EMPA-REG	Moderate	Good	1.2	-11.9	Yes, placebo	Yes	No	No	No	No	Yes
4	CANVAS	Moderate	Good	-2.0	-10.5	Yes, placebo	Yes	No	No	No	No	Yes
5	CARMELINA	Poor	Good	-6.4	-16.2	No	Yes	No	No	No	No	No
6	TECOS	Poor	Moderate	-6.8	-18.1	No	Yes	No	No	No	No	No
7	SAVOR-TIMI	Poor	Good	-3.8	-13.7	No	Yes	No	No	No	No	No
8	LEAD-2	Good	Moderate	-2.0	-6.0	Yes, both groups	No	No	Yes	No	No	Yes
9	TRITON-TIMI 38	Good	Good	3.4[f]	4.9	No	No	Yes	Yes	No	No	No
10	PLATO	Good	Good	-3.3[f]	-4.1	No	No	Yes	Yes	No	No	No
11	ISAR-REACT 5	Good	Good	5.6	0.9	No	No	Yes	Yes	No	No	No
12	ARISTOTLE	Good	Good	-6.1	-16.7	No	No	No	No	Yes[g]	No	Yes
13	RE-LY	Good	Good	-4.7	-5.9	No	No	No	No	Yes[g]	No	Yes
14	ROCKET-AF	Good	Good	-4.5	-14.9	No	No	No	No	Yes[g]	No	Yes
15	EINSTEIN DVT	Good	Moderate	-14.7	-17.0	No	No	No	Yes	No	No	Yes
16	EINSTEIN PE	Good	Moderate	-8.2	-4.9	No	No	No	Yes	No	No	Yes
17	RE-COVER II	Good	Moderate	-13.5	-16.4	No	No	No	No	No	No	Yes
18	AMPLIFY	Good	Moderate	-0.6	-10.1	No	No	No	Yes	No	No	Yes
19	RECORD1	Good	Good	1.0	1.6	No	No	No	No	No	No	Yes
20	TRANSCEND	Moderate	Good	-4.0	-14.1	Yes, both groups	Yes	No	No	No	No	Yes
21	ON TARGET	Good	Good	-2.4	-27.2	Yes, both groups	No	No	Yes	No	No	Yes
22	HORIZON PFT	Moderate	Good	-1.0	0	No	Yes	No	No	No	Yes	No
23	VERO	Good	Moderate	1.1	0	No	No	No	No	No	Yes	No
24	DAPA-CKD	Moderate	Moderate	-5.5	-11.4	No	Yes	No	No	No	Yes	No
25	PARADIGM-HF	Moderate	Moderate	-4.7	-6.2	Yes, both groups	No	No	No	Yes	No	No
26	P04334	Good	Good	-11.2	1.9	Yes, 1 class	No	No	No	Yes	No	No
27	D5896	Good	Good	-3.3	-1.8	No	No	No	No	Yes	No	No
28	IMPACT	Good	Good	-4.0	-25.5	Yes, baseline prescription	No	No	No	Yes	No	No
29	POET-COPD	Good	Good	-7.5	-28.3	Yes, mixed	No	No	No	Yes	No	No
30	INSPIRE[h]	Good	Moderate	-1.5	-44.6	Yes, 1 class	No	No	No	Yes	No	No
31	CAROLINA[i]	Good	Good	-6.3	-12.3	Yes, placebo	No	No	Yes	No	No	Yes
32	PRONOUNCE1[i]	Good	Good	-3.0	0	No	No	No	Yes	No	No	Yes

[a] Good indicates that the trial had an active comparator; moderate, the placebo emulated by the drug was expected to be unrelated to the outcome and the cohort characteristics were well balanced or the active comparator had to be modified for feasibility reasons; and poor, the placebo emulated by the drug expected to be unrelated to the outcome and expectation of residual confounding from characteristics poorly measured in claims (eg, socioeconomic status).

[b] Good indicates the outcome emulation was assessed with high specificity; moderate, lower outcome specificity or high missingness.

[c] Placebo, indicates that the run-in was for only the placebo group; both groups, the run-in was for both exposure and comparator in sequence or both groups were run-in on a drug that was neither exposure nor comparator; baseline drugs, run-in for baseline maintenance therapy; 1 class, 1 class of therapy used as either exposure or comparator; and mixed, a mix of therapies according to protocol algorithm. For these trials, the baseline, 1 class, or mixed types of run-ins were expected to selectively include responders to run-in therapy.

[d] A crude measure assessed based on appearance of violation of proportional hazards in published trial figures.

[e] Binary composite indicator was based on the emulation markers listed in this figure (Supplement 1).

[f] Difference in median.

[g] Trials included a postrandomization washout window, therefore did not mix the effects of randomization with discontinuation of baseline therapy.

[h] Trial had coprimary comparisons. The first listed was the primary comparison in the real-world evidence emulation protocol.

[i] Trial was ongoing at the start of the emulation and analysis.

Closer emulation of RCT design in the database studies is indicated by blue; emulation means that none of the following characteristics were present and the comparator and outcome emulation were at least moderate with 1 or both classified as good: (1) start of follow-up in the hospital (hospital prescription data were not available in claims but may be available in linked data); (2) run-in type that selectively included responders to 1 treatment group; (3) mixing effect of randomization and discontinuation of baseline maintenance therapy; and (4) delayed effect. Orange indicates RCT-database study pairs with more substantial design emulation differences.

Dose Titration During Follow-up

Eleven of 32 trials (34%) were designed with a loading dose or involved dose titration. Using clinical practice data, few patients met the specified titration schedules; we therefore focused on comparing new initiators and assumed that physicians' doses of medications followed best practice.

Run-in Window

Thirteen trials (41%) included some form of run-in phase requiring stable standard of care, tolerance of the study drug, or discontinuation of maintenance medication, which could not be emulated in clinical practice data. For 2 trials investigating asthma and COPD treatment (studies 26 and 30 in Table 1, Table 2, and Figure 2), the RCTs selectively included participants who responded well to the run-in treatment, which was one of the randomized treatment groups producing results more favorable toward the treatment group.[29] In contrast, responders or nonresponders before cohort entry cannot be differentiated in clinical practice data.

Discontinuation of Maintenance Therapy at Randomization

Nine trials (28%) required that participants discontinue baseline maintenance therapy at the time of randomization. Three COPD and 2 asthma trials included participants who were receiving maintenance therapy (studies 26-30 in Table 1, Table 2, and Figure 2) but had to discontinue that therapy at randomization. Discontinuation of maintenance therapies may cause short-term increases in the outcomes of interest[30-32] in the trials.

Delayed Effect Over a Long Follow-up

Three trials (9%; studies 22-24 in Table 1, Table 2, and Figure 2) had cumulative incidence curves showing delayed or time-varying effects. Due to lower treatment adherence in clinical practice compared with trials with extensive procedures to maximize adherence, the median time patients were followed up for on-treatment analyses was substantially shorter in the database studies emulating these trials (4-18 months vs 24-36 months).

Confounding, Replicability, Robustness, and Other Issues

Case studies (eMaterials 3 and 4 in Supplement 1) provide an in-depth look at the different research questions that are addressed in 2 RCT-database pairs that were not classified as close emulations. In addition to divergence in results due to emulation differences, confounding and other issues are likely to have played a role in observed divergence.

Discussion

In this emulation of 32 highly selected RCTs using nonrandomized health care claims databases, we evaluated agreement in treatment effect estimates for RCT-database study pairs across a range of indications. In any trial emulation, incomplete emulation due to differences in PICOT-defined research question, potential bias, and random error can contribute to observed disagreement.[33] The relative contribution of each is difficult to disentangle. Similar issues affect RCT-RCT pairs. Prior studies have systematically identified reanalyses of RCTs, with 35% disagreement from the original,[34] and meta-analyses of RCTs have noted multiple clinical topics in which at least 2 trials observed divergent results.[35,36] In these studies, differences in the details of the PICOT-defined questions were identified as drivers of divergence in RCT results. As with database studies, chance, mixed with other factors, may contribute to disagreement in RCT results. This is demonstrated by sister trials with virtually identical designs but discordant results.[37-40]

Although there was modest agreement overall for 16 trials that could be emulated closely in terms of their design using a post hoc exploratory indicator, higher agreement was observed between the RCT-database study pairs. In the remaining 16 studies, there were more substantial differences in observational analogues to the trial design for multiple reasons, including patient selection during run-in phases and treatment patterns counter to clinical practice. This led to weaker agreement in findings. The differences in treatment effect sizes could have occurred because the database study targeted a different study question, due to residual bias or due to chance.

To reduce potential bias from results-driven design choices, all attributes of the database studies were predefined and protocols were registered at ClinicalTrials.gov before inferential drug-outcome analyses started.[13] Given the aim of this study, which was to independently emulate RCT design rather than replicate the RCT population, post hoc population modeling to make population characteristics distributions similar beyond applying the same inclusion criteria was not performed. Countering the potential criticism that knowing the RCT result would enable the investigators to tailor the database study design toward the expected finding, we have started predicting results of 7 ongoing phase 4 trials. Two of those trials were subsequently completed and showed close alignment in findings.

This report includes all findings from predefined primary analyses across 32 RCT emulations. Numerous follow-on sensitivity analyses are ongoing, including an evaluation of methods such as reweighting to align distributions of RCT and database study patient characteristics,[41] double negative control[42] methods to address residual bias from poor placebo proxies,[43] different approaches for handling follow-up and intercurrent events,[44-46] and meta-regression to measure the performance of alternative analytic choices against the benchmark set by RCTs.

Bias is the major concern of decision-makers and is often quoted to explain differences between RCTs and database studies. This project demonstrated the ability of database studies to come to similar conclusions as RCTs when RCT design is closely emulated.

In post hoc analyses, results of database studies were closely concordant with those of RCTs when the trial's design and measurements could be closely emulated; however, database studies are not a substitute for RCTs. RCTs remain the standard for evidence generation on the efficacy of medical products for good reason. However, database studies can provide valuable complementary evidence by answering important questions on treatment effects in clinical practice that are not answered by RCTs (case studies 1 and 2 in eMaterials 3 and 4 in Supplement 1).

Database studies can address questions in cases when, for the lack of incentives, RCTs are unlikely to be conducted or completed expeditiously, such as evaluating the effect of combining 2 drugs marketed by different manufacturers or studying older, younger, or more diverse populations.

Limitations

Our project has several limitations. First, we assumed that the findings from a single RCT were internally valid, which is not guaranteed. Second, apparent agreement between RCT and database study results could occur if the effects of multiple factors (chance, emulation differences, bias) cancel each other out. Third, the results of 32 database studies that emulate RCTs may have limited generalizability due to the multistep selection process and feasibility requirements. Fourth, although our goal with this project was to calibrate database study results against RCTs, in practice, many highly controlled trials cannot be emulated with database studies and many questions of interest may never have trials to calibrate against. The principles applied in this project remain fundamental for the interpretation of database study results; namely, to specify a hypothetical trial that would answer the study question and to assess robustness through thoughtful sensitivity analyses.

Conclusions

In conclusion, we observed similar findings between highly selected, nonrepresentative RCTs and nonrandomized database studies. In the absence of RCT evidence, database studies can complement RCT evidence to enhance understanding of how medications work in clinical practice.

ARTICLE INFORMATION

Accepted for Publication: March 4, 2023.

The RCT-DUPLICATE Initiative Authors: Jessica M. Franklin, PhD; Rishi J. Desai, PhD; William Feldman, MD, DPhil; Elizabeth M. Garry, PhD, MPH; Robert J. Glynn, PhD, ScD; Kueiyu Joshua Lin, MD, ScD; Julie Paik, MD, ScD; Elisabetta Patorno, MD, DrPH; Samy Suissa, PhD; Elvira D'Andrea, MD, MPH, PhD; Dureshahwar Jawaid, MPH; Hemin Lee, MD, MPH; Ajinkya Pawar, PhD; Sushama Kattinakere Sreedhara, MBBS, MSPH; Helen Tesfaye, PharmD, ScM; Lily G. Bessette, MS; Luke Zabotka, BA; Su Been Lee, BA; Nileesa Gautam, MS; Cassie York, BS; Heidi Zakoul, BA; John Concato, MD, MS, MPH; David Martin, MD, MPH; Dianne Paraoan, MD, MPH; Kenneth Quinto, MD, MPH.

Affiliations of The RCT-DUPLICATE Initiative Authors: Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts (Franklin, Desai, Feldman, Glynn, Lin, Paik, Patorno, D'Andrea, Jawaid, H. Lee, Pawar, Sreedhara, Tesfaye, Bessette, Zabotka, S. B. Lee, Gautam, York, Zakoul); Now with Optum, Boston, Massachusetts (Franklin); Aetion, Inc, New York, New York (Garry); McGill University, Montreal, Quebec, Canada (Suissa); Now with AbbVie Inc, Washington, DC (D'Andrea); Office of Medical Policy, US Food and Drug Administration, Silver Springs, Maryland (Concato, Martin, Paraoan, Quinto); Now with Moderna, Cambridge, Massachusetts (Martin).

Author Contributions: Dr Wang had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Wang, Schneeweiss, Franklin, Martin, Paraoan, Quinto.
Acquisition, analysis, or interpretation of data: Wang, Schneeweiss, Franklin, Desai, Feldman, Garry, Glynn, Lin, Paik, Patorno, Suissa, D'Andrea, Jawaid, H. Lee, Pawar, Kattinakere Sreedhara, Tesfaye, Bessette, Zabotka, S. Lee, Gautam, York, Zakoul, Concato, Martin.
Drafting of the manuscript: Wang, Schneeweiss.
Critical revision of the manuscript for important intellectual content: Wang, Schneeweiss, Franklin, Desai, Feldman, Garry, Glynn, Lin, Paik, Patorno, Suissa, D'Andrea, Jawaid, H. Lee, Pawar, Kattinakere Sreedhara, Tesfaye, Gautam, Concato, Martin, Paraoan, Quinto.
Statistical analysis: Wang, Schneeweiss, Garry, Glynn, D'Andrea, Jawaid, H. Lee, Pawar, Kattinakere Sreedhara, Tesfaye, Bessette.
Obtained funding: Schneeweiss, Franklin, Wang, Martin.
Administrative, technical, or material support: Wang, Schneeweiss, H. Lee, D'Andrea, Jawaid, H. Lee, Pawar, Kattinakere Sreedhara, Tesfaye, Bessette, Zabotka, S. Lee, Gautam, York, Zakoul, Concato, Paraoan, Quinto.
Supervision: Wang, Franklin, Schneeweiss, Desai, Feldman, Garry, Lin, Paik, Patorno, Suissa, D'Andrea, Jawaid, H. Lee, Pawar, Kattinakere Sreedhara, Tesfaye, Bessette, Concato, Martin.

Conflict of Interest Disclosures: Dr Schneeweiss reported serving as a consultant for Aetion Inc and receiving grants from UCB Pharma and Boehringer Ingelheim. Dr Desai reported receiving grants from Bayer, Novartis, and Vertex. Dr Feldman reported receiving personal fees from Alosa Health, Aetion, and Blue Cross Blue Shield of Massachusetts and serving as an expert witness in litigation against inhaler manufacturers. Dr Glynn reported receiving grants from Amarin, Kowa, Novartis, and Pfizer. Dr Patorno reported receiving grants from the National Institute on Aging, the Patient Centered Outcomes Research Institute, and an investigator grant from Boehringer-Ingelheim to Brigham and Women's Hospital. Dr Suissa reported serving on the advisory boards of AstraZeneca, Atara, Bristol Myers Squibb, Novartis, and Panalgo and receiving speakers fees from Boehringer-Ingelheim and Novartis and consulting fees from Merck, Pfizer, and Seqirus. Dr Gautam reported owning equity in Aetion. No other disclosures were reported.

Funding/Support: This study was funded by contracts HHSF223201710186C and HHSF223201810146C from the US Food and Drug Administration to the Brigham and Women's Hospital (Drs Schneeweiss and Wang). Drs Wang and Schneeweiss were further supported by grants R01HL141505, R01AG053302, and R01AR080194 from the National Institutes of Health.

Role of the Funder/Sponsor: Neither the FDA nor the NIH had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the FDA.

Data Sharing Statement: See Supplement 2.

Additional Contributions: We thank our external advisor group for important input and review of study protocols: Wayne Ray, PhD, (Vanderbilt University), Steve Goodman, MD, MHS, PhD (Stanford University), and Alan Brookhart, PhD (Duke University). We thank FDA colleagues Mark Levenson, PhD, and Robert Temple, MD, for input on the design of the RCT DUPLICATE initiative; Jacqueline Corrigan-Curay, JD, MD, Jenny Li, PhD, and Marie Bradley, PhD, (all from the FDA) for thoughtful input on the interpretation of results; and members of multiple FDA review divisions for helpful comments on trial emulations. In addition, we wish to thank Danielle Isaman, MS, Liza Gibbs, MS, and Katherine Gilpin, MPH (Aetion Inc) for implementation support using the Aetion Evidence Platform.

REFERENCES

1. Friedman LM, Furberg C, DeMets DL. *Fundamentals of Clinical Trials*. Mosby Year Book; 1996.

2. US Food and Drug Administration. *Framework for FDA's Real World Evidence Program*. December 2018. Accessed January 31, 2019. https://www.fda.gov/media/120060/download

3. Eichler H-G, Pignatti F, Schwarzer-Daum B, et al. Randomized controlled trials versus real world evidence: neither magic nor myth. *Clin Pharmacol Ther*. 2021;109(5):1212-1218. doi:10.1002/cpt.2083

4. Suissa S. Reduced mortality with sodium-glucose cotransporter-2 inhibitors in observational studies: avoiding immortal time bias. *Circulation*. 2018;137(14):1432-1434. doi:10.1161/CIRCULATIONAHA.117.032799

5. Retraction—Mehra MR, Desai SS, Ruschitzka F, Patel AN. Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet*. 2020;395(10240):1820. doi:10.1016/S0140-6736(20)31324-6

6. Chan KA, Andrade SE, Boles M, et al. Inhibitors of hydroxymethylglutaryl-coenzyme A reductase and risk of fracture among older women. *Lancet*. 2000;355(9222):2185-2188. doi:10.1016/S0140-6736(00)02400-4

7. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*. 2000;342(25):1887-1892. doi:10.1056/NEJM200006223422507

8. Forbes SP, Dahabreh IJ. Benchmarking observational analyses against randomized trials: a review of studies assessing propensity score methods. *J Gen Intern Med*. 2020;35(5):1396-1404. doi:10.1007/s11606-020-05713-5

9. Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? a systematic comparison of studies on acute coronary syndromes. *Eur Heart J*. 2012;33(15):1893-1901. doi:10.1093/eurheartj/ehs114

10. 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*. 2016;352:i493. doi:10.1136/bmj.i493

11. Haynes RB. *Clinical Epidemiology: How to Do Clinical Practice Research*. Lippincott Williams & Wilkins; 2012.

12. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clin Pharmacol Ther*. 2017;102(6):924-933. doi:10.1002/cpt.857

13. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin Pharmacol Ther*. 2019;105(4):867-877. doi:10.1002/cpt.1351

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*. 2020;107(4):817-826. doi:10.1002/cpt.1633

15. 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(10):1002-1013. doi:10.1161/CIRCULATIONAHA.120.051718

16. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21(3):383-388. doi:10.1097/EDE.0b013e3181d61eeb

17. Arnold BF, Ercumen A. Negative control outcomes: a tool to detect bias in randomized trials. *JAMA*. 2016;316(24):2597-2598. doi:10.1001/jama.2016.17700

18. Desai JR, Hyde CL, Kabadi S, et al. Utilization of positive and negative controls to examine comorbid associations in observational database studies. *Med Care*. 2017;55(3):244-251. doi:10.1097/MLR.0000000000000640

19. Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, Bartels DB. Transparency and reproducibility of observational cohort studies using large healthcare databases. *Clin Pharmacol Ther*. 2016;99(3):325-332. doi:10.1002/cpt.329

20. Lodi S, Phillips A, Lundgren J, et al; INSIGHT START Study Group and the HIV-CAUSAL Collaboration. Effect estimates in randomized trials and observational studies: comparing apples with apples. *Am J Epidemiol*. 2019;188(8):1569-1577. doi:10.1093/aje/kwz100

21. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135-160. doi:10.1177/096228029900800204

22. Lopes RD, Higano CS, Slovin SF, et al; PRONOUNCE Study Investigators. Cardiovascular safety of degarelix versus leuprolide in patients with prostate cancer: the primary results of the PRONOUNCE randomized trial. *Circulation*. 2021;144(16):1295-1307. doi:10.1161/CIRCULATIONAHA.121.056810

23. Rosenstock J, Kahn SE, Johansen OE, et al; CAROLINA Investigators. Effect of linagliptin vs glimepiride on major adverse cardiovascular outcomes in patients with type 2 diabetes: the CAROLINA randomized clinical trial. *JAMA*. 2019;322(12):1155-1166. doi:10.1001/jama.2019.13772

24. Cavallari I, Patti G. Efficacy and safety of oral anticoagulation in elderly patients with atrial fibrillation. *Anatol J Cardiol*. 2018;19(1):67-71. doi:10.14744/AnatolJCardiol.2017.8256

25. Jones WS, Hellkamp AS, Halperin J, et al. Efficacy and safety of rivaroxaban compared with warfarin in patients with peripheral artery disease and non-valvular atrial fibrillation: insights from ROCKET AF. *Eur Heart J*. 2014;35(4):242-249. doi:10.1093/eurheartj/eht492

26. Wiviott SD, Braunwald E, McCabe CH, et al; TRITON-TIMI 38 Investigators. Prasugrel versus clopidogrel in patients with acute coronary syndromes. *N Engl J Med*. 2007;357(20):2001-2015. doi:10.1056/NEJMoa0706482

27. Wallentin L, Becker RC, Budaj A, et al; PLATO Investigators. Ticagrelor versus clopidogrel in patients with acute coronary syndromes. *N Engl J Med*. 2009;361(11):1045-1057. doi:10.1056/NEJMoa0904327

28. Schüpke S, Neumann F-J, Menichelli M, et al; ISAR-REACT 5 Trial Investigators. Ticagrelor or prasugrel in patients with acute coronary syndromes. *N Engl J Med*. 2019;381(16):1524-1534. doi:10.1056/NEJMoa1908973

29. Suissa S. Run-in bias in randomised trials: the case of COPD medications. *Eur Respir J*. 2017;49(6):1700361. doi:10.1183/13993003.00361-2017

30. Suissa S, Drazen JM. Making sense of triple inhaled therapy for COPD. *N Engl J Med*. 2018;378(18):1723-1724. doi:10.1056/NEJMe1716802

31. Suissa S, Ariel A. US Food and Drug Administration-mandated trials of long-acting β-agonists safety in asthma: will we know the answer? *Chest*. 2013;143(5):1208-1213. doi:10.1378/chest.12-2881

32. Suissa S, Ariel A. Triple therapy trials in COPD: a precision medicine opportunity. *Eur Respir J*. 2018;52(6):1801848. doi:10.1183/13993003.01848-2018

33. 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*. 2020;107(4):735-737. doi:10.1002/cpt.1793

34. Ebrahim S, Sohani ZN, Montoya L, et al. Reanalyses of randomized clinical trial data. *JAMA*. 2014;312(10):1024-1032. doi:10.1001/jama.2014.9646

35. Jane-wit D, Horwitz RI, Concato J. Variation in results from randomized, controlled trials: stochastic or systematic? *J Clin Epidemiol*. 2010;63(1):56-63. doi:10.1016/j.jclinepi.2009.02.010

36. Horwitz RI. Complexity and contradiction in clinical trial research. *Am J Med*. 1987;82(3):498-510. doi:10.1016/0002-9343(87)90450-5

37. Büller HR, Prins MH, Lensin AW, et al; EINSTEIN–PE Investigators. Oral rivaroxaban for the treatment of symptomatic pulmonary embolism. *N Engl J Med*. 2012;366(14):1287-1297. doi:10.1056/NEJMoa1113572

38. The Einstein Investigators. Oral Rivaroxaban for symptomatic venous thromboembolism. *N Engl J Med*. 2010;363(26):2499-2510. doi:10.1056/NEJMoa1007903

39. De Soyza A, Aksamit T, Bandel T-J, et al. RESPIRE 1: a phase III placebo-controlled randomised trial of ciprofloxacin dry powder for inhalation in non-cystic fibrosis bronchiectasis. *Eur Respir J*. 2018;51(1):1702052. doi:10.1183/13993003.02052-2017

40. Aksamit T, De Soyza A, Bandel T-J, et al. RESPIRE 2: a phase III placebo-controlled randomised trial of ciprofloxacin dry powder for inhalation in non-cystic fibrosis bronchiectasis. *Eur Respir J*. 2018;51(1):1702053. doi:10.1183/13993003.02053-2017

41. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. *Stat Med*. 2020;39(14):1999-2014. doi:10.1002/sim.8426

42. Shi X, Miao W, Nelson JC, Tchetgen EJT. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *J R Stat Soc Series B Stat Methodol*. 2020;82(2):521-540. doi:10.1111/rssb.12361

43. Huitfeldt A, Hernan MA, Kalager M, Robins JM. Comparative effectiveness research using observational data: active comparators to emulate target trials with inactive comparators. *EGEMS (Wash DC)*. 2016;4(1):1234. doi:10.13063/2327-9214.1234

44. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med*. 2017;377(14):1391-1398. doi:10.1056/NEJMsm1605385

45. Hernán MA, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. *Clin Trials*. 2012;9(1):48-55. doi:10.1177/1740774511420743

46. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R). International Council on Harmonization. November 20, 2019. Accessed March 29, 2022. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf