bmjmedicine

Check for updates

Design differences and variation in results between randomised trials and non-randomised emulations: meta-analysis of RCT-DUPLICATE data

Rachel Heyard [1], Leonhard Held [1], Sebastian Schneeweiss [2], Shirley V Wang [2]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjmed-2023-000709).

[1]Center for Reproducible Science, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland
[2]Division of Pharmacoepidemiology, Brigham and Womens Hospital Harvard Medical School, Boston, Massachusetts, USA
Correspondence to: Dr Rachel Heyard, Center for Reproducible Science, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland; rachel.heyard@uzh.ch

Cite this as: *BMJMED* 2024;3:e000709. doi:10.1136/bmjmed-2023-000709

Received: 17 July 2023
Accepted: 27 December 2023

ABSTRACT

OBJECTIVE To explore how design emulation and population differences relate to variation in results between randomised controlled trials (RCT) and non-randomised real world evidence (RWE) studies, based on the RCT-DUPLICATE initiative (Randomised, Controlled Trials Duplicated Using Prospective Longitudinal Insurance Claims: Applying Techniques of Epidemiology).

DESIGN Meta-analysis of RCT-DUPLICATE data.

DATA SOURCES Trials included in RCT-DUPLICATE, a demonstration project that emulated 32 randomised controlled trials using three real world data sources: Optum Clinformatics Data Mart, 2004-19; IBM MarketScan, 2003-17; and subsets of Medicare parts A, B, and D, 2009-17.

ELIGIBILITY CRITERIA FOR SELECTING STUDIES Trials where the primary analysis resulted in a hazard ratio; 29 RCT-RWE study pairs from RCT-DUPLICATE.

RESULTS Differences and variation in effect sizes between the results from randomised controlled trials and real world evidence studies were investigated. Most of the heterogeneity in effect estimates between the RCT-RWE study pairs in this sample could be explained by three emulation differences in the meta-regression model: treatment started in hospital (which does not appear in health insurance claims data), discontinuation of some baseline treatments at randomisation (which would have been an unusual care decision in clinical practice), and delayed onset of drug effects (which would be under-reported in real world clinical practice because of the relatively short persistence of the treatment). Adding the three emulation differences to the meta-regression reduced heterogeneity from 1.9 to almost 1 (absence of heterogeneity).

CONCLUSIONS This analysis suggests that a substantial proportion of the observed variation between results from randomised controlled trials and real world evidence studies can be attributed to differences in design emulation.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Real world evidence studies can complement randomised controlled trials by providing insights on the effectiveness of a medical treatment in clinical practice

⇒ Concerns about confounding have limited the use of real world evidence studies in clinical practice and policy decisions

WHAT THIS STUDY ADDS

⇒ This study suggests that heterogeneity among pairs of randomised controlled trials and their non-randomised emulations can be explained by differences in design emulation

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE, OR POLICY

⇒ These results could inform researchers and clinicians on the degree to which apparent divergence in results between randomised controlled trials and real world evidence studies can be driven by differences in study design and the research question

Introduction

Real world evidence (RWE) has been defined as evidence on the effects of medical products that are derived from the analysis of real world data, which includes different sources of patient health data, particularly data collected as part of routine clinical practice, including electronic health records and insurance claims data.[1] Interest in the use of real world evidence from real world data to support clinical practice and policy decisions has been increasing.[2–5] Concerns remain, however, about the validity of this evidence compared with the traditional randomised controlled trial (RCT).[5–7]

These concerns come from a misleading dichotomy that sets randomised controlled trials against database studies instead of viewing them as providing complementary information that informs a better understanding of the effects of drugs.[8] Results from databases and randomised controlled trials have been compared, and some have found high concordance, supporting the ability of well designed database studies to generate valid causal conclusions.[9–13] Others have used observed differences in results to criticise database studies as intractably confounded.[7 14–17]

The RCT-DUPLICATE initiative (Randomised Controlled Trials Duplicated Using Prospective Longitudinal Insurance Claims: Applying Techniques of Epidemiology) is one effort comparing randomised controlled trials with database studies.[10 18–20] RCT-DUPLICATE set out to emulate 32 trials by prospectively designing a series of insurance claims database studies to match each design of the randomised controlled trial as closely as possible within the confines and limitations of using data that were not collected for research purposes. Because of the nature of using routinely collected data from clinical practice, some elements of the trial design

Table 1 | Categorical emulation differences with possible levels, reference category, and description. All characteristics are binary

Characteristics	Levels	Reference category	Description
Comparator emulation	Good, moderate-poor	Moderate-poor	Good: randomised controlled trial had active comparator; moderate: placebo was emulated with treatment unrelated to outcome; poor: placebo was emulated with treatment potentially related to outcome
Outcome emulation	Good, moderate	Moderate	Good: outcome assessed with high specificity; moderate: outcome assessed with low specificity and high number of missing data
Run-in period to one treatment arm	Yes, no	No	Yes: randomised controlled trial included a run-in phase that selectively included responders to one treatment arm before randomisation
Placebo control	Yes, no	No	Yes: randomised controlled trial involved placebo comparator which was emulated with active comparator
Treatment started in hospital	Yes, no	No	Yes: treatment started in hospital which cannot be captured in real world data
Dose titration during follow-up	Yes, no	No	Yes: randomised controlled trial designed with a dose titration during follow-up
Discontinuation of maintenance treatment without washout	Yes, no	No	Yes: randomised controlled trial required participants to discontinue some baseline treatments without allowing for a washout period after randomisation
Delayed effect	Yes, no	No	Yes: treatment had a delayed effect possibly causing lower adherence in clinical practice recorded in real world data

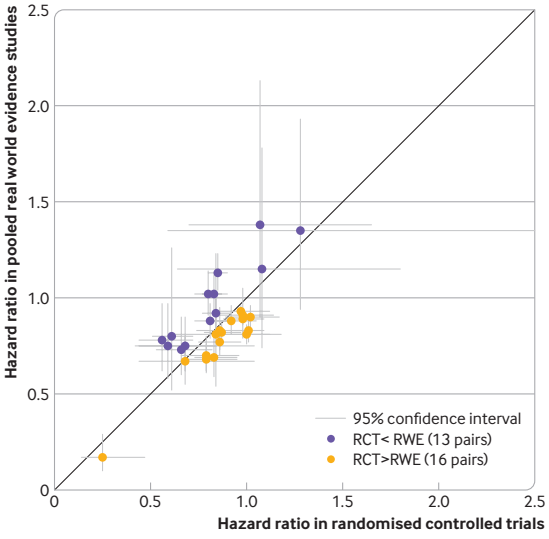

Figure 1 | Hazard ratios (95% confidence intervals) estimated in randomised controlled trials and real world evidence studies (pooled for all data sources). Diagonal line represents perfect emulation; all trials with points on the right side of the diagonal have an effect size estimated in the randomised controlled trial (RCT) that is larger than the effect size estimated in the pooled real world evidence study (RWE)

could not be exactly emulated (eg, measures to ensure prolonged adherence over long follow-up periods). These emulation differences can be summarised as differences in outcome measurements, demographics of the included patients, treatment implementation in clinical practice, and lack of placebo in clinical practice. Design emulation and population differences change the question or estimand being looked at in the randomised controlled trial compared with the database study.[21 22]

Our aim was to use the RCT-DUPLICATE collection of emulated trials to assess how design emulation and population differences relate to variation in results between randomised controlled trials and real world evidence database studies that were designed to emulate them. We explored whether the characteristics of design emulation and population differences can reduce and therefore explain the residual heterogeneity in differences in effect size in a meta-regression analysis.

Methods

Our analysis was exploratory rather than confirmatory, meaning that the data used for the analysis

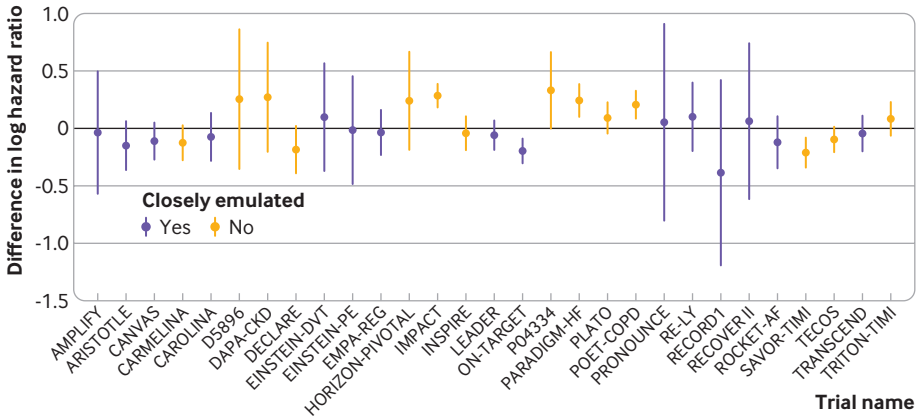

Figure 2 | Difference in effect estimates (log hazard ratio with 95% confidence interval) between the randomised controlled trials and pooled real world evidence studies, depending on whether the study was closely emulated or not closely emulated. Horizontal line represents no difference between real world evidence and randomised controlled trial estimates (upplementary table E.2 shows more information on each study)

were collected for another purpose. The conclusions drawn from our analysis might therefore help to formulate hypotheses to be tested in a subsequent confirmatory study. Our aim was to better understand emulation differences and how this affects variation in results between RCT-RWE study pairs.

RCT-DUPLICATE

The selection process for the RCT-DUPLICATE initiative is described in detail elsewhere.[18 23] In summary, the RCT-DUPLICATE consortium emulated 32 randomised controlled trials that were relevant to regulatory decision making and were potentially feasible to emulate based on insurance claims data because key study parameters, such as the primary outcome, treatment strategies, and inclusion and exclusion criteria were measurable. The selected trials included a mix of superiority and non-inferiority trials, trials with large and small effect sizes, and a mix of trials with active comparators and placebo added to active standard of care treatments. The consortium used three real world data sources to implement the database studies that emulated the randomised controlled trials: Optum Clinformatics Data Mart, 2004-19; IBM MarketScan, 2003-17; and subsets of Medicare parts A, B, and D (data from 2011 to 2017 including all patients with a diagnosis of diabetes or heart failure, and data from 2009 to 2017 including all patients who had been prescribed an oral anticoagulant). Whenever possible, the emulations of the randomised controlled trials were implemented in more than one of the data sources with a while on-treatment analysis (chosen because of the shorter duration of drug use in clinical practice whereas adherence to treatment is generally longer in randomised controlled trials) and the final analyses were based on estimates resulting from a fixed effects meta-analysis of the implementations in all databases.

In this study, only trials where the primary analysis resulted in a hazard ratio were used. The LEAD2 trial with continuous outcome was excluded. For two trials (ISAR-REACT5 and VERO) a χ^2 test indicated that the results were heterogeneous across databases so that the meta-analysis could not be performed to obtain a pooled real world evidence estimate for the hazard ratio[19] and these trials were also excluded. Online supplemental file, section A has a summary of the 29 trials included in the analysis. We evaluated hazard ratios that were adjusted for confounding by 1:1 nearest neighbour propensity score matching on prespecified risk factors (chosen in discussion with clinical experts), as described in Franklin et al,[19] for the RCT-RWE comparisons.

Design emulation and population differences identified in RCT-DUPLICATE

Emulation differences were recorded as covariates in RCT-DUPLICATE. Differences in age and sex distributions were captured as numerical variables representing the difference in mean age or percentage of women (the value in the

Table 2 | Model intercept and coefficient values (with 95% confidence intervals), and heterogeneity between real world evidence studies and randomised controlled trials, depending on model used. Heterogeneity close to 1 represents homogeneous effect size differences between study pairs

Model	Intercept (95% CI)	Coefficient (95% CI)	Heterogeneity
Simple	0.002 (−0.061 to 0.066)	—	1.905
Adjusted for close emulation	0.061 (−0.011 to 0.134)	−0.159 (−0.278 to −0.040)	1.725

CI=confidence interval.

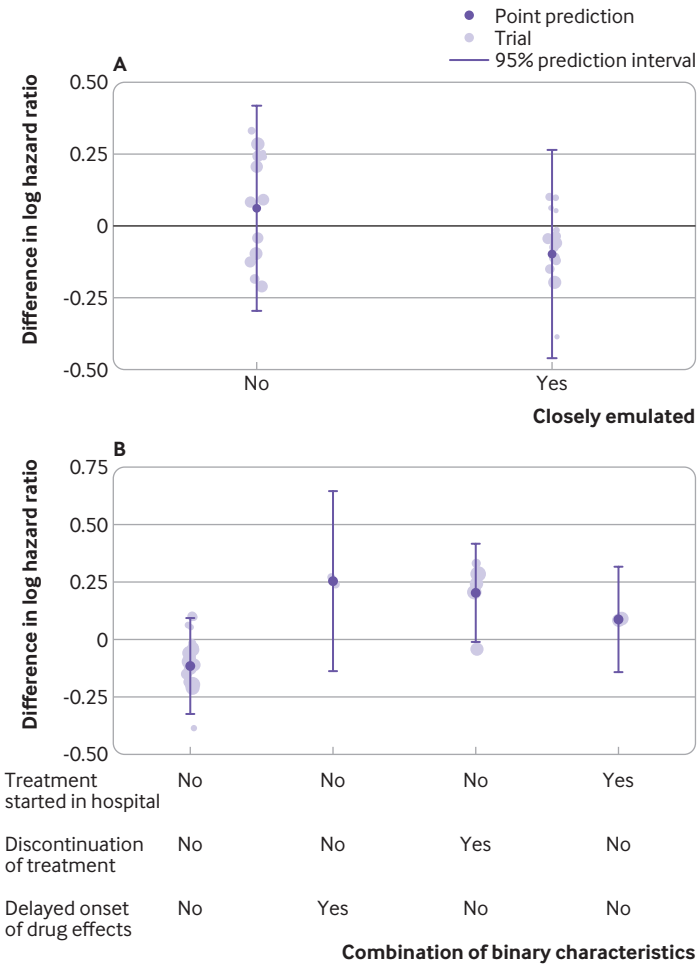

Figure 3 | Bubble plots showing associations of the log hazard ratio between the randomised controlled trial and pooled real world evidence with (top graph) whether study pairs were closely emulated (yes or no), and with (bottom graph) possible combinations of three binary characteristics (treatment started in hospital, discontinuation of maintenance treatment without washout, and delayed onset of effects of drugs). The larger the bubble, the more precise the estimate or trial. Horizontal jitter has been applied on the bubbles to enhance visibility. 95% prediction intervals are compiled by the meta-regression, including the binary composite covariate

randomised controlled trial minus the value in the real world evidence pooled emulation). Table 1 shows the categorical emulation difference characteristics by reference category recorded in RCT-DUPLICATE. [18]

All characteristics in table 1 were summarised as a binary composite covariate, indicating if the RCT-RWE study pair was closely emulated or not closely emulated. More specifically, a study pair was considered closely emulated if the comparator and outcome emulations were at least moderate, and at least one of them was good, and if none of the following was true: follow-up started in hospital; run-in window that selectively included responders to one treatment arm; effects of randomisation and discontinuation of baseline treatment were mixed; and delayed effect over a long period of follow-up. The composite indicator was defined as part of the post hoc explorations by the RCT-DUPLICATE team[18] to evaluate concordance in the results for randomised controlled trial-database pairs, with closer versus less close emulation of the design and research question based on the randomised controlled trial PICOT (population, intervention comparator, outcome, time).

Statistical analysis

All statistical analyses required that the effect estimates from randomised controlled trials and real world evidence were approximately normally distributed. Hence log transformations were applied on hazard ratios. The standardised differences in the RCT-RWE study pairs were computed by dividing the difference in log hazard ratios by the standard error of the difference. The squared standardised difference is the Q statistic which was used to perform the Q test for heterogeneity between the randomised controlled trials and real world evidence studies.[24 25] The sum of all computed Q statistics was used as an overall test for heterogeneity between the RCT-RWE study pairs included in RCT-DUPLICATE.

Heterogeneity can be quantified as a multiplicative parameter,[26] which is an overdispersion parameter generally larger than one, inflating the model's standard errors. As described by Mawdsley et al,[27] multiplicative heterogeneity is estimated by fitting a weighted linear regression on the observed differences from all RCT-RWE study pairs against a constant, with weights defined as the inverse of the squared standard error of the differences. Multiplicative heterogeneity is then simply this model's standard error, and absence of heterogeneity is achieved if the parameter is equal to 1. The heterogeneity parameter is set to its lower bound of 1 if estimated to be <1.

Characteristics describing emulation differences are used to explain heterogeneity. With meta-regression methods (chapter 7 of *Handbook of meta-analysis*[28]), the characteristics of the emulation differences (ie, differences in age and sex distributions as well as each of the binary characteristics summarised in table 1 and the composite indicator) are added to the weighted linear regression models estimating multiplicative heterogeneity. If the extracted residual heterogeneity from the more complex, adjusted model is smaller than the heterogeneity measured with the simple model (with only a constant), part of the variation can be explained by the set of included emulation differences.

To reduce the complexity of the meta-regression, avoid overfitting, and choose only the most predictive of the p candidate characteristics, leave-one-out cross validated mean squared errors[29] were computed for all 2^p possible candidate models. Many of the included characteristics were suspected to be dependent. The simplest model, with a mean squared error of at most one standard error from the smallest mean squared error across all models, was selected.[30] The model coefficients for the included characteristics have to be interpreted with respect to the model's intercept, the difference in RCT-RWE effect estimates that remains when all binary characteristics of emulation differences and the centred continuous characteristics are set to their reference or zero, respectively. Online supplemental file, section B gives a detailed description of the statistical analyses. All analyses were performed in R version 4.3.2.[31] Code and data to reproduce the analyses and recompile this manuscript are available from https://gitlab.com/heyardr/hte-in-rwe and from Heyard and Wang.[32]

Patient and public involvement

As a reanalysis of publicly available data, no patients or members of the public were involved in the conception, development, analysis, interpretation, or reporting of the results of our study. There are no plans to disseminate the study findings to patient and public communities.

Results

Figure 1 shows the estimated hazard ratios from the randomised controlled trials against the hazard ratios estimated with the pooled real world evidence studies (with 95% confidence intervals). Estimates from perfectly emulated trials would scatter around the diagonal line. Although more than half of the pooled estimates from the real world evidence studies tended to be smaller than the estimates for the randomised controlled trials, many were also larger. This finding is different from the results seen in the large scale replication projects where the effect size estimated in the replication study was generally smaller than in the original study, which might be attributable to publication bias or other questionable research practices, unlikely operating in this study.[33] This phenomenon is referred to as shrinkage of effect size.[34] Also, an overall test of heterogeneity suggested strong evidence of variation between all study pairs in RCT-DUPLICATE (online supplemental figure C.1).

To better understand the variability in results in RCT-DUPLICATE, variation was quantified and its sources were investigated. Figure 2 represents the differences in log hazard ratio for each study pair depending on whether the study was closely emulated or not closely emulated. Trials categorised as not closely emulated based on the indicator tended towards positive differences. The average difference in log hazard ratio over all included trials was estimated to be slightly negative (−0.015, 95% confidence interval −0.084 to 0.054), suggesting that, on average, the hazard ratio estimated with the real world data was larger than in the randomised controlled trial.

Table 2 shows the estimated multiplicative heterogeneity comparing the pooled real world evidence studies with the randomised controlled trials, together with the model intercept and coefficient values (with 95% confidence intervals). The simple model refers to the weighted regression with only a constant whereas the second model is a meta-regression adjusted for the binary characteristic, close emulation. Including close emulation in the weighted linear regression model reduced heterogeneity from 1.905 to 1.725, indicating that part of the observed variation between estimates in RCT-RWE study pairs can be attributed to the composite covariate. Although the intercept of the simple model was close to zero, the intercept of the adjusted model (difference in log hazard ratio for trials that were not closely emulated) tended to be positive. Closely emulated trials had, on average, slightly negative differences (figure 3).

We explored the use of a set of explanatory characteristics instead of the composite covariate, close emulation. Table 3 shows the univariate coefficients, respective model intercept, and residual heterogeneity. Some of the characteristics reduced heterogeneity more than others; for example, adding the

Table 3 | Univariate coefficients (with 95% confidence intervals) for each candidate characteristic, ordered by increasing heterogeneity. For each row (each characteristic) a separate model was fitted, resulting in separate intercept and residual heterogeneity. The closer residual heterogeneity is to 1, the more the characteristic explains part of the variations. Residual heterogeneity and R^2 values were added to further explain the proportion of variation for each covariate

Characteristics	Intercept (95% CI)	Univariate coefficient (95% CI)	Residual heterogeneity	R^2
Discontinuation of maintenance treatment without washout (yes)	−0.079 (−0.129 to −0.029)	0.282 (0.189 to 0.3754)	1.260	0.578
Run-in period to one treatment arm (yes)	−0.056 (−0.114 to 0.002)	0.249 (0.130 to 0.368)	1.511	0.393
Placebo control (yes)	0.073 (0 to 0.146)	−0.172 (−0.286 to −0.059)	1.675	0.254
Comparator emulation (good)	−0.061 (−0.150 to 0.029)	0.117 (−0.005 to 0.239)	1.819	0.120
Dose titration during follow-up (yes)	0.024 (−0.053 to 0.100)	−0.070 (−0.208 to 0.068)	1.904	0.037
Difference in mean age (centred)	0.009 (−0.057 to 0.074)	−0.010 (−0.032 to 0.012)	1.909	0.031
Treatment started in hospital (yes)	−0.008 (−0.076 to 0.060)	0.095 (−0.111 to 0.301)	1.909	0.031
Delayed effect (yes)	0 (-0.065 to 0.064)	0.254 (−0.370 to 0.878)	1.916	0.024
Difference in percentage of women (centred)	0.005 (-0.066 to 0.076)	0.001 (−0.005 to 0.007)	1.938	0.002
Outcome emulation (good)	−0.003 (-0.137 to 0.131)	0.007 (−0.146 to 0.159)	1.939	0.000

characteristic, discontinuation of maintenance treatment without washout, gave the largest decrease in heterogeneity, from 1.905 to 1.260. The intercept in table 3 can be interpreted as the difference in log hazard ratio for the respective reference category of the binary characteristics or no difference in the distribution for the two continuous characteristics, age and percentage of women. Then all possible candidate models (2^{10}=1024), depending on which of the 10 characteristics are included, were fitted and the models' leave-one-out mean squared errors were computed. The final model was the simplest model with leave-one-out mean squared errors smaller than the minimum mean squared errors plus one standard error (online supplemental figure D.2). With this tuning parameter, three characteristics would be included. Table 4 shows the coefficient estimates of the models with the best performance for each number of included characteristics. The models summarised in table 4 resulted in the model performance and heterogeneity illustrated in online supplemental figure D.2.

The best model with three design emulation differences includes delayed onset of effect of drugs, discontinuation of maintenance treatment without washout, and treatment started in hospital. This model's residual heterogeneity was 1.003. Figure 3 shows the association between the combination of these finally selected characteristics and outcome (difference in log hazard ratio). Only the prediction intervals for the combinations with observations are displayed; for example, none of the trials in RCT-DUPLICATE had more than one of the three emulation differences set to yes. The three included characteristics were mutually exclusive and together were better in reducing observed heterogeneity than close emulation. Hence the remaining characteristics only added noise to the indicator for close emulation, or cancelled each other out.

Discussion

Principal findings

Based on data from the RCT-DUPLICATE initiative, comparing results from RCT-RWE study pairs, we found that the study emulation characteristics delayed effect of treatment, discontinuation of treatment during run-in period, and treatment started in hospital explained most of the observed variation beyond chance in this sample. In this collection of RCT-RWE study pairs, most of the observed variation in effect estimates could be explained by these three emulation characteristics. The results suggest that, on average, the hazard ratios estimated with real world data tended to be slightly larger than the hazard ratios estimated in the randomised controlled trials.

Surprisingly little variation was explained by placebo comparator, which was thought to be an emulation challenge, in the absence of placebo in clinical practice, and a source of confounding bias. This result might have been influenced by the quality of the placebo proxy that was used in emulation of placebo controlled trials for RCT-DUPLICATE. Although all of the included studies focused on a hazard ratio for the primary result, the proposed analysis can be applied to studies investigating other outcome measures (ie, risk ratios or risk differences). The meta-regression analyses, however, required that the estimates for all studies were on the same scale. Appropriate transformations could be applied to include studies whose primary analyses used a different scale.

Table 4 | Model selection. Coefficient estimates for the best model with respect to leave-one-out mean squared errors for each number of characteristics included

				Characteristics of design emulation and population differences									
No of characteristics	Residual heterogeneity	LOO MSE	Intercept	Comparator emulation (good)	Outcome emulation (good)	Difference in mean age (centred)	Difference in % of women (centred)	Run-in period to one treatment arm (yes)	Placebo control (yes)	Treatment started in hospital (yes)	Dose titration during follow-up (yes)	Discontinuation of maintenance treatment without washout (yes)	Delayed effect of treatment (yes)
0	1.905	0.033	0.002	—	—	—	—	—	—	—	—	—	—
1	1.260	0.024	−0.079	—	—	—	—	—	—	—	—	0.282	—
2	1.218	0.016	−0.084	—	—	—	—	—	—	—	—	0.287	0.338
3	1.003	0.015	−0.115	—	—	—	—	—	—	0.203	—	0.318	0.369
4	1.000	0.012	−0.11	—	—	−0.011	—	—	—	0.245	—	0.309	0.38
5	1.000	0.011	−0.087	—	—	−0.01	0.005	—	—	0.152	—	0.357	0.318
6	1.000	0.011	−0.098	—	—	−0.011	0.006	—	—	0.116	0.042	0.375	0.326
7	1.000	0.012	−0.119	—	—	−0.011	0.006	—	0.024	0.127	0.054	0.395	0.324
8	1.000	0.013	−0.16	—	0.051	−0.011	0.006	—	0.03	0.127	0.046	0.402	0.331
9	1.000	0.014	−0.112	−0.025	0.029	−0.011	0.008	0.127	—	0.089	0.06	0.308	0.313
10	1.000	0.017	−0.1	−0.037	0.029	−0.011	0.008	0.138	−0.012	0.089	0.059	0.296	0.313

LOO MSE=leave-one-out mean squared errors.

Heterogeneity and LOO MSE for the same models are represented in online supplemental figure D.2.

Randomised controlled trials are seen as the standard in establishing the efficacy of medical products, but these studies might not be free of flaws in their implementation and might not always represent clinical practice. The results of multiple clinical trials that look at similar questions, even identical twin trials, can vary in their findings.[35–39] Discordant results between randomised controlled trials and real world evidence studies that investigate similar use of drugs and outcomes should not necessarily discredit the real world evidence study before considering emulation differences that might result in assessing a slightly different causal question. Therefore, the emphasis should be on understanding where these differences come from, and the clinical or research question that is being asked by each study type.

Limitations of this study

Our study had some limitations. We have presented the results of an exploratory analysis with a limited sample size from 29 RCT-RWE study pairs, non-randomly selected from the RCT-DUPLICATE initiative. Therefore, we could only include a limited number of explanatory emulation characteristics in our models. Other emulation differences could further reduce residual heterogeneity. A follow-up study designed for purpose could derive and investigate other emulation characteristics that might be informative in the meta-regression.

The trials included in RCT-DUPLICATE were selected as having a high probability of being feasible to emulate with insurance claims data. Therefore, our results provide an understanding of how concordance in results between randomised controlled trials and database studies are influenced by concordance in design, but the specific coefficients should not be interpreted as generalisable because of the highly selected sample of trials. Also, the design emulation and population differences recorded in this study might not be a comprehensive list of all of the important emulation challenges that could be considered. Different emulation differences might be more or less relevant for different clinical areas, and the direction of the effect of these differences are context dependent, limiting the generalisability of our empirical findings. Furthermore, the emulations were conducted with insurance claims data. Emulated randomised controlled trials from registry data or data from electronic health records might have other design emulation and population differences (eg, challenges to defining observable time when data from fragmented healthcare systems are used).

Conclusion

Overall, our study showed that a substantial proportion of heterogeneity between the results of randomised controlled trials and real world evidence studies can be attributed to differences in design emulation. Furthermore, our study showed how meta-regression can be used to define a more nuanced understanding of emulation differences.

Contributors All authors were involved in the design of the study. SW and SS provided access to the data. RH and LH developed the methodology. RH performed the statistical analyses with input from LH. All authors contributed to the interpretation of the results. RH and SW wrote the first draft of the manuscript. All authors approved the final and revised versions of the manuscript. RH and SW are the guarantors of this work. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Transparency statement: The lead author (RH) affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Funding This study was funded by contracts from the US Food and Drug Administration (HHSF223201710186C and HHSF223201810146C) to the Brigham and Women's Hospital (PI SS and SW). SS and SW were also supported by funding from the National Institutes of Health (RO1HL141505, R01AG053302, and R01AR080194). The content is solely the responsibility of the authors and does not necessarily represent the official views of the US Food and Drug Administration. The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from the US Food and Drug Administration and the National Institutes of Health for the submitted work; SS is principal investigator of the FDA Sentinel Innovation Center, funded by the FDA, and co-principal investigator of an investigator initiated grant to the Brigham and Women's Hospital from Boehringer Ingelheim and UCB Pharma unrelated to the topic of this study; SS is a consultant for Aetion, a software manufacturer of which he owns equity; SS interests were declared, reviewed, and approved by the Brigham and Women's Hospital and MGB HealthCare System in accordance with their institutional compliance policies; SW has been a consultant for Veracity Healthcare Analytics, Exponent, and MITRE, a federally funded research and development centre for Centers for Medicare and Medicaid Services, for unrelated work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Code and data to reproduce the analyses and recompile this manuscript are available from https://gitlab.com/heyardr/hte-in-rwe.

ORCID iDs

Rachel Heyard http://orcid.org/0000-0002-7531-4333

Leonhard Held http://orcid.org/0000-0002-8686-5325
Sebastian Schneeweiss http://orcid.org/0000-0003-2575-467X
Shirley V Wang http://orcid.org/0000-0001-7761-7090

REFERENCES

1 US Food and Drug Administration. Framework for FDA's real-world evidence program. 2018. Available: https://www.fda.gov/media/120060/download
2 Eichler H, Baird L, Barker R, *et al*. From adaptive licensing to adaptive pathways: delivering a flexible life-span approach to bring new drugs to patients. Clin Pharmacol Ther 2015;97:234–46. 10.1002/cpt.59 Available: https://ascpt.onlinelibrary.wiley.com/toc/15326535/97/3
3 Ball R, Robb M, Anderson S, *et al*. The FDAs sentinel initiative-a comprehensive approach to medical product surveillance. Clin Pharmacol Ther 2016;99:265–8. 10.1002/cpt.320 Available: https://ascpt.onlinelibrary.wiley.com/toc/15326535/99/3
4 Sun X, Tan J, Tang L, *et al*. Real world evidence: experience and lessons from China. BMJ 2018;360:j5262. 10.1136/bmj.j5262
5 Makady A, Ham RT, de Boer A, *et al*. Policies for use of real-world data in health technology assessment (HTA): A comparative study of six HTA agencies. Value Health 2017;20:520–32. 10.1016/j.jval.2016.12.003
6 Hampson G, Towse A, Dreitlein WB, *et al*. Real-world evidence for coverage decisions: opportunities and challenges. J Comp Eff Res 2018;7:1133–43. 10.2217/cer-2018-0066
7 Collins R, Bowman L, Landray M, *et al*. The magic of randomization versus the myth of real-world evidence. N Engl J Med 2020;382:674–8. 10.1056/NEJMsb1901642
8 Eichler H, Pignatti F, Schwarzer-Daum B, *et al*. Randomized controlled trials versus real world evidence: neither magic nor myth. Clin Pharmacol Ther 2021;109:1212–8. 10.1002/cpt.2083 Available: https://ascpt.onlinelibrary.wiley.com/toc/15326535/109/5
9 Crown W, Dahabreh IJ, Li X, *et al*. Can observational analyses of routinely collected data emulate randomized trials? Design and feasibility of the observational patient evidence for regulatory approval science and understanding disease project. Value in Health 2023;26:176–84. 10.1016/j.jval.2022.07.003
10 Franklin JM, Patorno E, Desai RJ, *et al*. Emulating randomized clinical trials with nonrandomized real-world evidence studies. Circulation 2021;143:1002–13. 10.1161/CIRCULATIONAHA.120.051718
11 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878–86. 10.1056/NEJM200006223422506
12 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–92. 10.1056/NEJM200006223422507
13 Moneer O, Daly G, Skydel JJ, *et al*. Agreement of treatment effects from observational studies and randomized controlled trials evaluating hydroxychloroquine, lopinavir-ritonavir, or dexamethasone for COVID-19: meta-epidemiological study. BMJ 2022;377:e069400. 10.1136/bmj-2021-069400
14 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. 10.1136/bmj.i493
15 Matthews AA, Szummer K, Dahabreh IJ, *et al*. Comparing effect estimates in randomized trials and observational studies from the same population: an application to percutaneous coronary intervention. J Am Heart Assoc 2021;10:e020357. 10.1161/JAHA.120.020357
16 Kumar A, Guss ZD, Courtney PT, *et al*. Evaluation of the use of cancer registry data for comparative effectiveness research. JAMA Netw Open 2020;3:e2011985. 10.1001/jamanetworkopen.2020.11985
17 Soni PD, Hartman HE, Dess RT, *et al*. Comparison of population-based observational studies with randomized trials in oncology. J Clin Oncol 2019;37:1209–16. 10.1200/JCO.18.01074
18 Wang SV, Schneeweiss S, RCT-DUPLICATE Initiative, *et al*. Emulation of randomized clinical trials with nonrandomized database analyses. JAMA 2023;329:1376–85. 10.1001/jama.2023.4221
19 Franklin JM, Glynn RJ, Martin D, *et al*. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. Clin Pharmacol Ther 2019;105:867–77. 10.1002/cpt.1351 Available: https://ascpt.onlinelibrary.wiley.com/toc/15326535/105/4
20 Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials Clin Pharmacol Ther 2017;102:924–33. 10.1002/cpt.857 Available: https://ascpt.onlinelibrary.wiley.com/toc/15326535/102/6
21 Franklin JM, Glynn RJ, Suissa S, *et al*. Emulation differences vs. biases when calibrating real-world evidence findings against randomized controlled trials. Clin Pharmacol Ther 2020;107:735–7. 10.1002/cpt.1793
22 Lodi S, Phillips A, Lundgren J, *et al*. Effect estimates in randomized trials and observational studies: comparing apples with apples. Am J Epidemiol 2019;188:1569–77. 10.1093/aje/kwz100
23 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;107:817–26. 10.1002/cpt.1633 Available: https://ascpt.onlinelibrary.wiley.com/toc/15326535/107/4
24 Cochran WG. The combination of estimates from different experiments. Biometrics 1954;10:101. 10.2307/3001666
25 Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58. 10.1002/sim.1186
26 Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: A comparison of methods. Stat Med 1999;18:2693–708. 10.1002/(sici)1097-0258(19991030)18:20<2693::aid-sim235>3.0.co;2-v
27 Mawdsley D, Higgins JPT, Sutton AJ, *et al*. Accounting for heterogeneity in meta-analysis using a multiplicative model-an empirical study. Res Synth Methods 2017;8:43–52. 10.1002/jrsm.1216
28 Schmid CH, Stijnen T, White IR, *et al*. Handbook of meta-analysis. In: Handbook of meta-analysis. First edition. | Boca Raton : Taylor and Francis, [2020] | Series: Chapman & Hall/CRC handbooks of modern statistical methods: Chapman; Hall/CRC, 2020. 10.1201/9781315119403
29 Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303–25. 10.1002/sim.4780091109
30 Hastie T, Tibshirani R, Wainwright M. *Statistical learning with sparsity*. Statistical learning with sparsity: The lasso and generalizations. CRC Press, 2015. 10.1201/b18401
31 R Core Team. R: A language and environment for statistical computing. In: R foundation for statistical computing 2022. Vienna, Austria, Available: https://www.R-project.org/
32 Heyard.R, Wang SV. Design differences explain variation in results between randomized trials and their non-randomized emulations. *Zenodo* 2024. doi:10.5281/zenodo.10451851
33 Open science collaboration. Estimating the reproducibility of psychological science. Science 2015;349. 10.1126/science.aac4716
34 Held L, Micheloud C, Pawel S. The assessment of replication success based on relative effect size. Ann Appl Stat 2022;16. 10.1214/21-AOAS1502
35 The EINSTEIN investigators. Oral rivaroxaban for symptomatic venous thromboembolism. N Engl J Med 2010;363:2499–510. 10.1056/NEJMoa1007903
36 The EINSTEIN investigators. Oral rivaroxaban for the treatment of symptomatic pulmonary embolism. N Engl J Med 2012;366:1287–97. 10.1056/NEJMoa1113572
37 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;51:1702052. 10.1183/13993003.02052-2017
38 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. 10.1183/13993003.02053-2017
39 Tampi RR, Forester BP, Agronin M. Aducanumab: evidence from clinical trial data and controversies. Drugs Context 2021;10:1–9. 10.7573/dic.2021-7-3

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjmed-2023-000709).