BMC Medical Research Methodology

RESEARCH

Open Access

Assessing the replicability of RCTs in RWE emulations

Jeanette Köppe[1*†], Charlotte Micheloud[2†], Stella Erdmann[3†], Rachel Heyard[2] and Leonhard Held[2]

Abstract

Background The standard regulatory approach to assess replication success is the two-trials rule, requiring both the original and the replication study to be significant with effect estimates in the same direction. The sceptical *p*-value was recently presented as an alternative method for the statistical assessment of the replicability of study results.

Methods We review the statistical properties of the sceptical *p*-value and compare those to the two-trials rule. We extend the methodology to non-inferiority trials and describe how to invert the sceptical *p*-value to obtain confidence intervals. We illustrate the performance of the different methods using real-world evidence emulations of randomized controlled trials (RCTs) conducted within the RCT DUPLICATE initiative.

Results The sceptical *p*-value depends not only on the two *p*-values, but also on sample size and effect size of the two studies. It can be calibrated to have the same Type-I error rate as the two-trials rule, but has larger power to detect an existing effect. In the application to the results from the RCT DUPLICATE initiative, the sceptical *p*-value leads to qualitatively similar results than the two-trials rule, but tends to show more evidence for treatment effects compared to the two-trials rule.

Conclusion The sceptical *p*-value represents a valid statistical measure to assess the replicability of study results and is useful in the context of real-world evidence emulations.

Keywords Randomized clinical trial, Real-world evidence, Replication, Sceptical *p*-value, Two-trials rule

Background

Randomized controlled trials (RCTs) represent the gold standard for proofing the efficacy of a new therapy [1–3]. To reduce bias and guarantee a high internal validity of the results, an RCT is standardized to the highest degree and has strict in- and exclusion criteria. However, older, multi-morbid patients, women (particularly those lactating, pregnant, or able to become pregnant) or non-consenting patients are often excluded from RCTs [4, 5] and as a result, a broad generalizability is sometimes missing [6, 7]. Thus, some vulnerable patient groups are underrepresented in RCTs, resulting in a lack of evidence for treatment decisions.

Real-world evidence (RWE) plays an increasing role in research and for regulatory decision making, particularly due to an increasing number of potential data sources [8–10]. RWE is given as the clinical evidence based on so-called real-world data (RWD), i.e. data outside of RCTs given as data relating to patient health status or data routinely collected from different sources [11]. With the increase in purely virtual data sources (such as

†Jeanette Köppe, Charlotte Micheloud and Stella Erdmann contributed equally to this work.

*Correspondence:
Jeanette Köppe
jeanette.koeppe@ukmuenster.de
[1] Institute of Biostatistics and Clinical Research, University of Muenster, Münster, Germany
[2] Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zürich, Switzerland
[3] Institute of Medical Biometry, University of Heidelberg, Heidelberg, Germany

"digital twins" or "Metaverse©"), the term "RWD" is more and more criticized, as all study data come from the "real" world, leading to an increased use of the term "routine practice data" [12]. For the sake of clarity, however, the term RWD will be used as a synonym for routine practice data in the following and includes all type of data outside of RCTs.

Studies based on RWD can be used for post-market analysis of treatments in actual real-world care to ensure patient safety [6, 7, 9], especially for research questions or patients groups that are difficult to address otherwise [9, 13–15]. In the literature known as efficacy-effectiveness gap, some studies reported a low performance of drugs and treatment in daily clinical practice [7, 16], but systematic analyses failed to proof a general significant difference between the effect estimate in observational studies and in RCTs [3, 17–20]. It is therefore of interest to determine if and when RCTs and well-designed RWE studies, closely aligned with the original RCT, reach the same conclusions. The RCT DUPLICATE initiative (Randomized, Controlled Trials Duplicated Using Prospective Longitudinal Insurance Claims: Applying Techniques of Epidemiology) consists of the emulation of 32 RCTs with RWD to check if and under which circumstances RWD studies explicitly designed to emulate RCTs are useful to draw the same causal conclusions [11, 21]. A standardized and transparent study protocol for each RWE study was developed, with pre-defined patient selection and definition of primary analyses and study measures. Moreover, all emulation studies were a-priori registered before analyses were done [11, 21]. All emulations were based on data of up to three different US claims data and propensity score matching with 120 pre-defined possible confounders was performed [11, 21].

It is common to distinguish direct and conceptual replications [22]. A direct replication is an attempt to reproduce a previously observed result based on new data, collected with a protocol as close as possible to the original study. Only the study sample size can be different and is often larger in replication studies to ensure sufficient power to confirm the original finding. In a conceptual replication, the aim is to show the robustness of a finding with a different study sample or study conditions. An emulation of an RCT is a conceptual replication, as it tries to confirm the result from an RCT with observational RWD [23]. The possible problem of confounding through the lack of randomisation is usually addressed with statistical adjustments such as propensity score matching [21].

Whether direct or conceptual, the evaluation of the replicability of two studies (e.g. RCT vs replication given by the RWE emulation) is not straightforward and different statistical methods are available [24, 25]. The regulatory agreement (also known as the two-trials rule, see e.g. Held, [26]) is a commonly used method to assess replicability of the two studies [11, 25]. Regulatory agreement is fulfilled if the original effect and replication effect go in the same direction and are both significant. However, the pure assessment of the significance of the original trial and the replication study (two-trials rule – TTR) has the disadvantage that it does not directly take into account the effect sizes of the two studies [27].

An alternative approach is to perform a meta-analysis of the RCT and RWE results. The two effect sizes are combined into an overall effect size and then significance is assessed. However, it is important to note that studies entering a meta-analysis are assumed to be interchangeable, which is sometimes a questionable assumption, since the studies were often conducted under different standards. In the case of the replication of an RCT with RWE in particular, this assumption is not fulfilled.

Another approach – the so-called "sceptical" p-value – was recently developed [24, 27, 28] and represents an alternative to the two-trials rule. It depends on the p-values of both studies and also on the variance ratio of both effect sizes. The original trial and the replication study are thus no longer interchangeable, a property which is particularly important with regards to the emulation of RCTs using RWD.

The aim of this paper is to investigate how the sceptical p-value performs in this context. To do so, the sceptical p-value will be used to evaluate the agreement of RWE emulation of RCTs given from the RCT DUPLICATE initiative [11, 21] and the results will be compared to the two-trials rule. First, we describe the data from the RCT DUPLICATE initiative in the "Data source: RCT DUPLICATE initiative" section. The "Methods" section reviews the sceptical p-value and extends it to non-inferiority studies. Type-I error control and power for replication studies are discussed and confidence intervals for the combined effect are derived. In the "Results" section, the sceptical p-value is used to evaluate the replicability of emulations of RCTs given by the data set of the RCT DUPLICATE initiative [11, 21] and compared to the two-trials rule. The results are discussed in the "Discussion" section; especially in comparison with the conclusions from the original RCT DUPLICATE initiative.

Data source: RCT DUPLICATE initiative

For the study at hand, data of 32 RCTs and related RWE emulations from the RCT DUPLICATE initiative were available [11, 21]. Detailed information about the RCT DUPLICATE initiative can be found in the literature [11, 21, 29, 30]. In short, published and ongoing RCTs were selected if the trials were relevant to regulatory decision

making and were potentially replicable using RWD [11, 29]. RCTs were considered to be potentially replicable if they fulfilled critical aspects concerning study protocol, primary outcome, inclusion and exclusion criteria. RCTs were chosen to get a mix of superiority and non-inferiority trials with large and small magnitudes of effect sizes and trials with active comparators as well as trials with active treatment and placebo added to active standard of care therapies [29]. All RCTs were required to be sufficiently large and well powered to guarantee reliable results.

For all emulation studies, the data of up to three different US claims data-sources were used to emulate the related RCTs: *Optum Clinformatics* (from 01/2004 to 03/2019), *IBM Market Scan* (from 2003 to 2017) and a subset of *Medicare Part 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). Information about other insurance holders are not available for this database. As a result, 13 of the 32 RCTs were not emulated on the Medicare dataset (TRITON-TIMI, PLATO, ISAR-REACT5, TRANSCEND, ON-TARGET, HORIZON-PIVOTAL, VERO, P04334, D5896, IMPACT, POET-COPD and INSPIRE). The development of all emulation studies was standardized and transparent, with prospectively defined patient selection, primary analyses and registration [11, 21]. Within a pre-defined time period, all available patient information in the claims data were included to the studies, with inclusion and exclusion criteria as closely as possible adopted from original RCTs. Moreover, a 1:1 nearest-neighbor propensity score matching with caliper width of 0.01 to control for more than 120 possible confounders was performed for all replication studies. Confounders were selected *a priori* and measured during 6 months before drug initiation. Detailed information about confounder selection can be found in Franklin et al. [11].

For the analyses presented here, the pooled estimates from fixed effects meta-analysis of up to three different data sources were used for all emulation studies. These pooled estimates have also been analyzed in Heyard et al. [30] with focus on differences between RCT and RWD effect size estimates due to design and population differences. For the original trials with a non-inferiority design, the margins were retrieved from the original protocols. For the sake of clarity, only original trials with hazard ratios (HRs) as the primary effect size were included, resulting in the exclusion of the LEAD2 trial due to a continuous outcome. However, in general, the sceptical p-value can also be used for continuous or binary outcomes. In addition, two further trials (ISAR-REACT5 and VERO trial) were excluded, because a meta-analysis was not performed due to significant heterogeneity between the estimates of the three US claims data detected by Cochran's Q-test. Thus, $N = 29$ pairs of original RCTs and related emulation studies (pooled data) were included for further analyses, 11 with a superiority and 18 with a non-inferiority design. The effect estimates of 28 trials were in the intended direction (HR < 1), out of which 26 were statistically significant. The two non-significant trials (INSPIRE and TRANSCEND) had one-sided p-values of $p = 0.34$ and $p = 0.10$, respectively. One trial (PRONOUNCE) had an effect estimate in the opposite direction (HR > 1), and is one of the two trials (with CAROLINA) for which the emulation was performed before the results of the RCT were made public. Hazard ratios with 95% confidence intervals of all 29 analyzed trials from the RCT DUPLICATE initiative are presented in Fig. 1 and Appendix 2 Table 3.

Methods

The sceptical *p*-value

The lack of a standard approach to define replication success motivated Held [24] to develop a new method: the sceptical p-value, which combines two principles: 'the Analysis of Credibility' by Matthews [31, 32] and a prior-data conflict assessment proposed by Box [33].

Assume that the effect estimate from the RCT is in the anticipated direction (HR < 1, so $\hat{\theta}_{\text{RCT}} = \log(\text{HR}) < 0$) and is significant at level α, i.e. one-sided p-value $p_{\text{RCT}} < \alpha$. This result from the RCT is challenged by a sceptical prior that renders the significant finding no longer convincing (lower limit of the posterior credible interval fixed at zero). The corresponding variance τ^2 of the sceptical prior decreases with increasing (one-sided) p-value p_{RCT}, so RCTs with a larger p-value will lead to a wider sceptical prior. Note that the sceptical p-value for the alternative HR < 1 is not available, if the effect estimate from the RCT is already in the opposite direction (HR > 1), such as in the PRONOUNCE study.

In the second step, the sceptical prior needs to be compared with the RWE effect estimate $\hat{\theta}_{\text{RWE}}$ with standard error σ_{RWE}. Held [24] proposed to assess prior-data conflict using the prior-predictive distribution of $\hat{\theta}_{\text{RWE}}$ (Spiegelhalter et al. [34], Section 5.8). This results in the test statistic $t_{\text{Box}} = \hat{\theta}_{\text{RWE}} / \sqrt{\tau^2 + \sigma^2_{\text{RWE}}}$, which is compared against a standard normal N(0,1) distribution to obtain the tail probability $p_{\text{Box}} = \Phi(t_{\text{Box}})$ where $\Phi(.)$ denotes the standard normal cdf and smaller values of p_{Box} indicate a larger conflict between the replication effect $\hat{\theta}_{\text{RWE}}$ and the sceptical prior. Replication success at level α is achieved if there is significant conflict between the prior and the RWE estimate, i.e. $p_{\text{Box}} \leq \alpha$.

The approach needs some adaptions for non-inferiority trials. If δ denotes the non-inferiority margin on the log

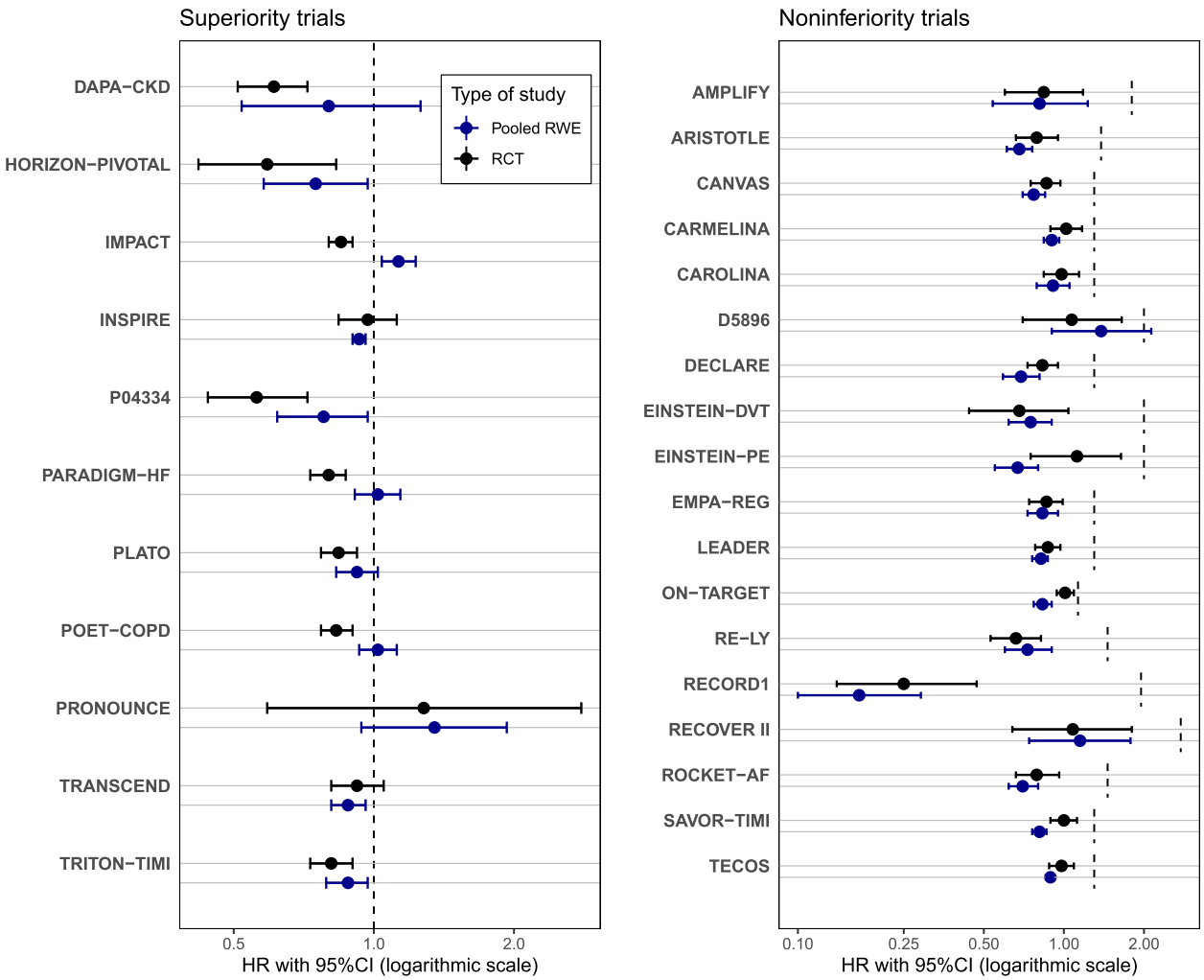

Fig. 1 Forest plot comparing RCT and RWE effect estimates (HR with 95% CI) from the RCT DUPLICATE initiative [11, 21]. Individual non-inferiority margins were added for non-inferiority trials. Three trials were excluded due to a continuous outcome (LEAD2) and significant heterogeneity between the estimates of the three US claims data (ISAR-REACT5 and VERO)

hazard ratio scale, then the sceptical prior is now centered at δ. We then obtain the modified test statistic

$$t_{\text{Box}} = (\hat{\theta}_{\text{RWE}} - \delta)/\sqrt{\tau^2 + \sigma^2_{\text{RWE}}}. \quad (1)$$

The original formulation

The value of p_{Box} depends on the significance level α, thus rendering the interpretation difficult. The sceptical p-value p_S is the smallest level α for which replication success can be achieved, so no longer depends on α. Some algebra shows that the sceptical p-value depends on the (one-sided) original and replication p-value p_{RCT} and p_{RWE} and the variance ratio $c = \sigma^2_{\text{RCT}}/\sigma^2_{\text{RWE}}$, here σ_{RCT} denotes the standard error of the RCT effect estimate. For non-inferiority trials, the p-values are calculated for the null hypothesis that the true effect is equal to the non-inferiority margin δ.

The sceptical p-value can be interpreted quantitatively: Smaller p_S values indicate a higher degree of replication success. The sceptical p-value is one-sided to ensure that replication success can only occur if the replication effect estimate has the same sign as the original effect estimate. Further properties of the sceptical p-value are studied in Held [24]. First, the approach can be shown to be more stringent than the two-trials rule: if replication success is achieved ($p_S \leq \alpha$), then two-trials rule is also fulfilled ($p_{\text{TTR}} = \max\{p_{\text{RCT}}, p_{\text{RWE}}\} \leq \alpha$) but not necessarily the other way around. As a consequence, the overall Type-I

error rate of the sceptical p-value (assuming that both the original and the replication effect estimate is zero) is smaller than the Type-I error rate α^2 of the two-trials rule. Second, for fixed original and replication p-value, the sceptical p-value increases with increasing variance ratio c. The variance ratio can be rewritten as the relative sample size $c = n_{\text{RWE}}/n_{\text{RCT}}$ (replication to original), if we assume that the unit variance in both studies is the same. This implies that a replication study with a large sample size will show a smaller degree of replication success than a replication study with the same replication p-value, but a smaller sample size. The method hence requires replication studies with large sample sizes to be more convincing than those with small sample sizes to achieve the same degree of replication success.

Exact Type-I error control

The sceptical p-value has been recently re-calibrated to achieve exact overall Type-I error control across both studies [28] for every value of c. The new version has still improved project power compared to the two-trials rule but can now lead to replication success even if one of the two studies is not significant. In what follows we will report the re-calibrated "controlled" version of the sceptical p-value.

The exact overall calibration at level α^2 implies that we can invert the sceptical p-value to obtain one-sided confidence intervals for the combined effect estimate. The upper limit θ_u of a 97.5% CI for the log hazard ratio is therefore obtained as the value of $\delta = \theta_u$, where the modified test statistic (1) gives a squared sceptical p-value of 0.025. Exponentiation gives an upper limit on the hazard ratio scale, which can directly be compared to the non-inferiority (or superiority) margin. This approach provides an alternative to standard meta-analytic pooling, which is not compatible to the sceptical p-value nor the two-trials rule.

A comparison

The (controlled) sceptical p-value has exact overall Type-I error control (across both studies) at level α^2 for all values of the variance ratio c. The two-trials rule also offers exact Type-I error control at level α^2. However, the corresponding p-value $p_{\text{TTR}} = \max\{p_{\text{RCT}}, p_{\text{RWE}}\}$ cannot be smaller than p_{RCT} nor p_{RWE}. Figure 2 compares the behaviour of the sceptical p-value p_S and the corresponding p-value p_{TTR} of the two-trials rule as a function of the relative sample size for fixed original p-values. The calculations assume that the relative effect size $\hat{\theta}_{\text{RWE}}/\hat{\theta}_{\text{RCT}}$ is fixed at 1 (top) and 0.75 (bottom), respectively. Note that for fixed relative effect size, the replication p-value depends on the assumed relative effect size and the relative sample size.

Figure 2 shows that the sceptical p-value behaves very similar to the two-trials rule if the relative sample size is small. However, it decreases smoothly towards zero and can get smaller than p_{RCT} if the relative sample size c becomes large, whereas the two-trials rule p-value has an edge and then stays constant, as it cannot get smaller than p_{RCT}.

Power of the replication study

The power of the replication study, *i.e.* the probability that a replication study with a certain sample size leads to replication success, has been extensively studied [28, 35–38]. The power is usually calculated as the probability to reach a significant replication result *conditional* on the effect estimate from the original trial. This concept is illustrated in Fig. 3 using the TRITON-TIMI trial from the RCT DUPLICATE initiative. The conditional power of the two-trials rule at the one-sided level $\alpha = 0.025$ is plotted as a function of the effect estimate from the RCT (log hazard ratio) assuming that the variance is the same in the original trial and the emulation study ($c = 1$). The original effect estimate in the TRITON-TIMI example is $\hat{\theta}_{\text{RCT}} = -0.21$, so with $c = 1$, a conditional power of 98.1% is achieved. However, the original effect estimate comes with uncertainty, represented by the density in the top axis, and the 95% confidence interval $[-0.31, -0.11]$ around the original effect estimate. The conditional power for the values inside this confidence interval greatly vary, from 52.6% to 100%. One way to incorporate this uncertainty is to compute the *predictive* power [39] by averaging the conditional power with respect to the distribution of the original effect estimate. The density of the conditional power is shown in the right axis of Fig. 3 and the average is 92.6%, so considerably smaller than 98.1%.

If the power to detect the effect from the original study is not too large (< 95%), the conditional Type-I error rate (the probability of replication success if the replication effect estimate is zero) of the sceptical p-value is shown to be bounded and never larger than 2α (Micheloud et al. [28], Section 3.4). Such a two-fold increase in Type-I error rate is the price to be paid for increased power, but considered to be acceptable even in the regulatory setting [40].

Power calculations are also used to determine which sample size is required in the replication study to reach a certain power, usually between 80% and 90%. While replication sample size calculation is not relevant in the RCT DUPLICATE study, where all available patient information in the claims data was used, it is of interest to know the power of a RWE study with the available data. It is therefore advisable to determine the power for an existing dataset before calculating the effect

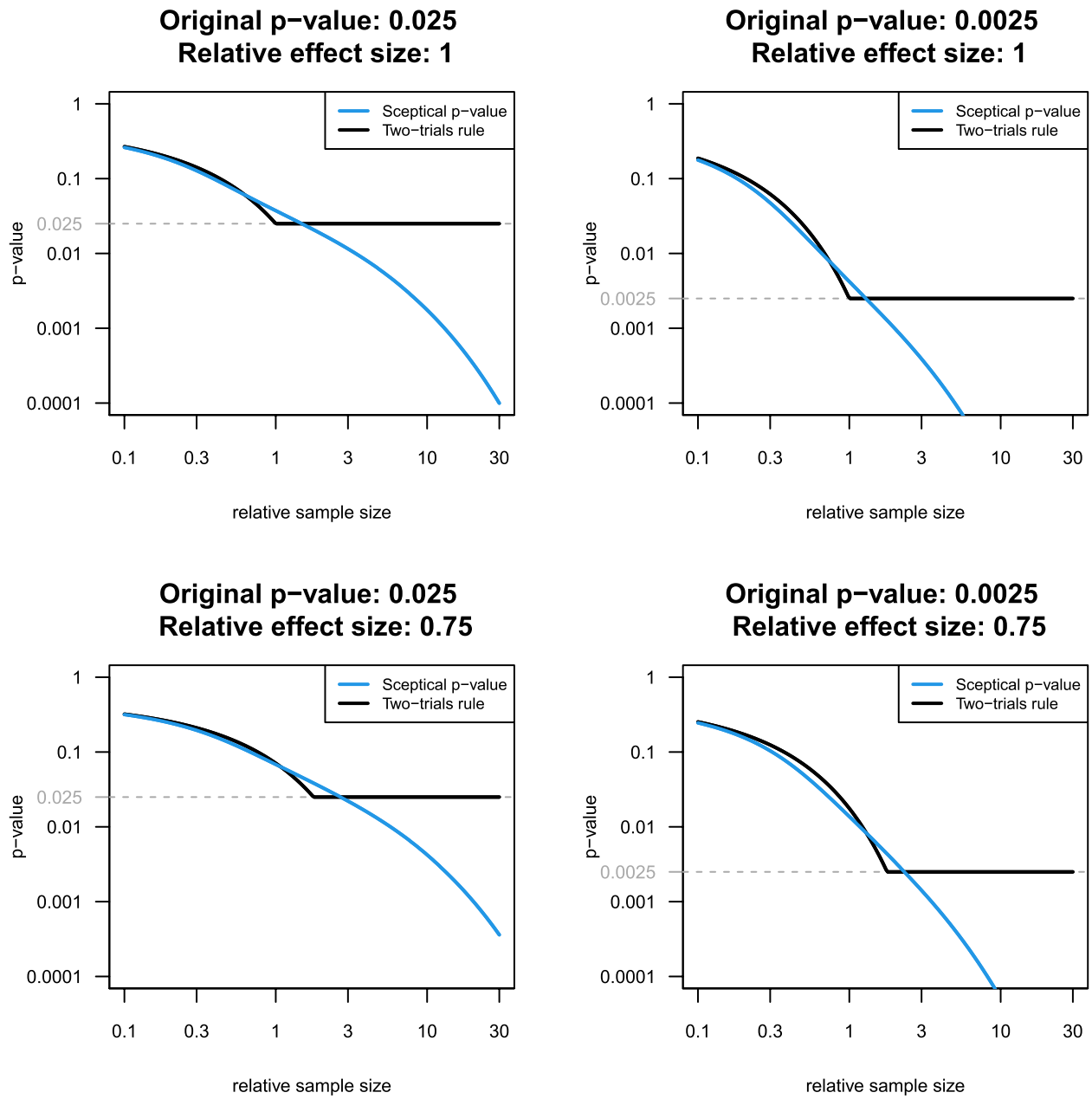

Fig. 2 Sceptical *p*-value (square root) and *p*-value from the two-trials rule shown as a function of relative sample size (RWE to RCT) for fixed *p*-value of the original study (dashed grey line) and fixed relative effect size (first row: 1, second row: 0.75)

estimates for the emulation, to ensure that the available data are indeed sufficient. Conversely, calculating the minimum necessary sample size can also be beneficial in the context of RWE, for example, to understand the cohort size required after propensity score matching. If the power is insufficient or if the sample size is inadequate after matching, the chosen dataset is likely not suitable for the planned emulation. In the "Results" section, we calculate the conditional and predictive power of the two-trials rule and of the sceptical *p*-value for all the RWE emulations [28, 38].

Results

Replication of RCTs

Results of the replication success assessment with the two methods are presented in Table 1. The variance ratio c is larger than 10 for two studies (TECOS, INSPIRE), see also Fig. 1. TECOS is a non-inferiority study, where

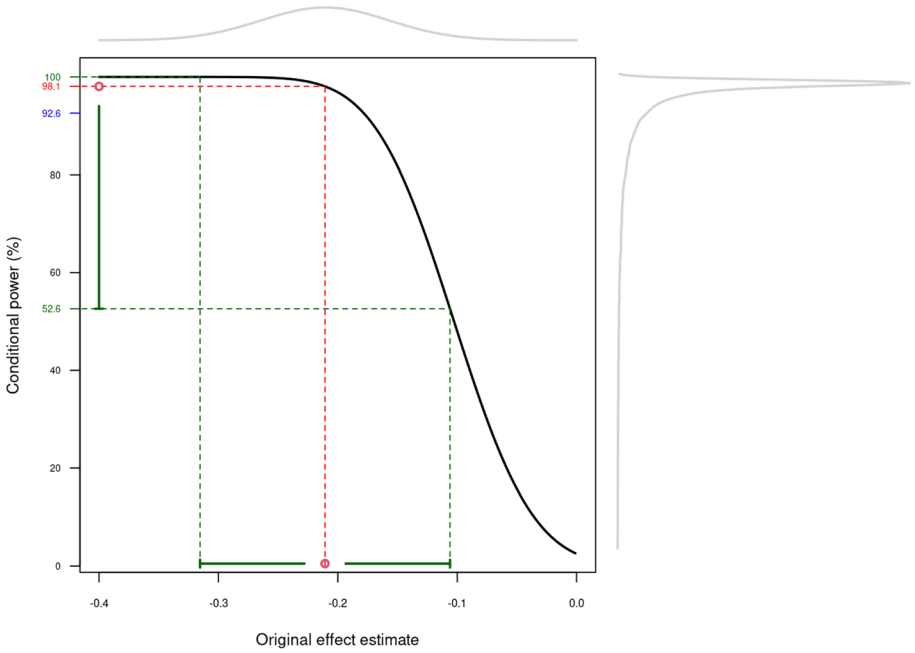

Fig. 3 Conditional power as a function of the original effect estimate. The original effect estimate $\hat{\theta}_{\text{RCT}} = -0.21$ of the TRITON-TIMI trial and the corresponding conditional power of 98.1% are colored in red. The grey lines in the top and right axis are the distributions of the original effect estimate and the corresponding conditional power, respectively. The average conditional power of 92.6% (in blue) is the predictive power. The green intervals represent the range given by the 95% confidence interval of the original effect estimate

both the RCT and the RWE study showed non-inferiority with respect to a HR margin of 1.4. INSPIRE is a superiority study, where both effect estimates have been close to 1, but the RWE estimate is slightly smaller and more much precise. The sceptical p-value is smaller or equal to the p-value from the two-trials rule for the vast majority of the studies (27/29 studies). Note also that in three cases (CARMELINA, TRANSCEND, INSPIRE) the sceptical p-value is smaller than p_{RCT} and in seven cases (TRITON-TIMI, PLATO, HORIZON-PIVOTAL, DAPA-CKD, P04334, D5896, and PRONOUNCE) it is smaller than p_{RWE}, both is impossible for p_{TTR}.

At the one-sided level $\alpha = 0.025$, the same conclusion is drawn with the two-trials rule and the sceptical p-value: The same 20 out of 29 (69%) emulations successfully replicate the corresponding original RCT results. However, it was noticeable that more studies failed to replicate the original effect, if no Medicare data were available for the pooled effect. If Medicare data were available, 16 of 19 (84%) emulations successfully replicated the original effect, compared to the lower rate (5 of 10, 50%) of successfully replications, when Medicare data was not available. This finding was independent from the used metric, i.e. the sceptical p-value and the p-value from the two-trials rule lead to the same rates of success.

Figure 4 shows $\hat{\theta}_{\text{RCT}}$ - log(margin) versus $\hat{\theta}_{\text{RWE}}$ - log(margin) for the 29 study pairs in the RCT DUPLICATE initiative. The margin is 1 in the superiority studies. Points below the diagonal represent cases where the effect estimate from the RWE study is shrunken as compared to the effect estimate from the original trial. This shrinkage is towards 0 for superiority studies and towards the margin for non-inferiority studies. Replication effect estimates usually present a large amount of shrinkage as compared to their original counterpart in recent large-scale replication projects (e.g. in Open Science Collaboration, Camerer et al. [41–43]). This is not the case here.

Replication power

Results are displayed in Table 2. The average conditional power [44, 45] is 85.9% with the two-trials rule and 87.0% with the sceptical p-value. These values reduce to 83.5%, and 85.0%, respectively, for predictive power. If the true underlying effect is the same in the original trial and in the emulation study and both were conducted with high standards, we would expect the average predictive power to be equal to the proportion of success, which is smaller in the RCT DUPLICATE study (69% with both methods). The proportion of replication success as well as the

Table 1 One-sided *p*-values p_{RCT} of the original trial and p_{RWE} of the emulation, as well as variance ratio *c* and one-sided *p*-values p_{TTR} of the two-trials rule and p_S of the sceptical *p*-value

	Study	p_{RCT}	p_{RWE}	*c*	p_{TTR}	p_S
1	LEADER	< 0.0001	< 0.0001	2.6	< 0.0001	< 0.0001
2	DECLARE	< 0.0001	< 0.0001	0.7	< 0.0001	< 0.0001
3	EMPA-REG	< 0.0001	< 0.0001	1.2	< 0.0001	< 0.0001
4	CANVAS	< 0.0001	< 0.0001	1.8	< 0.0001	< 0.0001
5	CARMELINA	0.0003	< 0.0001	4.2	0.0003	< 0.0001
6	TECOS	< 0.0001	< 0.0001	14.3	< 0.0001	< 0.0001
7	SAVOR-TIMI	< 0.0001	< 0.0001	3.5	< 0.0001	< 0.0001
8	TRITON-TIMI	< 0.0001	0.007	1.0	0.007	0.003
9	PLATO	< 0.0001	0.056	0.7	0.056	0.031
10	ARISTOTLE	< 0.0001	< 0.0001	2.7	< 0.0001	< 0.0001
11	RE-LY	< 0.0001	< 0.0001	1.2	< 0.0001	< 0.0001
12	ROCKET-AF	< 0.0001	< 0.0001	2.2	< 0.0001	< 0.0001
13	EINSTEIN-DVT	< 0.0001	< 0.0001	5.3	< 0.0001	< 0.0001
14	EINSTEIN-PE	0.0018	< 0.0001	4.4	0.0018	< 0.0001
15	RECOVER II	0.0002	< 0.0001	1.4	0.0002	0.0002
16	AMPLIFY	< 0.0001	< 0.0001	0.7	< 0.0001	< 0.0001
17	RECORD1	< 0.0001	< 0.0001	1.3	< 0.0001	< 0.0001
18	TRANSCEND	0.10	0.002	2.3	0.10	0.064
19	ON-TARGET	0.001	< 0.0001	0.9	0.001	< 0.0001
20	HORIZON-PIVOTAL	0.001	0.014	1.8	0.014	0.01
21	DAPA-CKD	< 0.0001	0.16	0.2	0.16	0.15
22	PARADIGM-HF	< 0.0001	0.63	0.6	0.63	0.65
23	P04334	< 0.0001	0.015	1.2	0.015	0.004
24	D5896	0.002	0.046	1.0	0.046	0.03
25	IMPACT	< 0.0001	1.00	0.5	1.00	1.00
26	POET-COPD	< 0.0001	0.66	0.7	0.66	0.68
27	INSPIRE	0.34	< 0.0001	19.9	0.34	0.25
28	CAROLINA	0.0001	< 0.0001	1.2	0.0001	< 0.0001
29	PRONOUNCE	0.73	0.95	4.7	0.95	NA

average predictive power in the RCT DUPLICATE initiative for other values of α are shown in Appendix 1 Fig. 5.

The three RCTs with a non-significant original *p*-value (PRONOUNCE, INSPIRE and TRANSCEND) have a power of 0 with all three methods. By definition, the power based on a non-significant original trial is always 0 with the two-trials rule. There is no such requirement on p_{RCT} for the sceptical *p*-value but large relative sample sizes are required to reach an acceptable power level for non-significant original studies.

For the 26 other studies, the power of the two-trials rule and the sceptical *p*-values is very similar, with slightly higher values with the sceptical *p*-values. This is because the original *p*-values are relatively small in every case, which is favored by the sceptical *p*-values (see Held et al. and Micheloud et al. [27, 28], Sections 3.1 and 3.4, respectively).

Confidence intervals

We have also calculated confidence intervals for the hazard ratio based on inversion of the sceptical *p*-value and compared those to the ones obtained from applying a fixed effect meta-analysis to the results from the RCT and RWD. The results are given in Table 3 in Appendix 2. The upper limit of both confidence intervals are similar, the upper limit of the sceptical CI being often slightly larger than the meta-analytic one. The differences get larger when the uncertainty of the effect estimates is substantially different, for example for DAPA-CKD or INSPIRE, where the meta-analytic pooled effect estimate is dominated by the larger study. This is also the case for the three superiority studies PARADIGM-HF, IMPACT and POET-COPD, where the meta-analytic confidence interval indicates significance (at the standard two-sided 95% level) with

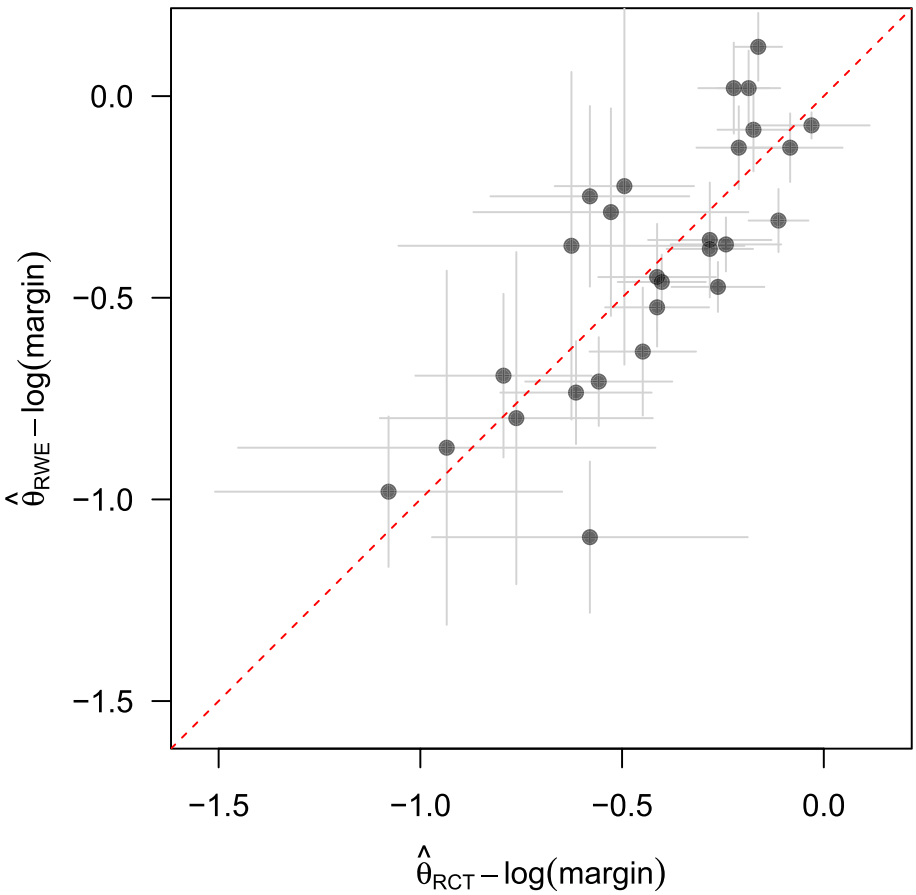

Fig. 4 $\hat{\theta}_{RCT}$ - log(margin) versus $\hat{\theta}_{RWE}$ - log(margin) with their respective 95% confidence interval for the studies in the RCT DUPLICATE initiative (without RECORD1, where differences are larger)

an upper CI limit below 1, although the RWE effect estimate is above 1 in all three cases. In contrast, the upper limit of the sceptical CI is always above one, in line with the large sceptical *p*-values for those studies reported Table 1. This illustrates that meta-analytic pooling can still flag a positive (significant) finding, even if one of the two studies has an effect in the wrong direction [26].

Discussion

A target trial emulation is a conceptual emulation of a "hypothetical" RCT, using a two-step process [46, 47]. In the first step, the causal questions are defined in a protocol for the hypothetical RCT, which has to specify the key elements to define the causal estimands (i.e. eligibility criteria, treatment strategies, treatment assignment, the start and end of follow-up, outcomes, causal contrasts) and the data analysis plan [46]. The RCT described in the protocol is defined as the target study for causal inference, which will be emulated using observational data in the second step [46, 47]. Furthermore, several studies used an existing RCT instead of a hypothetical RCT, which will be emulated by the target trial emulation [48–56]. However, even with the existing RCT as a template for the target trial emulation, the effect estimates of the RWD emulation are not always compared with those of the original trial [48, 52–55]. A statistical comparison of the emulation with the original RCT further substantiates the results and increase the resulting evidence. As seen in the data presented, the sceptical *p*-value can be an excellent extension to evaluate the agreement and thus complements the evidence from target trial emulations of existing RCTs.

The emulations of the 32 evaluated RCTs were previous implemented and analyzed by the RCT DUPLICATE Initiative [21]. In this project, Wang et al. [21] used different binary agreement metrics to check the replication success of the 32 emulations using claims data of three different US insurance companies. First, named *full statistical significance agreement*, is fulfilled, if the estimates (log(HR))

Table 2 Conditional and predictive power (in %) at the one-sided level $\alpha = 0.025$ with the two-trials rule and the sceptical *p*-value

			Conditional power		Predictive power	
	Study	p_{RCT}	**two-trials rule**	**sceptical *p***	**two-trials rule**	**sceptical *p***
1	LEADER	< 0.0001	100.0	100.0	100.0	100.0
2	DECLARE	< 0.0001	100.0	100.0	99.7	99.9
3	EMPA-REG	< 0.0001	100.0	100.0	99.8	99.9
4	CANVAS	< 0.0001	100.0	100.0	100.0	100.0
5	CARMELINA	0.0003	100.0	100.0	98.8	99.3
6	TECOS	< 0.0001	100.0	100.0	100.0	100.0
7	SAVOR-TIMI	< 0.0001	100.0	100.0	99.9	99.9
8	TRITON-TIMI	< 0.0001	98.1	99.2	92.6	95.3
9	PLATO	< 0.0001	91.2	95.0	84.8	89.3
10	ARISTOTLE	< 0.0001	100.0	100.0	100.0	100.0
11	RE-LY	< 0.0001	100.0	100.0	100.0	100.0
12	ROCKET-AF	< 0.0001	100.0	100.0	100.0	100.0
13	EINSTEIN-DVT	< 0.0001	100.0	100.0	100.0	100.0
14	EINSTEIN-PE	0.0018	100.0	100.0	100.0	97.0
15	RECOVER II	0.0002	98.7	99.4	92.4	94.9
16	AMPLIFY	< 0.0001	95.2	97.6	90.1	93.7
17	RECORD1	< 0.0001	100.0	100.0	100.0	100.0
18	TRANSCEND	0.10	0.0	0.0	0.0	0.0
19	ON-TARGET	0.001	80.5	85.6	73.4	77.9
20	HORIZON-PIVOTAL	0.001	98.0	98.9	89.3	91.8
21	DAPA-CKD	< 0.0001	59.1	65.6	58.5	64.6
22	PARADIGM-HF	< 0.0001	97.3	98.7	93.5	96.1
23	P04334	< 0.0001	99.9	100.0	98.2	99.1
24	D5896	0.002	81.2	85.7	73.5	77.5
25	IMPACT	< 0.0001	96.7	98.3	93.4	95.9
26	POET-COPD	< 0.0001	97.6	98.9	93.4	96.0
27	INSPIRE	0.34	0.0	0.0	0.0	0.0
28	CAROLINA	0.0001	97.3	98.8	90.6	93.7
29	PRONOUNCE	0.73	0.0	0.0	0.0	0.0
30	Average		85.9	87.0	83.5	85.0

and the related 95% CIs are on the same side of the null (without consideration of the predefined non-inferiority margins in the case of non-inferiority trials). 23 (72%) of the 32 emulations met statistical significance agreement [21]. This binary agreement metric should be equivalent to the two-trials rule. However, for two studies, i.e. D5896 and PRONOUNCE, the *p*-value of the two-trials rule delivered a different conclusion of the replication success. Wang et al. [21] reported that statistical agreement was fulfilled for all of the two emulations, whereby in our analysis, $p_{\mathrm{TTR}} > 0.025$ holds. In the case of PRONOUNCE, the original RCT had a non-significant treatment effect, resulting in an non-significant p_{TTR} due to the definition of the *p*-value. Moreover, the formalism of p_{TTR} is not meaningful in the case of non-significant effects of the original trial. In the case of the D5896 trial, Wang et al. [21] did not use the non-inferiority margins for the definition of full statistical agreement, as it was simply defined as having point estimates and confidence intervals on the same side of null. Using sceptical *p*-value, it is possible to account for the non-inferiority margins, which is one major advantage of this formalism compared to the binary agreement metrics.

As the second binary agreement metric, Wang et al. [21] analysed estimated agreement, whereby the estimate of the emulation has to be an element of the 95% CI of the original effect, which was met for 66% of the emulations. As the last one, Wang et al. [21] evaluated, if the absolute standardized differences of the estimates are smaller then 1.96, which was fulfilled in 72% of the emulations. Using the sceptical *p*-value, 69% of the emulations were evaluated as successful replications. Beside the emulations with non-significant original effects, the

findings of the study at hand are in accordance to findings of Wang et al. [21].

However, in general, it is not always possible to replicate the results of an RCT, especially if there are fundamental design differences between the RCT and RWE [21, 30]. Wang et al. [21] stated that the conceptual emulation of the original RCT leads to an independent study that emulate the design of the original trial instead of replicating the population of the RCT. Potential differences of the two population (e.g. different distribution of comorbidities) may result in differences in the effect size estimates [21]. Moreover, Heyard et al. [30] analyzed differences between the original RCTs and the related RWE emulations using meta-regression models. They concluded that most of the heterogeneity between RCT and RWD emulation could be explained by delayed effect of treatment, discontinuation of treatment during run-in period, and treatment started in hospital [30]. The sceptical p-value itself cannot distinguish between replication failure due to design differences, incorrect model specification or differences due to population differences. Instead, the sceptical p-value evaluates and quantifies the agreement between two effect estimators. A successful replication of an RCT using RWD, indicated by a significant sceptical p-value, suggests the external validity of the RCT results. Conversely, if the replication is unsuccessful, the p-value does not provide insight into the reasons. These issues must be addressed using other methods. In studies based on RWD, it is essential to adequately address data heterogeneity methodologically, for example, using propensity score methods. An important aspect is achieving sufficient balance between the treatment groups. However, it is possible that the treatment effect observed in an RCT cannot be replicated in RWD due to fundamentally different cohorts. This is less a failure of RWE and may indicate that external validity is limited.

There are other studies available that emulate RCTs using RWD, where the binary metrics regulatory agreement, estimated agreement and standardized difference agreement were often used to evaluate the agreement of the effect estimates from the original trials and the related emulation study [57–62]. However, the two-trials rule and the sceptical p-values are good measures to evaluate the agreement between an original RCT and the corresponding emulation study and have benefits compared to binary metrics. Besides the advantages discussed above, both methods allow to calculate the power of the emulation study based on the effect estimate found in the RCT, with and without uncertainty incorporation. For all 29 studies considered here, the conditional and predictive power is larger with the sceptical p-values than with the two-trials rule. Furthermore, the replication rate is in general slightly higher with the sceptical p-value.

In this paper, we focused on the controlled version of the sceptical p-value, to ensure a fair comparison with the two-trials rule in terms of overall Type-I error control. An alternative to the controlled re-calibration is the golden re-calibration of the sceptical p-value [27]. The golden version is less stringent than the original formulation, because, to establish replication success, original trial and replication study do not both necessarily need to be significant at level α. However, a borderline significant original trial ($p_{\text{RCT}} = \alpha$) cannot lead to replication success, if there is shrinkage of the replication effect estimate. As shrinkage of the effect estimate from the emulation study as compared to effect estimate from the original trial seems to not be an issue in this context, one could favor the controlled sceptical p-value over the golden sceptical p-value. Moreover, a comparison of the golden version with the two-trials rule is more difficult, as the two methods have different Type-I error rates.

The presented work had several limitations. No formalism to assess the replication success of an original null effect using the sceptical p-value was available [63]. Recently Micheloud and Held [64] provided an extension of the sceptical p-value to equivalence studies. However, a trial that was initially planned for superiority, but fails to show a significant treatment effect, should not be interpreted as an equivalence study and, thus, the formalism for equivalence studies should not be assigned in this case. Instead of performing a second study with the wrong design, a new study with the aim of proofing the equivalence should be initialized (and then replicated). Moreover, as stated above, fundamental design differences between RCT and RWD emulation may result in bias, which will also affect the effect estimate of the RWD study.

Conclusion

The sceptical p-value represents a valid statistical measure to evaluate the replicability of the study results and is useful in the context of real-world evidence.

Appendix

1 Replication power

Figure 5 shows the proportion of replication success as well as the average predictive power as a function of the one-sided level α. The proportion of success is for some values of α slightly smaller with the two-trials rule than with the sceptical p-value, but the two methods are in general very close to each other.

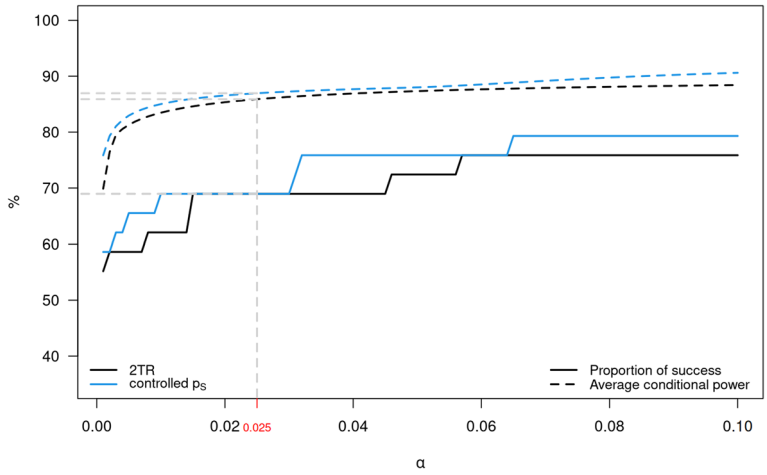

Fig. 5 Proportion of success and average predictive power as a function of the one-sided level α

2 Confidence intervals for combined effects

Table 3 Confidence intervals (CIs) for the hazard ratio (HR) from each RCT and pooled RWE. Combined confidence intervals based on fixed effect meta-analysis and sceptical *p*-value are also reported, as well as the non-inferiority (NI) margin (1.00 for superiority studies)

	Study	NI margin	RCT HR (95% CI)	pooled RWE HR (95% CI)	meta analysis HR (95% CI)	sceptical *p*-value (one-sided 97.5% CI)
1	LEADER	1.30	0.87 (0.78; 0.97)	0.82 (0.76; 0.87)	0.83 (0.79; 0.88)	(0; 0.91)
2	DECLARE	1.30	0.83 (0.73; 0.95)	0.69 (0.59; 0.81)	0.77 (0.70; 0.85)	(0; 0.88)
3	EMPA-REG	1.30	0.86 (0.74; 0.99)	0.83 (0.73; 0.95)	0.84 (0.76; 0.93)	(0; 0.92)
4	CANVAS	1.30	0.86 (0.75; 0.97)	0.77 (0.70; 0.85)	0.80 (0.74; 0.87)	(0; 0.91)
5	CARMELINA	1.30	1.02 (0.89; 1.17)	0.90 (0.84; 0.96)	0.92 (0.87; 0.98)	(0; 1.07)
6	TECOS	1.30	0.98 (0.88; 1.09)	0.89 (0.86; 0.91)	0.90 (0.87; 0.92)	(0; 1.01)
7	SAVOR-TIMI	1.30	1.00 (0.89; 1.12)	0.81 (0.76; 0.86)	0.85 (0.80; 0.90)	(0; 1.04)
8	TRITON-TIMI	1.00	0.81 (0.73; 0.90)	0.88 (0.79; 0.97)	0.84 (0.79; 0.91)	(0; 0.92)
9	PLATO	1.00	0.84 (0.77; 0.92)	0.92 (0.83; 1.02)	0.87 (0.82; 0.93)	(0; 0.96)
10	ARISTOTLE	1.38	0.79 (0.66; 0.95)	0.68 (0.61; 0.76)	0.71 (0.64; 0.78)	(0; 0.85)
11	RE-LY	1.46	0.66 (0.53; 0.82)	0.73 (0.60; 0.90)	0.70 (0.60; 0.81)	(0; 0.80)
12	ROCKET-AF	1.46	0.79 (0.66; 0.96)	0.70 (0.62; 0.80)	0.73 (0.65; 0.81)	(0; 0.85)
13	EINSTEIN-DVT	2.00	0.68 (0.44; 1.04)	0.75 (0.62; 0.90)	0.74 (0.62; 0.88)	(0; 0.86)
14	EINSTEIN-PE	2.00	1.12 (0.75; 1.64)	0.67 (0.55; 0.80)	0.74 (0.62; 0.87)	(0; 1.29)
15	RECOVER II	2.75	1.08 (0.64; 1.80)	1.15 (0.74; 1.78)	1.12 (0.80; 1.57)	(0; 1.48)
16	AMPLIFY	1.80	0.84 (0.60; 1.18)	0.81 (0.54; 1.23)	0.83 (0.64; 1.07)	(0; 1.03)
17	RECORD1	1.95	0.25 (0.14; 0.47)	0.17 (0.10; 0.29)	0.20 (0.13; 0.30)	(0; 0.33)
18	TRANSCEND	1.00	0.92 (0.81; 1.05)	0.88 (0.81; 0.96)	0.89 (0.83; 0.96)	(0; 0.97)
19	ON-TARGET	1.13	1.01 (0.94; 1.09)	0.83 (0.77; 0.90)	0.92 (0.87; 0.97)	(0; 1.04)
20	HORIZON-PIVOTAL	1.00	0.59 (0.42; 0.83)	0.75 (0.58; 0.97)	0.69 (0.56; 0.84)	(0; 0.84)
21	DAPA-CKD	1.00	0.61 (0.51; 0.72)	0.80 (0.52; 1.26)	0.63 (0.54; 0.74)	(0; 0.99)
22	PARADIGM-HF	1.00	0.80 (0.73; 0.87)	1.02 (0.91; 1.14)	0.88 (0.82; 0.94)	(0; 1.07)

	Study	NI margin	RCT HR (95% CI)	pooled RWE HR (95% CI)	meta analysis HR (95% CI)	sceptical *p*-value (one-sided 97.5% CI)
23	P04334	1.00	0.56 (0.44; 0.72)	0.78 (0.62; 0.97)	0.67 (0.57; 0.79)	(0; 0.86)
24	D5896	2.00	1.07 (0.70; 1.65)	1.38 (0.90; 2.13)	1.21 (0.90; 1.65)	(0; 1.68)
25	IMPACT	1.00	0.85 (0.80; 0.90)	1.13 (1.04; 1.23)	0.93 (0.89; 0.98)	(0; 1.17)
26	POET-COPD	1.00	0.83 (0.77; 0.90)	1.02 (0.93; 1.12)	0.90 (0.85; 0.96)	(0; 1.06)
27	INSPIRE	1.00	0.97 (0.84; 1.12)	0.93 (0.90; 0.96)	0.93 (0.90; 0.96)	(0; 1.01)
28	CAROLINA	1.30	0.98 (0.84; 1.14)	0.91 (0.79; 1.05)	0.94 (0.85; 1.05)	(0; 1.05)
29	PRONOUNCE	1.00	1.28 (0.59; 2.79)	1.35 (0.94, 1.93)	1.34 (0.96; 1.85)	(0; 1.85)

Abbreviations

CI Confidence interval
HR Hazard ratio
RCT Randomized controlled trial
RWD Real world data
RWE Real world evidence
TTR Two-trials rule

Acknowledgements
We thank the RCT DUPLICATE team for their effort and for making their data publicly available.

Authors' contributions
J. K.: Conceptualization, Methodology, Formal analysis, Software, Writing - Original draft, Writing - Review & Editing, Visualization. C. M.: Methodology, Formal analysis, Software,Writing - Original draft,Writing - Review & Editing, Visualization. S. E. : Conceptualization, Methodology, Formal analysis, Review & Editing. R. H.: Writing - Review & Editing. L. H.: Conceptualization, Methodology, Formal Analysis, Writing -Original draft, Writing - Review & Editing, Visualization.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Data availability
The R package ReplicationSuccess, available on CRAN at https://CRAN.R-project.org/package=ReplicationSuccess, has been used for the computation of the sceptical p-value and all power calculations. The data can be found at https://gitlab.com/heyardr/hte-in-rwe/-/tree/main/data, and the code to reproduce all the Figures and Tables at https://github.com/CharlotteMicheloud/RWE.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
JK reports research funding from the German Society for Trauma Surgery sponsored by Stryker, outside the submitted work. SE, CM, RH and LH have no conflict of interest to declare.

Received: 23 December 2024 Accepted: 7 May 2025
Published online: 24 May 2025

References

1. Byar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellenberg JH, et al. Randomized Clinical Trials - Perspectives on Some Recent Ideas. N Engl J Med. 1976;295(2):74–80. https://doi.org/10.1056/NEJM197607082950204.
2. Abel U, Koch A. The role of randomization in clinical studies: myths and beliefs. J Clin Epidemiol. 1999;52(6):487–97. https://doi.org/10.1016/s0895-4356(99)00041-4.
3. Concato J, Shah N, Horwitz RI. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. N Engl J Med. 2000;342(25):1887–92. https://doi.org/10.1056/NEJM200006223422507.
4. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals - A Systematic Sampling Review. JAMA. 2007;297(11):1233–40. https://doi.org/10.1001/jama.297.11.1233.
5. Food and Drug Administration, HHS. Federal Register, Evaluating Inclusion and Exclusion Criteria in Clinical Trials; Public Meeting (Docket number: FDA-2018-N-0129). 2018. https://www.federalregister.gov/d/2018-01643e. Accessed 1 Dec 2024.
6. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005;58(4):323–37. https://doi.org/10.1016/j.jclinepi.2004.10.012.
7. Eichler HG, Abadie E, Breckenridge A, Flamion B, Gustafsson LL, Leufkens H, et al. Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. Nat Rev Drug Discov. 2011;10(7):495–506. https://doi.org/10.1038/nrd3501.
8. US Food and Drug Administration. Real-World Evidence. 2023. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed 1 Dec 2024.
9. 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):867–77. https://doi.org/10.1002/cpt.1351.
10. Orsini LS, Berger M, Crown W, Daniel G, Eichler HG, Goettsch W, et al. Improving Transparency to Build Trust in Real-World Secondary Data Studies for Hypothesis Testing—Why, What, and How: Recommendations and a Road Map from the Real-World Evidence Transparency Initiative. Value Health. 2020;23(9):1128–36. https://doi.org/10.1016/j.jval.2020.04.002.
11. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies. Circulation. 2021;143(10):1002–13. https://doi.org/10.1161/CIRCULATIONAHA.120.051718.
12. Institute for Quality and Efficiency in Health Care. Concepts for the generation of routine practice data and their analysis for the benefit assessment of drugs according to §35a Social Code Book V (SGB V). 2020. https://www.iqwig.de/download/a19-43_routine-practice-data-for-the-benefit-assessment-of-drugs_extract-of-rapid-report_v1-0.pdf. Accessed 1 Dec 2024.

13. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol. 2022;22(1):287. https://doi.org/10.1186/s12874-022-01768-6.
14. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther. 2017;102(6):924–33. https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/cpt.857. Accessed 1 Dec 2024.
15. Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. Nat Rev Clin Oncol. 2019;16:312–25. https://doi.org/10.1038/s41571-019-0167-7.
16. Bhandari M, Tornetta P, Ellis T, Audigé L, Sprague S, Kuo J, et al. Hierarchy of evidence: Differences in results between non-randomized studies and randomized trials in patients with femoral neck fractures. Arch Orthop Trauma Surg. 2004;124:10–6. https://doi.org/10.1007/s00402-003-0559-z.
17. Golder S, Loke YK, Bland M. Meta-analyses of Adverse Effects Data Derived from Randomised Controlled Trials as Compared to Observational Studies: Methodological Overview. PLoS Med. 2011;8(5):1–13. https://doi.org/10.1371/journal.pmed.1001026.
18. Kuss O, Legler T, Börgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. J Clin Epidemiol. 2011;64(10):1076–84. https://doi.org/10.1016/j.jclinepi.2011.01.005.
19. Lonjon G, Boutron I, Trinquart L, Ahmad N, Aim F, Nizard R, et al. Comparison of Treatment Effect Estimates From Prospective Nonrandomized Studies With Propensity Score Analysis and Randomized Controlled Trials of Surgical Procedures. Milbank Q. 2014;259(1):18–25. https://doi.org/10.1097/SLA.0000000000000256.
20. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014;14(4). https://doi.org/10.1002/14651858.MR000034.pub2.
21. Wang SV, Schneeweiss S, Initiative RD. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. JAMA. 2023;329(16):1376–85. https://doi.org/10.1001/jama.2023.4221.
22. Nosek BA, Errington TM. Making sense of replications eLife. 2017;6:e23383. https://doi.org/10.7554/eLife.23383.
23. Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, et al. Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. Pharmacoepidemiol Drug Saf. 2017;26(9):1018–1032. https://doi.org/10.1002/pds.4295.
24. Held L. A new standard for the analysis and design of replication studies (with discussion). J R Stat Soc Ser A. 2020;183:431–69.
25. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, et al. Investigating the replicability of preclinical cancer biology. eLife. 2021;10:e71601. https://doi.org/10.7554/eLife.71601.
26. Held L. Beyond the two-trials rule. Stat Med. 2024;43(26):5023–42. https://doi.org/10.1002/sim.10055.
27. Held L, Micheloud C, Pawel S. The assessment of replication success based on relative effect size. Ann Appl Stat. 2022;16(2):706–20. https://doi.org/10.1214/21-AOAS1502.
28. Micheloud C, Balabdaoui F, Held L. Assessing replicability with the sceptical *p*-value: Type-I error control and sample size planning. Statistica Neerlandica. 2023;77(4):573–91. https://doi.org/10.1111/stan.12312.
29. Franklin JM, Pawar A, Martin D, Glynn RJ, Levenson M, Temple R, et al. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. Clin Pharmacol Ther. 2020;107(4):817–26. https://doi.org/10.1002/cpt.1633.
30. Heyard R, Held L, Schneeweiss S, Wang SV. Design differences and variation in results between randomised trials and non-randomised emulations: meta-analysis of RCT-DUPLICATE data. BMJ Med. 2024;3(1):e000709. https://doi.org/10.1136/bmjmed-2023-000709.
31. Matthews RAJ. Methods for assessing the credibility of clinical trial outcomes. Drug Inf J. 2001;35:1469–78.
32. Matthews RAJ. Why should clinicians care about Bayesian methods? (with discussion). J Stat Plan Infer. 2001;94:43–71.
33. Box GEP. Sampling and Bayes' Inference in Scientific Modelling and Robustness (with discussion). J R Stat Soc Ser A. 1980;143:383–430.
34. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester: John Wiley & Sons; 2004.
35. Anderson SF, Maxwell SE. Addressing the "replication crisis": using original studies to design replication studies with appropriate statistical power. Multivar Behav Res. 2017;52(3):305–24. https://doi.org/10.1080/00273171.2017.1289361.
36. Anderson SF, Kelley K. Sample size planning for replication studies: The devil is in the design. Psychol Methods. 2022. https://doi.org/10.1037/met0000520.
37. van Zwet EW, Goodman SN. How large should the next study be? Predictive power and sample size requirements for replication studies. Stat Med. 2022;41(16):3090–101. https://doi.org/10.1002/sim.9406.
38. Micheloud C, Held L. Power Calculations for Replication Studies. Stat Sci. 2022;37(3). https://doi.org/10.1214/21-sts828.
39. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. Stat Med. 1986;5(1):1–13. https://doi.org/10.1002/sim.4780050103.
40. Rosenkranz G. Is it possible to claim efficacy if one of two trials is significant while the other just shows a trend? Drug Inf J. 2002;36(1):875–9. https://doi.org/10.1177/009286150203600416.
41. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716. https://doi.org/10.1126/science.aac4716.
42. Camerer CF, Dreber A, Forsell E, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016;351(6280):1433–6. https://doi.org/10.1126/science.aaf0918.
43. Camerer CF, Dreber A, Holzmeister F, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat Hum Behav. 2018;2(9):637–44. https://doi.org/10.1038/s41562-018-0399-z.
44. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305–7. https://doi.org/10.1038/d41586-019-00857-9.
45. McShane BB, Böckenholt U, Hansen KT. Average Power: A Cautionary Note. Adv Methods Pract Psychol Sci. 2020;3(2):185–99. https://doi.org/10.1177/2515245920902370.
46. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016;183(8):758–64. https://doi.org/10.1093/aje/kwv254.
47. Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. JAMA. 2022 12;328(24):2446–2447. https://doi.org/10.1001/jama.2022.21383.
48. Deng Y, Polley EC, Wallach JD, Dhruva SS, Herrin J, Quinto K, et al. Emulating the GRADE trial using real world data: retrospective comparative effectiveness study. BMJ. 2022;379. https://doi.org/10.1136/bmj-2022-070717.
49. Merola D, Young J, Schrag D, Lin K, Robert N, Schneeweiss S. Oncology Drug Effectiveness from Electronic Health Record Data Calibrated Against RCT Evidence: The PARSIFAL Trial Emulation. Clin Epidemiol. 2022;14:1135–44. https://doi.org/10.2147/CLEP.S373291.
50. Leening MJ, Boersma E. The perpetual need of randomized clinical trials: challenges and uncertainties in emulating the REDUCE-AMI trial. Eur J Epidemiol. 2024;39:343–7. https://doi.org/10.1007/s10654-024-01127-3.
51. Walker B, Ray HE, Shao P, D'Ambrosio C, White C, Walker MS. Comparing prospectively assigned trial and real-world lung cancer patients. J Comp Eff Res. 2024;13(7):e230176. https://doi.org/10.57264/cer-2023-0176.
52. Petito LC, García-Albéniz X, Logan RW, Howlader N, Mariotto AB, Dahabreh IJ, et al. Estimates of Overall Survival in Patients With Cancer Receiving Different Treatment Regimens: Emulating Hypothetical Target Trials in the Surveillance, Epidemiology, and End Results (SEER)–Medicare Linked Database. JAMA Netw Open. 2020;3(3):e200452–e200452. https://doi.org/10.1001/jamanetworkopen.2020.0452.
53. McGrath LJ, Nielson C, Saul B, Breskin A, Yu Y, Nicolaisen SK, et al. Lessons Learned Using Real-World Data to Emulate Randomized Trials: A Case Study of Treatment Effectiveness for Newly Diagnosed Immune Thrombocytopenia. Clin Pharmacol Ther. 2021;110(6):1570–8. https://doi.org/10.1002/cpt.2399.
54. Khosrow-Khavar F, Desai RJ, Lee H, Lee SB, Kim SC. Tofacitinib and Risk of Malignancy: Results From the Safety of Tofacitinib in Routine Care Patients With Rheumatoid Arthritis (STAR-RA) Study. Arthritis Rheumatol. 2022;74(10):1648–59. https://doi.org/10.1002/art.42250.
55. Gallivan MD, Garcia KM, Torres AZ, Lum F, Li C, Mbagwu M, et al. Emulating VIEW 1 and VIEW 2 Clinical Trial Outcome Data Using the American

Academy of Ophthalmology IRIS Registry. Ophthalmic Surg Lasers Imaging Retina. 2023;54(1):6–14. https://doi.org/10.3928/23258160-20221214-01.
56. Yoon D, Jeong HE, Park S, You SC, Bang SM, Shin JY. Real-world data emulating randomized controlled trials of non-vitamin K antagonist oral anticoagulants in patients with venous thromboembolism. BMC Med. 2023;21(375). https://doi.org/10.1186/s12916-023-03069-1.
57. Jang HY, Kim IW, Oh JM. Using real-world data for supporting regulatory decision making: Comparison of cardiovascular and safety outcomes of an empagliflozin randomized clinical trial versus real-world data. Front Pharmacol. 2022;13. https://doi.org/10.3389/fphar.2022.928121.
58. Jin X, Ding C, Hunter DJ, Gallego B. Effectiveness of vitamin D supplementation on knee osteoarthritis - A target trial emulation study using data from the Osteoarthritis Initiative cohort. Osteoarthr Cartil. 2022;30(11):1495–505. https://doi.org/10.1016/j.joca.2022.06.005.
59. Merola D, Campbell U, Gautam N, Rubens A, Schneeweiss S, Wang SV, et al. The Aetion Coalition to Advance Real-World Evidence through Randomized Controlled Trial Emulation Initiative: Oncology. Clin Pharmacol Ther. 2023;113(6):1217–22. https://doi.org/10.1002/cpt.2800.
60. Antoine A, Pérol D, Robain M, Bachelot T, Choquet R, Jacot W, et al. Assessing the real-world effectiveness of 8 major metastatic breast cancer drugs using target trial emulation. Eur J Cancer. 2024;213: 115072. https://doi.org/10.1016/j.ejca.2024.115072.
61. Signori A, Ponzano M, Kalincik T, Ozakbas S, Horakova D, Kubala Havrdova E, et al. Emulating randomised clinical trials in relapsing-remitting multiple sclerosis with non-randomised real-world evidence: an application using data from the MSBase Registry. J Neurol Neurosurg Psychiatry. 2024;95(7):620–5. https://doi.org/10.1136/jnnp-2023-332603.
62. D'Andrea E, Schneeweiss S, Franklin JM, Kim SC, Glynn RJ, Lee SB, et al. Efficacy Versus Effectiveness: The HORIZON-Pivotal Fracture Trial and Its Emulation in Claims Data. Arthritis Rheumatol. 2024;n/a(n/a). https://doi.org/10.1002/art.42968.
63. Pawel S, Heyard R, Micheloud C, Held L. Replication of null results: absence of evidence or evidence of absence? eLife. 2024;12:RP92311. https://doi.org/10.7554/eLife.92311.
64. Micheloud C, Held L. The replication of equivalence studies. Biom J. 2024. https://doi.org/10.1002/bimj.202300232.

Publisher's Note