A Benchmark, Expand, and Calibration (BenchExCal) Trial Emulation Approach for Using Real-World Evidence to Support Indication Expansions: Design and Process for a Planned Empirical Evaluation

Shirley V. Wang[1,*], Massimiliano Russo[1,2], Robert J. Glynn[1], Marie C. Bradley[3], Jiwei He[4], John Concato[5] and Sebastian Schneeweiss[1]

Real-world evidence involving healthcare database studies is well established for making causal inferences in post-market drug safety studies and methods, data, and research infrastructure for evaluating effectiveness have advanced in recent years. The rapidly expanding field of etiologic research using insurance claims and electronic health records databases is being evaluated for supporting effectiveness claims. One such use case to support regulatory decision-making on effectiveness is for expanding indications beyond existing effectiveness claims. Confidence in the validity of findings from cohort studies conducted using databases (hereafter "database study") to support indication expansions could be increased through a structured benchmarking process of an initial database study against RCT evidence followed by calibration of a subsequent database study based on differences in results observed in the initial RCT-database pair. This paper proposes a benchmark, expand, and calibration (BenchExCal) approach to trial emulation and describes the design and process for evaluating the performance of the approach through both simulation studies; five planned empirical examples are also described. The project will provide insights regarding how a first-stage benchmarking emulation of a completed trial for an existing indication can be used to calibrate, increase confidence, and improve interpretation of the results for a second-stage emulation of a hypothetical trial that could potentially provide evidence for an expanded indication. Although the examples have been selected to provide a variety of learnings, five use cases do not address all clinical and data scenarios that may be encountered when seeking a supplemental indication for a marketed drug.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THIS TOPIC?

☑ Real-world evidence involving healthcare database studies is well established for making causal inferences in post-market drug safety studies and methods, data, and research infrastructure for evaluating effectiveness have advanced in recent years.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ This paper proposes a benchmark, expand, and calibration (BenchExCal) approach to trial emulation and describes the design and process for evaluating the performance of the approach through both simulation studies; five planned empirical examples are also described.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ Confidence in the validity of findings from cohort studies conducted using databases (hereafter "database study") to support indication expansions could be increased through a structured benchmarking process of an initial database study against RCT evidence followed by calibration of a subsequent database study based on differences in results observed in the initial RCT-database pair.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ The project will provide insights regarding how a first-stage benchmarking emulation of a completed trial for an existing indication can be used to calibrate, increase confidence, and improve interpretation of the results for a second-stage emulation of a hypothetical trial that could potentially provide evidence for an expanded indication. Although the planned examples have been selected to provide a variety of learnings, five use cases do not address all clinical and data scenarios that may be encountered when seeking a supplemental indication for a marketed drug.

Received October 10, 2024; accepted February 10, 2025. doi:10.1002/cpt.3621

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA; [2]Ohio State University, Columbus, Ohio, USA; [3]Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland, USA; [4]Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA; [5]Department of Medicine, Yale School of Medicine, New Haven, Connecticut, USA. *Correspondence: Shirley V. Wang (swang1@bwh.harvard.edu)

Healthcare database studies are well established for making causal inferences in post-market drug safety studies, and methods, data, and research infrastructure for evaluating effectiveness have advanced in recent years.[1,2] Advances in causal analysis, aided by the emulation of a hypothetical target trial[3] and analytic techniques such as optimized propensity score estimation,[4] address concerns regarding potential biases in non-interventional study designs using secondary healthcare data. These advances help to address biases frequently cited in critical reviews of non-randomized studies.[5,6]

Use cases were identified for healthcare database studies to potentially support regulatory decision-making on effectiveness, including supplemental indications or other labeling expansions after a successful initial new drug application.[7] One approach to promote rigor in the study design, measurements of key variables, analysis, and ultimately findings from a database study to support labeling expansions would involve benchmarking the results of a database study against the results of a previous trial. For example, investigators can design a database study to emulate an RCT that was conducted for a prior indication, implement the study using the proposed database, and then benchmark the database study results against the RCT.

Benchmarking database studies against RCTs that they are designed to emulate is not a new concept.[8,9] Benchmarking can involve anything from formal pre-specified criteria for comparison to informal post hoc deep dives into assessment of the potential impact of design differences, bias, and clinical context.[10–13] However, prior work on benchmarking and transportability of results between RCTs and database studies has required several assumptions that may be untenable in many practical settings. For example, in addition to the usual assumptions of conditional exchangeability (e.g., no unmeasured confounding) between treatment arms within the RCT and database study, consistency, and positivity, existing benchmarking and transportability methods have required an assumption that trial participation does not have an effect on the outcome other than through treatment.[8] In practice, participation in a drug trial often involves frequent follow-ups designed to maximize adherence. This, as well as the Hawthorne effect,[14] contributes to the so-called efficacy vs effectiveness gap.[15] Another assumption that is unlikely to be met in practical settings is that there are no differences in the measurement of key study variables between the RCT and database study.[8] An exception may be trials nested within the same healthcare data used for the observational analysis, but the rare occurrence of such trials limits its application.[13]

We describe a benchmark, expand, and calibrate (BenchExCal) approach that can potentially inform decisions on expanding indications for marketed medications that allows for some variation in measurement, follow-up, or other design differences that cannot be aligned between an RCT and a database study that emulates it. The first step of BenchExCal is a demonstration of the ability to emulate the trial(s) used for an initial indication reasonably closely, for benchmarking. Second, learnings from emulation of the trial(s) for the initial indication are used to plan, execute, and interpret database studies conducted within the same database(s), using highly similar measurements, design, and analytic strategy to inform decisions on expanded populations, subgroups, or outcomes. Next, a calibration sensitivity analysis is applied to integrate knowledge of the divergence observed in the initial RCT and database study into the observed results of the second database study.

Here, divergence is defined as the observed net difference in results between an RCT-database study pair, stemming from multiple causes including residual confounding, misclassification of the outcome, population differences, reduced adherence, etc. The approach combines confidence gained from (successfully) benchmarking the data and analytic methods against an RCT with sensitivity analyses conducted to calibrate and interpret the second database study results within the context of prior information on any small divergence observed in a similar RCT-database study pair from the initial step.

A key advantage of BenchExCal is that the interpretation of the evidence for the expanded indication is informed by learnings from a first-stage database study emulation of an actual RCT for the drug under consideration. The BenchExCal approach, as with quantitative bias analysis[16,17] and negative control methods,[18,19] attempts to quantify the direction, magnitude, and uncertainty around systematic sources of error. Importantly, however, BenchExCal focuses on quantifying the net effect of systematic differences, stemming not only from biases within the database study, but also from differences in participation, design, and measurement between an RCT and the database study designed to emulate it.

This paper describes the planned design and process for evaluating the performance of BenchExCal through both simulation and empirical examples.

METHODS

BenchExCal approach to trial emulation

The RCT-DUPLICATE demonstration project[11,12,20–23] increased our understanding of when and how database studies can come to similar causal conclusions on treatment effects as randomized clinical trials (RCT). In a pre-defined process, the study compared the results of 32 RCT-database study pairs where the database studies were explicitly designed to emulate RCTs that were actually conducted.[11] The results of RCT-DUPLICATE support the premise that when RCT study designs can be closely emulated and key confounding factors and endpoints measured reliably, results from database studies highly correlate with those of RCTs ($r = 0.93$).[11,21,22,24]

Nonetheless, some aspects of RCT designs are difficult to emulate with secondary data collected as part of routine clinical care.[11,21,24] While RCTs that study the efficacy of drugs are often designed to isolate the treatment experiment from particular practice patterns, database studies evaluate a mix of the drug effect and the system in which it is used.[21] Even when database studies use the same pre-defined inclusion/exclusion criteria, including age range and sex, RCT participants are usually younger, and fewer are female, which can have implications if the effects of the drug vary based on these characteristics.[22] Some exclusions are difficult to emulate

Table 1 Selected examples of RCT emulation challenges encountered in RCT-DUPLICATE

Emulation challenge	Specific issues	Trial example
Differences in study population when treatment effect modification is present	Despite the same sex and age range criteria, we still have different distributions of these characteristics	Most pre-approval trials
Variations in co-medications in multinational trials	Data sources limited to the US	PLATO
Long duration of use paired with time-varying hazards (induction periods)	Medication adherence in clinical practice is often substantially shorter	HORIZON-Pivotal, DAPA-CKD, VERO
Dose titration during follow-up	Lack of granular data; in-hospital start	TRITON-TIMI, PLATO, ISAR-REACT5
Outcome assessment	Low specificity often leads to bias toward to null	PARADIGM-HF
Run-in phases with selective drop out in RCTs and low adherence in clinical practice	Almost impossible in non-experimental settings	Most COPD trials
Placebo controls	How to define time zero? Are non-users too different to possibly balance outcome predictors?	Most pre-approval trials

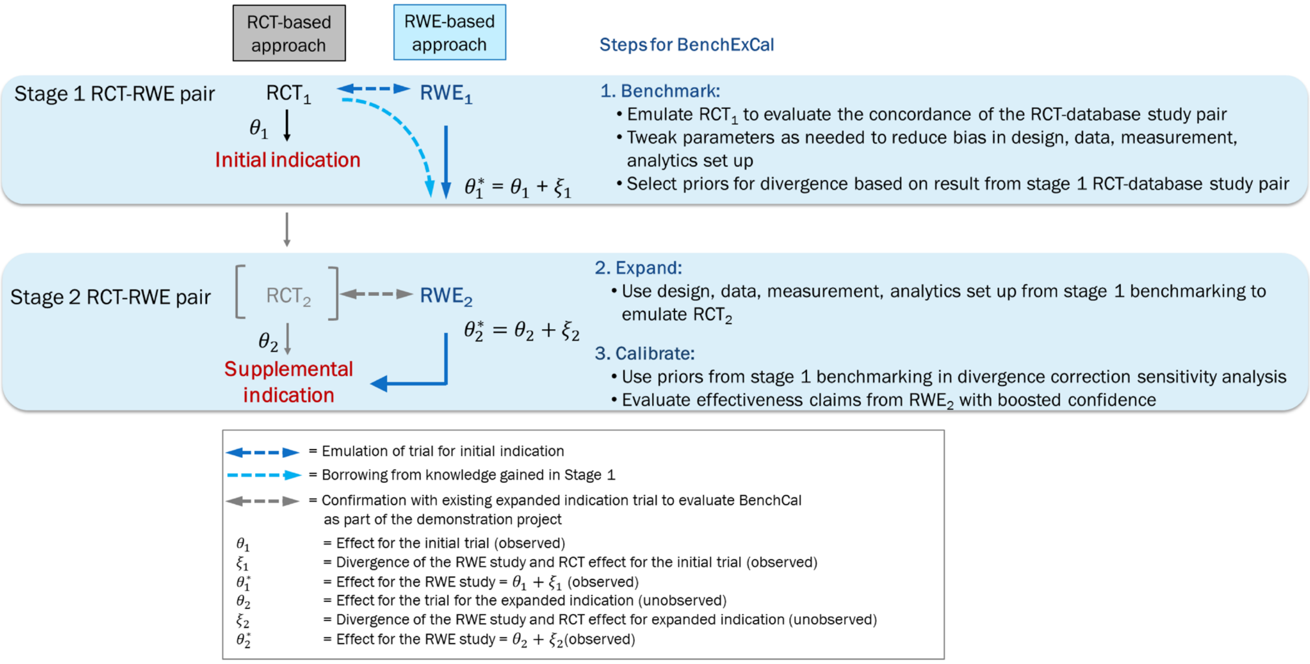
RCT-based approach
RWE-based approach
Steps for BenchExCal
Stage 1 RCT-RWE pair
RCT_1
RWE_1
θ_1
Initial indication
$\theta_1^* = \theta_1 + \xi_1$
1. Benchmark:
• Emulate RCT_1 to evaluate the concordance of the RCT-database study pair
• Tweak parameters as needed to reduce bias in design, data, measurement, analytics set up
• Select priors for divergence based on result from stage 1 RCT-database study pair
Stage 2 RCT-RWE pair
RCT_2
RWE_2
$\theta_2^* = \theta_2 + \xi_2$
θ_2
Supplemental indication
2. Expand:
• Use design, data, measurement, analytics set up from stage 1 benchmarking to emulate RCT_2
3. Calibrate:
• Use priors from stage 1 benchmarking in divergence correction sensitivity analysis
• Evaluate effectiveness claims from RWE_2 with boosted confidence
= Emulation of trial for initial indication
= Borrowing from knowledge gained in Stage 1
= Confirmation with existing expanded indication trial to evaluate BenchCal as part of the demonstration project
θ_1 = Effect for the initial trial (observed)
ξ_1 = Divergence of the RWE study and RCT effect for the initial trial (observed)
θ_1^* = Effect for the RWE study = $\theta_1 + \xi_1$ (observed)
θ_2 = Effect for the trial for the expanded indication (unobserved)
ξ_2 = Divergence of the RWE study and RCT effect for expanded indication (unobserved)
θ_2^* = Effect for the RWE study = $\theta_2 + \xi_2$(observed)

Figure 1 BenchExCal approach.

due to the lack of granular data or because the RCT criteria are difficult to operationalize, for example, "unlikely to die in the next year." Additionally, the reality of suboptimal drug adherence in clinical practice often makes it difficult to emulate treatment effects that occur after a prolonged induction period. **Table 1** summarizes some of the emulation challenges observed in prior work.

It has been argued that if a causal study design and analytic approach were applied in a reliable, and relevant data source to emulate a previously conducted RCT for the drug of interest and found to closely replicate results, this setup could then be used to evaluate other closely related effectiveness questions of interest.[10] Potential examples include evaluation of a clinical rather than biomarker endpoint, a different clinical endpoint, or expansion of the population (e.g., children, more/less severely ill patients).

The proposed BenchExCal approach involves sensitivity analyses that are conducted to evaluate the robustness of results from a primary frequentist analysis for an observational database study. The approach is illustrated in **Figure 1**. Let θ_1 denote the true treatment effect of interest in stage 1 (e.g., log hazard ratio), and $\hat{\theta}_1$ the estimate observed in a stage 1 RCT. $\hat{\theta}_1$ is a consistent estimator of θ_1 ensured by randomization in an RCT. Let $\hat{\theta}_1^*$ denote the estimate observed in a stage 1 database study. We expect the quantity targeted by $\hat{\theta}_1^*$, denoted by θ_1^*, to deviate from θ_1 due to residual confounding or other emulation differences. We denote this divergence by ξ_1 such that $\xi_1 = \theta_1^* - \theta_1$, and its estimate $\hat{\xi}_1 = \hat{\theta}_1^* - \hat{\theta}_1$. Similarly, let θ_2 denote the true treatment effect of interest in stage 2, θ_2^* the quantity targeted by the database study in stage 2, and ξ_2 the divergence in stage 2 ($\xi_2 = \theta_2^* - \theta_2$). $\hat{\theta}_2^*$ can be estimated from the database study in stage 2. Our goal is to account for the effect of the divergence ξ_2 in estimating θ_2 in the absence of a stage 2 RCT, using information from the database study in stage 2 and the divergence in stage 1.

If there are differences in results between the first-stage RCT-database pair ($\hat{\xi}_1$ is larger than expected from sampling error), then investigators have the opportunity to revise aspects of design in a way that improves the database study emulation of the stage 1 RCT and minimizes the divergence, ξ_1. Once the study design and analytic framework are settled in stage 1, any remaining observed divergences in results between the RCT-database study pair could be used to inform the analysis and interpretation of the results for the second stage database study.

Calibration step

Divergences in stage 1 could stem from differences in the emulation of the study design and its measurements, systematic bias, or random error. These sources of variation in results between an RCT and database study often cannot be fully disentangled empirically.[11,21,24] However, under the assumption that such divergences would similarly affect a similarly designed second-stage database study for a new indication, the divergences in a first-stage database study emulation of a published RCT ($\widehat{\xi_1}$) can be used to calibrate the stage 2 estimate. Specifically, such differences may be used heuristically to elicit prior distributions for the divergences in stage 2 (ξ_2).

Given that the rate of outcomes and effect sizes may differ in magnitude between the stage 1 and stage 2 questions, it will be important to appropriately scale the estimated divergence parameter when moving from stage 1 to stage 2. Ideally, the observed divergence in stage 1 would be standardized dividing by the standard deviation and we would assume that the standardized divergence in stage 1 $\delta_1 = \xi_1 / \sqrt{var(\theta_1) + var(\theta_2^*)}$ ≈ the standardized divergence in stage 2 $\delta_2 = \xi_2 / \sqrt{var(\theta_2) + var(\theta_2^*)}$. However, typically the $var(\theta_2)$ is unavailable because the stage 2 trial has not been implemented. Therefore, to express the divergence of stage 2 on an appropriate scale, we use the equality $\xi_2 = \sqrt{var(\theta_2^*)} / \sqrt{var(\theta_1^*)}\xi_1$, which only involves observed quantities. Thus, the observed standardized divergence in log hazard ratio from stage 1 can be used to estimate the standardized divergence for stage 2, via $\widehat{\xi_2} = \sqrt{var\left(\widehat{\theta_2^*}\right)} / \sqrt{var\left(\widehat{\theta_1^*}\right)}\widehat{\xi_{1.}}$

It will also be important to specify a probability distribution for the stage 2 standardized divergence. For the purposes of this demonstration project, we will use a normal distribution for the divergence in stage 2, where the mean is the scaled standardized divergence from the stage 1 RCT-database study pair $\widehat{\xi_2}$. The variance of the normal distribution used for the prior distribution will be equal to the pooled variance of the divergence in the stage 1 RCT-database study pair multiplied by the ratio of the variances for the stage 1 and stage 2 database studies $\left(\left(var(\theta_1) + var(\theta_1^*)\right) \frac{var(\theta_2^*)}{var(\theta_1^*)}\right)$. With these assumptions the distributions of the adjusted stage 2 estimator is normal with mean $= \widehat{\theta_2^*} + \widehat{\xi_2}$ and variance $= var\left(\widehat{\theta_2^*}\right) + \left(var\left(\widehat{\theta_1^*}\right) + var\left(\widehat{\theta_1}\right)\right)\frac{var\left(\widehat{\theta_2^*}\right)}{var\left(\widehat{\theta_1^*}\right)}$.

Prior distributions for the divergence in the stage 2 setting could then be used in Bayesian sensitivity analyses that calibrate the main frequentist second-stage database study estimate under different assumptions.

Using a range of estimates for ξ_2, including the estimate informed by the stage 1 analysis, we can identify a "tipping point"[25] at which the regulatory conclusion would be changed and consider the plausibility of the priors around the tipping point from both clinical and methodologic perspectives. This range of priors could include the 5th, 25th, 50th, 75th, 95th levels of the confidence interval around the estimated ξ_2, to show the potential range in estimates after accounting for divergences observed between a similar prior RCT and database study. Also of interest may be the tipping point value of ξ_2 that would produce a result that would lead the stakeholders to make a different decision. If such a tipping point value is far away from the priors for ξ_2 informed by the stage 1 study, this may boost confidence in claims of effectiveness for the secondary indication even in the absence of a second-stage trial. Although P-value thresholds and other such cutoffs can be somewhat arbitrary,[26] decision-makers are often faced with the need to make a binary decision. Therefore, decision-changing thresholds must be defined in conjunction with the relevant decision-making organization and aligned with the specific context being considered.[25,27,28]

Transparent and reproducible approach for evaluating BenchExCal

Following the established multi-step RCT-DUPLICATE process for transparent and reproducible trial emulation, we will develop database study protocols for a sample of first-stage trial emulations. We will conduct a series of data checks at each pre-specified step in the process, which include an assessment of power (required to be at least equal to the trial) and balance on key confounders (as captured by standardized differences and c-statistics after propensity score matching). At each step in this project, the interim protocol and data checks will be reviewed before moving on to the next step.[29] At no point during this process will outcomes be evaluated stratified by exposure status. Once a protocol is complete, it will be pre-registered on clinicaltrials.gov with details of the data checks completed up until that point. After registration, the pre-specified primary, secondary, and sensitivity analyses (including the BenchExCal sensitivity analyses) will be conducted. If additional changes to the design or analyses seem warranted after seeing the results, these changes will be first documented with an amendment to the registered protocol before they will be executed.

The observed net difference in results between the stage 1 RCT-database study pair will be used to select Bayesian priors for the divergence in stage 2, under the assumption that the clinical setting and the design and measurement issues of the emulations are similar enough that the expected divergence is comparable in the two stages. We will then follow the same multi-step process to emulate the second stage trial, using learnings from the first stage to set up the design and analysis plan. Both the empirical results of the second-stage trial emulation and the divergence-adjusted results will be presented.

Simulation

We plan to evaluate BenchExCal through simulations where the true treatment effects and the relative contributions of design emulation differences and bias are known. We will use prior trials from the RCT-DUPLICATE library of 32 trial emulations as empirical data to form the basis of these simulations. The simulations will include varying degrees of unmeasured confounding and emulation differences in design and measurement, which collectively contribute to the observed divergence in results between the simulated RCT-database study pairs.

Empirical evaluation

We will explore the empirical performance of BenchExCal through evaluation of the concordance of RCT-database pairs in selected case studies, choosing sets of stage 1/stage 2 trials that evaluate the same drug but are conducted to support effectiveness claims for separate indications.

To evaluate the performance of BenchExCal, we will include sets of trials where both the first- and second-stage trials have been completed. This approach will allow us to evaluate the method and how well the adjustment for divergences affects the concordance of results in the second stage for case studies where the second-stage trial results have been published.

In practice, when implementing BenchExCal, the second stage trial may or may not be conducted at some point in the future. Therefore, we will also include a pair of trials where the second stage trial is ongoing at the time the second stage database study is implemented. This strategy will allow us to evaluate how well the approach facilitates the prediction of the result for the second stage trial.

Table 2 Selection of example RCTs

	DOACs		Glucose-lowering drugs	
	Excluded	Remaining	Excluded	Remaining
Trials identified in search		152		419
Comparator not measurable	67	85	57	362
Outcome not measurable	7	78	163	199
Inclusion–exclusion not measurable	2	76	36	163
Other exclusion	10	66	5	158

DOAC, direct oral anticoagulant.

Table 3 Initial set of selected trials for planned empirical evaluation of BenchExCal trial emulation approach

(A) RCTs of direct oral anticoagulants (DOACs)

						Indication change from Stage 1 to 2	
NCT number	Trial Acronym	Type	Indication	Exposure	Comparator	Change in surgical site: Hip to knee	Change in population: from surgical patients to patients with prior VTE
NCT05083455	RECORD1	Sup	Reduced post-operative (hip replacement) VTE	Rivaroxaban	Enoxaparin	Stage 1	Stage 1 (pooled)
NCT00361894	RECORD3	Sup	Reduced post-operative (knee replacement) VTE	Rivaroxaban	Enoxaparin	Stage 2	
NCT00440193	EINSTEIN-DVT	NI	In patients with DVT: prevent a recurrent VTE	Rivaroxaban	Warfarin		Stage 2

(B) RCTs of glucose-lowering drugs

						Indication change from Stage 1 to 2		
NCT number	Trial	Type	Indication	Exposure	Comparator	Change in drug formulation: Injectable to oral		Change in population: from patients with HFpEF to HFrHF
NCT01720446	SUSTAIN 6	NI	MACE	Semaglutide injectable	Placebo	Stage 1	Stage 1	
NCT02692716	PIONEER 6	NI	MACE	Semaglutide oral	Placebo	Stage 2		
NCT03914326	SOUL[a]	Sup	Prevention of MACE	Semaglutide oral	Placebo		Stage 2	
NCT03057977	EMPEROR preserved	Sup	Prevention of HHF in pts with HFpEF	Empagliflozin	Placebo			Stage 1
NCT03057951	EMPEROR reduced	Sup	Prevention of HHF in pts with HFrEF	Empagliflozin	Placebo			Stage 2

HFpEF, Heart failure with preserved ejection fraction; HFrEF, Heart failure with reduced ejection fraction; HHF, hospitalization for heart failure; MACE, non-fatal myocardial infarction, non-fatal stroke, cardiovascular death; NI, non-inferiority hypothesis testing; PE, pulmonary embolism; Sup, superiority hypothesis testing; VTE, venous thromboembolism. [a]Ongoing trial, estimated completion July 2024.

Trial selection

We conducted a search of clinicaltrials.gov in two therapeutic areas, direct oral anticoagulants (DOAC) and antidiabetic agents. Our search terms are provided in the **Data S1**. Based on this search, we identified 152 DOAC trials and 419 diabetes trials (**Table 2**). We excluded trials where a comparator (or active comparator proxy for placebo) could not be identified, where the outcome was not measurable in claims data, where a key eligibility criterion was not measurable, or where other trial elements did not fit our criteria (e.g., < 100 participants, trial conducted solely outside of the United States, pediatric or single-arm study, drug removed from the market, or drug approved too recently for uptake in claims databases).

From the remaining trials, we examined the drugs, comparators, and indications to identify groups of trials that could be used to create stage 1 and stage 2 pairs. The research team discussed potential learnings to be gained from the set of potential trial pairs, and after preliminary feasibility counts, we selected the top five pairs of interest (**Table 3**). What we hope to learn from this set of trials is how net-difference-calibrated results could potentially inform regulatory decision-making for a supplemental indication, building on accumulated evidence for anticipated divergences from prior, similar RCT-database study pairs. If one or more sets from the initially selected trials do not pass the data checks, we will document the reason for dropping the set and consider developing protocols and data checks for another set.

Concordance metrics

Assessment of concordance in results between RCT-database study pairs will use the metrics previously implemented[12] in the RCT-DUPLICATE initiative,[11] namely statistical significance agreement, estimate agreement, and standardized difference agreement. In this project, "full" statistical significance agreement was defined by estimates and confidence intervals on the same side of null; estimate agreement was defined by whether estimates for the trial emulation fell within the 95% CI for the trial results; standardized difference agreement between treatment effect estimates from trials and emulations was defined by standardized differences $|Z| < 1.96$ ($Z = \frac{\widehat{\Theta}_{RCT} - \widehat{\Theta}_{RWE}}{\sqrt{\widehat{\sigma}^2_{RCT} + \widehat{\sigma}^2_{RWE}}}$ where $\widehat{\Theta}$ are the treatment effect estimates and the $\widehat{\sigma}^2$ are associated variances). In addition, "partial" significance agreement was defined as meeting pre-specified non-inferiority criteria even though a more highly powered database study may have indicated superiority.

As part of the BenchExCal approach, if we do not meet at least one of the binary agreement metrics described above in the first-stage emulation of a trial, then we will determine that we do not have sufficient confidence in the design and measurement setup for that clinical scenario and will not proceed to the second stage emulation.

Considerations on effect measure modification

Differences may exist in the distribution of risk factors for the outcome in a trial population and the patients in a database study that emulates the trial design. Re-weighting and other transportability methods have been proposed to align the distributions of risk factors for the outcome in an RCT and database study prior to benchmarking the results.[8] Such methods are particularly useful in nested RCT-database study settings where the measurement of key risk factors is the same in both studies. However, in most practical settings, RCTs and the database studies that emulate them will be using data that are not captured in the same way (e.g., primary vs. secondary data collection). This means that there will often be notable differences in terms of measurement characteristics such as sensitivity, specificity, positive, and negative predictive values for risk factors measured in RCTs vs. databases. Transportability methods performed on data with differentially measured factors may provide misleading assurance of similarity in distributions of risk factors.

Additionally, while effect measure modification on the absolute scale (e.g., risk differences) can be present due to differences in risk factor distribution, in the absence of effect measure modification for the effect of treatment on the relative scale (e.g., hazard ratios), such differences in risk factor distribution will not influence estimates of divergence or recalibrated database study estimates for relative risk or hazard ratios of treatment in second stage sensitivity analyses. Therefore, for the purposes of this demonstration project, we focus on relative measures of effect for both stage 1 and stage 2 trials for this empirical demonstration project and will implement methods to align the distribution of only age and sex as sensitivity analyses because these demographic variables are strong risk factors for cardiovascular outcomes and are likely to be well captured regardless of data source. In our primary analyses, we assume that there is no effect measure modification on the relative scale.

Patient and public involvement statement

Patients were not involved in the conception of the proposed BenchExCal approach or the evaluation plan described in this paper. The study research questions, design, and outcome measures for the 5 empirical examples were selected from previously completed or ongoing clinical trials after applying a systematic search and filtering process through collaborative discussion with members of the FDA. The results of this methods demonstration and evaluation study will be disseminated to the public through a public workshop, presentations at scientific conferences, and publication in peer-reviewed journals.

RESULTS

The initial selected set of DOAC RCTs to target include four trials with published results: RECORD1,[30] RECORD3,[31] EINSTEIN-PE,[32] and EINSTEIN-DVT[33] (**Table 3A**). Three of these trials were previously emulated as part of RCT-DUPLICATE. From this set of trials, the first and second stage pairs involve a change in surgical site and a change in population and comparator.

Change in surgical site

Moving from RECORD1 as a stage 1 trial which evaluated rivaroxaban vs. enoxaparin for prevention of venous thromboembolism after hip surgery 1 to the same comparison for knee surgery as a stage 2 supplemental indication is a relatively small leap in population, where the only major difference is surgical site yet most clinical parameters remain unchanged.[34] In this setting, could a net-difference-adjusted second-stage database study provide sufficient evidence of effectiveness?

Change in population and comparator

Moving from RECORD1 and RECORD3 which evaluated venous thromboembolism post hip and knee surgery to evaluating rivaroxaban for prevention of recurrent venous thromboembolism, we would investigate the issue of how the 1-stage findings might bolster confidence in a second-stage database study to potentially inform a supplemental indication when there is a bigger change in population/indication. This example is one where the assumption of similarity in divergences between stage 1 and stage 2 is more difficult to make because of the relatively large jump in clinical settings and risk factors from incident post-surgical venous thromboembolism to prevention of recurrent thromboembolism. Additionally, not only does the patient population change for this example, but the comparator therapy differs between stage 1 and stage 2 as well.

The initial set of diabetes RCTs includes five trials: SUSTAIN6,[35] PIONEER6,[36] SOUL,[37] EMPEROR reduced,[38] and EMPEROR preserved[39] (**Table 3B**). None of these trials were previously emulated by the RCT-DUPLICATE team. Four of the trials are completed and have published results; SOUL is an ongoing trial with a targeted primary completion date in July 2024. We use these examples to examine different types of changes in indication, including a change in the route of administration and a change in population.

Change in route of administration

Moving from SUSTAIN6 as a stage 1 non-inferiority trial that evaluated injectable semaglutide vs. placebo on the risk of a composite outcome of non-fatal myocardial infarction, non-fatal stroke, cardiovascular death (MACE) to evaluation of oral semaglutide and the risk of MACE changes the formulation for the drug of interest from stage 1 (injectable vs. placebo) to stage 2 (oral vs. placebo). As second-stage trials, we will use PIONEER6, a non-inferiority trial, and SOUL, an ongoing (as of July 2024) trial that targets superiority which evaluated oral semaglutide on the risk of MACE. These trials will be used to evaluate the utility of using the BenchExCal approach to incorporate prior knowledge about concordance in the results of an RCT-database study pair that evaluated injectable semaglutide, to increase confidence in a second, similarly designed database study to potentially support claims of effectiveness for oral semaglutide.

Change in population

Moving from EMPEROR preserved as a stage 1 trial, which compared empagliflozin to placebo on the risk of hospitalization for heart failure in patients with reduced ejection fraction (HFrEF), to the same comparison in patients with preserved ejection fraction (HfpEF) would change the population. A challenge with this set of trials is the potentially different confounding structure for patients who have preserved vs. reduced ejection fraction, as physicians may preferentially prescribe empagliflozin differently in ways related to baseline risk factors for heart failure hospitalization. This scenario may challenge the assumption that the divergence between RCT-database study pair results in stage 1 and stage 2 would be similar. Another challenge will be classification of patients as having preserved vs. ejection fraction because claims databases do not contain this information directly (although validated algorithms have been developed[40] which correctly classified reduced and preserved ejection fraction in 83% of the sample), and many commercial EHR databases currently have limited capture of extracted ejection fraction values in patient cohorts that would be trial eligible (unpublished data queries from multiple vendors). For this set of trials, we will focus on the diabetic subgroups reported in the trial results. This approach was selected because, after FDA approval of empagliflozin to prevent hospitalization for heart failure in 2022, there is likely a lag in the uptake of empagliflozin use in the heart failure patient population without diabetes that is compounded by the lag in refreshed data available from our healthcare databases. Accordingly, there has not been enough time to accrue data on patients with heart failure but without diabetes who are treated with empagliflozin.

DISCUSSION

The RCT-DUPLICATE initiative demonstrated the feasibility of generating similar conclusions from database studies and trials when the design and measurements are closely aligned.[11] Such alignment in design and measurements may not be possible in some clinical settings and the magnitude of divergence in results between a hypothetical RCT and database study due to design emulation differences and bias is not knowable absent a completed RCT.

The BenchExCal approach may offer insights into the potential magnitude of divergence in results between a similarly designed database study and RCT when planning a database study for a supplemental indication. The results of a first-stage trial emulation are intended to increase confidence in the second-stage emulation of a hypothetical target trial for an expanded indication by providing assurance that the design, data, and measurements were sufficient to closely emulate the first-stage trial. Or the results may highlight inadequacies of the selected design, measurements, or database to emulate a trial, which could result in not pursuing the second stage study using RWD for that particular clinical question. However, the first-stage results may also be used in a Bayesian sensitivity analysis that calibrates the second-stage database study results based on priors informed by the divergences observed in the first-stage trial emulation. “Tipping point” sensitivity analyses can be performed with different priors to aid in the evaluation of the degree to which plausible divergences between a hypothetical RCT and database study designed to emulate it might affect interpretation or decision-making for a second stage database study designed to provide evidence in support of effectiveness claims.

There are several limitations to the approach and this demonstration study. First, binary metrics for agreement between the results of an RCT-database pair are simplistic and cannot capture the many nuances that a deep dive into the context, design emulation differences, measurement differences, and consideration of sources or direction of potential residual confounding could reveal. Second, it will often not be possible to empirically disentangle the effects of design differences, biases, and chance, which means that the priors in BenchExCal will be based on divergences stemming from multiple factors, any one of which may be more aligned or less aligned between the first and second stage database study. The BenchExCal approach assumes that the net effect of the design differences and bias is similar for the stage 1 completed RCT-database pair and the stage 2 hypothetical RCT-database pair. Therefore, analogous to negative control[19] and quantitative bias analysis[16] approaches to sensitivity analyses, it would be important to evaluate a range of plausible priors in BenchExCal sensitivity analyses.

As a related issue, the appropriateness of the assumption that the divergence is similar between a stage 1 and stage 2 setting will depend on how “close” the clinical context is. The more similar the clinical context between a first- and second-stage question, the more likely that such an assumption will be met. Therefore, the parameters and interpretation of the stage 2 sensitivity analyses will require thoughtful justification and consideration from both clinical and methodological perspectives. When there is knowledge from validation studies or subset linkage to richer clinical data regarding the potential magnitude of misclassification for exposure or outcome or the magnitude of residual confounding within the initial and second-stage database studies, then BenchExCal may be used in conjunction with traditional QBA methods[16,17] that address biases within an observational study.

Third, a challenge we face in our selected use cases that emulate second-stage trials where RCT results have already been publicly reported is that there may be changes in how physicians prescribe drugs over the time frame that trial-eligible patients can be identified in healthcare databases due to the availability of the second stage trial results. Finally, the BenchExCal approach involves a bias-variance trade-off. The approach adjusts for expected divergences between an RCT and a database study but also makes the variance larger by incorporating uncertainty regarding the magnitude of potential divergences. The larger variance after the incorporation of this uncertainty complements the interpretation of tighter confidence intervals from the primary frequentist analysis that does not account for expected divergences.

That said, to our knowledge, this project is novel in terms of the plan to systematically demonstrate and evaluate the BenchExCal approach to trial emulation through both simulation and empirical examples. The project will provide insights regarding how a first-stage emulation of a completed trial for an initial indication can be used to increase confidence and inform the interpretation of a second-stage emulation of a hypothetical trial designed to provide evidence for a supplemental indication. Although examples have been selected to provide a variety of learnings, five use cases will not be able to address all clinical and data scenarios that may be encountered when seeking a supplemental indication for a marketed drug. Similar follow-on work will add to this library of learnings.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

FUNDING

This study was funded by a contract from the US Food and Drug Administration (75F40122C00154) to the Brigham and Women's Hospital (PI Dr. Wang and Schneeweiss). The content is solely the responsibility of the authors and should not be construed to represent the views or policies of the US Food and Drug Administration.

CONFLICTS OF INTEREST

Dr. Wang has consulted *ad hoc* for Cytel Inc., Exponent Inc., and MITRE a federally funded research center for the Centers for Medicare and Medicaid Services on unrelated work. Dr. Schneeweiss is the principal investigator of the FDA Sentinel Innovation Center funded by the FDA, 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. He is a consultant to Aetion Inc., a software manufacturer of which he owns equity. His interests were declared, reviewed, and approved by the Brigham and Women's Hospital and MGB HealthCare System in accordance with their institutional compliance policies. Dr. Concato is an Adjunct Professor of Medicine at Yale University School of Medicine. All other authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

All authors wrote the manuscript; S.V.W., M.R., R.J.G. and S.S. designed the research; S.V.W. and M.R. performed the research; S.V.W. analyzed the data.

1. Ball, R., Robb, M., Anderson, S.A. & Dal Pan, G. The FDA's sentinel initiative—a comprehensive approach to medical product surveillance. *Clin. Pharmacol. Ther.* **99**, 265–268 (2016).
2. Kurz, X., Perez-Gutthann, S. & ENCePP Steering Group. Strengthening standards, transparency, and collaboration to support medicine evaluation: ten years of the European network of Centres for Pharmacoepidemiology and pharmacovigilance (ENCePP). *Pharmacoepidemiol. Drug Saf.* **27**, 245–252 (2018).
3. Hernan, M.A., Sauer, B.C., Hernandez-Diaz, S., Platt, R. & Shrier, I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J. Clin. Epidemiol.* **79**, 70–75 (2016).
4. Schneeweiss, S., Rassen, J.A., Glynn, R.J., Avorn, J., Mogun, H. & Brookhart, M.A. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20**, 512–522 (2009).
5. Schneeweiss, S. & Patorno, E. Conducting real-world evidence studies on the clinical outcomes of diabetes treatments. *Endocr. Rev.* **42**, 658–690 (2021).
6. Suissa, S.S. & Suissa, S. Advanced approaches to addressing confounding and bias in pharmacoepidemiologic studies. In *Pharmacoepidemiology* 5th edn. (eds. Strom, B., Hennessey, S. & Kimmel, S.) 868–891 (Hoboken, NJ: John Wiley & Sons, 2012).
7. US_Food_and_Drug_Administration. Framework for FDA's real-world evidence program. <https://www.fda.gov/media/120060/download?attachment> (2018). Accessed February 28, 2025.
8. Dahabreh, I.J., Matthews, A., Steingrimsson, J.A., Scharfstein, D.O. & Stuart, E.A. Using trial and observational data to assess effectiveness: trial emulation, transportability, benchmarking, and joint analysis. *Epidemiol. Rev.* **46**, 1–16 (2024).
9. Dahabreh, I.J., Robins, J.M. & Hernan, M.A. Benchmarking observational methods by comparing randomized trials and their emulations. *Epidemiology* **31**, 614–619 (2020).
10. Matthews, A.A. *et al*. Benchmarking observational analyses before using them to address questions trials do not answer: an application to coronary thrombus aspiration. *Am. J. Epidemiol.* **191**, 1652–1665 (2022).
11. Wang, S.V. *et al*. Emulation of randomized clinical trials with nonrandomized database analyses. *JAMA* **329**, 1376–1385 (2023).
12. Franklin, J.M. *et al*. Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. *Clin. Pharmacol. Ther.* **107**, 817–826 (2020).
13. Matthews, A.A., Dahebreh, I.J., MacDonald, C.J. *et al*. Prospective benchmarking of an observational analysis in the SWEDEHEART registry against the REDUCE-AMI randomized trial. *Eur. J. Epidemiol.* **39**, 349–361 (2024).
14. McCambridge, J., Witton, J. & Elbourne, D.R. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J. Clin. Epidemiol.* **67**, 267–277 (2014).
15. Nordon, C. *et al*. The "efficacy-effectiveness gap": historical background and current conceptualization. *Value in Health* **19**, 75–81 (2016).
16. Lash, T.L., Fox, M.P., Cooney, D., Lu, Y. & Forshee, R.A. Quantitative bias analysis in regulatory settings. *Am. J. Public Health* **106**, 1227–1230 (2016).
17. Lash, T.L., Fox, M.P., MacLehose, R.F., Maldonado, G., McCandless, L.C. & Greenland, S. Good practices for quantitative bias analysis. *Int. J. Epidemiol.* **43**, 1969–1985 (2014).
18. Arnold, B.F. & Ercumen, A. Negative control outcomes: a tool to detect bias in randomized trials. *JAMA* **316**, 2597–2598 (2016).
19. Lipsitch, M., Tchetgen Tchetgen, E. & Cohen, T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* **21**, 383–388 (2010).
20. Franklin, J.M., Glynn, R.J., Martin, D. & S, S. Evaluating the use of nonrandomized real world data analyses for regulatory decision making. *Clin. Pharmacol. Ther.* **105**, 867–877 (2019).
21. Franklin, J.M., Glynn, R.J., Suissa, S. & Schneeweiss, S. Emulation differences vs. biases when calibrating real-world evidence findings against randomized controlled trials. *Clin. Pharmacol. Ther.* **107**, 735–737 (2020).
22. Franklin, J.M. *et al*. Emulating randomized clinical trials with nonrandomized real-world evidence studies. *Circulation* **143**, 1002–1013 (2021).
23. Franklin, J.M. & Schneeweiss, S. When and how can real world data analyses substitute for randomized controlled trials? *Clin. Pharmacol. Ther.* **102**, 924–933 (2017).
24. Rachel, H., Leonhard, H., Sebastian, S. & Shirley, V.W. Design differences and variation in results between randomised

trials and non-randomised emulations: meta-analysis of RCT-DUPLICATE data. *BMJ Med.* **3**, e000709 (2024).

25. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. <https://www.fda.gov/media/71512/download> (2010). Accessed May 22, 2024.
26. Wasserstein, R.L., Schirm, A.L. & Lazar, N.A. Moving to a World Beyond "$p<0.05$". *Am. Stat.* **73**(suppl 1), 1–19 (2019).
27. Thorlund, K. *et al*. Quantitative bias analysis for external control arms using real-world data in clinical trials: a primer for clinical researchers. *J. Comp. Eff. Res.* **13**, e230147 (2024).
28. Deloughery, E.P. & Prasad, V. If the IMPROVE-IT trial was positive, as reported, why did the FDA denied expanded approval for ezetimibe and simvastatin? An explanation of the tipping point analysis. *J. Gen. Intern. Med.* **33**, 1213–1214 (2018).
29. Wang, S.V. & Schneeweiss, S. Data checks before registering study protocols for health care database analyses. *JAMA* **331**, 1445 (2024).
30. Eriksson, B.I. *et al*. Rivaroxaban versus enoxaparin for thromboprophylaxis after hip arthroplasty. *N. Engl. J. Med.* **358**, 2765–2775 (2008).
31. Lassen, M.R., Ageno, W., Borris, L.C. *et al*. Rivaroxaban versus enoxaparin for thromboprophylaxis after total knee arthroplasty. *N. Engl. J. Med.* **358**, 2776–2786 (2008).
32. Büller, H.R., Prins, M.H., Lensin, A.W. *et al*. Oral rivaroxaban for the treatment of symptomatic pulmonary embolism. *N. Engl. J. Med.* **366**, 1287–1297 (2012).
33. EINSTEIN Investigators *et al*. Oral rivaroxaban for symptomatic venous thromboembolism. *N. Engl. J. Med.* **363**, 2499–2510 (2010).
34. Douketis, J.D. & Mithoowani, S. Prevention of venous thromboembolism in adults undergoing hip fracture repair or hip or knee replacement. <https://www.uptodate.com/contents/prevention-of-venous-thromboembolism-in-adults-undergoing-hip-fracture-repair-or-hip-or-knee-replacement>. UpToDate. 2024. Accessed May 2, 2024.
35. Marso, S.P. *et al*. Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *N. Engl. J. Med.* **375**, 1834–1844 (2016).
36. Husain, M. *et al*. Oral Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *N. Engl. J. Med.* **381**, 841–851 (2019).
37. A heart disease study of semaglutide in patients with type 2 diabetes (SOUL). <https://classic.clinicaltrials.gov/ct2/show/NCT03914326> (2023). Accessed September 22, 2023.
38. Packer, M. *et al*. Cardiovascular and renal outcomes with Empagliflozin in heart failure. *N. Engl. J. Med.* **383**, 1413–1424 (2020).
39. Anker, S.D. *et al*. Empagliflozin in heart failure with a preserved ejection fraction. *N. Engl. J. Med.* **385**, 1451–1461 (2021).
40. Desai, R.J., Lin, K.J., Patorno, E. *et al*. Development and preliminary validation of a Medicare claims-based model to predict left ventricular ejection fraction class in patients with heart failure. *Circ. Cardiovasc. Qual. Outcomes* **11**, e004700 (2018).