A Process for Emulating Comparative Oncology Trials with Real-world Evidence Studies

Authors: Janick Weberpals[1], Kenneth L. Kehl[2], Donna R. Rivera[3], Pallavi Mishra-Kalyani[3], Georg Hahn[1], Priyanka Anand[1], Yanina Natanzon[4], Andrew J. Belli[5], Ching-Kun Wang[5], Jenna Collins[6], JonathanKish[6], Janet Espirito[7], Nicholas J Robert[7], Robert J. Glynn[1], Sebastian Schneeweiss[1], Shirley V. Wang[1]

Author affiliations:

[1] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[2] Dana-Farber Cancer Institute, Boston, MA, USA

[3] Oncology Center of Excellence, US Food and Drug Administration, Silver Spring, MD, USA

[4] ConcertAI, Cambridge, MA, USA

[5] COTA, Inc., New York, NY, USA

[6] Flatiron Health, Inc., New York, NY, USA

[7] Ontada, Boston, MA, USA

Correspondence:

Shirley V. Wang, PhD

Division of Pharmacoepidemiology and Pharmacoeconomics,

Department of Medicine, Brigham and Women's Hospital, Harvard Medical School,

1620 Tremont Street, Suite 3030-R, Boston, MA 02120, USA

Phone: +1 617-278-0932

Fax: + 1 617-232-8602

Email: SWANG1@BWH.HARVARD.EDU

Article type: Review

Manuscript word count: 4,131 words / 8,000 words

Abstract word count: 248 words / 250 words

Tables: 3

Figures: 3

Supplementary material: Supplementary figures and material

Short running title: Emulation of Comparative Oncology Trials with Real-world Evidence (ENCORE)

Keywords: Oncology, Real-World Evidence, Trial emulation, EHR

Funding Statement: This project was supported by an FDA BAA contract with contract number 75F40122C00181.

Competing Interests Statement: Dr. Weberpals is now an employee of AstraZeneca and owns stocks in AstraZeneca. Dr. Kehl has received research funding from Meta, Inc. to his institution. Drs. Espirito and Robert are employees of McKesson and own McKesson stock. Dr. Wang has consulted ad hoc for Exponent Inc. and MITRE a federally funded research center for the Centers for Medicare and Medicaid Services on unrelated work. Dr. Schneeweiss is participating in investigator-initiated grants to the Brigham and Women's Hospital from Bayer and UCB unrelated to the topic of this study. He consults for and owns equity in Aetion Inc., a software manufacturer. He is an advisor to Temedica GmbH, a patient-oriented data generation company. His interests were declared, reviewed, and approved by the Brigham and Women's Hospital in accordance with their institutional compliance policies.

Data sharing statement: No data was analyzed as part of this project.

Analytic code sharing statement: Code to reproduce this manuscript, including all figures and tables, can be found at https://github.com/janickweberpals/encore-process-manuscript.

Proposed target journals: *CPT => JCO CCI => …*

Manuscript last updated: 2025-03-04 18:43:42.372054

Abstract

Real-world evidence (RWE) studies are increasingly used to complement evidence from randomized controlled trials (RCTs), contextualizing the effectiveness and safety of medical interventions as delivered in clinical practice. Advancements in the curation and accessibility of electronic health record data (EHR) have presented the opportunity to investigate disease domains such as oncology, where administrative healthcare claims databases alone are not fit-for-purpose. The RCT DUPLICATE initiative has previously enhanced understanding of when RWE studies come to the same causal conclusions by emulating trials in non-oncology indications. Here, we present the Emulation of Comparative Oncology Trials with Real-world Evidence (ENCORE) project, which aims to extend this work to oncology. ENCORE will emulate 12 RCTs in four oncology-specialized EHR databases across four different cancer indications, specifically non-small cell lung cancer, breast cancer, colorectal cancer, and multiple myeloma. It will place a special emphasis on systematic evaluation of fitness of data in relation to the study design and statistical analysis for a particular research question, and pre-registration of study protocols prior to analysis. Agreement of treatment effect estimates between RCTs and their respective emulations will be assessed using pre-specified criteria. Through extensive sensitivity analyses benchmarked against RCT results, the ENCORE project will inform understanding of how measurement, design, and analytic decisions influence the validity of oncological RWE studies.

Background

Randomized controlled trials (RCTs) have been the go-to methodology for establishing the efficacy and safety of medical products. Under the 21st Century Cures Act directive,[1] the Food and Drug Administration (FDA) established a framework to increasingly consider real-world evidence (RWE) generated from routine-care health data such as electronic health records (EHR) to evaluate and contextualize the comparative safety and effectiveness of novel cancer therapies.[2] Accounting for 21% of all approvals, oncology was the disease area with the most FDA drug approvals in 2023.[3] RWE has particularly important potential to complement evidence coming from RCTs in the field of precision oncology, where potential use cases include the assessment of effectiveness in specific patient populations that are underrepresented in RCTs, the construction of external control arms in single-arm trials where active recruitment to a randomized trial may not be feasible, or the discovery of biomarkers among pan-tumor populations that harbor specific genomic and immuno-pathological signatures.

However, the validity and transportability of results derived between RWE studies and RCTs depends on multiple factors. Frequently referenced limitations include, lack of baseline randomization, missing data, small study sizes, data discontinuity,[4, 5] adoption of changes in guidelines in real-world care and the inability to measure and emulate common eligibility criteria, including prognostic factors, and standardized response assessments in real-world data (RWD).[6] While there are already published examples of oncology trial emulations,[6–8] a systematic and scaled approach to emulate a diverse set of different oncology trials with multiple heterogeneous databases is necessary to gain confidence in the validity of RWE studies and to provide context as to which questions can be validly answered.

The RCT DUPLICATE initiative[9] increased our understanding of when RWE studies can come to causal conclusions on treatment effects by benchmarking results against RCTs under the assumption that each RCT finding reflects a causal treatment effect. In settings where the RCT designs could be emulated well, RWE studies came to the same conclusions.[10] However, prior work from RCT-DUPLICATE has focused primarily on emulating trials in non-oncology settings using claims databases.

The *Emulation of Comparative Oncology Trials with Real-world Evidence* (ENCORE) project[11] aims to extend this work to oncology. Studies in oncology come with their own unique set of challenges which must be systematically explored and understood. Building on a process co-developed with the FDA through RCT DUPLICATE,[9] this expansion to oncology will emulate 12 randomized clinical trials using multiple specialty oncology EHR data sources. The process will emphasize transparency and include documented data fitness assessment[12, 13] for each RWD source with respect to each trial emulation as well as extensive sensitivity analyses to assess robustness of findings.

The objectives of this project are to 1) develop state-of-the-art methodological approaches and 2) to apply them to create insights that may provide guidance on the potential use of RWE for regulatory science in oncology. To achieve these objectives, this demonstration project

will systematically emulate 12 oncology trials across four cancers and assess the agreement of treatment effect estimates between RCTs and their respective emulations.

Here, we describe the design and process for the selection of the 12 oncology RCTs, the assessment of the database quality and selection, protocol development and pre-registration, study design and statistical analysis, and pre-specified agreement metrics to evaluate the concordance between RCTs and emulations.

Methods

A visual summary of the entire systematic process from trial selection to final results is provided in Figure 1.

Trial selection

The focus of ENCORE is to maximize potential learnings on when RWE studies can or cannot yield similar results compared to RCTs. To that end, the emphasis of the project is on trials of therapies for the most common cancers for which there has been substantial therapeutic development in recent years. After review and discussion with clinical and regulatory experts, four cancer indications were identified: non-small cell lung cancer, breast cancer, colorectal cancer and multiple myeloma. For each cancer type, we aim to conduct three trial emulations using multiple databases accessible for the scope of this project (i.e., the total number of emulations will equal 12 trials x n databases which are found fit-for-purpose for each trial).

We used a semi-automated process for trial selection where the eligibility criteria are documented in CONSORT diagrams showing reasons for excluding RCTs for each cancer type. The search was conducted using the *Aggregated Analysis of ClinicalTrials.gov* (AACT) database which is a publicly available relational database developed and maintained by the Clinical Trials Transformation Initiative (CTTI) that contains all information (protocol and result data elements) about every study registered on ClinicalTrials.gov.[14] To identify eligible trials, we used a combined search query strategy of the National Library of Medicine (NLM)-controlled *Medical Subject Headings (*MeSH) term and a free keyword search for the respective cancer indication in the *conditions*, *studies* and *detailed_descriptions* fields of each trial entry on ClinicalTrials.gov.

Eligible trials had to fulfill the following basic criteria:

- Interventional
- Randomized
- Intervention model: parallel assignment
- Industry-sponsored

- Trial start in 2011 or later
- Primary purpose was to study treatment effects
- Overall survival must be one of the endpoints reported (either as hazard ratio or median overall survival time)
- Recruitment status: 'Completed' or 'Active, not recruiting'
- Feasibility and clinical relevance (latter was defined as treatment or paradigm-changing trials or such that challenged existing treatment policies)

The rationale and operationalization of each criterion is listed in detail in Table 1. We will mainly consider pivotal interventional, randomized trials after 2011 because treatment guidelines among included cancer indications have undergone significant changes in recent years. Due to the rapid adoption of new breakthrough therapies in routine care, it is unlikely to find patients who may be still treated with outdated treatment regimens in current clinical practice. Conversely, we excluded trials with results published too recently. To allow for enough data and follow-up time accrual in the databases used for this project. Although there have been substantial methodological advancements to increase our understanding on the emulation and comparison of real-world progression-free survival (PFS) and objective response rates (ORR) to a RECISTv1.1[15]-based PFS and ORR assessment in RCTs[16, 17], imaging-based evaluations still hold a level of granularity which may not be necessarily reflected in chart-abstracted assessments of a patient's progression in routine care.[18, 19] The timing and cadence of intervals between progression assessments can differ between RCTs and routine care which may result in measurement error and bias.[20] Given this, and the large number of other methodological challenges like missing data, small sample sizes, data discontinuity and rapidly changing guideline treatments, we focus on the emulation of overall survival (OS) as the endpoint of interest. Therefore, we include only trials that have reported overall survival (OS) as one of the pre-specified endpoints in the protocol.

While most trial eligibility criteria could be operationalized in an automated fashion, our final criterion was an assessment of emulation feasibility and clinical relevance. This criterion involved extensive human review. The critical points considered in this step include an initial feasibility assessment of the data fitness,[19] including assessment of whether critical eligibility criteria (e.g., biomarker status) and prognostic factors (e.g., ECOG performance score) are measurable and whether preliminary study size counts are reasonable. Lastly, trial candidates were ranked and shortlisted into primary and runner-up candidates based on their clinical and regulatory relevance.

A list of tentative, shortlisted primary candidates is presented in Table 2 and the corresponding selection process is illustrated in the CONSORT diagrams (Supplementary Figures 1-4). Naturally, the majority of trials cover advanced (locally advanced, inoperable, recurrent/progressive disease) or metastatic cancer populations because a large proportion of drug development efforts start in these settings. A key objective that we aim to foster with the shortlisted trials is to achieve a better understanding how different disease settings (early, late), line settings

([neo]adjuvant, first line, advanced lines of therapy), therapy protocols (monotherapy, combination therapy) and population characteristics (simple versus complex genetic or immunological signatures) can be emulated using RWD. If more thorough feasibility assessments suggest that the threat of bias from mis-measurement of key study parameters or residual confounding remains high, or that the study size is not sufficient, then runner-up candidates will be considered instead.

Databases

The ENCORE project will use data from four oncology-specific electronic health records (EHR)-derived data sources (in alphabetical order): ConcertAI, COTA, Flatiron Health, and Ontada/McKesson. All available databases draw from a comprehensive national sample of patients with cancer in the US with the detailed EHR-based data necessary to study medication effectiveness in oncology. A detailed description and methodology on how patients are sampled will be provided with each trial emulation protocol. For ENCORE, not all databases will be available for each cancer indication and the names of the databases will be blinded and referred to as ENCORE DataBase (EDB) 1, 2, 3 and 4 for the final reporting of results (the numbering does not coincide with the above order of mention of the databases). If more than one database is considered fit-for-purpose for a respective trial emulation, the best possible analytic model will be employed for each database separately and final treatment effect estimates will be pooled using a meta-analytic approach using fixed effects (weights reflect the inverse of the variance of each database study) and random effects (weights reflect the inverse of the variance of each database study plus an additional variance term that quantifies the assumed heterogeneity between databases) models.[21]

Protocol development

For each shortlisted and selected RCT, a detailed protocol, pre-specifying key elements of the trial emulation, will be developed using the HARPER protocol template[22] recommended for regulatory submissions of RWE studies[23] and will be registered on ClinicalTrials.gov after review by a clinical and FDA regulatory expert panel. Following the target trial emulation framework, we will provide an explicit statement and rationale on how each element will be emulated including database selection, covariate measurement, operationalization of key eligibility criteria, study design, data analysis and causal contrasts of interest.[24, 25] Since it is common that oncology RCTs update OS estimates periodically based on accrued follow-up time, the protocol will give a brief summary of each emulated RCT and specify which target OS estimates will be used to compare agreement metrics to (see Section). All eligibility criteria will be extracted based on publications, publicly available protocols and statistical analysis plans of the selected RCT.

Emulation feasibility

Fit-for-purpose data

Real-world data fitness and emulation feasibility for each shortlisted candidate trial will be assessed in multiple steps based on guidance of the oncology quality, characterization, and assessment of real-world data (Oncology QCARD) Initiative.[19] The first step assesses if relevant variables like exposure/line of therapy, outcomes, and covariates are generally available, measured and operationalizable in routine-care. Since many oncological RCTs in recent years have focused on selected, biomarker-defined populations, subtleties in measurement and operationalizability of specific biomarkers must be reflected to ensure a representative and large enough study population. For example, immune checkpoint inhibitors have significantly changed the cancer treatment landscape since the first approvals of the CTLA-4 inhibitor ipilimumab in 2011 and the PD-1 inhibitors pembrolizumab and nivolumab in 2014 in the United States. With many trials that have followed thereafter, the operationalization of the expression of the PD-L1 biomarker in RCTs (e.g., as a percent staining, tumor proportion score or combined positive score) has also evolved, and PD-L1 '*positivity*' may have different definitions across calendar years based on different cut-off values.

According to the *Structured Process to Identify Fit-For-Purpose Data* (SPIFD) framework[13], the next step will outline tables that describe how eligibility criteria will be ascertained using a color-coded heatmap that will indicate the level of confidence on how well each criterion can be emulated in each selected database. As there are general eligibility criteria in oncology trials which either will not be possible to emulate (e.g., physician-assessed survival prognosis of x months) or that are clinically not relevant for the emulation of the trial (e.g., male patients should be willing to use barrier contraception), the study team will decide on key eligibility criteria for the emulation of the trial.

We will additionally provide a definition on how exposures, outcomes and covariates are exactly defined and operationalized in each respective database. There will be special emphasis on how exposure, in context of their respective disease and line of therapy settings, and the OS outcome will be emulated. For all considered databases, the OS endpoint is typically a composite that, depending on the underlying database, can be derived from different sources comprising EHR documentation, social security death index, obituary and other linkages. Given that not all relevant sources that provide mortality data are synchronized and updated uniformly, sensitivity analyses with more conservative (i.e., earlier) censoring dates will be considered for each trial emulation to mitigate the potential impact of ghost-time bias.[26]

Descriptives and data exploration

Critical aspects when emulating oncology trials are the choice and estimation of the appropriate estimand of interest.[27] Particularly when emulating pivotal trials of paradigm-changing treatments, multiple aspects need to be considered such as the contemporaneity of the (historical) control cohort, the adoption rate of the novel intervention in routine care, the magnitude

of the clinical treatment benefit and the rate in which (particularly patients in the control arm) discontinue or cross-over to the interventional treatment, which could lastly bias emulated treatment effects towards the null. To that end, comprehensive data explorations will be performed as part of the protocol development to contextualize these parameters and (if reported) draw comparisons to the emulated trial. Examples for such standard diagnostics are visualized in Figure 2. All exploratory analyses will be conducted blinded towards the outcome to avoid influencing study design and analytic choices.

The distribution of patient characteristics, stratified by exposure status, will be examined in Table 1's before and after applying eligibility criteria and contrasted with the distributions of patient characteristics of the original RCT. Initial propensity score matching or weighting methods will be applied to ensure that measured pre-exposure covariates can be balanced, exposure cohorts are conditionally exchangeable at baseline and resulting sample sizes are still sufficient after matching or weighting. At this stage, all exploratory analyses will be conducted blinded towards the outcome to not bias any study design and analytic choices based on known outcome information.

Statistical power considerations

Causal analyses of observational data may not have the same pre-requisites in terms of formal hypothesis testing and statistical power as RCTs since the number of 'recruited' patients is given and cannot be influenced.[28] For this project, however, the emulations in EHR data must have at least equal power to the relevant trial for interpretation of the pre-specified agreements metrics between RCT and RWD results. Since the main outcome of interest is defined as time to all-cause mortality (OS), the estimation of the statistical power is driven by the number of events rather than the number of patients. To assess if the overall number of events, unstratified by exposure, is sufficient such that a significant difference can be detected based on the original RCT-reported hazard ratio (HR), the statistical power (1-β) will be estimated using Schoenfeld's sample-size formula for the proportional-hazards regression model.[29]

Agreement metrics

To formally compare treatment effects between RCTs and their respective emulations, we will adapt the approach of the RCT-DUPLICATE project.[9, 30] That is, for the primary endpoint of interest (HR for OS and corresponding 95% confidence intervals), we will derive three qualitative agreement metrics: statistical agreement, estimate agreement and agreement based on the standardized mean difference (SMD). Examples are illustrated Table 3.

Statistical significance agreement: agreement between RCT and emulated trial treatment effect with regards to directionality and statistical significance.

Estimate agreement: agreement that the estimated RWE treatment effect is within the 95% CI of the RCT treatment effect estimate. Provided that for some emulations, the power

of the RWE study may be larger than that of the original RCT, this could lead to situations where there is no statistical significance agreement although the treatment effect estimates are highly overlapping but with the RCT estimate crossing the null (or vice versa in case the RCT has a larger power than the RWE emulation).

SMD agreement: quantification of the agreement between the emulated RWE and RCT treatment effect estimate. The SMD is calculated as

$$SMD = \frac{\hat{\theta}_{RCT} - \hat{\theta}_{RWE}}{\sqrt{\text{Var}(\hat{\theta}_{RCT}) + \text{Var}(\hat{\theta}_{RWE})}}$$

where $\hat{\theta}_{RCT}$ and $\hat{\theta}_{RWE}$ are the treatment effect estimates (hazard ratios or median survival time differences) and $\text{Var}(\hat{\theta}_{RCT})$ and $\text{Var}(\hat{\theta}_{RWE})$ are the corresponding variances for RCT and RWE, respectively. The resulting SMDs will be interpreted such that with an SMD of 1.00, the effect estimate from the RCT and the RWE emulation are 1 standard deviation apart. For an α-level of 0.05, the null hypothesis of no difference would be rejected whenever $|Z| > 1.96$.

For the secondary endpoints of interest (e.g., median survival time or survival probabilities) only the SMD agreement metric will be applicable.

Study design and statistical analysis

The study design for each trial emulation will be visualized as part of the protocol using a graphical depiction of the exact measurement windows of eligibility criteria, washout periods and covariates relative to the cohort entry time.[31]

Missing data

To establish an analytic cohort, key eligibility criteria will be applied in which patients with missing values in eligibility criteria are considered eligible in the respective attrition steps to allow for thorough missing data investigations. These missing data investigations will empirically assess assumptions on potential underlying missingness mechanisms according to Rubin's classification of missing data (i.e., missing completely at random [MCAR], missing at random [MAR] and missing not at random [MNAR]).[32] To that end, we will adopt a principled process on missing data that empirically evaluates different aspects across partially observed covariates based on three group diagnostics.[33, 34] These diagnostics cover (1) comparisons of patients characteristics with and without an observed level of the partially observed covariate, (2) ability to predict missingness given observed data, and (3) assessments if outcomes between patients with a missing value are systematically different. Together with expert domain knowledge and assumptions about the underlying missing data structure through canonical causal diagrams,[35] this will inform decisions regarding the in- or exclusion of patients with missing

values in key eligibility criteria and potential sensitivity analyses to assess the robustness of these decisions.[36]

While the MAR assumption is a strong assumption to hold across all considered covariates, it was shown that especially in the context of partially observed covariate data (as opposed to missing exposure and outcome data), only mechanisms in which a covariate causes its own missingness leads to critical bias (MNAR).[35] Hence, methodologies which retain patients and give the potential to adjust for a broader set of prognostic factors (e.g., multiple imputation[37] or doubly robust methods[38]) may be preferred over complete case analyses.

Endpoints and propensity score analyses

Due to its ubiquity in oncology trials, the primary parameter of interest in ENCORE will be defined as the marginal hazard ratio (HR) coefficient for the treatment comparison for time to all-cause mortality (OS).[39] We will also consider alternative endpoints on an absolute risk scale such as median survival times, survival probabilities at pre-defined time points during follow-up[40] and restricted mean survival times as secondary endpoints of interest.

For the estimation of marginal treatment effects, we will employ propensity score methods to adjust for measured confounding between treatment arms. The selection of important prognostic covariates will be based on expert clinical knowledge and published literature on prognostic scores in oncology.[41] The implementation of propensity scores in combination with multiple imputation will follow the '*within*' methodology as described by Leyrat et al.[42, 43] That is, propensity score matching or weighting will be applied to each imputed dataset. The marginal treatment effect will then be estimated in each imputed and matched or weighted dataset separately and pooled into a final estimate following Rubin's rule.[44, 45] This approach has been shown to lead to unbiased estimates across different simulated scenarios with a sufficient estimation of the variance.[42]

To asses the balance of pre-exposure covariates after matching or weighting on each imputed dataset, the average SMD and corresponding minimum and maximum SMD range will be visualized (see example in Figure 3). Covariate balance is typically considered at a SMD $<$ 0.1.[46] Further, we will compute the average post-matching or post-weighting C-statistics.[47] In addition, we will use a published prognostic score for OS[41] as a balance measure to visually assess if the prognostic score is balanced between treatment arms after propensity score matching or weighting as this approach was described to show the highest correlations with bias compared with other balance measures and not affected by model misspecification.[48]

Similarly, survival probabilities for individual time points will be estimated in each imputed and propensity score matched or weighted dataset according to the Kaplan-Meier method.[40] Since survival probabilities typically do not follow normal distributions which are required to apply Rubin's rule, these will be transformed through a complementary log-log transformation $log(-log(1 - pr(surv)))$ with $pr(surv)$ denoting the survival probability at a given

time during follow-up.[49, 50] The transformed survival probabilities are then pooled across imputed datasets and individual time points following Rubin's rule and back-transformed via $1 - exp(-exp(qbar))$ with $qbar$ denoting the pooled survival probability. The median survival time can be finally determined by extracting the time point during follow-up at which the survival probability drops below 0.5 for the first time.

Sensitivity analyses

With the goal to better understand which factors could influence a difference between RCT results and emulated trial results, a range of sensitivity analyses will be conducted. This can comprise decisions on the considered databases, calendar time period, covariate measurement (e.g., measurement windows or trade-offs on sensitivity versus specificity in measurements), approaches to missing data, selection of covariates for imputation and propensity score models and decisions on when to censor patients. All sensitivity analyses will be pre-specified in the study protocol and reported using appropriate visualizations such as forest plots.

Reproducibility and consistency across emulation

To ensure a transparent, reproducible and consistent way of deriving analytic cohorts and performing statistical analyses across trial emulations, we developed an internal R package `encore.io` with parameterized functions. Detailed documentation can be found in the Supplementary Material.

Discussion

Building on an established approach developed by the RCT DUPLICATE project[9, 30], the ENCORE project aims to emulate 12 oncology trials using four oncology EHR data sources to inform the potential use of RWE for regulatory science in oncology. The project focuses on four common cancers and will evaluate the agreement of treatment effect estimates between RCTs and their respective emulations. Historically, administrative health claims databases have been the backbone of most research in RWE. With increasing access to EHR data and a maturing set of methodological approaches to draw causal inferences from such data, the ENCORE project is one of the first of its kind to evaluate when and how RWD can be used to deliver similar causal conclusions compared to RCTs in the field of oncology.

The RCT DUPLICATE initiative[9, 30] identified multiple emulation challenges which we expect to also encounter in ENCORE. One particular aspect that we may not always be able to emulate is the exact distribution of patient characteristics of the trial population. This can be due to the lack of data granularity and comprehensiveness to emulate relevant eligibility criteria or the fact that large pivotal trials in oncology are typically conducted in multiple countries worldwide, which we will not be able to mirror given that all considered databases reflect the US only. This may be a critical factor especially given that the pathophysiology, prognosis and factors that drive heterogeneous treatment outcomes of certain cancers differ between countries. For example, some cancer indications (e.g., GI cancers) or some genetic mutations (e.g., EGFR mutations in lung adenocarcinoma) are much more prevalent in Asian countries compared to the US or Europe.

Another common challenge in the emulation of oncology trials is the estimation of an "intention-to-treat" (ITT) analogous estimand which is usually the primary estimand reported in oncology RCTs. Due to intercurrent events, such as non-adherence, crossover of a high proportion of patients from the control to the intervention arm, or differences in subsequent therapy lines there may be estimand differences between trial and emulation.[51] Although this is also a common challenge in the analysis of RCTs,[27] treatment protocols in routine care are often observed to be less stringent compared to trials.[8] In order to derive comparable estimands it is therefore crucial to understand and contextualize the proportion and timing of treatment switching and discontinuation in both the trial and its emulation.

This issue may be even augmented with more complex treatment protocols like combination regimens (e.g., CheckMate9LA) as compared to monotherapies (e.g., CheckMate057) since the ascertainment of the exposure needs to happen over a pre-defined time window which is often difficult to calibrate in RWD. As a result, exposure misclassification (due to overly ascertainment windows) and selection bias (due to overly long ascertainment windows) are common trade-offs in such scenarios. Alternative analytic approaches which target a per-protocol estimand, such as the clone-censor-weight design,[52] may be viable options if the ITT estimand cannot be estimated due to the aforementioned parameters, provided that necessary covariate measurements are available to account for the artificial censoring introduced with these methods.

Conclusions

RWE based on fit-for-purpose data and principled, well-designed and reproducible studies may complement evidence coming from RCTs. Through a systematic benchmarking approach, the ENCORE project will provide insights as to how measurement, design and analytic decisions influence bias and validity in oncological RWE studies.

References

1. Framework for FDA's real-world evidence program (last accessed 11/28/2024) [Internet], 2018Available from: https://www.fda.gov/media/120060/download?attachment

2. Purpura CA, Garry EM, Honig N, et al: The role of real-world evidence in FDA-approved new drug and biologics license applications. Clinical Pharmacology & Therapeutics 111:135–144, 2022

3. Senior M: Fresh from the biotech pipeline: Record-breaking FDA approvals. Nature Biotechnology, 2024

4. Merola D, Schneeweiss S, Schrag D, et al: An algorithm to predict data completeness in oncology electronic medical records for comparative effectiveness research [Internet]. Annals of Epidemiology 76:143–149, 2022Available from: http://dx.doi.org/10.1016/j.annepidem.2022.07.007

5. Joshua Lin K, Jin Y, Gagne J, et al: Longitudinal data discontinuity in electronic health records and consequences for medication effectiveness studies. Clinical Pharmacology & Therapeutics 111:243–251, 2022

6. Rider JR, Wasserman A, Slipski L, et al: Emulations of oncology trials using real-world data: A systematic literature review. American journal of epidemiology kwae346, 2024

7. Merola D, Campbell U, Gautam N, et al: The aetion coalition to advance real-world evidence through randomized controlled trial emulation initiative: oncology. Clinical Pharmacology & Therapeutics 113:1217–1222, 2023

8. Merola D, Campbell U, Lenis D, et al: Calibrating observational health record data against a randomized trial. JAMA Network Open 7:e2436535–e2436535, 2024

9. Wang SV, Schneeweiss S, Franklin JM, et al: Emulation of randomized clinical trials with nonrandomized database analyses: Results of 32 clinical trials. Jama 329:1376–1385, 2023

10. Heyard R, Held L, Schneeweiss S, et al: Design differences and variation in results between randomised trials and non-randomised emulations: Meta-analysis of RCT-DUPLICATE data [Internet]. BMJ medicine 3, 2024Available from: https://doi.org/10.1136/bmjmed-2023-000709

11. Calibrating real-world evidence studies in oncology against randomized trials: ENCORE (last accessed 11/28/2024) [Internet], 2024Available from: https://www.fda.gov/about-fda/oncology-center-excellence/calibrating-real-world-evidence-studies-oncology-against-

randomized-trials-encore

12. Rivera DR, Eckert JC, Rodriguez-Watson C, et al: The oncology QCARD initiative: Fostering efficient evaluation of initial real-world data proposals. Pharmacoepidemiology and Drug Safety 33:e5818, 2024

13. Gatto NM, Campbell UB, Rubinstein E, et al: The structured process to identify fit-for-purpose data: A data feasibility assessment framework. Clinical Pharmacology & Therapeutics 111:122–134, 2022

14. Tasneem A, Aberle L, Ananth H, et al: The database for aggregate analysis of ClinicalTrials. Gov (AACT) and subsequent regrouping by clinical specialty. PloS one 7:e33677, 2012

15. Eisenhauer EA, Therasse P, Bogaerts J, et al: New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European journal of cancer 45:228–247, 2009

16. Ton TGN, Pal N, Trinh H, et al: Replication of overall survival, progression-free survival, and overall response in chemotherapy arms of nonâ small cell lung cancer trials using real-world data [Internet]. Clinical Cancer Research 28:2844–2853, 2022Available from: https://doi.org/10.1158/1078-0432.CCR-22-0471

17. McKelvey BA, Garrett-Mayer E, Rivera DR, et al: Evaluation of real-world tumor response derived from electronic health record data sources: A feasibility analysis in patients with metastatic non–small cell lung cancer treated with chemotherapy. JCO Clinical Cancer Informatics 8:e2400091, 2024

18. Chen L, Davis R, Lee J, et al: Comparison of response from RECIST1.1 and abstraction in real-world patients with lung cancer. [Internet]. Journal of Clinical Oncology 41:e21194–e21194, 2023Available from: https://ascopubs.org/doi/abs/10.1200/JCO.2023.41.16_suppl.e21194

19. Rivera DR, Henk HJ, Garrett-Mayer E, et al: The friends of cancer research real-world data collaboration pilot 2.0: Methodological recommendations from oncology case studies. Clinical Pharmacology & Therapeutics 111:283–292, 2022

20. Ackerman B, Gan RW, Meyer CS, et al: Measurement error and bias in real-world oncology endpoints when constructing external control arms. Frontiers in Drug Safety and Regulation 4:1423493, 2024

21. Nikolakopoulou A, Mavridis D, Salanti G: How to interpret meta-analysis models: Fixed effect and random effects meta-analyses. BMJ Ment Health 17:64–64, 2014

22. Wang SV, Pottegård A, Crown W, et al: HARmonized protocol template to enhance reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force [Internet]. Value in Health 25:1663–1672, 2022Available from: https://doi.org/10.1016/j.jval.2022.09.001

23. ICH guideline. General principles on plan, design and analysis of pharmacoepidemiological studies that utilize real-world data for safety assessment of medicines M14. Available from: Https://database.ich.org/sites/default/files/ICH_M14_Step3_DraftGuideline_2024_0521.pdf (last accessed 01/07/2025)

24. Hernán MA, Wang W, Leaf DE: Target trial emulation: A framework for causal inference from observational data. Jama 328:2446–2447, 2022

25. Hernán MA, Sauer BC, Hernández-Díaz S, et al: Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. Journal of clinical epidemiology 79:70–75, 2016

26. Meyer A-M, Davies J, Taylor M, et al: Open cohorts and ghost-time bias in real world data, in PHARMACOEPIDEMIOLOGY AND DRUG SAFETY. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2020, pp 426–426

27. Rufibach K: Treatment effect quantification for time-to-event endpoints–Estimands, analysis strategies, and beyond [Internet]. Pharmaceutical Statistics 18:145–165, 2018Available from: http://dx.doi.org/10.1002/pst.1917

28. Hernán MA: Causal analyses of existing databases: No power calculations required. Journal of clinical epidemiology 144:203–205, 2022

29. Schoenfeld D: The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 68:316–319, 1981

30. Franklin JM, Pawar A, Martin D, et al: Nonrandomized real-world evidence to support regulatory decision making: Process for a randomized trial replication project. Clinical Pharmacology & Therapeutics 107:817–826, 2020

31. Schneeweiss S, Rassen JA, Brown JS, et al: Graphical depiction of longitudinal study designs in health care databases. Annals of internal medicine 170:398–406, 2019

32. Rubin DB: Inference and missing data. Biometrika 63:581–592, 1976

33. Weberpals J, Raman SR, Shaw PA, et al: Smdi: An r package to perform structural missing data investigations on partially observed confounders in real-world evidence studies. JAMIA open 7:ooae008, 2024

34. Weberpals J, Raman SR, Shaw PA, et al: A principled approach to characterize and analyze partially observed confounder data from electronic health records [Internet]. Clinical Epidemiology 16:329–343, 2024Available from:https://www.tandfonline.com/doi/abs/10.2147/CLEP.S436131

35. Moreno-Betancur M, Lee KJ, Leacy FP, et al: Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. American journal of epidemiology 187:2705–2715, 2018

36. Tompsett DM, Leacy F, Moreno-Betancur M, et al: On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. Statistics in medicine 37:2338–2353, 2018

37. Weberpals J, Shaw PA, Lin KJ, et al: High-dimensional multiple imputation (HDMI) for partially observed confounders including natural language processing-derived auxiliary covariates [Internet]. American Journal of Epidemiology, 2025Available from: https://arxiv.org/abs/2405.10925

38. Shaw CK P., Williamson BD: Assessing treatment effects in observational data with missing confounders: A comparative study of practical doubly-robust and traditional missing data methods [Internet]. GitHub repository, 2024Available from: https://github.com/PamelaShaw/Missing-Confounders-Methods

39. Cox DR: Regression models and life-tables. Journal of the royal statistical society. Series B (Methodological) 34:187–220, 1972

40. Kaplan EL, Meier P: Nonparametric estimation from incomplete observations. Journal of the American statistical association 53:457–481, 1958

41. Becker T, Weberpals J, Jegg A, et al: An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. Annals of Oncology 31:1561–1568, 2020

42. Leyrat C, Seaman SR, White IR, et al: Propensity score analysis with partially observed covariates: How should multiple imputation be used? Statistical methods in medical research 28:3–19, 2019

43. Pishgar F, Greifer N, Leyrat C, et al: MatchThem:: Matching and weighting after multiple imputation. arXiv preprint arXiv:200911772, 2020

44. Rubin DB: Multiple imputation, in Flexible imputation of missing data, second edition. Chapman; Hall/CRC, 2018, pp 29–62

45. Van Buuren S, Groothuis-Oudshoorn K: Mice: Multivariate imputation by chained equations in r. Journal of statistical software 45:1–67, 2011

46. Austin PC: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in medicine 28:3083–3107, 2009

47. Franklin JM, Rassen JA, Ackermann D, et al: Metrics for covariate balance in cohort studies of causal effects. Statistics in medicine 33:1685–1699, 2014

48. Stuart EA, Lee BK, Leacy FP: Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. Journal of clinical epidemiology 66:S84–S90, 2013

49. Marshall A(Andrea), Billingham LJ, Bryan S: Can we afford to ignore missing data in cost-effectiveness analyses? [Internet]. European Journal of Health Economics Vol.10:1–3, 2009Available from: http://dx.doi.org/10.1007/s10198-008-0129-y

50. Morisot A, Bessaoud F, Landais P, et al: Prostate cancer: Net survival and cause-specific survival rates after multiple imputation. BMC medical research methodology 15:1–14, 2015

51. Manitz J, Kan-Dobrosky N, Buchner H, et al: Estimands for overall survival in clinical trials with treatment switching in oncology. Pharmaceutical Statistics 21:150–162, 2022

52. Gaber CE, Ghazarian AA, Strassle PD, et al: De-mystifying the clone-censor-weight method for causal research using observational data: A primer for cancer researchers. Cancer Medicine 13:e70461, 2024

Tables

Figures

Figure 1: Systematic process to understand effectiveness claims of oncology trials using real-world evidence.

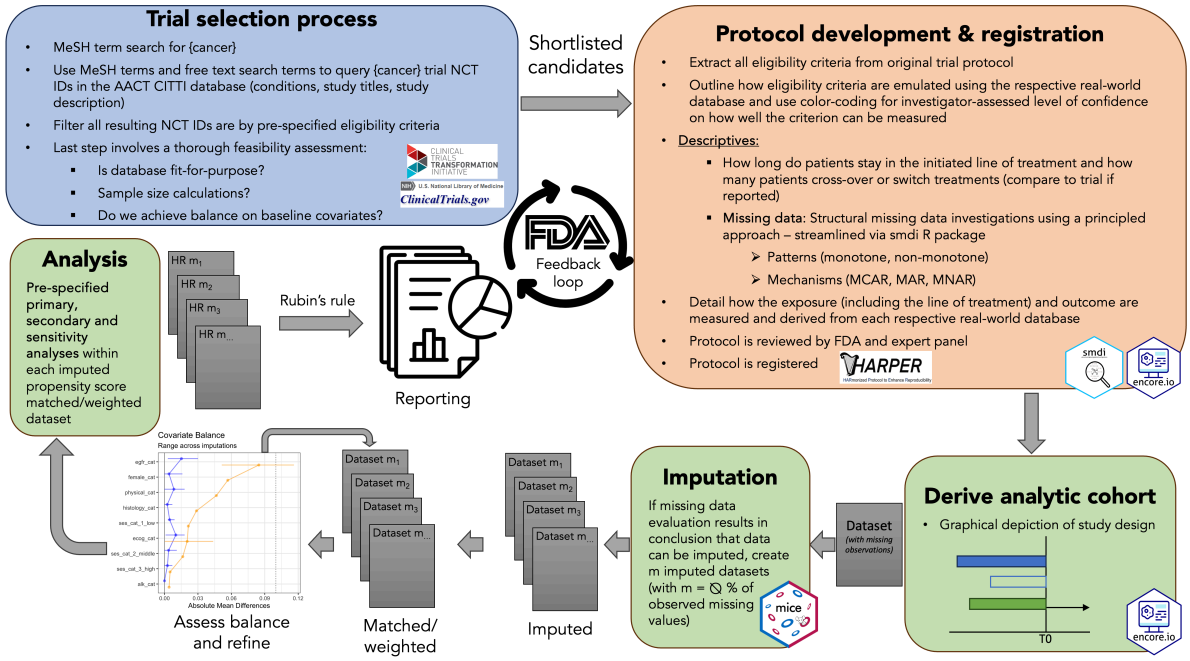

Figure 2: Example visualization of descriptive drug utilization analyses displaying a) initiation trends between compared regimens based on calendar time, b) cumulative rate of patients switching to another line of treatment.

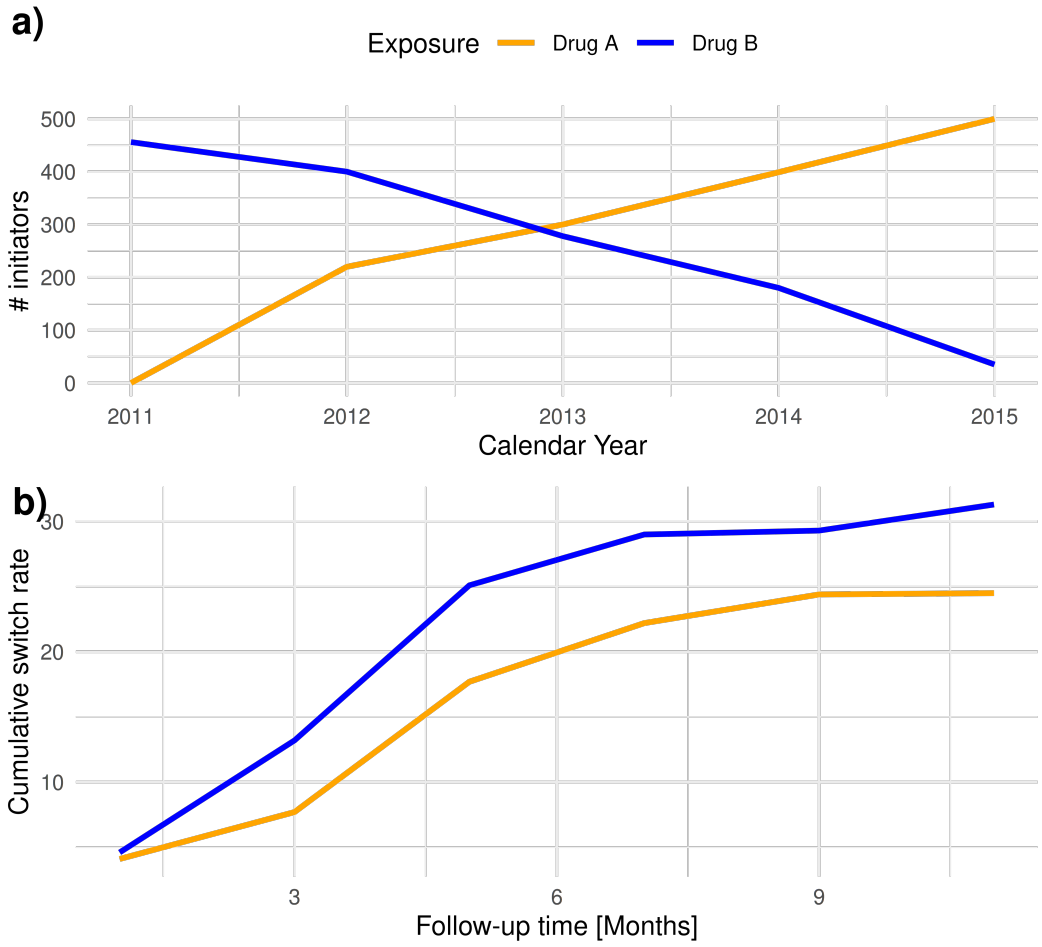

Figure 3: Assessment of a) covariate balance and b) distributional balance of a prognostic score for overall survival before and after propensity score matching or weighting across multiple imputed datasets.

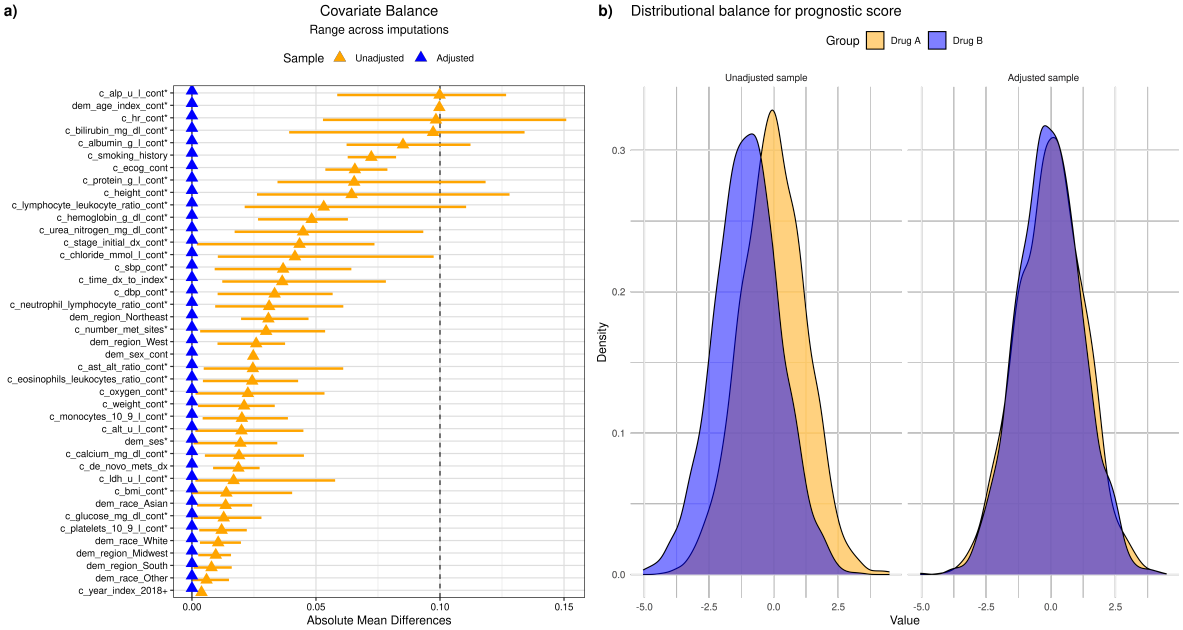

Table 1: Criteria to select eligible trials for emulation in ENCORE.

Criteria	Definition	Eligible
Interventional study	The nature of the investigation or investigational use for which clinical study information is being submitted	Interventional (clinical trial): Participants are assigned prospectively to an intervention or interventions according to a protocol to evaluate the effect of the intervention(s) on biomedical or other health related outcomes.
Randomized allocation	The method by which participants are assigned to arms in a clinical trial.	Randomized: Participants are assigned to intervention groups by chance
Interventional study model	The strategy for assigning interventions to participants.	Parallel: Participants are assigned to one of two or more groups in parallel for the duration of the study
Sponsor/source	The entity (for example, corporation or agency) that initiates the study	Industry
Study start date	The estimated date on which the clinical study will be open for recruitment of participants, or the actual date on which the first participant was enrolled.	2011 or later
Primary purpose	The main objective of the intervention(s) being evaluated by the clinical trial.	Treatment: One or more interventions are being evaluated for treating a disease, syndrome, or condition.
Primary outcome	A description of each primary outcome measure (or for observational studies, specific key measurement[s] or observation[s] used to describe patterns of diseases or traits or associations with exposures, risk factors or treatment).	Primary or secondary outcome needs to include overall survival
Overall Recruitment Status	The recruitment status for the clinical study as a whole, based upon the status of the individual sites. If at least one facility in a multi-site clinical study has an Individual Site Status of "Recruiting," then the Overall Recruitment Status for the study must be "Recruiting."	Completed: The study has concluded normally; participants are no longer receiving an intervention or being examined (that is, last participant's last visit has occurred) OR Active, not recruiting: Study is continuing, meaning participants are receiving an intervention or being examined, but new participants are not currently being recruited or enrolled
Feasibility and clinical relevance	Are all key variables available to emulate the clinical trial at hand and is the clinical trial considered clinically relevant?	Trials for which there is reasonable believe that key study parameters can be emulated and there is a high enough clinical relevance (e.g., paradigm-changing trials)

Table 2: Tentative list of randomized controlled trials (RCTs) considered for emulation.

NCTID	Acronym	Clinical setting	Line of therapy	Treatment comparison
Non-small cell lung cancer				
NCT02296125	FLAURA	Advanced/metastatic EGFRm+	1L	osimertinib versus erlotinib or gefitinib
NCT01673867	CheckMate017/057	Metastatic squamous/non-squamous	2L	nivolumab versus docetaxel
NCT03215706	CheckMate9LA	Metastatic	1L	nivolumab, ipilimumab, chemotherapy versus chemotherapy alone
Breast cancer				
NCT01740427	PALOMA-2	Advanced postmenopausal ER-positive and HER2-negative	1L	palbociclib, letrozole versus letrozole
NCT02819518	KEYNOTE-355	Locally recurrent inoperable or metastatic triple negative	1L	pembrolizumab, chemotherapy vs. placebo, chemotherapy
NCT01772472	KATHERINE	HER2-positive	Adjuvant	trastuzumab emtansine versus trastuzumab
Colorectal cancer				
NCT04737187	SUNLIGHT	Refractory metastatic	3L	trifluridine, tipiracil, bevacizumab versus trifluridine, tipiracil
NCT01374425	MAVERICC	Metastatic	1L	bevacizumab, mFOLFOX6 versus bevacizumab, FOLFIRI
NCT02563002	KEYNOTE-177	Metastatic microsatellite instability-high (MSI-H) or mismatch repair deficient (dMMR)	2L+	pembrolizumab versus standard of care
Multiple Myeloma				
NCT01568866	ENDEAVOR	Relapsing or progressing disease	2L/3L	carfilzomib, dexamethasone versus bortezomib, dexamethasone
NCT02252172	MAIA	Newly diagnosed	1L	daratumumab, lenalidomide, dexamethasone versus lenalidomide, dexamethasone
NCT01239797	ELOQUENT - 2	Relapsed or refractory	2L+	elotuzumab, lenalidomide, dexamethasone versus lenalidomide, dexamethasone

Table 3: Visualization of agreement metrics example.

Trial	HR (95% CI)		Statistical significance agreement	Estimate agreement	SMD
	RCT	RWE			
Trial 1	0.75 (0.61 - 0.91)	0.80 (0.51 - 1.10)	No	Yes	Yes (-0.29)
Trial 2	0.62 (0.51 - 0.71)	0.65 (0.55 - 0.69)	Yes	Yes	Yes (-0.46)
Trial 3	0.71 (0.67 - 0.80)	0.51 (0.41 - 0.61)	Yes	No	No (2.98)
Trial 4	0.90 (0.81 - 0.99)	1.20 (1.09 - 1.34)	No	No	No (-3.92)

Abbreviations: CI = Confidence interval, HR = Hazard ratio, RCT = Randomized controlled trial, RWE = Real-world evidence, SMD = standardized mean difference (based on log hazard ratios)