Emulating Comparative Oncology Trials with Real-world Evidence Studies (ENCORE): Process Development and Methodological Considerations for Oncology Real-World Data

Authors: Janick Weberpals[1], Sebastian Schneeweiss[1], Kenneth L. Kehl[2], Donna R. Rivera[3*], Pallavi Mishra-Kalyani[3], Catherine C. Lerro[3*], Erin Larkins[4], Preeti Narayan[4], Richard Curley[3], Georg Hahn[1], Priyanka Anand[1], Yanina Natanzon[5], Andrew J. Belli[6], Ching-Kun Wang[6], Jenna Collins[7], Jonathan Kish[7], Janet Espirito[8], Nicholas J Robert[8], Robert J. Glynn[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] Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

[5] ConcertAI, Cambridge, MA, USA

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

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

[8] Ontada, Boston, MA, USA

*At the time this research was conducted

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: 4,541 words / 8,000 words

Abstract word count: 224 words / 250 words

Tables: 3

Figures: 3

Supplementary material: Supplementary figures, tables and material

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

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

FDA Disclaimer: This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

Funding Statement: This project was supported by the Oncology Center of Excellence, Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a contract [75F40122C00181]. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by FDA/HHS, or the U.S. Government.

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. Glynn has received support for investigator-initiated grants to the Brigham and Women's Hospital from Amarin, AstraZeneca, Kowa, Novartis, and Pfizer unrelated to the current 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: Simulated examples and code to implement analytic workflows described in this manuscript are illustrated at https://janickweberpals.github.io/imputation-ps-workflows/ and can be reproduced via the `encore.analytics` R package (https://github.com/janickweberpals/encore.analytics/).

Manuscript last updated: 2025-10-22 19:41:51.622791

Abstract

Real-world evidence (RWE) is increasingly used to complement findings from randomized controlled trials (RCTs), contextualizing the effectiveness and safety of medical interventions as delivered in routine clinical practice. Advances in the curation and accessibility of electronic health record data (EHR) present the opportunity to utilize real-world data (RWD) to investigate therapeutic areas including oncology, where administrative healthcare claims databases alone are often not fit-for-purpose. The RCT DUPLICATE initiative has previously evaluated when RWE can most appropriately draw causal conclusions by emulating trials for non-oncology indications. Here, we present the design and trial selection for the Emulation of Comparative Oncology Trials with Real-world Evidence (ENCORE) project, which extends this work to oncology. ENCORE is designed to 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 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 initiation and analysis. Pre-specified criteria will assess agreement of treatment effect estimates between RCTs and their respective emulations. Through extensive sensitivity analyses benchmarked against RCT results, the ENCORE project aims to inform understanding of how measurement, design, and analytic decisions influence the interpretation of results from emulated oncology trials using RWD.

Background

Randomized controlled trials (RCTs) are the evidentiary gold standard methodology for establishing the efficacy and safety of medical products. Guided by the 21st Century Cures Act,[1] the Food and Drug Administration (FDA) established a framework to factor considerations for use of real-world evidence (RWE) generated from real-world data (RWD) such as electronic health records (EHR) to support approval of new indications or satisfy postmarketing requirements.[1] Accounting for approximately 30% of new FDA drug approvals, oncology was the disease area with the most new approvals in 2023[2] as well as several indication expansions, and has many areas of high unmet medical need to treat serious conditions. Therefore, RWE can have a particularly important potential to complement evidence from RCTs in the field of oncology. Potential uses include the assessment of effectiveness in specific patient populations that are not adequately represented in RCTs, or the precision oncology-focused discovery of biomarkers among pan-tumor populations that harbor specific genomic and immuno-pathological signatures.

However, the comparability and transportability of results derived between non-interventional studies and RCTs depends on multiple factors. Frequently referenced limitations include lack of baseline randomization and the imbalance in prognostic factors resulting thereof, missing data, unmeasured confounding, small study sizes, data discontinuity,[3,4] adoption of changes in treatment guidelines over time in routine clinical care and the inability to measure and emulate common eligibility criteria, including prognostic factors, and standardized endpoint assessments in real-world data (RWD).[5] While examples of oncology trial emulations have been published,[5–7] a systematic and scaled approach to emulate a diverse set of oncology trials with multiple heterogeneous databases would enhance increase confidence in the interpretability of non-interventional studies, evaluate regulatory considerations, and to provide context as to which questions can be validly answered.

The RCT DUPLICATE initiative[8] evaluated when non-interventional studies can come to causal conclusions on treatment effects by benchmarking results against RCTs under the assumption that well-designed and conducted RCT findings reflect causal treatment effects. In settings where the RCT designs could be emulated well, non-interventional studies were able to reach similar conclusions.[9] However, this prior work from RCT-DUPLICATE focused on emulating trials in non-oncology settings using claims databases.

The *Emulation of Comparative Oncology Trials with Real-world Evidence* (ENCORE) project[10] aims to extend this work to oncology. Clinical studies in oncology come with unique methodological challenges which must be systematically explored and understood. Building on a process co-developed with the FDA through RCT DUPLICATE,[8] this expansion to oncology will emulate 12 randomized clinical trials using multiple specialty oncology EHR data sources. The process will emphasize transparency and include a data fitness-for-use assessment[11,12] 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) apply these methods to gain insights into the potential use of RWE to enhance regulatory and clinical decision-making in oncology. To achieve these objectives, this demonstration project will systematically emulate 12 oncology trials across four cancer types 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, assessment of the database quality and selection, protocol development and pre-registration, study design and statistical analysis plans, 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 clarify when non-interventional studies can or cannot yield similar results compared to oncology RCTs. Therefore, the project emphasizes trials of therapies for common cancers for which there has been substantial therapeutic development in recent years. After review and collaboration with clinical and regulatory experts, four cancer types 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 accessible databases (i.e., the total equaling 12 trial emulations x n databases which are found fit-for-purpose for each trial).

We used a systematic process for trial selection where the eligibility criteria are documented in CONSORT diagrams showing reasons for excluding RCTs by 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 information (protocol and result data elements) about studies registered on ClinicalTrials.gov.[13] 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: 2011 or later
- Primary purpose: Study treatment efficacy
- Endpoint(s): 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 trials that challenged existing treatment policies)

The operationalization for each criterion is listed in detail in Table 1. We mainly considered pivotal large late-phase RCTs that were initiated after 2011 because treatment guidelines for included cancer indications have undergone significant changes in recent years. Due to the rapid adoption of newly-approved therapies in routine care, patients are less likely to receive outdated treatment regimens in clinical practice. Conversely, we excluded trials with results published too recently to allow for substantive medical product uptake and follow-up time accrual in real-world databases used for this project. We did not define a global cut-off as the requirements for follow-up time are different for each cancer type and population (e.g., advanced NSCLC versus early-stage breast cancer) and the decisions were made on a clinically relevant basis.

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[14]-based PFS and ORR assessment in RCTs[15,16], imaging-based evaluations still require a level of granularity which is not well reflected in chart-abstracted assessments of a patient's progression in routine care.[17,18] The timing and cadence of intervals between progression assessments can differ between RCTs and routine care which may result in measurement error and bias.[19] Given this, and the large number of other methodological challenges including 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 OS as a pre-specified endpoint 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 fundamental points considered in this step include an initial feasibility assessment of the data fitness,[18] involving an assessment of whether critical eligibility criteria (e.g., biomarker status) including prognostic factors (e.g., Eastern Cooperative Oncology Group [ECOG] performance score) are measurable, and whether preliminary study size based on a rough estimation of the number of patients observed in the data with the combination of treatment regimen and line of therapy is reasonable for shortlisting the trial.

Lastly, trial candidates were ranked and shortlisted as primary or runner-up based on their clinical and regulatory relevance.

A list of tentative, shortlisted primary candidates is presented in Table 2. The corresponding selection process is illustrated in the CONSORT diagrams (**Supplementary Figures 1-4**) and runner up candidate trials are listed in **Supplementary Table 1**. The majority of shortlisted trials represent indications for advanced (locally advanced, inoperable, recurrent/progressive disease) or metastatic cancer patients because a large proportion of drug development efforts have recently focused on these settings. A key objective that we aim to explore with the selected trials is to evaluate how stage (early, late), line of therapy ([neo]adjuvant, first line, advanced lines of therapy), therapy protocols (monotherapy, combination therapy) and clinical characteristics (all-comer versus complex genetic or immunological signatures) can be emulated using RWD. If more thorough feasibility assessments after applying all relevant eligibility criteria (including formal power calculations and assessment of sufficient balance after propensity score matching or weighting) 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 candidate trials will be considered instead.

Databases

The ENCORE project will use data from four U.S.-based oncology-specific EHR-derived RWD data sources (in alphabetical order): ConcertAI, COTA, Flatiron Health, and Ontada/McKesson. A detailed description and sampling methodology will be provided with each trial emulation protocol. For ENCORE, not all databases will be available for each cancer indication given the specificity of data elements required 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 (in randomly assigned order). If more than one database is considered fit-for-use for a respective trial emulation, the most suitable analytic model will be employed for each database separately and final treatment effect estimates will be pooled using a meta-analytic approach using random effects (primary; weights reflect the inverse of the variance of each database study plus an additional variance term that quantifies the assumed heterogeneity between databases) and fixed effects (secondary; weights reflect the inverse of the variance of each database study) models.[20]

Protocol development

For each selected RCT, a detailed protocol, pre-specifying key elements of the trial emulation, will be developed using the HARPER protocol template[21,22] and will be registered on ClinicalTrials.gov after review by an expert panel comprising clinicians and FDA reveiwers. Following the target trial emulation framework, we will provide an explicit description 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.[23,24] Since it is common that oncology RCTs update survival 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 for agreement metric evaluation (see Section). All eligibility criteria will be extracted based on publications, publicly available protocols, and statistical analysis plans of the selected RCT.

Emulating the RCT estimand

An important aspect when emulating oncology trials is the choice and estimation of the appropriate estimand of interest.[25] In prior RCT-DUPLICATE efforts, a "while on treatment" strategy was chosen for database studies that were designed to emulate trials where the primary trial analysis was intention-to-treat (e.g., a "treatment policy" estimand). This estimand was chosen to mimic the high adherence typically observed in large cardiovascular clinical trials in the context of low adherence in clinical practice. However, a "while on treatment" estimand would be inappropriate when emulating oncology trials, not only because of highly informative censoring, but also because OS outcome analyses in oncology are typically "treatment policy" estimands that allow for crossover or other post-progression antineoplastic therapy.[26] We plan to focus on "treatment policy" estimands for the primary analyses in our trial emulations. However, allowing crossover can dilute treatment effects, moving point estimates toward the null. Therefore, differences in cross-over and discontinuation rates between the trial and emulation may still contribute to differences in observed outcomes. Recognizing these practical challenges, we will carefully characterize the patterns of discontinuation or cross-over in both the trials and the database studies designed to emulate them.

Emulation feasibility

Fit-for-purpose data

Real-world data fitness and emulation feasibility for each candidate trial will be assessed in multiple steps based on the approach described by the Oncology Quality, Characterization, and Assessment of Real-World Data (Oncology QCARD) Initiative.[18] 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 oncology RCTs in recent years have focused on biomarker-defined populations, nuances in measurement and operationalizability of specific biomarkers must be understood to ensure a representative and adequately sized 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. However, the operationalization of the expression of the PD-L1 biomarker in RCTs (e.g., percent staining, tumor proportion score, or combined positive score) has evolved, and as such PD-L1 '*positivity*' may have different definitions by year based on different cut-off values.

According to the *Structured Process to Identify Fit-For-Purpose Data* (SPIFD) framework[12], the next step outlines 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 eligibility criteria in oncology clinical trials which either are infeasible to emulate (e.g., physician-assessed survival prognosis quantification) or that do not impact clinical care for the emulation of the trial (e.g., male patients should be willing to use barrier contraception), the study team will determine key eligibility criteria for the emulation of the trial based on consensus.

Additionally, we will provide conceptual and operationalized definitions on how exposures, outcomes and covariates are defined in each respective database. There will be special emphasis on how exposure, in context of disease and line of therapy settings, and the survival outcomes are emulated. For all considered databases, real-world OS (rwOS) is typically a composite measure that, depending on the underlying database, can be derived from different mortality sources (i.e., EHR documentation, Social Security Death Index, 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.[27]

Descriptives and data exploration

Particularly when emulating pivotal trials of practice-changing treatments, multiple aspects need to be considered, such as the contemporaneity of the control cohort, the adoption rate of the novel medical product in routine care, the magnitude of the clinical treatment benefit, and the rate in which patients discontinue or cross-over, as they could influence effect estimates.

To that end, comprehensive data explorations will be performed as part of the protocol development to contextualize these parameters and (if reported) evaluate comparisons to the emulated trial. Examples for such standard diagnostics are visualized in Figure 2.

The distribution of patient characteristics, by exposure status, will be examined 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[28], 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 or analytic choices based on known outcome information.

Statistical power considerations

Causal analyses of non-interventional data are often not designed with respect to formal hypothesis testing and statistical power in the same manner as RCTs since the number of 'recruited' patients is limited by available data .[29] 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 agreement metrics between RCT and RWD results. Since the main outcome

of interest is defined as time from index treatment initiation to all-cause mortality (rwOS), the estimation of the statistical power is driven by the number of events rather than the number of patients. To assess whether 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 with a 2-sided α of 0.05) will be estimated using Schoenfeld's sample-size formula for the proportional-hazards regression model.[30]

Agreement metrics

To formally compare treatment effects between RCTs and their respective emulations, we will adapt the approach of the RCT-DUPLICATE project.[8,31] That is, for the primary endpoint of interest (hazard ratio [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 in Table 3.

Statistical significance agreement: agreement between RCT and emulated trial treatment effect with regards to directionality and statistical significance (nominal in the case of emulated trials).

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 (RCT estimate crossing the null although the treatment effect estimates are overlapping) 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 (log 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.[32]

Missing data

To establish an analytic cohort, missingness will be assessed across patients who meet all eligibility criteria, including those with missing values as a first iteration. 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]).[33] We will adopt a principled process on missing data that empirically evaluates different aspects across partially observed covariates based on three group diagnostics.[34,35] 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. Expert domain knowledge and assumptions about the underlying missing data structure through canonical causal diagrams[36] will additionally inform decisions regarding the inclusion or exclusion of individuals with missing values in key eligibility criteria and potential sensitivity analyses to assess the robustness of these decisions.[37]

While MAR is a strong assumption to hold across all considered covariates, it has been shown that especially in the context of partially observed covariate data (as opposed to missing exposure and outcome data), generally only mechanisms in which a covariate causes its own missingness lead to significant bias (MNAR).[36] Hence, methodologies which retain otherwise eligible patients and give the potential to adjust for a broader set of prognostic factors (e.g., multiple imputation[38] or doubly robust methods[39]) may be preferred over complete case analyses, although such tradeoffs must be carefully evaluated within the clinical and study contexts.

Outcome and propensity score analyses

Due to its frequency of use in oncology trials, the primary parameter of interest in ENCORE will be defined as the marginal HR coefficient for the treatment comparison for OS.[40] 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[41] 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 relevant prognostic

covariates will be based on expert clinical knowledge and published literature on prognostic scores in oncology.[42] The implementation of propensity scores in combination with multiple imputation will follow the '*within*' methodology as described by Leyrat et al.[43,44] As appropriate, 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.[45,46] This approach has been shown to lead to unbiased estimates across different simulated scenarios with a sufficient estimation of the variance.[43]

To assess 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 often considered reasonable at a SMD < 0.1.[28] 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[42] 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.[41] 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 follow-up time during.[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 differences between RCT results and emulated trial results, it is appropriate to conduct a range of sensitivity analyses. 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 censoring decisions. 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 two R packages. The first package, `encore.io`, streamlines the query of analytic cohorts and measurement of covariates and outcomes using parameterized functions. Due to the blinding of databases used for ENCORE, this package is not able to be publicly available; however a comprehensive and detailed documentation for each function are available in the Supplementary Material. A second open-source R package, `encore.analytics`, will then facilitate the complex multi-step workflows of multiple imputation, propensity score matching and weighting and estimating pooled marginal treatment effects and Kaplan-Meier curves along with other established statistical R packages.[51–56] Simulated examples with code that illustrate the analytic workflows described in this manuscript are available in the online Supplementary Material at https://janickweberpals.github.io/imputation-ps-workflows/ and can be reproduced via the `encore.analytics` R package (https://github.com/janickweberpals/encore.analytics/).

Discussion

Building on an established approach developed by the RCT DUPLICATE project[8,31], the ENCORE project aims to emulate 12 oncology trials using four oncology EHR data sources to inform the potential use of RWD for regulatory purposes 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 claims databases have been the predominant data source used to evaluate medical product safety and effectiveness in the postmarketing setting. With increasing access to EHR data and availability of granular clinical data elements often necessary to gather fit-for-use data in oncology, as well as a maturing set of methodological approaches for causal inference using such data, the ENCORE project is one of the first of its kind to evaluate contexts where RWD may be reliably used to draw similar conclusions compared to RCTs in the field of oncology.

The RCT DUPLICATE initiative[8,31] identified multiple emulation challenges that may be similar 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 (all considered databases reflect the US only). This may impact the results of this trial emulation given that the distribution of genetic alterations and biomarkers, prognosis, and factors that drive heterogeneous treatment outcomes of certain cancers can differ between countries. For example, certain cancers (e.g., GI cancers) or certain genetic mutations (e.g., EGFR mutations in lung cancer) are much more prevalent in certain geographic regions and populations compared to the US. Additionally, treatment landscapes such as the preference for specific regimens or compounds, their use in earlier or later lines of therapies, and supportive care practices can differ between geographic regions, too.

Another common challenge in the emulation of oncology trials is the estimation of an "treatment policy"/"intention-to-treat" 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 therapies there may be observed differences in outcomes between trial and emulation.[26] Although this is also a common challenge in the analysis of RCTs,[25] treatment in routine clinical practice might be less stringent compared to trials in terms of pre-specified dosing schedules, monitoring and surveillance.[7] 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 amplified 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 narrow 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,[57] 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

Principled, well-designed and reproducible studies using fit-for-purpose RWD to generate RWE, such as through trial emulation, may complement evidence from RCTs. Through a systematic benchmarking approach, the ENCORE project will provide insights as to how methodological, design, and analytic decisions influence quantifiable bias and interpretability of non-interventional studies in oncology.

References

1. Framework for FDA's real-world evidence program (last accessed 11/28/2024). (2018).at <https://www.fda.gov/media/120060/download?attachment>
2. Mullard, A. 2023 FDA approvals. *Nat Rev Drug Discov* **23**, 88–95 (2024).
3. Merola, D., Schneeweiss, S., Schrag, D., Lii, J. & Lin, K. J. An algorithm to predict data completeness in oncology electronic medical records for comparative effectiveness research. *Annals of Epidemiology* **76**, 143–149 (2022).
4. Joshua Lin, K. *et al.* Longitudinal data discontinuity in electronic health records and consequences for medication effectiveness studies. *Clinical Pharmacology & Therapeutics* **111**, 243–251 (2022).
5. Rider, J. R. *et al.* Emulations of oncology trials using real-world data: A systematic literature review. *American journal of epidemiology* kwae346 (2024).
6. Merola, D. *et al.* The aetion coalition to advance real-world evidence through randomized controlled trial emulation initiative: oncology. *Clinical Pharmacology & Therapeutics* **113**, 1217–1222 (2023).
7. Merola, D. *et al.* Calibrating observational health record data against a randomized trial. *JAMA Network Open* **7**, e2436535–e2436535 (2024).
8. Wang, S. V. *et al.* Emulation of randomized clinical trials with nonrandomized database analyses: Results of 32 clinical trials. *Jama* **329**, 1376–1385 (2023).
9. Heyard, R., Held, L., Schneeweiss, S. & Wang, S. V. Design differences and variation in results between randomised trials and non-randomised emulations: Meta-analysis of RCT-DUPLICATE data. *BMJ medicine* **3**, (2024).
10. Calibrating real-world evidence studies in oncology against randomized trials: ENCORE (last accessed 11/28/2024). (2024).at <https://www.fda.gov/about-fda/oncology-center-excellence/calibrating-real-world-evidence-studies-oncology-against-randomized-trials-encore>
11. Rivera, D. R. *et al.* The oncology QCARD initiative: Fostering efficient evaluation of initial real-world data proposals. *Pharmacoepidemiology and Drug Safety* **33**, e5818 (2024).
12. Gatto, N. M. *et al.* The structured process to identify fit-for-purpose data: A data feasibility assessment framework. *Clinical Pharmacology & Therapeutics* **111**, 122–134 (2022).
13. Tasneem, A. *et al.* The database for aggregate analysis of ClinicalTrials. Gov (AACT) and subsequent regrouping by clinical specialty. *PloS one* **7**, e33677 (2012).
14. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European journal of cancer* **45**, 228–247 (2009).
15. Ton, T. G. N. *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. *Clinical Cancer Research* **28**, 2844–2853 (2022).

16. McKelvey, B. A. *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).
17. Chen, L. *et al.* Comparison of response from RECIST1.1 and abstraction in real-world patients with lung cancer. *Journal of Clinical Oncology* **41**, e21194–e21194 (2023).
18. Rivera, D. R. *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).
19. Ackerman, B. *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).
20. 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).
21. Wang, S. V. *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. *Value in Health* **25**, 1663–1672 (2022).
22. 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_Draft-Guideline_2024_0521.pdf (last accessed 01/07/2025).
23. Hernán, M. A., Wang, W. & Leaf, D. E. Target trial emulation: A framework for causal inference from observational data. *Jama* **328**, 2446–2447 (2022).
24. Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R. & Shrier, I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of clinical epidemiology* **79**, 70–75 (2016).
25. Rufibach, K. Treatment effect quantification for time-to-event endpoints–Estimands, analysis strategies, and beyond. *Pharmaceutical Statistics* **18**, 145–165 (2018).
26. Manitz, J. *et al.* Estimands for overall survival in clinical trials with treatment switching in oncology. *Pharmaceutical Statistics* **21**, 150–162 (2022).
27. Meyer, A.-M., Davies, J., Taylor, M. & Fruechtenicht, C. Open cohorts and ghost-time bias in real world data. In *PHARMACOEPIDEMIOLOGY AND DRUG SAFETY* **29**, 426–426 (WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2020).
28. Austin, P. C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine* **28**, 3083–3107 (2009).
29. Hernán, M. A. Causal analyses of existing databases: No power calculations required. *Journal of clinical epidemiology* **144**, 203–205 (2022).
30. Schoenfeld, D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68**, 316–319 (1981).

31. Franklin, J. M. *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).
32. Schneeweiss, S. *et al.* Graphical depiction of longitudinal study designs in health care databases. *Annals of internal medicine* **170**, 398–406 (2019).
33. Rubin, D. B. Inference and missing data. *Biometrika* **63**, 581–592 (1976).
34. Weberpals, J. *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).
35. Weberpals, J. *et al.* A principled approach to characterize and analyze partially observed confounder data from electronic health records. *Clinical Epidemiology* **16**, 329–343 (2024).
36. Moreno-Betancur, M. *et al.* Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. *American journal of epidemiology* **187**, 2705–2715 (2018).
37. Tompsett, D. M., Leacy, F., Moreno-Betancur, M., Heron, J. & White, I. R. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Statistics in medicine* **37**, 2338–2353 (2018).
38. Weberpals, J. *et al.* High-dimensional multiple imputation (HDMI) for partially observed confounders including natural language processing-derived auxiliary covariates. *American Journal of Epidemiology* (2025).at <https://arxiv.org/abs/2405.10925>
39. Shaw, C. K., P. & Williamson, B. D. Assessing treatment effects in observational data with missing confounders: A comparative study of practical doubly-robust and traditional missing data methods. *GitHub repository* (2024).at <https://github.com/PamelaShaw/Missing-Confounders-Methods>
40. Cox, D. R. Regression models and life-tables. Journal of the royal statistical society. *Series B (Methodological)* **34**, 187–220 (1972).
41. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53**, 457–481 (1958).
42. Becker, T. *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).
43. Leyrat, C. *et al.* Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical methods in medical research* **28**, 3–19 (2019).
44. Pishgar, F., Greifer, N., Leyrat, C. & Stuart, E. MatchThem:: Matching and weighting after multiple imputation. *arXiv preprint arXiv:2009.11772* (2020).
45. Rubin, D. B. Multiple imputation. In *Flexible imputation of missing data, second edition* 29–62 (Chapman; Hall/CRC, 2018).
46. Van Buuren, S. & Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in r. *Journal of statistical software* **45**, 1–67 (2011).
47. Franklin, J. M., Rassen, J. A., Ackermann, D., Bartels, D. B. & Schneeweiss, S. Metrics for covariate balance in cohort studies of causal effects. *Statistics in medicine* **33**, 1685–1699 (2014).

48. Stuart, E. A., Lee, B. K. & Leacy, F. P. 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, L. J. & Bryan, S. Can we afford to ignore missing data in cost-effectiveness analyses? *European Journal of Health Economics* **Vol.10**, 1–3 (2009).
50. Morisot, A. *et al.* Prostate cancer: Net survival and cause-specific survival rates after multiple imputation. *BMC medical research methodology* **15**, 1–14 (2015).
51. Pasek, J. Anesrake: ANES raking implementation. (2018).
52. Pishgar, F., Greifer, N., Leyrat, C. & Stuart, E. MatchThem:: Matching and weighting after multiple imputation. (2021).doi:10.32614/RJ-2021-073
53. Therneau, T. M. A package for survival analysis in r. (2024).at <https://CRAN.R-project.org/package=survival>
54. Weberpals, J. Encore.analytics: Functions and wrappers to streamline complex analytic workflows in real-world data studies based on the ENCORE trial emulation project. (2025).at <https://github.com/janickweberpals/encore.analytics>
55. Sjoberg, D. D., Whiting, K., Curry, M., Lavery, J. A. & Larmarange, J. Reproducible summary tables with the gtsummary package. **13**, 570–580 (2021).
56. Sjoberg, D. D., Baillie, M., Fruechtenicht, C., Haesendonckx, S. & Treis, T. Ggsurvfit: Flexible time-to-event figures. (2024).at <https://github.com/pharmaverse/ggsurvfit>
57. Gaber, C. E. *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

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)

Figures

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

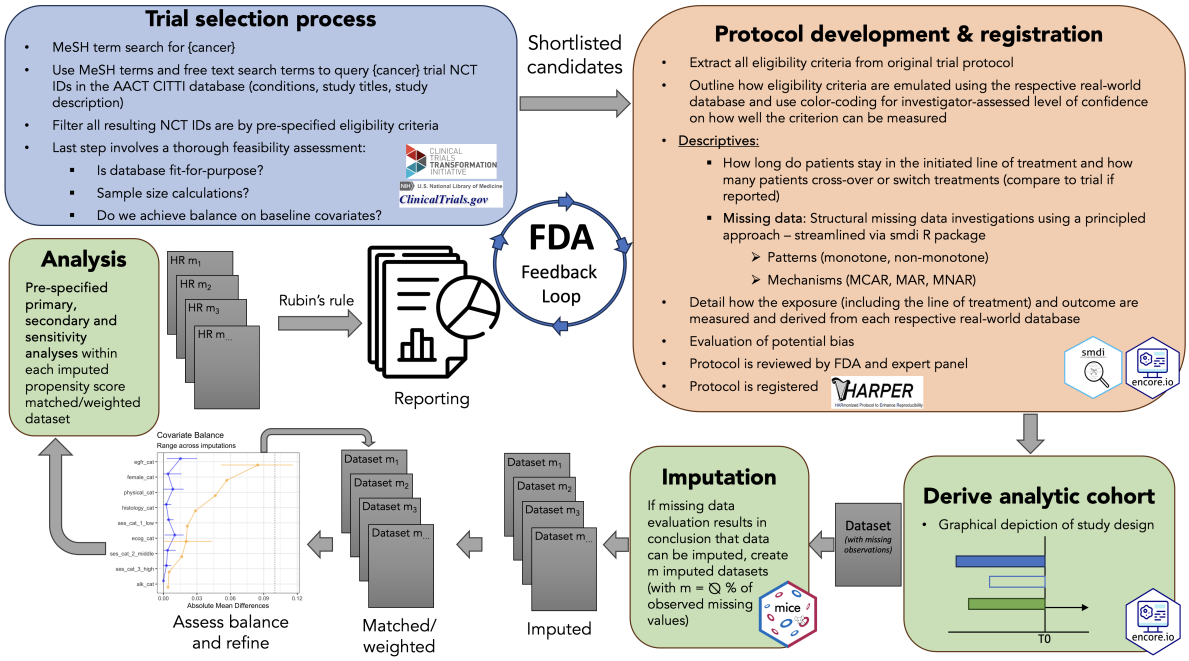

Figure 2: Example visualization of descriptive drug utilization analyses displaying a) initiation trends between compared regimens based on calendar time, b) cumulative rate of patients switching to another line of treatment.

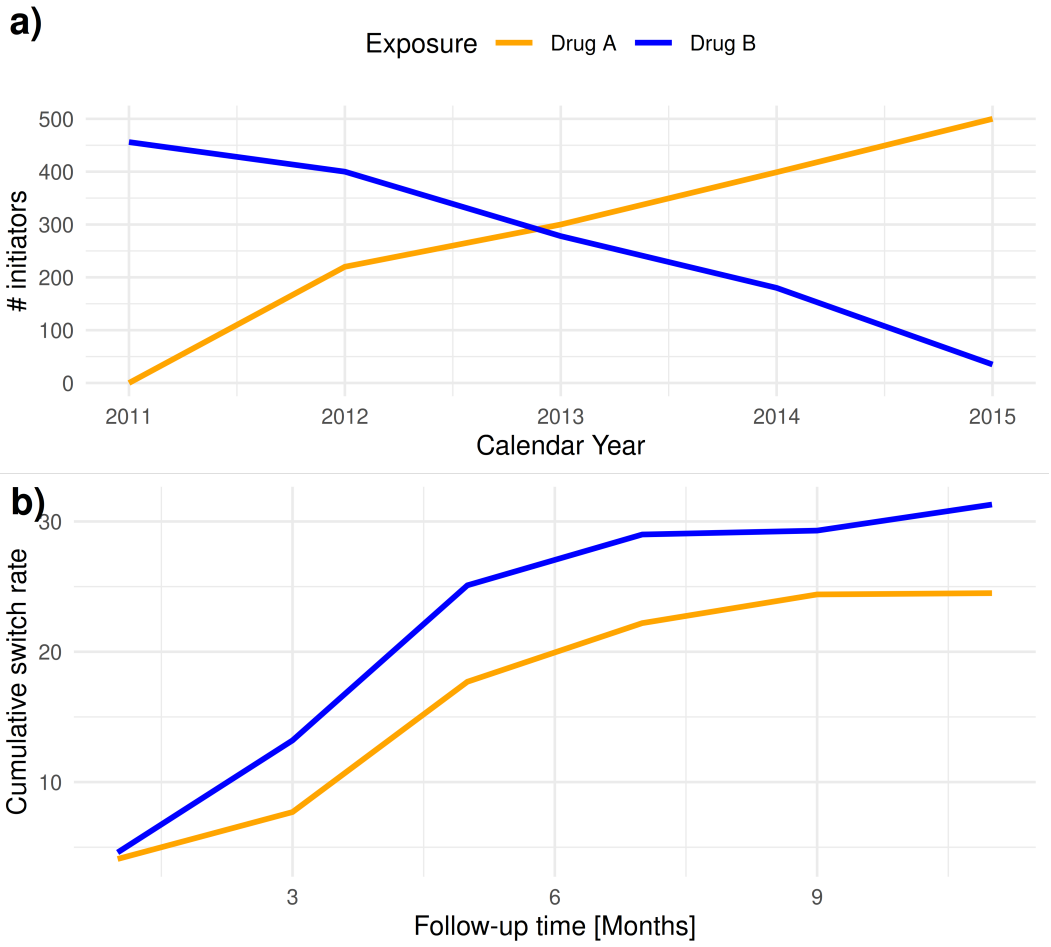

Figure 3: Assessment of a) covariate balance and b) distributional balance of a prognostic score for overall survival before and after propensity score matching or weighting across multiple imputed datasets.

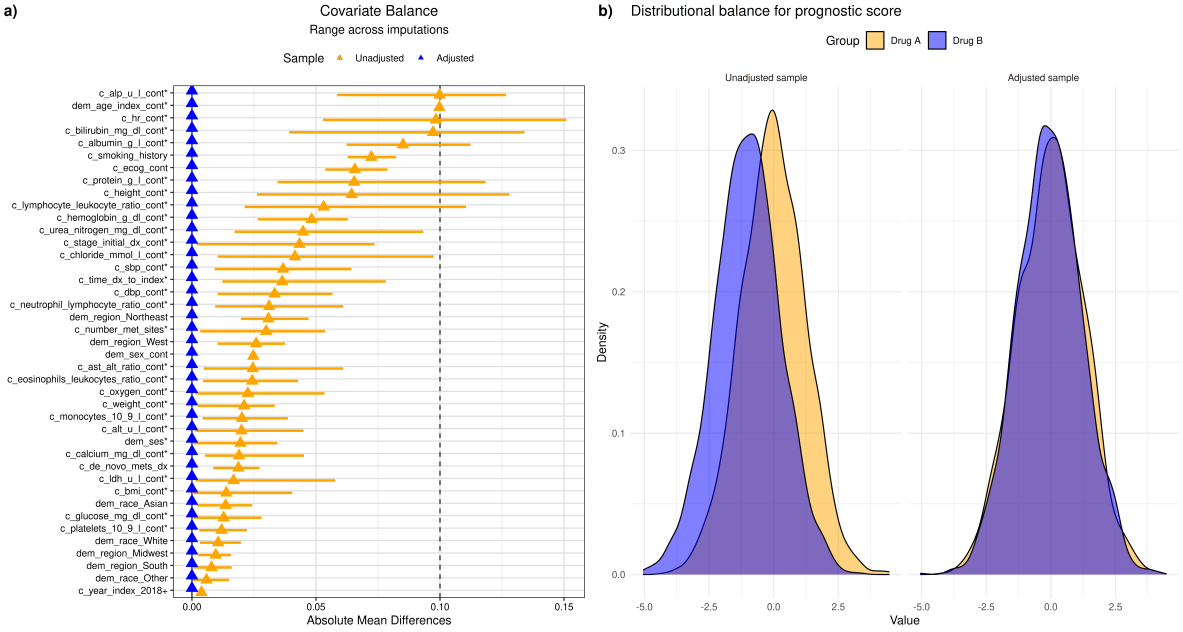

Table 2: Tentative list of randomized controlled trials (RCTs) considered for emulation.

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: Example visualization of agreement metrics.

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)