Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reproducing comparative therapeutic efficacy in relapsed refractory multiple myeloma using an external control arm derived from historical clinical trial data

Xiang Yin
Dassault Systèmes Company
Mehmet Burcu
Merck & Co., Inc
Mark D. Stewart
Friends of Cancer Research
Elizabeth Stuart
Johns Hopkins Bloomberg School of Public Health
Ruthanna Davi
ruthanna.davi@3ds.com

Dassault Systèmes Company

Research Article

Keywords: External controls, historical clinical trials, externally-controlled trials, methods, regulatory decision making

Posted Date: July 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9682017/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

External control arms (ECAs) are increasingly used to support single-arm oncology trials, but concerns remain about bias from measured and unmeasured confounding. We evaluated whether an ECA built from patient-level data from phase 3 historical clinical trials (HCTs) could reproduce the overall survival (OS) treatment effect observed in a randomized controlled trial (RCT) in relapsed/refractory multiple myeloma (RRMM).

Methods

Using de-identified patient-level data from dexamethasone control arms in historical RCTs and from a target RCT, we harmonized baseline covariates and applied propensity-score full optimal matching to create an ECA. Primary outcome was OS; time-to-event analyses included Cox proportional hazards models and Kaplan-Meier survival analyses with log-rank tests. We also performed sensitivity analyses (including tipping point analyses and E-value calculation) under a hypothetical context assuming the true effect was unknown, in order to evaluate the extent to which unmeasured confounder(s) would need to exert influence to change the analytical conclusion.

Results

After optimal full matching, the effective analytic sample comprised 290 patients in each cohort; 99% of the target RCT investigational arm patients were retained. Matching substantially improved covariate balance (propensity-score standardized mean difference decreased from 0.827 to 0.171) and all available baseline covariates had absolute standardized mean differences (SMD) ≤ 0.20 after matching. The target RCT comparison (investigational arm vs its randomized dexamethasone control from the same target RCT) showed improved OS for the investigational therapy (HR 0.74; 95% CI 0.60–0.92; log-rank p = 0.006). The externally controlled comparison (matched target RCT investigational arm vs matched ECA derived from other trials) produced a similar treatment estimate: HR 0.76 (95% CI 0.63–0.91; log-rank p = 0.0158). Sensitivity analyses indicated that only a moderately strong and imbalanced unmeasured confounder could overturn statistical significance.

Conclusions

An ECA constructed from harmonized historical trial data, analyzed with propensity-score-based full matching algorithm, produced an OS estimate closely matching the RCT findings. These findings support the potential utility of well-constructed ECAs to support oncology trials when high-quality patient-level

data are available from relevant prior trials, while underscoring the importance of rigorous confounding assessment.

INTRODUCTION

In drug development, randomized controlled trials (RCTs) have been considered the gold standard status in evaluating the safety and efficacy of medical treatments.[1] However, when single-arm trials are employed in oncology—often because deploying a concurrent control arm poses ethical, scientific, or feasibility challenges—use of non-randomized external control cohorts has become increasingly acceptable and common.[2–6] While single-arm trials can yield meaningful safety and efficacy signals and may support regulatory decision making in certain clinical contexts, well-conducted external controls may provide additional context and supplementary evidence.[7,8] Expanding the use of external controls could reduce patient burden and could also address situations where concurrent randomized control may be difficult to interpret, such as confirmatory trials of medical products made available through the accelerated approval pathway where the control arm may be compromised by early discontinuation or treatment crossover to the investigational therapy.[9]

Study designs that deviate from the traditional RCT, such as single arm or externally controlled trials, are considered in guidance for regulatory approval when their use is appropriately justified.[10–13] These designs may be warranted in settings where randomization is difficult or infeasible due to the rarity of the disease, limited patient availability, concerns about treatment switching/crossover, or ethical considerations related to withholding an investigational therapy. As described in regulation, external controls have generally been permitted in “special circumstances;” for example, “diseases with high and predictable mortality” and when, “effect of the drug is self-evident.”[10,11]

In addition to clinical, ethical, and regulatory considerations, the relatively restricted use is due, in part, to the perceived inability of external controls to be, “well assessed with respect to pertinent variables as can concurrent control populations,” as stated in FDA guidance and regulation. For example, as indicated in a recent landscape review,[14] while use of external controls is increasing, particularly in oncology submissions, some external control applications were deemed insufficient due to several limitations including unmeasured confounding and missing data issues inherent within real-world sources such as electronic medical/health records (EMR/EHRs). This underscores the need for additional methodological efforts and empirical demonstration projects that evaluate conditions under which external controls can provide reliable and interpretable comparative evidence including systematic assessments of data source characteristics, eligibility alignment, analytic adjustments, and sensitivity to unmeasured confounding.

We assessed the feasibility and reproducibility of constructing an external control arm using patient-level data from prior/historical clinical trials, which is a data source that can differ meaningfully from real-world data sources. Using historical trials to establish external controls may be advantageous due to standardized data collection in trial settings. Specifically, we developed an external control arm from

historical multiple myeloma clinical trials to mimic the control arm of a target RCT in this disease setting. The main objective was to assess whether the comparative treatment effect observed in the target RCT can be reproduced by comparing the investigational arm of the target RCT with an external control arm constructed from patient-level prior trial data by applying similar inclusion and exclusion criteria. In addition, we also performed sensitivity analyses (including tipping point analyses and E-value calculation) under a hypothetical context assuming the true effect was unknown, in order to evaluate the strength of a potential unmeasured confounder necessary to alter the observed treatment effect.[15–18]

METHODS

Data source

The study is based on analysis of patient level data from multiple historical randomized clinical trials in relapsed refractory multiple myeloma, accessible through Medidata's data insights sharing program. This platform includes more than 30,000 clinical trials, conducted by the pharmaceutical industry for drug or medical product development, with patient-level data recorded by electronic data capture, and are available through the Medidata Enterprise Data Store (MEDS).[19] MEDS is a collection of thousands of previous clinical trials with patient level data recorded through the Medidata electronic data capture system. Per the legal agreements with the sponsors of these historical clinical trials and Medidata, these data are available for use in deidentified (i.e., patients and original sponsor of the trial cannot be identified) and aggregated (i.e., every analysis must include data from two or more sponsors) form.

Characteristics of the included historical trials in multiple myeloma

The historical clinical trials included in this study were open-label phase 3 multinational randomized clinical trials that were conducted between 2010 and 2017 and included overall survival as an endpoint and dexamethasone as a control regimen. In terms of patient characteristics, at study inclusion (baseline), all patients in these trials were age ≥ 18 years, had relapsed or refractory multiple myeloma, received at least 2 prior lines of treatment, and received prior treatment with lenalidomide and bortezomib. The selection criteria were defined based on clinical importance, balancing the need to identify a homogenous set of historical clinical trial participants representative of a typical single indication in drug development of relapsed and refractory multiple myeloma while maximizing the volume of applicable historical data. There were a total of 946 patients (across trials) that were available after the selection process.

Construction process of the external control arm corresponding to a target trial

One of the largest RCT in the pool of eligible trials was selected as the 'target trial', and external control arm (ECA) patients were then selected based on eligibility requirements for the target trial. The 'target'

randomized trial had a 2:1 treatment assignment ratio and included 294 patients assigned to an investigational treatment and 149 patients assigned to dexamethasone as control. Patients for the ECA were selected from the remaining patients assigned to dexamethasone control in all other studies available within this project. Patients assigned to investigational therapies in trials except the target trial were not utilized in this study.

We used a two-step approach to (1) construct the ECA and (2) compare ECA and target trial patients with respect to overall survival endpoints. Overall survival was the endpoint of interest in this study and was measured as a key outcome in all historical trials that had similar study designs, such as the disease population and follow-up time. To construct the ECA, we standardized and appended patient-level historic clinical trial data to the target trial data and modeled treatment group membership as a function of prognostically important baseline demographic and disease variables.

Clinically important baseline covariates available across historical trials

Because the historical data in this study came from trials that had been conducted as part of clinical development programs, patient populations, study design, data collection methods, and endpoints were overall consistent across trials. Nevertheless, differences across studies in some variable definitions were present but have been reconciled as part of the data standardization process. Clinically important baseline covariates available across studies used in the creation of the ECA were as follows: age (continuous), sex (female vs. male), race (White vs. Others/unknown), region (Europe vs. Others/unknown), The Eastern Cooperative Oncology Group (ECOG) performance score (0 vs 1 vs 2 or 3), number of drug classes refractory (≥ 4 vs. <4), cytogenetic risk (high vs. standard/unknown), prior stem cell transplant (yes vs. no/unknown), days since last progression / relapse to first study dose (continuous), presence of bone lesions (yes vs. no/unknown), best response to last therapy (≥ partial response vs. <partial response), number of prior lines of therapy (continuous), years since diagnosis (continuous), and weight/body mass index (continuous).

Statistical and analytical methods

We used propensity scores to match ECA patients [derived from patients assigned dexamethasone control in all studies (except the target trial)] to patients assigned to the investigational treatment in the target clinical trial. Only baseline characteristics were utilized without knowledge of any post-randomization data to avoid selection bias and the index date for overall survival was consistently defined as the randomization date across all trials to avoid immortal time biases.[20,21]

A commonly used approach to propensity score matching, greedy 1–1 matching, was not possible for this case study due to the limited number of historical control patients available. Instead, we utilized a matching method called optimal full matching (often referred as full matching) which was introduced by

Rosenbaum and recommended recently by others.[22–24] In brief, full matching subdivides the subjects into strata of different sizes, consisting of either one target RCT investigational arm patient and at least one ECA patient, or, conversely, one ECA patient and at least one target RCT investigational arm patient. An attractive feature of this approach is that it can use most or even the entire set of all investigational treatment patients from the target trial in the analysis—in contrast with conventional matching approaches such as greedy matching where a portion of treated subjects cannot be matched and therefore are excluded from the final analysis. In summary, full matching aims to minimize the average differences within matched set in the propensity score between comparison cohorts, while avoiding potential bias due to incomplete matching, which can occur when some treated subjects are excluded from the matched sample. SAS PROC PSMATCH (SAS/STAT® 15.1) was used for matching, and the maximal number of ECA patients to be matched with each target trial patient and the maximal number of target trial patients matched to each ECA control were determined based on the ratio of number of subjects between the target RCT investigational arm patients and ECA patients[24] as well as the performance in terms of balancing baseline characteristics.

Post-matching covariate balance between matched cohorts was examined. Accordingly, baseline demographic and disease characteristics were summarized with descriptive statistics for the target RCT investigational arm and ECA patients. Standardized differences in covariate means before matching and after matching were computed and compared.[25] Absolute standardized difference of less than 0.20 has been taken to indicate a negligible difference in the mean or prevalence of a covariate between treatment groups. In addition, the matching process was evaluated by examining the distribution of propensity scores as well as individual baseline characteristics. Overall survival was compared between the comparison groups and assessed using a weighted Kaplan Meier curve, weighted log-rank test, and weighted Cox proportional hazard regression model for hazard ratio and corresponding 95% confidence interval.

Finally, under a hypothetical context assuming the true effect was unknown, we also conducted sensitivity analyses, including tipping point analysis and E-value calculation, in order to evaluate the extent to which an unmeasured confounder would need to exert influence to change the analytical conclusion. Utilizing methods proposed by Lin and colleagues,[16] we implemented the ‘tipping point analysis’ methodology. In brief, tipping point analysis is a special type of sensitivity analyses designed to assess how strong the association of an unobserved confounder with the treatment assignment and the outcome would have to be to change the study inference. We adjusted the observed treatment effect (HR and 95% confidence intervals) for overall survival to reflect the impact of a theoretical unobserved confounder. Additionally, as described by VanderWeele and colleagues,[18] we also conducted a quantitative bias analysis and calculated E-value. In brief, E-value is defined as the minimum strength of association that a hypothetical unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment–outcome association, conditional on the measured covariates. The calculation of the E-value for HR is based on the rarity of the outcome (common ≥ 15% or rare < 15%) and the observed hazard ratio. In our case, the outcome was common

where the observed overall death rate was 82%, therefore, the E-value for the estimated HR is calculated using the following formula:

RESULTS

Distribution of propensity scores

Before matching, the target RCT investigational arm included 294 patients and the control patients from historical trials included 201 patients; after optimal full propensity-score matching, the comparative analysis set comprised 290 target RCT investigational arm patients and weighted sample size of 290 ECA patients (Table 1). The Cloud Plot (**Supplementary Fig. 1**) shows the distribution of propensity scores for the target RCT investigational arm and all available control patients from other historical RCTs. Overlap in the distribution of propensity scores for the target RCT investigational arm and the controls from other historical RCTs was nearly complete. There were few patients for whom a match was not available; these are generally in the tails of the distributions where there were no analogous patients available in the opposite group. Near all (N = 290, 99%) in the target RCT investigational arm were successfully matched. The remaining 4 patients (1%) were not matched and were removed from further analysis.

Patient demographic and disease characteristics before and after optimal full propensity-score matching

Table 1 summarizes baseline demographics and disease characteristics for patients in the target RCT investigational compared with patients in control arms of historical RCTs, before and after optimal full propensity-score matching. Matching substantially improved covariate balance: the standardized mean difference for the propensity score decreased from 0.83 before matching to 0.17 after matching, and standardized mean differences for most key baseline variables (age, sex, race, ECOG performance status, number of drug classes refractory, cytogenetic risk, prior stem cell transplant, days since last progression/relapse, bone lesions, best response to last therapy, number of prior lines, years since diagnosis, and body weight) were reduced to ≤ 0.20 or brought close to zero (Table 1, Supplementary Fig. 2).

After matching, the two groups were highly comparable. Mean age was 63.5 (SD 9.40) years in the matched investigational arm and 64.3 (SD 9.58) years in the external control arm; approximately 60% of patients in both cohorts were male. The majority of patients were White (81.0% vs 83.1%) and were recruited from Europe (80.0% vs 77.2%). Baseline ECOG performance status distributions were similar (ECOG 0: 36.2% vs 31.4%; ECOG 1: 46.2% vs 49.7%; ECOG 2–3: 17.6% vs 19.0%), as were measures of disease burden and prior therapy: roughly 70% had bone lesions, approximately 72%–69% were refractory to ≥ 4 drug classes, most patients had standard/unknown cytogenetic risk (90.0% vs 86.6%), about 61% had a prior stem cell transplant, and approximately 78% had received ≥ 4 prior lines of therapy.

Median time from last progressive disease/relapse to first study dose was 44 days in both groups, and years since diagnosis and body weight were also similar (Table 1). These post-matching characteristics indicate that the matched investigational and external control groups are well balanced for subsequent comparative analyses.

Comparison of treatment effects across patient cohorts

In this methodology project, we aimed to examine whether the treatment effect based on an external control arm (i.e., matched target RCT investigational arm vs. ECA derived from control arm patients from other historical RCTs) can mimic the treatment effect from the target RCT. Figure 1 provides Kaplan-Meier curves of overall survival estimates in these four groups (i.e., target RCT investigational arm, target RCT control arm, matched investigational arm, matched ECA). There were few patients from the target RCT investigational arm that were not matched, and thus, Kaplan-Meier curves were presented separately for the target RCT investigational arm and for the *matched* target RCT investigational arm.

The target RCT demonstrated a positive treatment effect on overall survival, as evidenced by a separation of the Kaplan Meier curves representing the investigational and randomized control arms of the target RCT. In the target RCT, the HR for the investigational arm versus the randomized control is 0.74 with confidence interval that excludes 1 (95% CI: (0.60, 0.92)). This difference between groups is also supported by the log-rank test (p = 0.006).

The treatment effect utilizing ECA is very similar. The Kaplan Meier curve for the ECA visually overlaps and crosses with that of the target RCT control arm and the quantified differences between ECA and the matched target RCT investigational arm were very similar to the original RCT. The hazard ratio for the matched investigational arm versus the ECA is 0.76 with confidence interval that excludes 1 (95% CI: (0.63, 0.91). This difference between groups is also supported by the log-rank test (p = 0.0158).

Sensitivity Analyses for Potential Unmeasured Confounding

Tipping point analysis

As discussed in the methods section, using Lin's method,[16] the observed treatment effect can be adjusted for a hypothetical unmeasured confounder when its prevalence in each arm and its effect on outcome are specified. For example, if the unmeasured confounder is present in 10% of the investigational arm and 30% of the ECA, with a confounder effect for overall survival HR of 1.5, the adjusted treatment effect is HR = 0.83 (95% CI 0.69–0.99), which remains statistically significant and consistent with the unadjusted result. If the unmeasured confounder prevalence (in the ECA) is instead 35% (all else equal), the adjusted treatment effect HR becomes 0.85 (95% CI 0.71–1.02), losing statistical significance— a "statistical tipping point." A separate "clinical tipping point" occurs where the unmeasured confounder prevalence in ECA is 80% (all else equal), the adjusted treatment effect HR exceeds 1 and the direction of effect reverses.

To explore the full range of plausible scenarios, we calculated adjusted treatment effect HRs (and 95% CIs) across a grid of confounder prevalences (0–0.80 in 0.05 steps) (Table 2 and Supplemental Table 1). Table 2 presents results when moderate confounder effect is assumed (HR = 1.5), and Supplemental Table 1 presents results when a stronger confounder effect is assumed (HR = 2.0); each cell displays the adjusted treatment effect (HR and 95% CI) under those assumptions.

Quantitative bias analysis

In a quantitative bias analysis, we calculated the E-value using the formula for a common outcome; and based on the observed hazard ratio. The proportion of death was 81.76% in the overall population, with 84.46% in the investigational cohort vs. 79.05% in the ECA. The observed HR when comparing investigational arm matched to ECA was 0.76 (95% CI, 0.63, 0.91). Therefore, the point estimate for E-value was calculated as 1.71 as illustrated in the supplementary Fig. 3. This means an unmeasured confounder would need to be associated with a 1.71-fold increased risk for both exposure and outcome to nullify the observed 0.76 hazard ratio.

DISCUSSION

In this methodological case study, we evaluated whether an external control arm constructed from patient-level data from prior randomized trials could reproduce the treatment effect observed in a contemporaneous randomized control arm in relapsed/refractory multiple myeloma (RRMM). After harmonizing trial data, applying the target trial eligibility criteria, and implementing optimal full matching on propensity score, the matched investigational cohort and the external control arm were highly comparable across available baseline covariates. Importantly, the estimated treatment effect using the external control (HR 0.76; 95% CI 0.63–0.91) closely aligned with the randomized comparison in the target trial (HR 0.74; 95% CI 0.60–0.92). In addition, under a hypothetical assumption if the true effect was not known, our sensitivity analyses (tipping-point analysis and E-value) indicated that only a moderately strong and imbalanced unmeasured confounder could overturn statistical significance..

Our findings demonstrate that, under favorable conditions and depending on clinical and regulatory context, historical trial data can be leveraged to produce an external control that yields treatment effect estimates consistent with those from a randomized controlled comparison. Several factors likely contributed to the successful replication in this project. First, the historical trials used to build the external control were similar in population, design, covariate and outcome ascertainment, and follow-up to the target trial; these similarities reduced the risk of systematic differences in measurement and patient follow-up. Second, because the source data were patient-level clinical trial records collected through standardized electronic data capture, key prognostic baseline covariates were consistently available and of high quality. Third, the analytic approach—optimal full matching—maximized use of available treated subjects while achieving good covariate balance, thereby reducing bias that can arise when treated patients are excluded by more restrictive matching algorithms especially in low sample size situations.

Our work builds on existing methodological demonstrations,[26–31] but further empirical and systematic evaluations are also needed across different disease areas, endpoints (including intermediate endpoints, e.g., progression-free survival, response rate), data sources (clinical trials, registries, and EHRs across regions), and study designs (retrospective and prospective) to elucidate strengths and limitations of different approaches and lessons they yield under varied clinical and regulatory circumstances. Collaborative, multidisciplinary, cross-sector efforts with regulators to establish standards for acceptable external control construction, pre-specification, and sensitivity analyses would support appropriate use of these methods in regulatory and clinical decision making. Encouragingly, several initiatives are already advancing this work. Project SignifiCanT (Statistics in Cancer Trials) at the FDA Oncology Center of Excellence fosters stakeholder collaboration to improve the design and analysis of innovative cancer trial designs,[32] including those with hybrid control arms.[33] Similarly, Friends of Cancer Research is piloting an external-control project that uses multiple data sources to replicate a randomized clinical trial and convened a public meeting to explore ways to expand the use of external control arms in oncology development.[34]

Strengths and Limitations

Using de-identified patient-level data from multiple phase 3 trials allowed detailed harmonization of covariates and consistent application of the target trial's eligibility criteria. The use of optimal full matching preserved nearly all treated patients from the target trial. We assessed balance comprehensively with standardized differences and propensity-score distributions and applied transparent, pre-specified methods for outcome comparison. The validity of externally controlled comparisons hinges on the assumption that all important confounders are measured, or that any unmeasured confounding is weak or balanced. When generating evidence based on external control arms, it is critical to conduct sensitivity analyses. Therefore, to demonstrate the importance of conducting sensitivity analyses, under a hypothetical context assuming the true effect was unknown, we implemented sensitivity analyses (tipping-point and E-value) to quantify the robustness of the results to unmeasured confounding, rather than rely solely on covariate balance metrics. While sensitivity analyses can help quantify how strong a potential hypothetical unmeasured confounder would need to be to change statistical inference, it is also important to note that those analyses cannot identify whether such a confounder actually exists. Although tipping-point analysis is often defined to exploring missing-data effects,[17] the availability of investigational-arm data (from the target RCT) in our project dataset permitted estimation of hazard ratios across different unmeasured-confounding assumptions — a capability not feasible in prior work where randomized-control-based comparative results were unavailable.[26]

In RRMM, unmeasured factors such as subtle differences in prior therapies, frailty measures, supportive care, or temporal changes in practice patterns could influence survival and remain incompletely captured even within trial datasets. Although historical clinical trials offer advantages over many real-world data sources (e.g., standardized measurements, curated endpoints), they have their own selection biases.

Trials typically enroll patients who meet strict inclusion/exclusion criteria and may differ from routine clinical populations in ways that affect outcomes.[35] However, because our external control was constructed from trials similar to the target trial and we applied the same inclusion criteria, this source-selection bias was minimized in the matched comparison; nevertheless, residual differences in unrecorded eligibility nuances or site-level practices across sponsoring trials could remain. While the historical trials used here spanned a relatively narrow calendar period (2010–2017) and had common control regimens (dexamethasone), secular trends in supportive care, subsequent lines of therapy, or diagnostic approaches could introduce bias if the treatments or follow-up differed meaningfully between trials.[7] Our design reduced, but perhaps could not eliminate, the influence of such temporal or potential center-specific effects. Even with optimal full matching, matching is limited by the available pool of external controls. For example, although overlap in propensity scores was nearly complete in this project, applications to rarer subgroups or situations with fewer historical controls may not achieve similar balance or coverage of treated patients. Finally, the analysis focused on overall survival, a 'hard' endpoint uniformly collected across trials; externally controlled comparisons of more 'subjective' endpoints (e.g., progression-free survival, patient-reported outcomes) may be more vulnerable to ascertainment differences and measurement bias across studies.

Conclusions

In this RRMM case study, an external control arm constructed from harmonized historical randomized trial data and analyzed with optimal full matching produced an estimate of treatment effect on overall survival that closely matched the randomized control comparison. When historical trial sources are well aligned with the target trial and when rich, high-quality patient-level covariates are available, externally controlled analyses can provide credible comparative effectiveness evidence. Nevertheless, these analyses remain susceptible to unmeasured confounding and other biases; therefore, transparency, careful design, rigorous confounding assessment, and close consideration of the data generation context are essential before relying on external controls for regulatory or clinical decision making.

Declarations

Conflict of Interest: Dr. Yin is an employee and stockholder of Dassault Systèmes company, New York, USA. Dr. Burcu is an employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA, and owns stock in Merck & Co., Inc., Rahway, NJ, USA. Dr. Stuart receiving methodological research grants from the US National Institutes of Health, the Patient Centered Outcomes Research Institute, and Eli Lilly & Co. She also reports editorial service as Statistical Editor for JAMA Health Forum, consulting for Eli Lilly & Co., and legal consulting on topics related to genital talc exposure and cancer screening. Dr. Davi is an employee and stockholder of Dassault Systèmes company, New York, USA. The views and opinions expressed in this article are those of the authors and may not necessarily represent the views and opinions of any organization, employer, or affiliated group.

Funding: No direct monetary funding was provided. Data and related administrative support were provided by Medidata, a Dassault Systèmes Company, New York, NY, USA.

Author contributions: All authors have contributed to fulfilling the International Committee of Medical Journal Editors (ICMJE) authorship criteria. Each author had substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; and drafting the work or revising it critically for important intellectual content; and final approval of the version to be published; and agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the Gold Standard--Lessons from the History of RCTs. N Engl J Med 2016;374(22):2175-81. DOI: 10.1056/NEJMms1604593.
2. Agrawal S, Arora S, Amiri-Kordestani L, et al. Use of Single-Arm Trials for US Food and Drug Administration Drug Approval in Oncology, 2002-2021. JAMA Oncol 2023;9(2):266-272. DOI: 10.1001/jamaoncol.2022.5985.
3. Goring S, Taylor A, Muller K, et al. Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: a systematic review. BMJ Open 2019;9(2):e024895. DOI: 10.1136/bmjopen-2018-024895.
4. Izem R, Buenconsejo J, Davi R, Luan JJ, Tracy L, Gamalo M. Real-World Data as External Controls: Practical Experience from Notable Marketing Applications of New Therapies. Ther Innov Regul Sci 2022;56(5):704-716. DOI: 10.1007/s43441-022-00413-0.
5. Jahanshahi M, Gregg K, Davis G, et al. The Use of External Controls in FDA Regulatory Decision Making. Ther Innov Regul Sci 2021;55(5):1019-1035. DOI: 10.1007/s43441-021-00302-y.
6. Taylor L, Chen A, Pierre A. Utilization of real-world evidence in regulatory approvals for multiple myeloma therapies. Blood Cancer J 2025;15(1) (In English). DOI: ARTN 210 10.1038/s41408-025-01420-8.
7. Burcu M, Dreyer NA, Franklin JM, et al. Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. Pharmacoepidemiol Drug Saf 2020;29(10):1228-1235. DOI: 10.1002/pds.4975.
8. Mack C, Christian J, Brinkley E, Warren EJ, Hall M, Dreyer N. When Context Is Hard to Come By: External Comparators and How to Use Them. Ther Innov Regul Sci 2020;54(4):932-938. DOI: 10.1007/s43441-019-00108-z.
9. Schmidli H, Haring DA, Thomas M, Cassidy A, Weber S, Bretz F. Beyond Randomized Clinical Trials: Use of External Controls. Clin Pharmacol Ther 2020;107(4):806-816. DOI: 10.1002/cpt.1723.
10. CFR 314.126. "Adequate and well-controlled studies".
11. U.S. Food and Drug Administration. Guidance for Industry. E 10 Choice of Control Group and Related Issues in Clinical Trial. (https://www.fda.gov/media/71349/download).

12. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. (https://www.fda.gov/media/120060/download?attachment).
13. U.S. Food and Drug Administration. Guidance for Industry. Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products (https://www.fda.gov/media/164960/download).
14. Alipour-Haris G, Liu X, Acha V, Winterstein AG, Burcu M. Real-world evidence to support regulatory submissions: A landscape review and assessment of use cases. Clin Transl Sci 2024;17(8):e13903. DOI: 10.1111/cts.13903.
15. Ding P, VanderWeele TJ. Sensitivity Analysis Without Assumptions. Epidemiology 2016;27(3):368-77. DOI: 10.1097/EDE.0000000000000457.
16. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 1998;54(3):948-63. (https://www.ncbi.nlm.nih.gov/pubmed/9750244).
17. Liu Y, Zhou K, Sims KD. Tipping Point Analysis: Assessing the Potential Impact of Missing Data. JAMA 2025;334(3):265-266. DOI: 10.1001/jama.2025.7420.
18. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Ann Intern Med 2017;167(4):268-274. DOI: 10.7326/M16-2607.
19. Medidata Clinical Trial Data Platform. (https://www.medidata.com/en/clinical-trial-products/medidata-ai/real-world-data/).
20. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol 2008;167(4):492-9. DOI: 10.1093/aje/kwm324.
21. Yadav K, Lewis RJ. Immortal Time Bias in Observational Studies. JAMA 2021;325(7):686-687. DOI: 10.1001/jama.2020.9151.
22. Austin PC, Stuart EA. Optimal full matching for survival outcomes: a method that merits more widespread use. Stat Med 2015;34(30):3949-67. DOI: 10.1002/sim.6602.
23. Rosenbaum PR. A Characterization of Optimal Designs for Observational Studies. J R Stat Soc Series B Stat Methodol 1991;53(3):597-610.
24. Hansen BB. Full matching in an observational study of coaching for the SAT. J Am Stat Assoc 2004;99(467):609-618. (In English). DOI: 10.1198/016214504000000647.
25. Normand ST, Landrum MB, Guadagnoli E, et al. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. J Clin Epidemiol 2001;54(4):387-98. DOI: 10.1016/s0895-4356(00)00321-8.
26. Yin X, Mishra-Kalyan PS, Sridhara R, Stewart MD, Stuart EA, Davi RC. Exploring the Potential of External Control Arms created from Patient Level Data: A case study in non-small cell lung cancer. J Biopharm Stat 2022;32(1):204-218. (In English). DOI: 10.1080/10543406.2021.2011901.
27. 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-

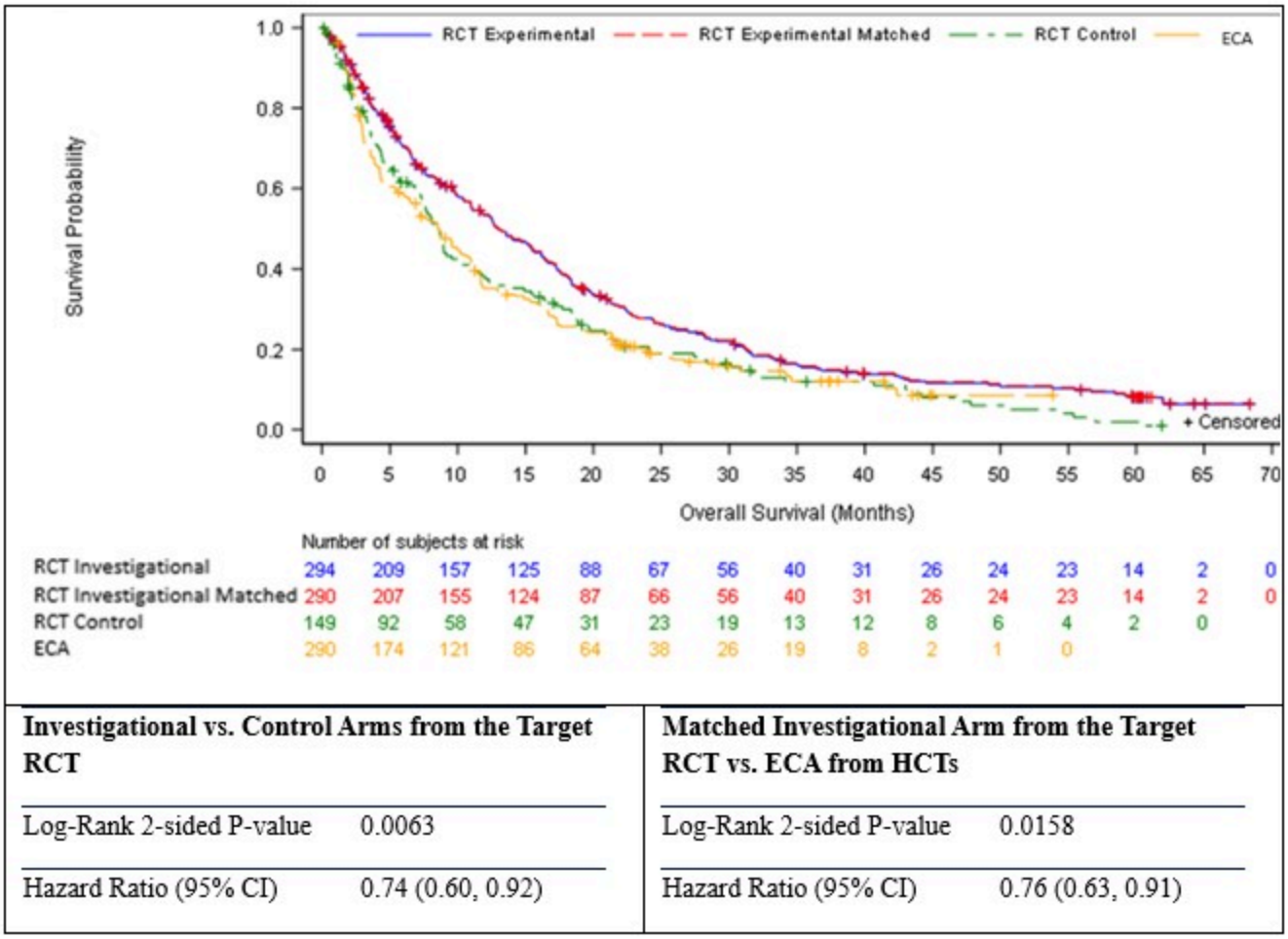
Small Cell Lung Cancer Treated With Chemotherapy. Jco Clin Cancer Info 2024;8 (In English). DOI: ARTN e2400091 10.1200/CCI.24.00091.

28. 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 2023;113(6):1217-1222. (In English). DOI: 10.1002/cpt.2800.
29. Natalie L, Paige S, Ulka C, et al. Oncology Trial Emulation Using Real-World Electronic Health Record Data: Results of the Coalition to Advance Real-World Evidence through Randomized Controlled Trial Emulation (Care) Initiative. Value Health 2025;28(6):S14-S14. (In English) (<Go to ISI>://WOS:001543678800034).
30. Weberpals J, Schneeweiss S, Kehl KL, et al. Emulating Comparative Oncology Trials With Real-World Evidence Studies (ENCORE): Process Development and Methodological Considerations for Oncology Real-World Data. Clinical Pharmacology & Therapeutics 2026 (In English). DOI: PMID 372741 10.1002/cpt.70208.
31. Jemielita T, Li XY, Piperdi B, Zhou W, Burke T, Chen C. Overall Survival With Second-Line Pembrolizumab in Patients With Non-Small-Cell Lung Cancer: Randomized Phase III Clinical Trial Versus Propensity-Adjusted Real-World Data. Jco Clin Cancer Info 2021;5:56-65. (In English). DOI: 10.1200/Cci.20.00099.
32. U.S. Food and Drug Administration. Project SignifiCanT: Statistics in Cancer Trials. Promoting collaboration in design and analysis of cancer clinical trials. (https://www.fda.gov/about-fda/oncology-center-excellence/project-significant-statistics-cancer-trials).
33. Ran D, Zhang F, Broglio K, Shahsavari S, Henderson A, Yu B. Statistical Guardrails for Hybrid-Controlled Trials: Robust to Confounding and Between-Study Heterogeneity. Ther Innov Regul Sci 2026. DOI: 10.1007/s43441-026-00932-0.
34. Friends of Cancer Research. ECA Pilot: Informing the Use of External Control Arms in Clinical Development and Regulatory Decision-Making. (https://friendsofcancerresearch.org/eca/).
35. Vo TT, Kim S, Jaksa A, et al. Diversity by Design: Real-World Data to Enhance Representation in Clinical Cancer Research. Clin Cancer Res 2025;31(21):4396-4404. (In English). DOI: 10.1158/1078-0432.Ccr-25-1255.

Tables

Tables 1 and 2 are available in the Supplementary Files section.

Figures



Investigational vs. Control Arms from the Target RCT		Matched Investigational Arm from the Target RCT vs. ECA from HCTs	
Log-Rank 2-sided P-value	0.0063	Log-Rank 2-sided P-value	0.0158
Hazard Ratio (95% CI)	0.74 (0.60, 0.92)	Hazard Ratio (95% CI)	0.76 (0.63, 0.91)

Figure 1

Kaplan-Meier Curve of Overall Survival (OS) in Comparison Groups

Legend—The following groups were compared: Experimental arm from the target RCT; Control arm from the target RCT; Post-matching experimental arm from the target RCT; and Post-matching ECA derived from historical RCTs

Abbreviations, HCT = historical clinical trials; RCT = randomized controlled trial; ECA = external control arm

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementalmaterials05052026.docx
- Maintables.docx