


Augmenting external control arms using Bayesian borrowing: a case study in first-line non-small cell lung cancer

Alessandria Struebing^{*,1} , Chelsea McKibbon², Haoyao Ruan², Emma Mackay², Natalie Dennis³, Russanthy Velummailum², Philip He⁴, Yoko Tanaka⁴, Yan Xiong⁴, Aaron Springford² & Mats Rosenlund^{1,5}

¹Daiichi Sankyo Europe, Munich, 81379, Germany

²Cytel Inc., Toronto, Ontario, M5J, 2P1, Canada

³Daiichi Sankyo Oncology, Rueil-Malmaison, 92500, France

⁴Daiichi Sankyo, Inc., Basking Ridge, NJ 07920, USA

⁵Department of Learning, Informatics, Management & Ethics (LIME), Karolinska Institutet, Stockholm, 171 77, Sweden

*Author for correspondence: alessandria.struebing@daiichi-sankyo.eu

Aim: This study aimed to improve comparative effectiveness estimates and discuss challenges encountered through the application of Bayesian borrowing (BB) methods to augment an external control arm (ECA) constructed from real-world data (RWD) using historical clinical trial data in first-line non-small-cell lung cancer (NSCLC). **Materials & methods:** An ECA for a randomized controlled trial (RCT) in first-line NSCLC was constructed using ConcertAI Patient360™ to assess chemotherapy with or without cetuximab, in the bevacizumab-inappropriate subpopulation. Cardinality matching was used to match patient characteristics between the treatment arm (cetuximab + chemotherapy) and ECA. Overall survival (OS) was assessed as the primary outcome using Cox proportional hazards (PH). BB was conducted using a static power prior under a Weibull PH parameterization with borrowing weights from 0.0 to 1.0 and augmentation of the ECA from a historical control trial. **Results:** The constructed ECA yielded a higher overall survival (OS) hazard ratio (HR) (HR = 1.53; 95% CI: 1.21–1.93) than observed in the matched population of the RCT (HR = 0.91; 95% CI: 0.73–1.13). The OS HR decreased through the incorporation of BB (HR = 1.30; 95% CI: 1.08–1.54, borrowing weight = 1.0). BB was applied to augment the RCT control arm via a historical control which improved the precision of the observed HR estimate (1.03; 95% CI: 0.86–1.22, borrowing weight = 1.0), in comparison to the matched population of the RCT alone. **Conclusion:** In this study, the RWD ECA was unable to successfully replicate the OS estimates from the matched population of the selected RCT. The inability to replicate could be due to unmeasured confounding and variations in time-periods, follow-up and subsequent therapy. Despite these findings, we demonstrate how BB can improve precision of comparative effectiveness estimates, potentially aid as a bias assessment tool and mitigate challenges of traditional methods when appropriate external data sources are available.

Plain language summary: Researchers and health agencies need accurate and reliable estimates of the relative effectiveness of treatments. In some cases, an 'external control group' can be created from pre-existing healthcare records, eliminating the need for a concurrent control and the associated patient burden. However, it can be challenging to find comparable patients with similar characteristics, and comparable information on health outcomes for those patients. Often, there are just not enough patients available in existing records to make a meaningful comparison. Bayesian borrowing can be used to include control arm information from historical studies, even if these studies differed somewhat from the present study.

In this article, we investigate the ability of external controls and Bayesian borrowing to replicate findings from a randomized controlled trial (RCT – the gold standard). We begin with an RCT and then remove the control group before constructing an external control group using real-world data from electronic health record data. We then use Bayesian borrowing to add information from another historical trial in the same indication to see whether the original RCT results can be replicated, and under what conditions. Our example illustrates common challenges of working with real-world data and provides

practical insights for the incorporation of additional data sources when comparing the effectiveness of treatments that have not been compared directly in an RCT.

Tweetable abstract: In #CancerResearch, #RealWorldData can provide external control arms (ECAs) for single arm trials. But are ECAs fit-for-purpose when comparing treatments? We show how historical trials and #BayesianBorrowing might be used to increase confidence in ECAs.

First draft submitted: 1 December 2023; Accepted for publication: 1 March 2024; Published online: 4 April 2024

Keywords: Bayesian borrowing • bias assessment • comparative effectiveness • external control arm • health technology assessment • real-world data

For regulatory decision making, randomized controlled trials (RCTs) are considered the gold standard for providing comparative effectiveness evidence, but they are not without limitations. Although well-designed RCTs can demonstrate the relative efficacy of an intervention, they have challenges for implementation: they can be expensive to carry out, patients may be difficult to recruit and volunteer trial patients may not generalize to the broader population of interest [1]. Single-arm trials, particularly in rare diseases and oncology, are well suited to alternative trial designs – the target patient populations can be small and patients difficult to recruit, the use of placebo raises ethical concerns in some cases, the trial designs are often complex and patient burden tends to be high [2]. For these reasons, single-arm trials are being increasingly accepted by regulatory bodies such as the US FDA and the European Medicines Agency (EMA) [3,4]. External control arms (ECAs) are an alternative to providing comparative effectiveness estimates and can be constructed using external information from real-world data (RWD) or data from previous clinical trials. In the absence of randomization, statistical methods such as propensity score matching (PSM) or weighting are generally used when constructing ECAs for single-arm trials in an attempt to limit the influence of confounding factors on estimates of relative effectiveness [5,6].

Much has been published on the considerations for conducting externally controlled studies and their performance in recent years [5,7–10]. Regulatory and reimbursement agencies are also actively developing recommendations for the use of external data sources and applications of ECAs [11–13]. A common challenge for ECAs is that the external data may have been collected for purposes other than comparative effectiveness analyses and may be missing important covariate information and be subject to an unstandardized evaluation of outcomes. These differences can result in reduced study power or residual bias – for example due to differences in routine care or data recording standards, and difficulties implementing eligibility criteria that match the single-arm trial [5,14–16]. Regulatory and reimbursement agencies most commonly criticize ECAs based on the choice of study populations (e.g., not representative of standard of care and/or non-generalizability), unmeasured or uncontrolled confounding (selection bias), immortal time bias, inconsistent definition of outcomes, data missingness and lack of statistical power [16,17]. Thus, when selecting an external data source, capture and coverage of relevant covariates, comparability of populations and data provenance all require careful consideration.

Bayesian borrowing (BB) methods use data from external sources such as historical control arms to bolster limited sample sizes and, if chosen appropriately, can increase the precision of estimates while mitigating type I error [10,18]. Power prior methods are a type of BB in which the amount of borrowing from the external data source is down-weighted relative to the concurrent data [19]. The amount of borrowing can be fixed, or it can be allowed to vary based on the compatibility of the internal and external data sources (i.e., if outcomes are consistent or inconsistent across data sources). If the assessment of comparative effectiveness is sensitive to the amount of borrowing, a tipping point analysis can determine how much weight on the external data is needed to meet a particular effectiveness condition [20]. Other BB applications to reduce control group allocation by borrowing information from external data are advantageous in instances where recruitment may be particularly challenging [21]. A recent example of this type of approach was the use of robust meta-analytic predictive (MAP) priors to augment a small concurrent control arm in a trial of the BI 1015550 phosphodiesterase 4 inhibitor for idiopathic pulmonary fibrosis [22]. The current study focuses on static borrowing via a fixed power prior in conjunction with tipping point analysis to contextualize the impact of the fixed borrowing choice.

Existing guidance on BB has mostly been limited to the regulatory space, starting with early receptiveness from the FDA to borrowing of information from historical controls in medical device trials by means of Bayesian priors [23] and from the EMA by proposing a framework for extrapolation between source and target populations

including the use of Bayesian methodology in trials with small patient populations [24–26]. Further precedent on the acceptance of BB methods in a regulatory setting has been established with the FDA approval of belimumab in pediatric patients with systemic lupus erythematosus, through borrowing of adult effectiveness data [27–29]. The FDA has since provided additional guidance on the use of Bayesian methods for borrowing information from external data sources for drugs and biologics, with particular emphasis on the need for selection of appropriate external data sources and consideration of the operating characteristics of these complex and innovative designs [30,31]. Bayesian approaches hold a significant degree of promise, but additional work is needed to outline potential applications and understand their relative advantages and disadvantages, especially within a health technology assessment (HTA) framework for assessing comparative effectiveness.

In this paper, we augment an ECA using BB to improve estimates of comparative effectiveness in first-line metastatic non-small-cell lung cancer (NSCLC). The treatment landscape for NSCLC is evolving rapidly since the discovery of driver mutations and development of targeted therapies and immune checkpoint inhibitors resulting in a significant change in standard of care for metastatic disease in recent years [32,33], increasing the need to generate sound comparative effectiveness evidence for regulatory and HTA bodies to use for decision-making. To compare the effectiveness of active treatment against a control in terms of overall survival (OS), we used the treatment arm from an open-access two-arm RCT, constructed an ECA using a RWD source and augmented it by borrowing additional control arm data from a previously conducted RCT using BB. Having a concurrent control arm available allowed us to compare results from the ECA plus BB against the RCT to evaluate the ability of BB to improve estimates of comparative effectiveness. In this study, we discuss the practical considerations for implementing ECAs and Bayesian borrowing analyses based on historical trials and RWD sources.

External control arm analysis

Methods

Our study design required data from an RCT with both the active treatment and control arm available to provide a benchmark for the estimated treatment effect (we denote the trial treatment arm as ‘TTA’ and the trial control arm as ‘TCA’), an RWD source for construction of the ECA (ConcertAI Patient360™) and a historical control group (denoted as ‘HC’). See section ‘Augmentation via Bayesian Borrowing’ for further detail on the construction and application of the HC. With a focus on metastatic, treatment-naïve NSCLC patients, the two clinical trials were selected based on sample size, availability of overall survival as an end point, and use of the same treatments in the control arm. The ECA analysis was conducted using the R programming language [34].

Reference RCT

For the TTA and TCA, patients were selected from NCT00946712, a multi-center, randomized, open-label, phase III study of cetuximab plus carboplatin and paclitaxel with or without bevacizumab versus carboplatin and paclitaxel with or without bevacizumab in patients with stage IV NSCLC who had not received any prior chemotherapy treatment [35]. The trial was conducted across multiple sites (US and Mexico), initiated in August 2009, and reached primary completion in August 2017. For this study, we restricted the sample to the subset of patients who were bevacizumab ineligible. No restrictions were placed on epidermal growth factor receptor (EGFR) fluorescence *in situ* hybridization (FISH) status. The individual patient data (IPD) and accompanying documentation used in our study was acquired via Project Data Sphere (PDS) (<http://www.projectdatasphere.com>), an open-source repository of individual-level patient data from phase IIB/III oncology trials. Additionally, the research project was approved by the National Cancer Institute for acquisition of the NCT00946712 dataset.

Overview of ECA construction

The ECA (paclitaxel + carboplatin) was constructed using patient-level RWD from the ConcertAI Patient360™ database. The ConcertAI Patient360™ product is a large, representative de-identified oncology database of human-curated comprehensive RWD, sourced from US academic and community electronic health records (EHR) with full data provenance. This database was selected as a representative RWD source providing coverage of the NSCLC population across the US with pre-abstracted clinical and patient characteristics necessary to meet the requirements of the analysis. Deidentified data were acquired for patients between January 2001 and April 2022, with a data cut of 30 June 2022 (Q4’22 release). Additionally, updated overall survival (OS) data was provided by ConcertAI Patient360™, in the Q1’23 release.

Target trial emulation principles were employed to attempt to replicate the efficacy results of the NCT00946712 RCT and minimize identifiable sources of bias when performing the comparison. Target trial emulation provides a framework to prespecify a protocol for a hypothetical RCT to emulate as closely as possible using nonrandomized data by considering patient eligibility, outcome definitions and time periods [36]. The analysis was prespecified in a statistical analysis plan, unless otherwise noted, and broadly included the following steps to construct the ECA:

- The NCT00946712 trial treatment and control arms were restricted to the subgroup of patients who did not receive bevacizumab in conjunction with chemotherapy with or without cetuximab, which we refer to as the TTA and TCA groups, respectively.
- The inclusion/exclusion criteria for the bevacizumab-inappropriate subset of the NCT00946712 trial were applied to the RWD cohort to mirror the criteria as closely as possible (Supplementary Tables 1–3).
- Matching methods were applied to construct a matched subset of patients with a similar distribution of baseline characteristics to the trial (the ‘ECA’). This step was blinded to patient outcomes by removing outcomes from the data table prior to analysis.

OS was compared between the TTA and ECA to estimate the relative effectiveness between the two treatments.

Eligibility criteria

The RWD cohort was constructed by selecting metastatic NSCLC patients from ConcertAI Patient360™ who initiated carboplatin and paclitaxel between January 2001 and April 2022 and applying the eligibility criteria of the NCT00946712 for the bevacizumab-ineligible subgroup. An initial feasibility assessment was performed on the RWD dataset to assess the ability to implement key eligibility criteria, data coverage (i.e., missingness), and ensure sufficient patient counts. Full eligibility criteria were applied to the extent feasible based on structured and pre-abstracted variables available at the time of analysis (Supplementary Tables 4 & 5). No restriction on time period was applied to maximize the RWD cohort available for consideration.

Matching

Matching covariates of interest were considered based on demographics and characteristics of the RCT publication [35], and potential prognostic factors identified in the literature in metastatic NSCLC, such as cancer stage, performance status, histology, biomarkers, previous therapy, health related quality of life and smoking history [37,38]. Based on availability in both data sources, the following baseline covariates were considered for matching between the TTA and RWD cohort: age, sex (male, female), race (Asian, Black, White, Native American and other), smoking status (former, current or never), Eastern Cooperative Oncology Group (ECOG) (0, 1), histology (squamous, nonsquamous) and disease stage (M1A, M1B).

After application of the eligibility criteria, patient baseline characteristics were balanced between TTA and RWD cohort using matching methods. The analysis plan originally proposed the use of PSM to construct an ECA using an average treatment effect on the treated (ATT) estimand [39]. However, due to the small sample size of the RWD cohort and baseline differences in patient characteristics, the approach was amended to use cardinality matching (CM) to better preserve sample size [40,41]. The CM approach yields estimates of the treatment effect in the largest matched subsets of the TTA and ECA that satisfy the specified balance criteria and may differ from the treatment effect in the overall trial population depending on the degree of overlap in patient characteristics between the TTA and post-inclusion/exclusion RWD cohorts.

A two-step CM approach was applied to obtain well-balanced subgroups in the post-inclusion/exclusion RWD (i.e., the ECA), TTA and TCA. Covariates were considered in balance between treatment groups if the absolute standardized mean difference (ASMD) was within 0.1. ASMDs are a measure of imbalance in patient characteristics which are insensitive to sample size [42]. The TTA and ECA were matched first by the CM algorithm to meet the balance constraint so that all the ASMDs of the baseline characteristics in the matched sample fell within 0.1. The resulting matched TTA subgroup was then used in the second CM step at the same ASMD tolerance level to pair up with subjects in the TCA to construct the benchmark RCT effectiveness estimate for the CM subpopulation. Matching the TCA to the TTA subgroup is required in case the treatment effect differs in the selected subpopulation of TTA/TCA patients.

Effectiveness estimates from the ECA analysis were compared against the determined gold standard RCT effectiveness estimates obtained by comparing the treatment and control arms in the CM subpopulation of the

TTA/TCA. Kaplan–Meier (KM) survival curves were generated, and Cox proportional hazard (PH) models were used to estimate hazard ratios (HR) with 95% confidence intervals (CIs) for the matched populations. Robust standard errors were used for the CM subpopulations.

Evaluation of performance

We assessed whether the ECA analysis was able to replicate the intention-to-treat (ITT) effect estimates from the TTA versus TCA in the matched subpopulation (the ‘RCT estimates’). HR estimates with 95% CI were produced for the unadjusted comparison of TTA versus post-inclusion/exclusion RWD (prior to matching), the comparison of the post-matching TTA versus ECA and TTA versus TCA estimates in the matched subpopulation. KM curves were also reported to provide a visual comparison.

Augmentation via Bayesian borrowing

As a next step, we implemented BB to augment the ECA using information borrowed from another trial with a similar ‘historical’ control arm – the HC.

Historical control data

NCT00540514 is a multi-center, randomized, open-label, phase III study of albumin-bound paclitaxel plus carboplatin versus paclitaxel plus carboplatin in patients with treatment-naive stage IIIB or IV NSCLC [43]. The international trial initiated in November 2007 and completed follow-up in February 2013. The paclitaxel plus carboplatin control arm was used as the HC for BB. OS data was reconstructed from the primary publication [43] by digitizing the published KM survival curves using an established algorithm [44]. This allowed for the construction of pseudo-IPD in the absence of available IPD (IPD was initially obtained from PDS but was found to be missing key variables needed to implement the analysis, necessitating an amendment to the protocol and SAP to instead use pseudo-IPD derived from published information). The NCT00540514 trial was selected as a candidate HC alongside the NCT00946712 when screening for available open-access clinical trial datasets.

Methods

BB methods can augment an existing trial arm or ECA using an external data source. Power priors provide a convenient approach to borrowing of external information in which the external data can be down-weighted to account for differences between the external arm and the concurrent arm [19]. Discounting is done using a weighting parameter which can be fixed *a priori* or can be assigned a prior distribution and updated dynamically, allowing for the external data to be more heavily down-weighted when outcomes are inconsistent between data sources. We opted to use a fixed power prior where the discount parameter was varied to assess the stability of estimates and to identify whether a tipping point exists. If very little weight on the external data is needed to reach a decision threshold (e.g., reject the null hypothesis of no difference in efficacy) then our conclusions are less sensitive to the contributions and suitability of the external data. Tipping point analysis in conjunction with BB has some precedent [20,29]. An overview of the BB application is presented in [Supplementary Figure 3 and Appendix 1](#). The BB analysis was conducted using the R statistical programming language [34] and Stan probabilistic programming language [45].

Borrowing into the ECA

Data from the HC was used as the borrowing source to augment the ECA described above. No sample restrictions were placed on the HC prior to borrowing. Consequently, appropriate selection of the external data source is crucial as notable differences in inclusion/exclusion criteria and patient baseline characteristics could confound the comparison and introduce bias into the treatment effect estimates, which would only partially be mitigated through down-weighting of the external data.

BB via a power prior was implemented using the following steps:

1. A joint likelihood was specified for the matched ECA and TTA under a Weibull PH parameterization (i.e., both arms share a common Weibull shape parameter but differ in their scale parameters, yielding a constant hazard ratio).
2. A power prior for the Weibull shape and scale parameters was specified by forming a likelihood for the HC data under a Weibull parameterization with common scale parameter to the ECA and common shape parameter to

both the ECA and matched TTA. This likelihood was then raised to the power of a discount/weight parameter between 0 and 1 to down-weight the external data, where a weight of 0 yields no borrowing of information from the HC and a weight of 1 yields complete pooling.

We amended our analysis to include additional borrowing scenarios to mimic a ‘ideal ECA’ scenario as well as two small sample size ECA scenarios. The scenarios tested included:

- Scenario 1: Borrow into the previously constructed ECA (see ‘External Control Arm Analysis’ section)
- Scenario 2: Borrow into the matched TCA (mimics a hypothetical ‘ideal ECA’ scenario)
- Scenario 3a: Borrow into an $n = 60$ subset of the ECA
- Scenario 3b: Borrow into an $n = 60$ subset of the matched TCA

For scenarios 3a and 3b, we opted to draw random samples of $n = 60$ without replacement based on our experience with prior ECA analyses. These sample sizes are roughly one-third of the size of the ECA we constructed and are intended to inform the application of BB as a practical means to augment ECAs with limited sample sizes. These scenarios also allow us to examine how BB via a fixed power prior may allow us to mitigate or assess risk of bias under an unrepresentative ECA scenario and improve precision under an ‘ideal ECA’ scenario (albeit at a potential cost of increased bias if the external data source is not suitable as a stand-in for a concurrent control).

Posterior inference was performed using Markov chain Monte Carlo (MCMC) implemented using Stan. We considered 11 values (increments of 0.1 from 0 to 1) for the power prior weight parameter and, for each value, ran the MCMC algorithm with four chains of 10,000 iterations each with a burn-in length of 2000 MCMC iterations. Convergence was assessed via trace plots and R-hat statistics and the number of independent draws from the posterior for the log-HR was estimated to be at least 10,000 for all scenarios.

HR estimates were summarized using posterior medians and 95% credible intervals (CrI). The fit of the Weibull PH model was assessed visually and was determined to be a sufficiently good approximation to the data for all four scenarios.

Results

Patient characteristics

Prior to matching, 365, 374 and 181 patients fulfilling all inclusion and exclusion criteria were available in the TTA, TCA and post-inclusion/exclusion RWD cohort respectively (Supplementary Table 6). Post-CM, a total of 178 patients were present within each cohort. Patient and disease characteristics for all cohorts of interest are shown in Table 1. Prior to matching, differences were observed between the TTA and RWD cohort, particularly in terms of ECOG score. For the HC, 531 patients were available and used in the analyses. In contrast to the matched TTA and TCA, and ECA cohort, the HC included more patients with an ECOG score of 2 (0.4%) and stage IIIB disease (21%). Within the randomly sampled TCA subset, patients’ characteristics were similar to the TCA, with the exception of histology, which had a higher proportion of nonsquamous patients. Similarly, the randomly sampled ECA subset demonstrated similar patient characteristics except for ECOG score and smoking history, which demonstrated a higher proportion of patients with an ECOG score of 1, and a slightly larger proportion of patients who never smoked, in comparison to the original ECA.

ECA results

Following CM, the ASMD showed substantial reduction in the magnitude of differences between step 1 and step 2 for most baseline characteristics (Supplementary Figure 2). All variables considered fell within 0.1 ASMD.

The primary result of the ECA analysis was to compare the OS between the matched TTA and the ECA group. The observed median survival time in the TTA was shorter both before (9.17 months; 95% CI: 8.18–10.94 months), and after CM adjustment (10.91 months; 95% CI: 8.38–13.27 months) than in the ECA (before-matching median: 13.68 months (95% CI: 10.65–18.12); after-matching median: 13.97 months (95% CI: 10.75–18.05)). Although, the CM adjustment did narrow the gap in median OS between the two arms.

The Cox PH regression model was used to estimate and assess the treatment effect size based on the primary outcome (OS). The model results before and after CM adjustment, were both in favor of the ECA. Before matching, the naive HR for TTA versus ECA was 1.59 (95% CI: 1.30–1.95; $p < 0.001$), favoring the chemotherapy control treatment (Figure 1A). Similarly, the CM-adjusted HR for the TTA compared with ECA was 1.53 (95% CI:

Table 1. Baseline characteristics for relevant cohorts: matched trial treatment arm, matched trial control arm, external control arm, historical control, matched trial control arm subset and matched external control arm subset.

n (%)	Matched TTA (n = 178)	Matched TCA (n = 178)	ECA (n = 178)	HC [†] (n = 531)	Matched TCA subset (n = 60)	Matched ECA subset (n = 60)
Median age (range)	66.5 (29.5–83)	64.6 (34.6–82.6)	66 (24–86)	60.8 (24.9–84.7)	65 (34.6–80)	63.5 (40–84)
Sex						
Male	112 (62.9)	104 (58.4)	107 (60.1)	397 (74.8)	34 (56.7)	37 (61.7)
Female	66 (37.1)	74 (41.6)	71 (39.9)	134 (25.2)	26 (43.3)	23 (38.3)
ECOG						
0	59 (33.1)	54 (30.3)	51 (28.7)	113 (21.3)	20 (33.3)	25 (41.7)
1	119 (66.9)	124 (69.7)	127 (71.3)	416 (78.3)	40 (66.7)	35 (58.3)
2	0	0	0	2 (0.4)	0	0
Stage						
IIIB	0	0	0	110 (20.7)	0	0
IV	178 (100)	178 (100)	178 (100)	421 (79.3)	60 (100)	60 (100)
Histology						
Squamous	97 (54.5)	93 (52.2)	89 (50.0)	221 (41.6)	26 (43.3)	27 (45.0)
Nonsquamous	81 (45.5)	85 (47.8)	89 (50.0)	310 (58.4)	34 (56.7)	33 (55.0)
Smoking history						
Former	82 (46.1)	80 (44.9)	84 (47.2)	148 (27.9)	25 (41.7)	28 (46.7)
Current	70 (39.3)	78 (43.8)	64 (36.0)	234 (44.1)	27 (45.0)	20 (33.3)
Never	26 (14.6)	20 (11.2)	30 (16.9)	144 (27.1)	8 (13.3)	12 (20.0)

Summaries presented for matched TTA, matched TCA and ECA are based on the two-step CM approach.
[†] 5 patients were missing smoking history in the HC.
 ECA: External control arm; HC: Historical control; TCA: Trial control arm; TTA: Trial treatment arm.

1.21–1.93; $p < 0.001$), which again suggested that the investigational treatment yielded poorer survival outcomes (Figure 1B).

The Cox HR estimates for the TTA versus ECA are consistently higher than the gold standard RCT HR estimate of 0.91 (95% CI: 0.73–1.13; $p = 0.39$) (Supplementary Figure 1B) in the CM subpopulation, indicating that the ECA analysis was unsuccessful at replicating the results of the RCT. It should be noted that the HR estimates did not substantially change before versus after applying CM. For comparison, in the original trial publication, the HR for the non-bevacizumab population was reported as 0.90 (95% CI: 0.78–1.05; $p = 0.19$) [35].

Sensitivity analyses were conducted to assess the robustness of the ECA analysis results. These analyses included limiting the time period for eligibility in the RWD cohort, exploring the number and type of subsequent lines of therapy received by patients in the RWD cohort, an assessment of the potential impacts of EGFR FISH status, and nonrandom missingness in ECOG score. These analyses did not provide evidence of a specific key driver explaining the discrepancy between ECA analysis and RCT results.

Bayesian borrowing results

Results showed that the ECA had the longest time to event, particularly beyond 12 months, followed by the HC, the matched TTA and the matched TCA (Figure 2). OS in the HC was only slightly longer than the matched TTA and TCA, which were comparable.

Forest plots summarizing the results of the four BB scenarios show that as borrowing weights increase from 0 to 1, the HR decreases from 1.669 (95% CrI: 1.328–2.103) to 1.295 (95% CrI: 1.082–1.537) (Figure 3A). When borrowing more heavily from the HC, the point estimate and CrI are pulled closer to the RCT estimate. The shift in estimates with increased borrowing illustrates the sensitivity of the HR estimates to the relative weight put on both the ECA and HC and highlights the discordance between external data sources.

As the amount of borrowing from the HC into the ‘ideal ECA’ (matched TCA) increases, the HR estimates increase from 0.908 (95% CrI: 0.728–1.130) to 1.027 (95% CrI: 0.858–1.223) and the width of the 95% CrIs decreases slightly from 0.402 to 0.365 (Figure 3B). Due to the similarity of outcomes between the matched TCA

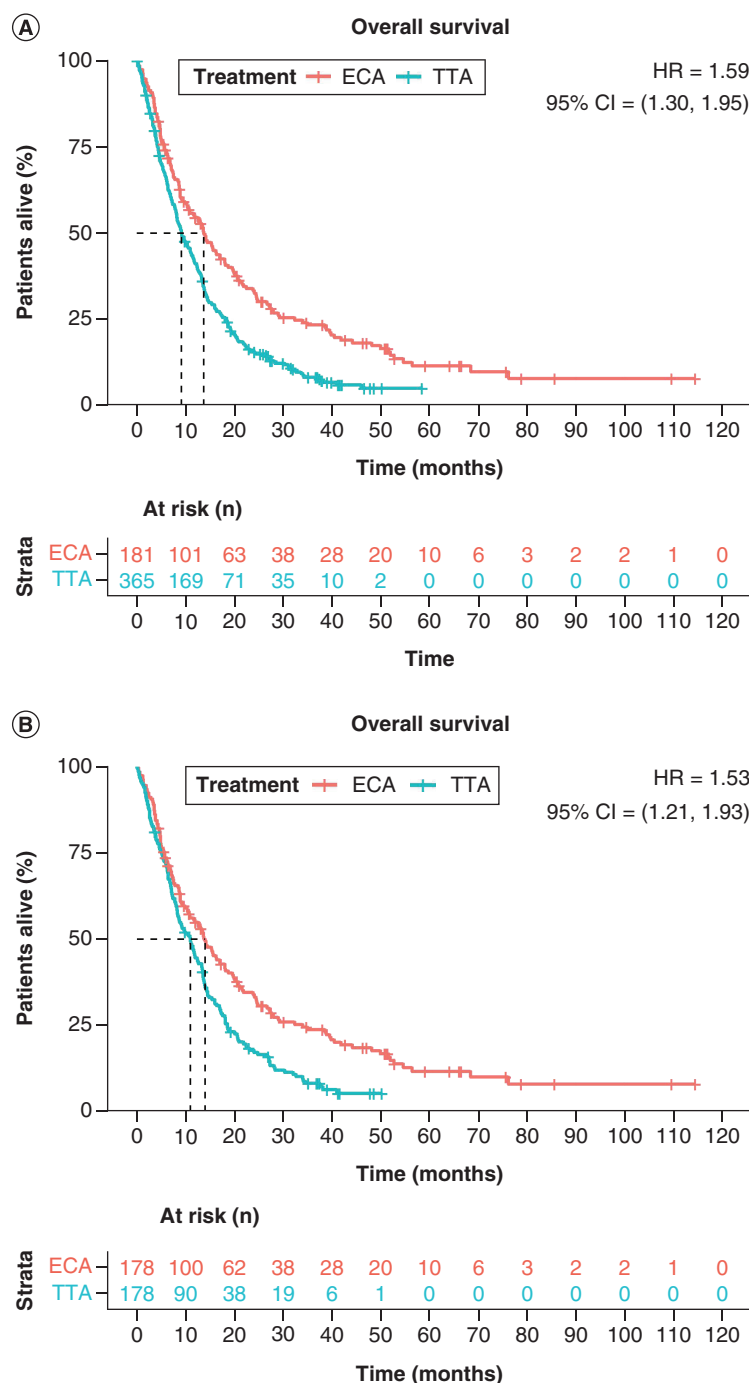


Figure 1. Kaplan–Meier curves of overall survival of trial treatment arm and external control arm. (A) Pre-matching and (B) Cardinality matching adjusted. CI: Confidence interval; ECA: External control arm; HR: Hazard ratio; TTA: Trial treatment arm.

and HC, the impact on the HR estimate is modest, however, with a sizeable 178 patients in the TCA, we do not see much impact of borrowing on the precision of the estimates.

When borrowing from the HC into a reduced $n = 60$ subset of the ECA, the HR decreases more noticeably – from 1.730 (95% CrI: 1.267–2.398) to 1.188 (95% CrI: 0.988–1.416) – as the amount of borrowing is increased (Figure 3C). Due to the very limited sample size in the ECA subset, borrowing from the HC aggressively pulls the HR estimates to a value closer to 1 and the width of the 95% CrI shrinks substantially from 1.131 to 0.428.

A similar $n = 60$ subset of the ‘ideal ECA’ (matched TCA) shows a more modest change in the HR from 0.978 (95% CrI: 0.722–1.350) to 1.066 (95% CrI: 0.888–1.275) as the amount of borrowing is increased (Figure 3D). However, the width of the 95% CrI substantially narrows from 0.628 to 0.387, demonstrating the potential of BB to improve precision when the external control arm is well constructed.

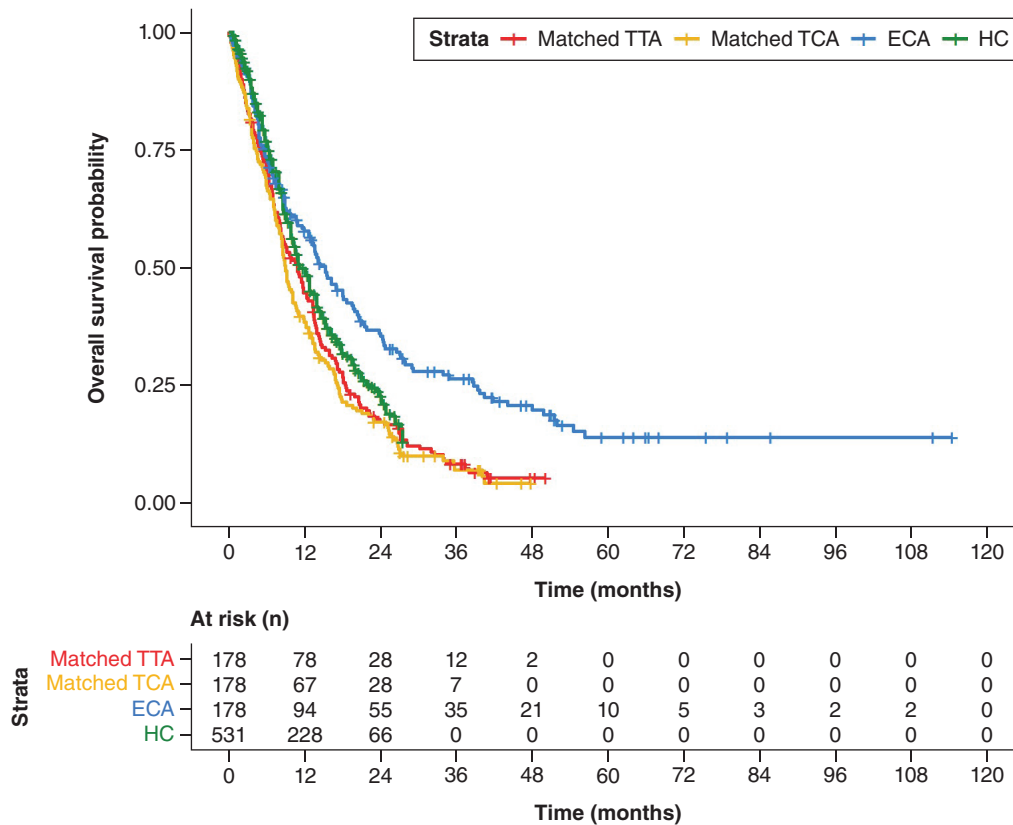


Figure 2. Kaplan–Meier survival curves for the matched trial treatment arm and trial control arm, external control arm and historical control.

ECA: External control arm; HC: Historical control; TCA: Trial control arm; TTA: Trial treatment arm.

Discussion

In this paper, we proposed a combined ECA and BB approach which would incorporate information from HCs to improve estimates from comparison against an ECA alone. We suggest that through the incorporation of HCs, BB could be applied more routinely to both contextualize heterogeneity across external data sources and improve the accuracy of comparisons made against ECAs alone. BB can provide an additional tool for sensitivity analysis to address challenges with the availability and comparability of key patient demographic and clinical characteristics encountered when conducting an ECA. Although we were unable to successfully replicate the treatment effect estimate observed in the RCT using an ECA, the incorporation of BB provided insight into the heterogeneity in outcomes between the ECA and historical trial data. When an additional reliable external data source exists, BB can help gauge the degree of agreement between external data sources – bolstering confidence and improving precision when outcomes are similar and helping to contextualize risk of bias when outcomes differ materially between external data sources (as was the case in this study). BB into both the TCA and the ECA resulted in improved precision, and we recommend the approach when appropriate external/historical data sources are available – especially in instances where sample sizes in the ECA are very small. In this study, we demonstrated with each BB scenario how researchers might employ BB methods to improve precision, improve confidence in results, or act as a sensitivity analysis to assess risk of bias in future applications, including HTA scenarios. These scenarios also illustrate the various strengths and limitations of the method.

Single-arm trials are increasingly being used for HTA assessment in rare oncology indications [3,16], but trying to estimate effectiveness relative to standard of care is challenging, as demonstrated by this study. However, because the patients in the trial and the patients in the ECA are not from the same population and were not assigned to treatment using any known mechanism, it is impossible to know whether differences in outcome between patients in the trial and the ECA are due to the active treatment or other factors. Restricting the patients included in the ECA to those who would be trial-eligible, and applying weighting or matching techniques can help to limit the

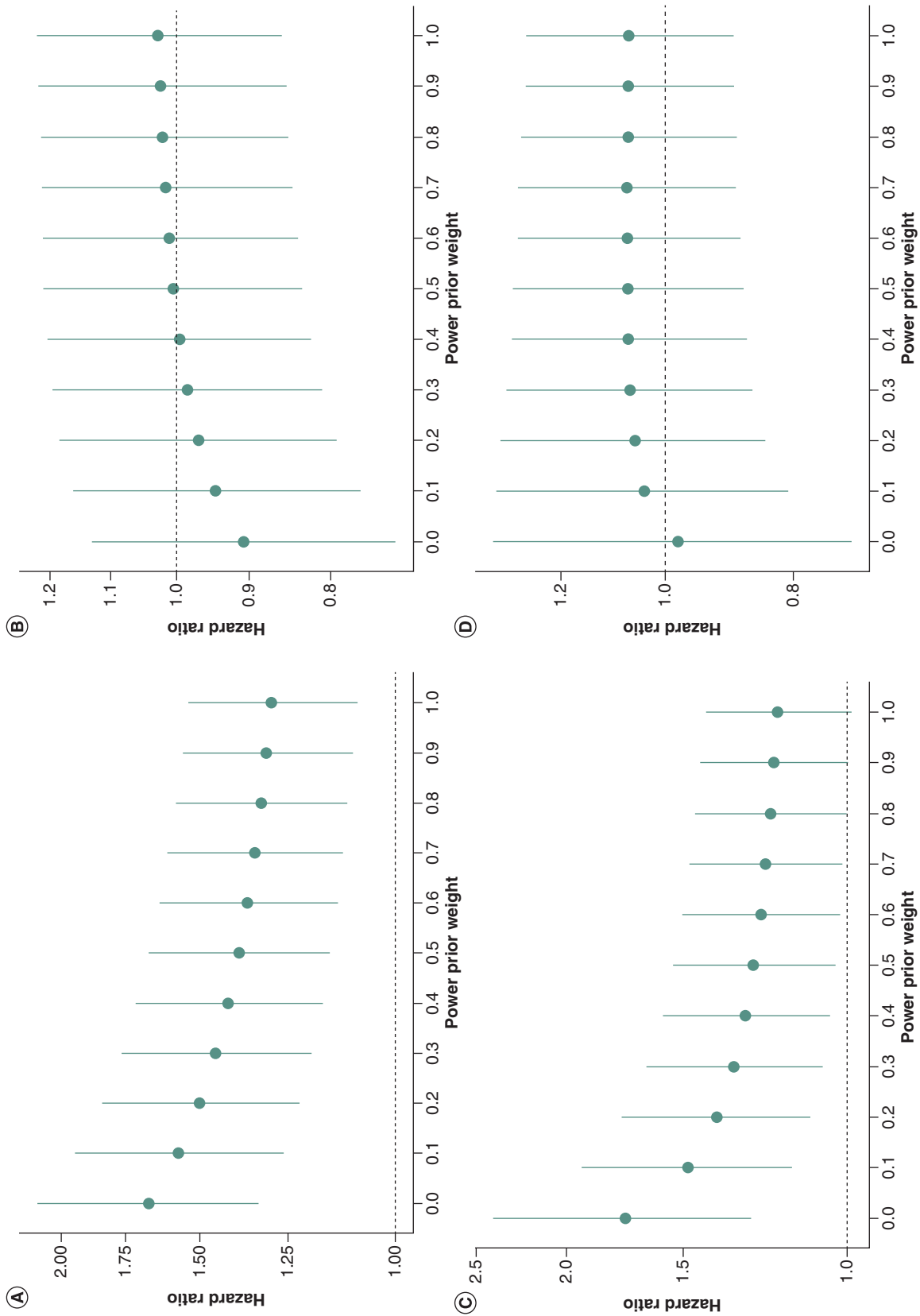


Figure 3. Posterior medians and 95% credible intervals for the hazard ratio for different Bayesian borrowing weights. (A) Borrowing into the (unrepresentative) external control arm (ECA; Scenario 1), **(B)** borrowing into the matched trial control arm (TCA; mimics a hypothetical “ideal ECA” scenario) (Scenario 2), **(C)** borrow into an $n = 60$ subset of the (unrepresentative) ECA (Scenario 3a) and **(D)** borrow into an $n = 60$ subset of the matched TCA (Scenario 3b).

influence of nontreatment factors when making comparisons, but may need to trade off sample size and precision for reduced bias in the estimate [5,46].

Contributing to the understanding and applicability of these approaches remains necessary as regulatory agencies continue to demonstrate an openness for the use of non-RCT data in submissions. In practice, selection of an appropriate external data source is crucial. Recent guidelines and case studies present an openness to the use of external controls and advanced methods, but noted in many situations the likelihood of successfully demonstrating the effectiveness against an external control arm remains low [11,47]. Although the intention of this study was to assess the feasibility and use of BB for ECA analysis, many learnings can still be ascertained regarding data selection and challenges of conducting an ECA to help inform future application of these methods in regulatory and HTA settings. In practice, we cannot be sure that a constructed ECA is a reliable stand-in for a concurrent control arm. Scenarios 1 and 3a illustrate how the use of BB with an HC can provide an independent check on the ECA by gauging the potential impact of heterogeneity in outcomes between external standard of care data sources. In scenarios 2 and 3b where the ECA is an appropriate stand-in for a concurrent control arm, BB can improve precision of estimates by augmenting limited ECA sample sizes and can bolster confidence if external data sources are reasonably homogenous.

BB methods provide an established approach to incorporate external data using a cohesive framework which allows for appropriate down-weighting of its influence [18,21,48]. A particular strength of the current study is the ability to down-weight the external data, to allow for borrowing to be limited when outcomes in the external data are inconsistent with the data that is being augmented (often a concurrent control arm but, in our case, an ECA). In our study we also illustrate how, in the absence of a concurrent control arm, BB with a sliding scale of borrowing weights can be informative for either bolstering our ECA analysis (when outcomes in the ECA and external data source are homogenous) or as a sensitivity analysis when outcomes are inconsistent between the ECA and external data source. We also show how BB can improve precision of estimates by augmenting a small ECA – a frequent challenge for ECAs in rare oncology indications when patient counts in RWD sources may be very limited [49]. The ability to downweight the external data used in BB is also advantageous due to the ability to partially discount the contributions of external data sources that suffer from specific limitations. This can be beneficial when limited data are available or when it is impractical or infeasible to conduct a full ECA with imposed eligibility criteria using matching or other methods – for example, in the present study we were unable to exclude all stage IIIB patients in the HC data.

The applications and challenges of working with RWD for ECA analyses are well understood and have been discussed extensively in the literature [9,49–51] – including many of the limitations encountered in this study. However, previous evaluations of the performance of ECA analyses at replicating RCTs in metastatic NSCLC have shown the method to be viable [14,49,51,52]. Although a notable limitation to the current study, the decision to refrain from the incorporation of a time restriction for the RWD cohort was necessary to ensure enough eligible patients remained for analysis. As such, changes in the standard of care for metastatic NSCLC since the end of the enrollment window of the NCT00946712 and NCT00540514 may have contributed to better OS observed in the RWD-derived ECA. Another potential limitation to the current study was the lack of consideration of laboratory values in the CM approach. In the DUPLICATE study, a large-scale project dedicated to replicating RCTs using RWD, assessing laboratory values was noted as a part of the methodology post-matching [14]. Another limitation regarding matching, was the inability to consider disease characteristics such as biomarkers, comorbidity disease burden or disease stage at diagnosis which are often useful markers of aggressiveness of disease as these were not available within both datasets. Similarly, additional unobserved characteristics of trial participants, such as social determinants of health, may differ from the average real-world patient, further contributing to the inability to replicate the trial findings. Although a general limitation of CM is that it provides treatment effect estimates in a subset of trial patients (and therefore provides treatment effect estimates for a matched subpopulation rather than the overall trial population), the method was able to achieve good covariate balance while preserving the limited sample size in the RWD. The ability of CM to achieve covariate balance without excessive sample size losses in the RWD cohort demonstrates the usefulness of this method for challenging ECA settings where preservation of limited sample sizes needs to be prioritized – albeit at the cost of rendering the target patient population less interpretable. Even in instances of well-established target trial emulation plans, successful emulation is not guaranteed, and performing sensitivity and quantitative bias analyses to assess the reliability of ECA approaches will be relevant to regulatory and HTA personnel when evaluating submissions.

Implementation of BB can be challenging. It can be difficult to determine an appropriate weight to place on the external data or, in the case of dynamic borrowing methods, how to formulate appropriate priors. Additionally, while down-weighting of external data sources can be used to mitigate bias introduced by incorporation of the external data, this does not alleviate the need for careful selection and evaluation of the appropriateness of the external data source, including historical RCT data.

While BB methods provide a potential approach to augmenting an ECA, frequentist alternatives also exist and may also be extended to augmentation of an ECA [53]. Additionally, while this study considered the application of static borrowing via a power prior, other BB methods such as dynamic borrowing via a power prior [19], MAP or robust MAP priors [54,55] would be worth exploring – especially where multiple external data sources are available from which to borrow. Last, it would be worth expanding on the potential for use of BB with tipping point analysis as a means of quantifying bias as a future research direction.

When implementing ECAs in practice, the ground truth is rarely known. The reliability of the ECA analysis hinges on the suitability of the data source used to construct the ECA, the credibility of the methods used to mitigate bias in the comparison, and careful consideration of the limitations and remaining risk of bias. We demonstrate how BB can be used in comparative effectiveness research for decision making to improve the precision of estimates, mitigate challenges of traditional methods and as a bias assessment tool in cases where appropriate external data sources are available.

Summary points

- Randomized controlled trials are the gold standard for comparative effectiveness evidence but are not always practical or feasible.
- External control arms (ECAs) are assembled from historical controls or real-world data (RWD) and can be used to assess comparative effectiveness in absence of a concurrent control arm.
- Regulatory and reimbursement agency critiques regarding ECAs have centered on unmeasured or uncontrolled confounding, missing data, choice of study population, outcome definitions and lack of statistical power.
- Bayesian borrowing (BB) methods incorporate external data sources and can improve precision of ECA outcome estimates when used appropriately.
- This study employs BB from a historical clinical trial to augment an ECA constructed from RWD to improve comparative effectiveness estimates in non-small-cell lung cancer.
- Despite challenges in replicating overall survival estimates through the ECA, BB was effective in identifying heterogeneity between the trials and improving the precision of overall survival estimates.
- Through this study, we demonstrate how BB can be used to mitigate bias and challenges of traditional methods and function as a bias assessment tool which can be valuable for future regulatory and reimbursement agency submissions.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: <https://bpl-prod.literatumonline.com/doi/10.57264/cer-2023-0175>

Author contributions

A Struebing, E Mackay and M Rosenlund were responsible for the study conception, study design and revision of the manuscript; authors C McKibbon and R Velummailum were responsible for evaluation and acquisition of data and drafting and revision of the manuscript; authors H Ruan, E Mackay and A Springford were responsible for the data analysis and drafting and revision of the manuscript; authors N Dennis, P He, Y Tanaka and Y Xiong were responsible for providing feedback on the study design and revision of the manuscript.

Financial disclosure

This study was funded by Daiichi Sankyo Europe. The authors have received no other financial and/or material support for this research or the creation of this work apart from that disclosed.

Competing interests disclosure

A Struebing and M Rosenlund are employed by and reported stock ownership in Daiichi Sankyo Europe. N Dennis is employed by Daiichi Sankyo Oncology France. P He, Y Tanaka and Y Xiong are employed by and reported stockownership in Daiichi Sankyo

Inc. C McKibbin, H Ruan, E Mackay, R Velummailum and A Springford are employees of Cytel which has received funding from Daiichi Sankyo Europe to conduct the study analyses, and consulting fees from various manufacturers unrelated to this research. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Writing disclosure

No writing assistance was utilized in the production of this manuscript.

Data sharing statement

The authors certify that this manuscript reports the secondary analysis of clinical trial data, and the use of this shared data is in accordance with the agreed upon data sharing policies, including review by the trial Sponsor and data provider. This manuscript was prepared using data from Dataset nci-data-470_NCT00946712 from the NCTN/NCORP Data Archive of the National Cancer Institute's (NCI's) National Clinical Trials Network (NCTN) obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere. Data were originally collected from clinical trial NCT no. NCT00946712, S0819: Carboplatin and Paclitaxel with or Without Bevacizumab and/or Cetuximab in Treating Patients with Stage IV or Recurrent non-small-cell Lung Cancer. All analyses and conclusions in this manuscript are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, the NCORP, the NCI or Project Data Sphere. The de-identified RWD supporting this study was acquired from ConcertAI and may be available by request and subject to a license agreement with ConcertAI. The data for the historical clinical trial was acquired from the publicly available publication.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>

References

Papers of special note have been highlighted as: ● of interest; ●● of considerable interest

1. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: study design: randomised controlled trials. *BJOG* 125(13), 1716 (2018).
2. Kempf L, Goldsmith JC, Temple R. Challenges of developing and conducting clinical trials in rare disorders. *Am. J. Med. Genet. A* 176(4), 773–783 (2018).
3. 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.* 9(2), 266–272 (2023).
4. Mulder J, Teerenstra S, Van Hennik PB *et al.* Single-arm trials supporting the approval of anticancer medicinal products in the European Union: contextualization of trial results and observed clinical benefit. *ESMO Open* 8(2), 101209 (2023).
5. Mack C, Christian J, Brinkley E *et al.* When context is hard to come by: external comparators and how to use them. *Therap. Innovat. Regul. Sci.* 54, 2168479019878672 (2019).
6. Gagne JJ, Thompson L, O'keefe K, Kesselheim AS. Innovative research methods for studying treatments for rare diseases: methodological review. *BMJ* 349, g6802 (2014).
7. Ghadessi M, Tang R, Zhou J *et al.* A roadmap to using historical controls in clinical trials – by drug information association adaptive design scientific working group (dia-adswg). *Orphanet J. Rare Dis.* 15(1), 69 (2020).
8. Thorlund K, Dron L, Park JJH, Mills EJ. Synthetic and external controls in clinical trials - a primer for researchers. *Clin. Epidemiol.* 12, 457–467 (2020).
- **This primer highlights the current landscape of the use of SCAs and provides an appraisal framework to assess future publications regarding the use of SCAs.**
9. Velummailum RR, McKibbin C, Brenner DR *et al.* Data challenges for externally controlled trials: viewpoint. *J. Med. Internet Res.* 25, e43484 (2023).
10. Gray CM, Grimson F, Layton D, Pocock S, Kim J. A framework for methodological choice and evidence assessment for studies using external comparators from real-world data. *Drug Saf.* 43(7), 623–633 (2020).
- **Provides an overview and evaluation framework for the appropriate selection of external data and design for use in regulatory settings.**
11. Food and Drug Administration. Considerations for the design and conduct of externally controlled trials for drug and biological products (2023). <https://www.fda.gov/media/164960/download>

12. Food and Drug Administration. Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drug and biological products, draft guidance for industry (2021). <https://www.fda.gov/regulatory-information/search-h-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making-drug>
13. National Institute for Health and Care Excellence. NICE real-world evidence framework (2022). <https://www.nice.org.uk/corporate/ecd9/chapter/overview>
14. Franklin JM, Patorno E, Desai RJ *et al.* Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT duplicate initiative. *Circulation* 143(10), 1002–1013 (2021).
15. Ramagopalan SV, Hernán MA, Pinilla P, Thorlund K, Kent S. in *ISPOR Europe 2022*. (2022).
16. Jaksa A, Louder A, Maksymiuk C *et al.* A comparison of seven oncology external control arm case studies: critiques from regulatory and health technology assessment agencies. *Value Health* 25, 1967–1976 (2022).
- **This research article highlights the common critiques from regulatory and health technology assessment agencies from oncology submissions which have incorporated the use of external control arms.**
17. Seeger JD, Davis KJ, Iannacone MR *et al.* Methods for external control groups for single arm trials or long-term uncontrolled extensions to randomized clinical trials. *Pharmacoepidemiol. Drug Saf.* 29(11), 1382–1392 (2020).
18. Viele K, Berry S, Neuenschwander B *et al.* Use of historical control data for assessing treatment effects in clinical trials. *Pharm. Stat.* 13(1), 41–54 (2014).
19. Ming-Hui C, Joseph GI. Power prior distributions for regression models. *Statist. Sci.* 15(1), 46–60 (2000).
20. Best N, Price RG, Pouliquen IJ, Keene ON. Assessing efficacy in important subgroups in confirmatory trials: an example using Bayesian dynamic borrowing. *Pharm. Stat.* 20(3), 551–562 (2021).
21. Dron L, Golchi S, Hsu G, Thorlund K. Minimizing control group allocation in randomized trials using dynamic borrowing of external control data – an application to second line therapy for non-small-cell lung cancer. *Contemp. Clin. Trials Commun.* 16, 100446 (2019).
22. Richeldi L, Azuma A, Cottin V *et al.* Trial of a preferential phosphodiesterase 4b inhibitor for idiopathic pulmonary fibrosis. *N. Engl. J. Med.* 386(23), 2178–2187 (2022).
23. Food and Drug Administration. Guidance for the use of Bayesian statistics in medical device clinical trials (2010). <https://www.fda.gov/media/71512/download>
- **This guidance document highlights the early precedent for the use of advanced methods such as Bayesian borrowing in medical device trials.**
24. Lim J, Walley R, Yuan J *et al.* Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: review of methods and opportunities. *Therap. Innov. Regul. Sci.* 52(5), 546–559 (2018).
25. European Medicines Agencies. Concept paper on extrapolation of efficacy and safety in medicine development (2013). https://www.ema.europa.eu/en/documents/scientific-guideline/concept-paper-extrapolation-efficacy-and-safety-medicine-development_en.pdf
26. European Medicines Agencies. Guideline on clinical trials in small populations (2006). https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-trials-small-populations_en.pdf
27. Brunner HI, Abud-Mendoza C, Viola DO *et al.* Safety and efficacy of intravenous belimumab in children with systemic lupus erythematosus: results from a randomised, placebo-controlled trial. *Ann. Rheum. Dis.* 79(10), 1340–1348 (2020).
28. Food and Drug Administration. Bla 125370/s-064 and bla 761043/s-007 multi-disciplinary review and evaluation Benlysta® (belimumab) for intravenous infusion in children 5 to 17 years of age with sle (2021). <https://www.fda.gov/media/127912/download>
- **This FDA acceptance demonstrates the acceptability and use of Bayesian borrowing in a regulatory setting.**
29. Food and Drug Administration. Pediatric postmarketing pharmacovigilance review - Belimumab (2022). <https://www.fda.gov/media/161020/download>
30. Food and Drug Administration. Adaptive designs for clinical trials of drugs and biologics (2018). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry>
31. Food and Drug Administration. Interacting with the FDA on complex innovative trial designs for drugs and biological products (2020). <https://www.fda.gov/media/130897/download>
32. Singh N, Temin S, Baker S *et al.* Therapy for stage IV non–small-cell lung cancer with driver alterations: ASCO living guideline. *J. Clin. Oncol.* 40(28), 3310–3322 (2022).
33. Daly ME, Singh N, Ismaila N *et al.* Management of stage III non–small-cell lung cancer: ASCO guideline. *J. Clin. Oncol.* 40(12), 1356–1384 (2021).
34. R Core Team. R Foundation for Statistical Computing (2022). <https://www.R-project.org>
35. Herbst RS, Redman MW, Kim ES *et al.* Cetuximab plus carboplatin and paclitaxel with or without bevacizumab versus carboplatin and paclitaxel with or without bevacizumab in advanced nscl (SWOG S0819): a randomised, Phase III study. *Lancet Oncol.* 19(1), 101–114 (2018).

36. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* 183(8), 758–764 (2016).
- **Highlights the best practices for emulating a target trial using real-world data to perform SCA analyses.**
37. Cuyún Carter G, Barrett AM, Kaye JA *et al.* A comprehensive review of nongenetic prognostic and predictive factors influencing the heterogeneity of outcomes in advanced non-small-cell lung cancer. *Cancer Manag. Res.* 6, 437–449 (2014).
38. Berghmans T, Paesmans M, Sculier JP. Prognostic factors in stage III non-small-cell lung cancer: a review of conventional, metabolic and new biological variables. *Ther. Adv. Med. Oncol.* 3(3), 127–138 (2011).
39. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat. Methods Med. Res.* 21(3), 273–293 (2012).
40. Zubizarreta JR, Paredes RD, Rosenbaum PR. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Annals Appl. Stat.* 8(1), 204–231 (2014).
- **Details the cardinality-matching method used in the current study to construct the SCA from real-world data.**
41. Fortin SP, Johnston SS, Schuemie MJ. Applied comparison of large-scale propensity score matching and cardinality matching for causal inference in observational research. *BMC Med. Res. Methodol.* 21(1), 109 (2021).
42. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* 28(25), 3083–3107 (2009).
43. Socinski MA, Bondarenko I, Karaseva NA *et al.* Weekly nab-paclitaxel in combination with carboplatin versus solvent-based paclitaxel plus carboplatin as first-line therapy in patients with advanced non-small-cell lung cancer: final results of a Phase III trial. *J. Clin. Oncol.* 30(17), 2055–2062 (2012).
44. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan–Meier survival curves. *BMC Med. Res. Methodol.* 12, 9 (2012).
45. Carpenter B, Gelman A, Hoffman MD *et al.* Stan: a probabilistic programming language. *J. Statist. Softw.* 76(1), 1–32 (2017).
46. Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J. Clin. Epidemiol.* 66(Suppl. 8), S84–S90.e81 (2013).
47. Food and Drug Administration. Cid case study: external control in diffuse b-cell lymphoma. <https://www.fda.gov/media/155405/download>
48. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Stat. Med.* 34(28), 3724–3749 (2015).
49. Popat S, Liu SV, Scheuer N *et al.* Addressing challenges with real-world synthetic control arms to demonstrate the comparative effectiveness of Pralsetinib in non-small-cell lung cancer. *Nat. Commun.* 13(1), 3500 (2022).
50. Hsu GG, Mackay E, Scheuer N, Ramagopalan SV. Keeping it real: implications of real-world treatment outcomes for first-line immunotherapy in metastatic non-small-cell lung cancer. *Immunotherapy* 13(18), 1453–1456 (2021).
51. Sengupta S, Ntambwe I, Tan K *et al.* Emulating randomized controlled trials with hybrid control arms in oncology: a case study. *Clin. Pharmacol. Ther.* 113(4), 867–877 (2023).
- **Demonstrates the application of Bayesian borrowing methods in a hybrid control arm setting.**
52. Ali MS, Prieto-Alhambra D, Lopes LC *et al.* Propensity score methods in health technology assessment: principles, extended applications, and recent advances. *Front. Pharmacol.* 10, 973 (2019).
53. Majumdar A, Davi R, Bexon M *et al.* Building an external control arm for development of a new molecular entity: an application in a recurrent glioblastoma trial for mdna55. *Stat. Biosci.* 14(2), 285–303 (2022).
54. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. *Clin. Trials* 7(1), 5–18 (2010).
55. Schmidli H, Gsteiger S, Roychoudhury S *et al.* Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 70(4), 1023–1032 (2014).