BMC Medical Research Methodology

RESEARCH

Open Access

Incorporating external controls in the design of randomized clinical trials: a case study in solid tumors

Emily M. Damone[1], Jiawen Zhu[2], Herbert Pang[2], Xiao Li[3], Yinqi Zhao[2], Evan Kwiatkowski[4], Lisa A. Carey[5] and Joseph G. Ibrahim[1*]

Abstract

Background The use of historical external control data in clinical trials has grown in interest and needs when considering the design of future trials. Hybrid control designs can be more efficient to achieve the same power with fewer patients and limited resources. The literature is sparse on appropriate statistical methods which can account for the differences between historical external controls and the control patients in a study. In this article, we illustrate the analysis framework of a clinical trial if a hybrid control design was used after determining an RCT may not be feasible.

Methods We utilize two previously completed RCTs in nonsquamous NSCLC and a nationwide electronic health record derived de-identified database as examples and compare 5 analysis methods on each trial, as well as a set of simulations to determine operating characteristics of such designs.

Results In single trial estimation, the Case Weighted Adaptive Power Prior provided estimated treatment hazard ratios consistent with the original trial's conclusions with narrower confidence intervals. The simulation studies showed that the Case Weighted Adaptive Power Prior achieved the highest power (and well controlled type-1 error) across all 5 methods with consistent study sample size.

Conclusions By following the proposed hybrid control framework, one can design a hybrid control trial transparently and accounting for differences between control groups while controlling type-1 error and still achieving efficiency gains from the additional contribution from external controls.

Keywords Hybrid clinical trials, Historical controls, Real world data, Clinical trials

*Correspondence:
Joseph G. Ibrahim
ibrahim@bios.unc.edu
[1]Department of Biostatistics, University of North Carolina, 3109 McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599, USA
[2]Department of Data Science, Product Development, Genentech, Inc, South San Francisco, CA, USA
[3]Department of Personalized Healthcare, Product Development, Genentech, Inc, South San Francisco, CA, USA
[4]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA
[5]Division of Oncology, University of North Carolina, Chapel Hill, NC, USA

Introduction

In February 2023, the US Food and Drug Administration (FDA) released a draft guidance document [1] that covers recommendations to consider when designing externally controlled trials. Its release is timely as the demand for the incorporation of external controls into studies at the design stage grows. Even though randomized controlled trials (RCT) have been the gold standard for evaluating the efficacy and safety of medical products and for regulatory decision-making, there are many situations where a RCT may not be feasible, such as in rare disease or cancer therapy development, or excessively large and expensive, such as noninferiority trials, which are increasingly important in this era of precision medicine. This makes the smaller, more nimble trial designs incorporating external controls more appealing.

An attractive recently developed clinical trial design that utilizes external controls to augment the internal controls in a randomized trial is the "hybrid control" design, which includes both patient-level external control data as well as randomized internal controls [2]. Hybrid control designs can be more efficient, as a study can achieve the desired power with fewer patients as well as potentially enrolling more patients to the treated group. However, the incorporation of external controls can also lead to biased results and inflation of type I error due to dissimilarity between the trial population and the external control population if not used appropriately [3]. Moreover, the methods and tools for these novel designs are also context dependent. But the promise of this kind of trial design approach means that oncologists, statisticians, and trialists need to understand the framework for hybrid designs in time-to-event settings which is particularly important in cancer research.

The key element that must be considered when designing and analyzing hybrid control trials is heterogeneity between internal and external controls. Multiple Bayesian methods have been developed to adjust for heterogeneity when dynamically borrowing information from external controls (EC). For example, a power prior method [4] was proposed for downweighting the contribution of EC by assigning a pre-fixed weight between 0 and 1 in the model for the historical data. Later, a modified power prior [5, 6] was developed to estimate the weight from the data choosing a prior distribution for the weight parameter. Further, to directly parametrize the commensurability of the external and internal control, Hobbs et al. [7, 8] developed the commensurate prior method to characterize the similarity between these two sources of controls via the estimation of the precision (inverse of the variance) of the parameter of interest, hence adaptively determining the borrowing strength. When multiple real world data (RWD) cohorts are available, the meta-analytic predictive (MAP) prior [9] and its variant, the robust MAP prior [10] can be leveraged.

Recently, Kwiatkowski et al. [11] developed a case weighted adaptive power prior method for hybrid analyses that assigns individual discounting weights for each external control where the degree of borrowing is determined based on similarity between the RCT and external control patients to account for systematic differences. For RWD, it is very likely that certain individuals are different from the RCT population. The case weighted adaptive power prior method will be able to handle such data and only downweight individuals found to be different from the RCT population.

The proposed Bayesian design for hybrid control studies requires a more complicated modeling set up than the traditional design for a RCT in order to account for potential differences between the internal and external controls. In this paper, we emulate situations where one may design hybrid control studies instead of the traditional RCT designs and provide a framework for such designs and analyses. We include the steps to pre-specify the parameters of the case weighted power prior, and evaluate hybrid trial designs assuming fewer patients are assigned to the control arm. Finally, we discuss the operating characteristics in comparison with propensity score matching.

Methods

Our analyses and simulations consider trials with time-to-event outcomes, specifically overall survival (OS), with an analysis goal of testing efficacy of treatment compounds with one-sided null hypotheses for the hazard ratio (HR) of treatment benefit compared to control. We use OS as the primary endpoint since it is a cleaner, better-established measure in external data compared to progression free survival. The case weighted adaptive power prior method can easily be adapted to other endpoints and alternative statistical hypotheses.

We used two previously published randomized controlled trials in non-squamous non-small cell lung cancer (NSCLC) to compare conventional trial designs with potential hybrid control trial methodologies. We utilize all patient-level data available from each trial and reanalyze each trial with the addition of external control data. We conduct a comparative analysis of 5 methods: Cox Proportional Hazards, Pooled Cox Proportional Hazards, Propensity Score Matching, Fixed-Weight Power Prior, and Case Weighted Adaptive Power Prior. Thus, we consider three frequentist methods and two Bayesian methods, respectively. Adjusted treatment hazard ratios and their corresponding 95% confidence or credible intervals are given for decision-making.

Clinical Data

We use two RCTs (Impower 132 and Impower150) to build two hypothetical hybrid control trials (Trial 1 and Trial 2) to evaluate analysis methods. Both RCTs were Phase III randomized trials that investigate anti-PD-L1 antibody (atezolizumab) in combination withchemotherapy (with or without bevacizumab) in stage IV non-squamous NSCLC patients (NCT02657434, NCT02366143) [12, 13]. Trial 1 was constructed based on Impower132, which failed to reject the null hypothesis on OS. The addition of atezolizumab among IMpower132 patients (N=578) produced an OS HR estimate of 0.86 (95% CI: 0.71–1.06). Trial 2 was derived based on Impower150 using two of the three arms comparing atezolizumab plus chemotherapy (ACP or ABCP) to chemotherapy BCP (N=1202). IMpower150 found statistically significant improvement in OS with the addition of atezolizumab (ABCP) versus BCP with an OS HR of 0.80 (95% CI: 0.67–0.95). We consider both trials to show consistency in conclusions of the trials when similar effective sample size are considered in the analysis.

Unique external control cohorts were constructed separately from the internal control arms of the Phase III trials using the nationwide Flatiron Health EHR-derived de-identified database [see supplementary] by following the steps provided by Carrigan and colleagues [14], to match the inclusion criteria of each trial. Applicable covariates of interest were identified, which were consistent with baseline covariates of interest from the RCT. Analysis datasets were determined by complete cases, that is, cases with no missing values within outcome or baseline covariate data. In general, vast amounts of data is available from external data EHR sources. As a result, we cannot control the number of external control patients available for a given study, especially when considering the time-period of data collection and all inclusion/exclusion criteria of the original trial. It is well known that censoring rates will differ between RCT and external control data. We assume here that censoring, while at different rates, is independent from OS in all analyses. It is important to note here that we are not generating censoring rates at any point in the main analysis, we utilize the true observed data including censoring time for the purpose of this paper.

Cox Proportional hazards Method

A standard Cox proportional hazards method [15] was applied to the RCT data from Trial 1 and Trial 2 to reproduce results found in the original trial and serve as a comparison for operating characteristics of each analysis method. No external control data was included in this first analysis. Baseline covariates of interest were determined via scientific knowledge, including consistency with the original trials as well as availability of external control data utilized in further analyses. The baseline covariates were consistent across all analysis methods in the estimation of the treatment hazard ratio.

Pooled Cox Proportional hazards Method

The most naive form of incorporating external controls is to treat the internal (RCT) controls and external controls as if they are from the same population. In this analysis, both the internal and external controls were pooled into one control group and a Cox proportional hazards model was applied as done previously. The control set increases in size but will likely increase the bias in the estimated hazard ratio since we are not accounting for differences between the control groups.

Propensity score (PS) Method

It is common for hybrid control trials to use propensity score methods to weight and match external controls based on their estimated propensity of being included in the trial. Thus, we considered a Covariate Balancing Propensity Score (CBPS) to match external control patients to the total set of RCT patients for both Trial 1 and Trial 2. The PS was estimated using CBPS [16] and matched by nearest neighbors to identify the external controls most likely to be included in the trial based on their propensity scores [17]. When there are fewer external controls than RCT patients, as in Trial 2, k-nearest neighbors matching is used with larger k to allow an external control patient to match to at most k RCT patients. This occurs with Trial 2, and we use k=2. The weights were not used in the final parameter estimation and were only used to identify which external controls most closely matched the patients in the RCTs. Finally, the baseline covariates used to estimate the PS were consistent with those used to estimate the subject-specific case weight in all methods.

Fixed Weight Power Prior Method

A power prior application to a hybrid control trial allows for the incorporation of external controls by weighting each external control subject by a single value. In this comparison, we weight each patient by 0.5, as this is the average weight expected from the subject-specific method described below. The weights are applied to a proportional hazards model with piecewise constant baseline hazards and baseline covariates consistent with the other analysis methods, which can be written as a case-weighted Poisson regression model [18]. The case-weighted Poisson regression model was then used to estimate the treatment hazard ratio.

Case Weighted Adaptive Power Prior Method

The Case Weighted Adaptive Power Prior (CWAPP) method begins by breaking up the time to event axis into disjoint intervals, in which we have an equal number of

events (i.e., deaths) in each interval. Increasing the number of intervals will give a higher sensitivity to changes in the hazard rate, but will increase computation time and complexity. The intervals should be based on the RCT data and contain an equal number of events, as is standard with piecewise baseline hazards models [18].

The weight of interest for each patient is a measure of the compatibility of the time at risk in a given interval relative to the predictive distribution of the RCT data. To fully estimate the weights, three models must be fit separately: (1) a model for the random censoring of the external data, (2) a model for the events in the external data, and (3) a model for the events in the RCT. All three of these models are proportional hazards models with piecewise constant baseline hazards (using the intervals determined above), which are fit using a Poisson regression [19, 20]. The baseline covariates across these models should be consistent, regardless of data source. The subject-interval specific weights and the probability of observing data as or more extreme than the observed external control data are estimated via Box's p-value [21]. The weights are transformed to control type-1 error as discussed in the supplemental material.

The estimated weights are applied to a generalization of the power prior applied to a proportional hazards model using the case-weighted Poisson regression model used to estimate the treatment hazard ratio. A visualization of the flow of a CWAPP based analysis can be found in Fig. 1.

Simulation Study

As a further illustration of the CWAPP method, we present a simulation study which may mimic the set-up of a trial where hybrid controls are useful for whichthe internal control set is smaller than the treatment set. Consider the following general set-up: A trial with 2:1 randomization to treatment and control, with an external dataset at least as large as the randomized control arm. In this case, we aim for the external dataset (with average case-weight approximately 0.5) to add to the effective sample size of the control data to a 1:1 comparison with the treatment arm to provide maximum statistical power.

Simulation Study - Sample size considerations

To emulate an accurate comparison with the two trials considered, we will keep the RCT treatment arm consistent with the original study. Our goal is to examine the augmentation of the control arm with external controls, and thus reduce the number of internal controls. We achieve this by reducing the internal control arm by half. To augment the control arm, we consider utilizing an external sample equivalent in size to the RCT control arm in the original trial or the maximum number of external controls available, whichever is larger. A flowchart for sample size considerations can be found in the Supplemental Materials, Fig. 2.

Considering Trial 1, we will sample 292 RCT treatment patients, 142 RCT control patients, and 284 external control patients, for a final sample size of N=718. Each dataset is analyzed with the same five analysis methods performed on the full data analysis outlined above. We consider 3000 simulated samples. A similar structure is given to the simulation study for Trial 2, with details outlined in the supplementary material. In this simulation set-up, ties are created in the dataset if a participant is selected more than once when sampled with replacement. All methods being compared utilize an Efron approximation to account for any ties in a given sample data set. Treatment assignments are kept consistent in the resampling scheme throughout the simulations for both Trial 1 and Trial 2. A simulation with resampling in

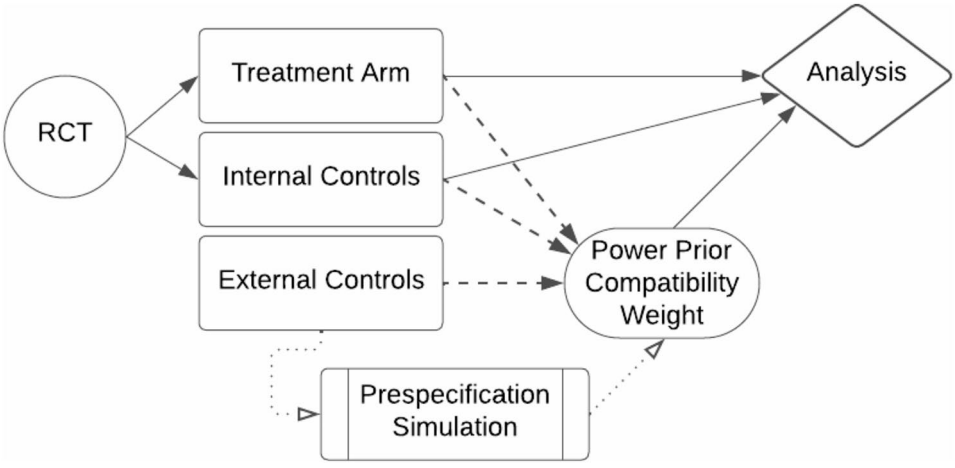

Fig. 1 Flowchart for trial analysis and design considerations when using a Case Weighted Adaptive Power Prior. The RCT treatment arms are given standard weight in the analysis, but are also used in calculating the subject specific weight which will contribute to the power prior in the final analysis

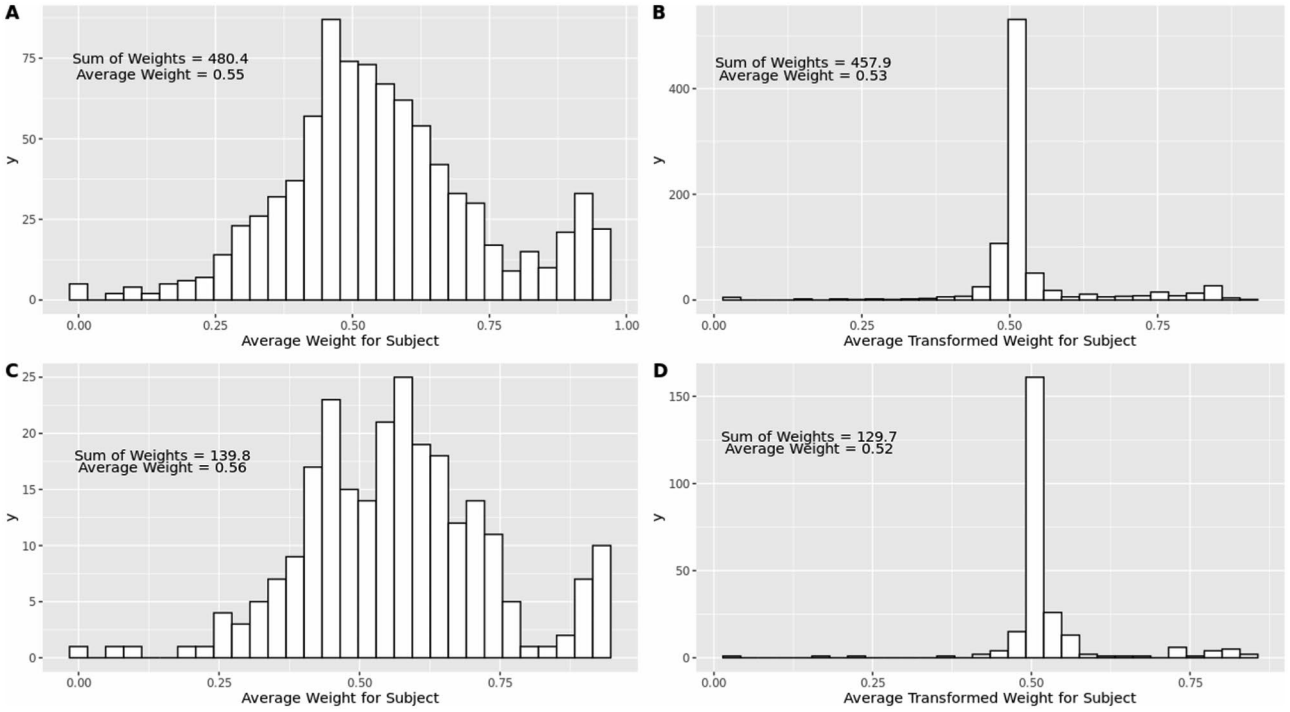

Fig. 2 Histograms of estimated average subject specific weights from Trials 1 (**A**&**B**) and 2 (**C**&**D**). Figure 2a and c depict the subject specific weights before transformation, and Fig. 2b and d depict the subject specific weights after transformation, where the tail weights are allowed to stay close to the extremes, but most weights are pushed towards 0.5. These transformations help control the inflation of the type 1 error

Table 1 Single trial analysis results for trial 1 and trial 2

	A. Trial 1, N_{RCT} = 576					B. Trial 2, N_{RCT} = 686				
Analysis Type	**N**	**N_{EC}**	**Treatment Hazard Ratio**	**95% CI***	**CI Width**	**N**	**N_{EC}**	**Treatment Hazard Ratio**	**95% CI***	**CI Width**
Cox - no external controls	576	0	0.888	(0.738, 1.088)	0.350	686	0	0.762	(0.610, 0.954)	0.344
Cox - pooled controls	1445	869	0.928	(0.788, 1.092)	0.304	934	248	0.748	(0.598, 0.935)	0.337
Propensity Score Matching	1152	576	0.938	(0.795, 1.108)	0.313	930	244	0.75	(0.599, 0.938)	0.339
Fixed Weight Power Prior	1445	8,691,445	0.928	(0.782, 1.101)	0.319	934	248	0.764	(0.621, 0.939)	0.318
Case Weighted Adaptive Power Prior	1445	869	0.922	(0.778, 1.093)	0.315	934	248	0.753	(0.614, 0.924)	0.310

Single Trial analysis results for Trial 1 and Trial 2, with estimated treatment hazard ratios, Confidence Intervals (CI), and Confidence Interval width. Note the changing sample size in the analyses, as each utilizes differing amounts of the available external control data, noted in the column N_{EC}. The number of RCT patients used remains consistent for each analysis method within a trial, as noted by N_{RCT}

which the original treatment arm assignment is blinded is discussed in the Supplemental Material.

Results

Trial 1 – non-statistically significant OS

Trial 1 consists of 576 patients from the RCT and a possible 869 external control patients for a complete case analysis. Baseline patient characteristics including age, ECOG status, race, and smoking status were similar across both groups. The external control group contained a higher percentage of females than the RCT, 51% vs. 33% respectively, as well as a higher rate of carboplatin chemotherapy, with the external control group containing 89% of the patients on carboplatin vs. 61% in the RCT. Prespecification simulation studies identified the appropriate weight transformations to control the type-1 error under both compatible external data and external data with shift confounding ($p=3$, $q=0$). Further details on prespecification of p and q can be found in the Ssupplemental Materials. Table 1A summarizes results across all five methods and the resulting estimate for the treatment hazard ratio for overall survival. Conclusions from hypothesis tests across all 5 methods are consistent with the originally published trial data [12]. We see all methods incorporating external control data produce a hazard

ratio estimate closer to 1 (the null hypothesis) and have confidence intervals with smaller width.

CWAPP subject-specific weights were estimated and have an average of 0.53 after the pre-specified weight transformation was applied. The total contribution from the external control data can be determined from the sum of weights across all confidence intervals and subjects. In the complete-case analysis, the contribution from the external controls is 457.9 subjects from the data set of 869 patients.

Trial 1: bootstrap analysis - trial emulation

The following simulation study was designed to emulate a trial where hybrid control designs would be beneficial, starting with a trial with 2:1 randomization of treatment: control. As seen in the complete case analysis, we expect the contribution of the external controls to be approximately half the number of subjects included in the analysis. Thus, to augment the control arm, we should include twice the number of subjects we would like to account for in the effective sample size. In trial 1, this leads us to analyze 292 RCT treatment patients, 142 RCT control patients, and 284 external control patients. Three thousand datasets of size $N=718$ are constructed by random sampling from the combined RCT and external data with replacement using stratified random sampling while keeping the treatment arm assignment and censoring rate consistent with what was observed in the original data set.

Averaging over three thousand samples, the average treatment hazard ratios, confidence intervals, and bootstrap-estimated type-1 error estimates are summarized in Table 2A. All 5 methods considered provide treatment hazard ratios consistent with findings of the original RCT with confidence intervals noticeably smaller for the Bayesian methods, with the CWAPP providing the most precise estimates, i.e. smaller average confidence interval widths. All methods control the bootstrap-estimated type-1 error at the nominal level, with CWAPP having a type-1 error rate of 0.03.

Trial 2 - statistically significant OS

Trial 2 consists of 686 patients from the RCT and a possible 248 external control patients for a complete case analysis. Baseline patient characteristics including age, ECOG status, and smoking status were similar across both groups. The external control group contained a higher percentage of females than the RCT, 46% vs. 37% respectively, as well as a lower percentage of non-white patients in the external control sample compared to the RCT, 61% vs. 86% respectively. Table 1B summarizes results across all five methods and the resulting estimates for the treatment hazard ratio based on overall survival. Conclusions from hypothesis tests across all 5 methods are consistent with the originally published trial data [13]. We see that three methods which incorporate external control data (pooled Cox, PS Matching, and CWAPP) produce an estimate further from the null hypothesis and all methods estimate the confidence intervals with smaller width than the analysis without external control data.

CWAPP subject-specific weights were estimated and have an average of 0.52 after transformation. The CWAPP subject-specific weights can be seen both before and after transformation in Fig. 2. The total contribution from the external control data can be determined from the sum of the weights across all intervals and subjects. In the complete-case analysis, the contribution from external controls is 129.7 from the data set of 248 patients.

Trial 2: bootstrap analysis - trial emulation

We conduct a similar simulation study as before, but the external control dataset is not large enough for us to augment the entire external control set as was done for Trial 1. Instead of oversampling external control patients to yield a sample size that we would consider ideal, we sample an external dataset equal in size to the total external data available. In trial 2, this leads us to analyze 353 RCT treatment patients, 166 RCT control patients, and 248 external control patients (ideally we would be able to sample 332). Three thousand datasets of size $N=767$ are sampled from the combined RCT and external data with

Table 2 Simulation study results (trial 1 & trial 2)

	A. Trial 1			B. Trial 2		
Analysis Type	**Average Treatment Hazard Ratio**	**Average CI Width**	**Bootstrap-Estimated Type 1 Error**	**Average Treatment Hazard Ratio**	**Average CI Width**	**Bootstrap-Estimated Power**
Cox - no external controls	0.885	0.443	0.025	0.765	0.424	0.471
Cox - pooled controls	0.907	0.454	0.002	0.741	0.411	0.695
Propensity Score Matching	0.915	0.458	0.042	0.708	0.393	0.800
Fixed Weight Power Prior	0.903	0.372	0.019	0.764	0.349	0.803
Case Weighted Adaptive Power Prior	0.894	0.365	0.033	0.749	0.338	0.892

Simulation Study results for Trial 1 and Trial 2, with estimated average treatment hazard ratios, average Confidence Interval (CI) width, and boostrap-estimated Type 1 error or power. Note Trial 1 estimates the type 1 error with the assumed truth of no treatment effect in the clinical data, whereas Trial 2 estimates power as the assumed truth is that there is a positive treatment effect in the clinical data

replacement using stratified random sampling, keeping the censoring rate consistent with what was observed in the original data .

Averaging over the the three thousand samples, treatment hazard ratios, confidence intervals, and power estimates are summarized in Table 2B. All 5 methods considered provide treatment hazard ratios consistent with findings of the original RCT with confidence intervals noticeably smaller for the Bayesian methods, with the CWAPP providing the most precise estimates in terms of confidence interval widths. This can most clearly be seen in Fig. 3 where the estimated treatment hazard ratios and confidence interval widths are plotted for each run of the simulation studies. The methods utilizing external controls have more power than the traditional Cox model (recall that this has half the number of control patients). Both Bayesian methods have higher power compared to the frequentist methods, with CWAPP having the highest power (0.89 vs. 0.80 of the fixed power prior method).

Discussion

Hybrid design clinical trials are being considered for situations where standard RCTs are not feasible, especially with the increased availability and quality of electronic health records. While acknowledging the utilization of external controls can lead to statistically biased effect estimates, when handled with the proper methods, one can achieve controlled type 1 error inflation and gains in efficiency and power when appropriate external controls are available.

We have illustrated the application of the CWAPP method to a hypothetical hybrid control trial with external controls from RWD sources. In the single-study

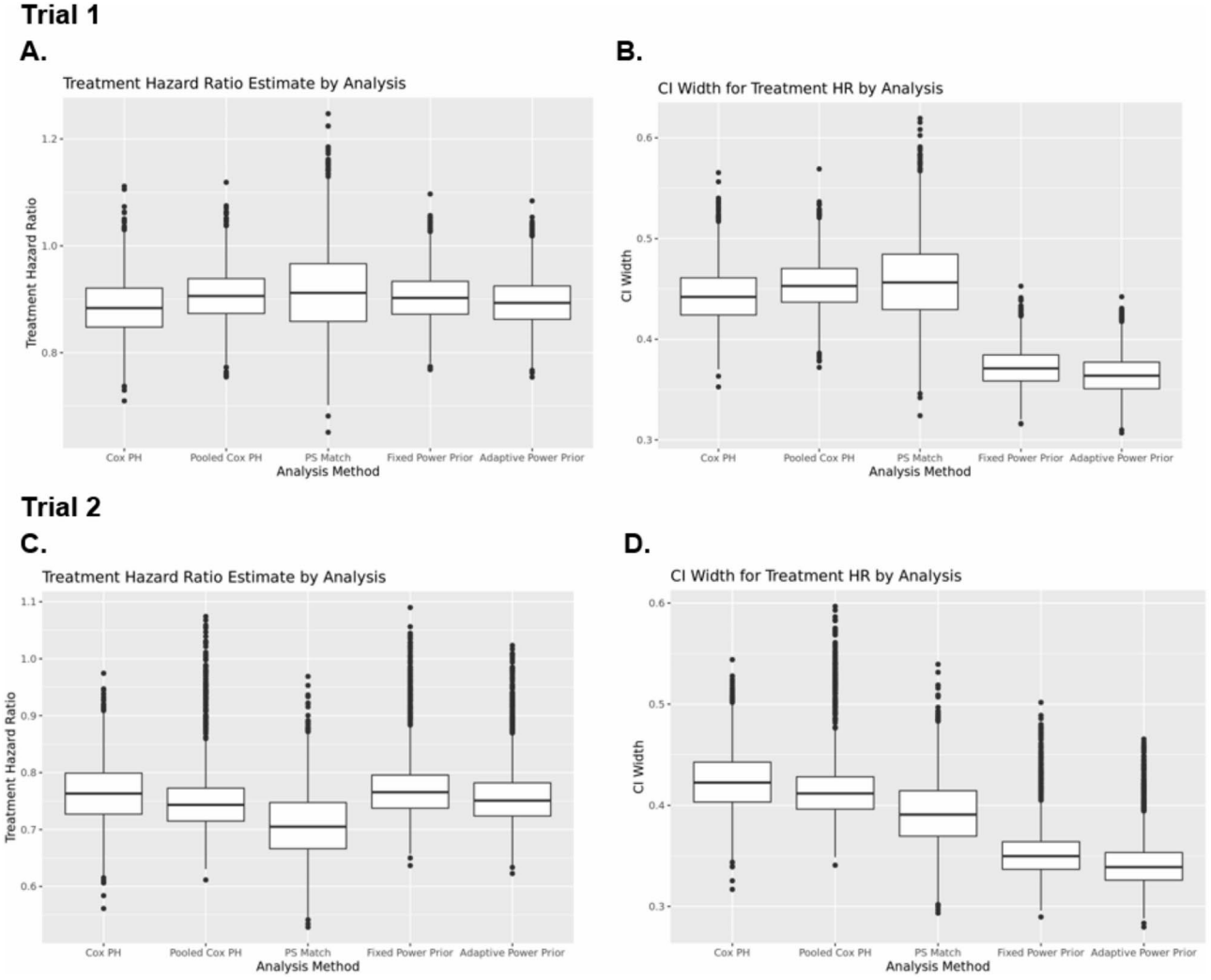

Fig. 3 Summary box plots from simulation studies run for both Trial 1 (**A** & **B**) and Trial 2 (**C** & **D**). Figure 3a and c depict the estimated treatment hazard ratios for all five methods. We see consistent for estimates across the methods in both trials. Figure 3b and d depict the confidence interval widths across all five methods, where we see the Bayesian methods with smallest CI widths, with CWAPP performing best in both trial simulations

application, both Trial 1 and Trial 2 provide CWAPP treatment hazard ratio estimates consistent with the original trial results without using external controls. We see a reduction in confidence interval (CI) widths from the original Cox analysis method, but also see a gain in precision of the CWAPP method compared to the fixed weight power prior (Trial 1 CI Widths: 0.315 vs. 0.319, Trial 2 CI Widths: 0.318 vs. 0.310).

Through the bootstrapped trial emulations, we have compared the analysis methods on 3000 datasets which mirror a trial for which external controls may be most useful. Through these simulations, we see consistent trial hazard ratio estimates when compared to the standard Cox model. In Trial 1, the CWAPP method has a 20% decrease in average CI width compared to the naive pooled control method. Similarly, an the CWAPP method has an 18% decrease in average CI width compared to the naïve pooled control method in Trial 2. We also see when the appropriate weight transformations prespecified, the boostrap estimated type 1 error rate is well controlled at the 0.05 level. With both the fixed power prior and CWAPP methods we see a gain in power at the given sample size from the frequentist Cox methods to either power prior method, with the CWAPP method providing the best bootstrap estimated power in simulations (89%). It is important to note the study from which Trial 2 was derived was designed to reach 87% power to detect an OS treatment HR of 0.75. With the CWAPP method, a similar power was achieved while reducing the internal control arm by one half, utilizing 25% less RCT patients than the original study.

We believe the CWAPP method provides a unique application when subject level data is available from external sources and we no longer have to weight the external data assuming a homogeneous block of patients separate from the RCT and instead can differentiate external control patients we believe to be similar to those in the RCT. The CWAPP method naturally down-weights patients substantially different from patients in the RCT analysis set. Along with the flexibility in weight applied to external controls, the weight provides an estimated effective sample size for the study by summing the average weight per patient. This helps quantify the proportion of the final analysis that is contributed by the prior assumptions, in this case the external control data. The case weight provides transparency on how much external data contributes to the analysis. This framework can be extended to other clinical studies applications with reasonable external controls.

It is important to note that while the weights contribute to this analysis, they do not change the estimates we use for decision making in the trial. In the analysis of each of these trials, we still report familiar treatment hazard ratios and can apply standard hypothesis tests. One must be careful to look at the estimated weights, which can provide an estimate for the contribution of the external patient set to the sample size of the analysis, but they do not change how we interpret the results from the trial.

While we have considered trials depending on time-to-event analysis in these examples, as they are most appropriate in many oncology trials, the case weighted adaptive power prior can easily be extended to other therapeutic areas and other (non-time-to-event) primary outcomes.

Limitations of these analyses and hybrid control approaches as a whole include the large quantities of missing covariate data in external datasets, especially when these data are sourced from real world data. Because of this, our current analysis uses only external controls with complete data, and large quantities of data were excluded from the final analysis. We believe existing missing data methods [22] which are applied to the standard power prior can be extended to apply in this method and this is a topic of future research.

Conclusion

Through two separate trial analyses and emulations, we have seen that hybrid control studies, and specifically the Case Weighted Adaptive Power Prior method, can provide consistent results with previously completed RCTs. When appropriate statistical methods are applied, the bias due to heterogeneity between the RCT and external control arm can be minimized, and consistent study conclusions can be achieved. Moreover, statistical power remains at levels used in the original study sample size calculations while reducing the RCT sample size by 25%.

Abbreviations

FDA	US Food and Drug Administration
RCT	Randomized Controlled Trial
EC	External Controls
RWD	Real World Data
MAP	Meta-analytic Predictive
OS	Overall Survival
HR	Hazard Ratio
NSCLC	Non-small Cell Lung Cancer
HER	Electronic Health Record
PS	Propensity Score
CBPS	Covariate Balancing Propensity Score
CWAPP	Case Weighted Adaptive Power Prior
CI	Confidence Interval

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02383-3.

Supplementary Material 1

Author contributions

EMD, JZ, HP, XL, and JGI conceived the project and designed the analysis in the manuscript. EMD and YZ conducted the analysis and simulation studies, with input from the co-authors. EMD, JZ, HP, XL, YZ and JGI drafted the manuscript.

EK and LAC provided critical input to the manuscript. All the co-authors read and approved the final version of this paper.

Funding
This work is partially supported by a grant from the FDA (HHS U01 FD007206, Ibrahim, Li, Pang, Zhu). The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS or the U.S. Government. EMD was partially supported by the NIH T32 training grant [T32ES007018].

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 December 2023 / Accepted: 21 October 2024
Published online: 01 November 2024

References

1. U.S. Food and Drug Administration. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Oncology Center of Excellence. (2023, February). *Considerations for the Design and Conduct of Externally Controlled Trial for Drug and Biological Products: Guidance for Industry*.
2. Pocock SJ. The combination of randomized and historical controls in clinical trials. J Chronic Dis. 1976;29(3):175–88.
3. Burger HU, Gerlinger C, Harbron C, Koch A, Posch M, Rochon J, Schiel A. The use of external controls: to what extent can it currently be recommended? Pharm Stat. 2021;20(6):1002–16.
4. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. Stat Med. 2015;34(28):3724–49.
5. Ibrahim JG, Chen MH. Power prior distributions for regression models. Stat Sci (2000): 46–60.
6. Duan Y, Smith EP, Ye K. Evaluating water quality using power priors to incorporate historical information. Environmetrics: Official J Int Environmetrics Soc. 2006;17(1):95–106.
7. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* 67.3 (2011): 1047–1056.
8. Lewis CJ, Sarkar S, Zhu J, Carlin BP. Borrowing from historical control data in cancer drug development: a cautionary tale and practical guidelines. Stat Biopharm Res. 2019;11(1):67–78.
9. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. Clin Trails. 2010;7(1):5–18.
10. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics. 2014;70(4):1023–32.
11. Kwiatkowski E, Zhu J, Li X, Pang H, Lieberman G, Psioda MA. Case Weighted Adaptive Power Priors for Hybrid Control Analyses with Time-to-Event Data. *arXiv preprint arXiv:2305.05913* (2023).
12. Nishio M, Barlesi F, West H, Ball S, Bordoni R, Cobo M, Longeras PD, Goldschmidt J Jr, Novello S, Orlandi F, Sanborn RE, Szalai Z, Ursol G, Mendus D, Wang L, Wen X, McCleland M, Hoang T, Phan S, Socinski MA. Atezolizumab Plus Chemotherapy for First-Line treatment of Nonsquamous NSCLC: results from the Randomized phase 3 IMpower132 trial. J Thorac Oncol. 2021;16(4):653–64. Epub 2020 Dec 14. PMID: 33333328.
13. Socinski MA, Nishio M, Jotte RM, Cappuzzo F, Orlandi F, Stroyakovskiy D, Nogami N, Rodríguez-Abreu D, Moro-Sibilot D, Thomas CA, Barlesi F, Finley G, Kong S, Lee A, Coleman S, Zou W, McCleland M, Shankar G. IMpower150 final overall survival analyses for Atezolizumab Plus Bevacizumab and Chemotherapy in First-Line Metastatic Nonsquamous NSCLC. J Thorac Oncol. 2021;16(11):1909–24. Epub 2021 Jul 24. PMID: 34311108.
14. Carrigan G, Whipple S, Capra WB, Taylor MD, Brown JS, Lu M, Arnieri B, Copping R, Rothman KJ. Using Electronic Health Records To Derive Control Arms for early phase single-arm lung Cancer trials: Proof-of-Concept in Randomized controlled trials. Clin Pharmacol Ther. 2020;107(2):369–77. Epub 2019 Oct 11. PMID: 31350853; PMCID: PMC7006884.
15. Cox DR. Regression models and life tables (with discussion). J R Statist Soc. 1972;34:187–220.
16. Fong C, Ratkovic M. and Imai K. CBPS: Covariate Balancing Propensity score. (2002) https://CRAN.R-project.org/package=CBPS
17. Ho DE, Kosuke I, King G, Stuart EA. MatchIt: nonparametric preprocessing for Parametric Causal Inference. J Stat Softw. 2011;428:1–28.
18. Ibrahim JG, Chen MH, Sinha D. Bayesian survival analysis. New York, NY: Springer Science + Business Media, LLC; 2001.
19. Holford TR. The analysis of rates and of survivorship using Log-Linear models. Biometrics. 1980;36(2):299.
20. Laird NM, Olivier DC. Covariance analysis of censored survival data using Log-Linear analysis techniques. J Am Stat Assoc. 1981;76(374):231.
21. Box GEP. Sampling and Bayes' inference in scientific modelling and robustness (with discussion). J Royal Stat Soc Ser A. 1980;143:383–430. https://doi.org/10.2307/2982063. MR0603745. doi.
22. Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: issues and methods. *J Clin Oncol*. 2012;30(26):3297 – 303. doi: 10.1200/JCO.2011.38.7589. Epub 2012 May 29. PMID: 22649133; PMCID: PMC3948388.

Publisher's note