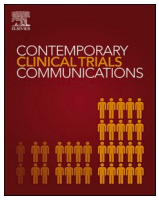
Contemporary Clinical Trials Communications 30 (2022) 101000

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc



Augmenting control arms with real-world data for cancer trials: Hybrid control arm methods and considerations

W. Katherine Tan [a,*], Brian D. Segal [a,1], Melissa D. Curtis [a,2], Shrujal S. Baxi [a], William B. Capra [b], Elizabeth Garrett-Mayer [c], Brian P. Hobbs [d], David S. Hong [e], Rebecca A. Hubbard [f], Jiawen Zhu [b], Somnath Sarkar [a], Meghna Samant [a]

[a] Flatiron Health, Inc., New York, NY, 10013, USA
[b] Genentech, South San Francisco, CA, 94080, USA
[c] American Society of Clinical Oncology Center for Research and Analytics (CENTRA), Alexandria, VA, 22314, USA
[d] Dell Medical School, University of Texas, Austin, TX, 78712, USA
[e] University of Texas M.D. Anderson Cancer Center, Houston, TX, 77230, USA
[f] University of Pennsylvania School of Medicine, Philadelphia, PA, 19104, USA

ARTICLE INFO

Keywords:
Hybrid control arms
External comparator cohorts
Real-world data

ABSTRACT

Background: Hybrid controlled trials with real-world data (RWD), where the control arm is composed of both trial and real-world patients, could facilitate research when the feasibility of randomized controlled trials (RCTs) is challenging and single-arm trials would provide insufficient information.

Methods: We propose a frequentist two-step borrowing method to construct hybrid control arms. We use parameters informed by a completed randomized trial in metastatic triple-negative breast cancer to simulate the operating characteristics of dynamic and static borrowing methods, highlighting key trade-offs and analytic decisions in the design of hybrid studies.

Results: Simulated data were generated under varying residual-bias assumptions (no bias: $HR_{RWD} = 1$) and experimental treatment effects (target trial scenario: $HR_{Exp} = 0.78$). Under the target scenario with no residual bias, all borrowing methods achieved the desired 88% power, an improvement over the reference model (74% power) that does not borrow information externally. The effective number of external events tended to decrease with higher bias between RWD and RCT (i.e. HR_{RWD} away from 1), and with weaker experimental treatment effects (i.e. HR_{Exp} closer to 1). All dynamic borrowing methods illustrated (but not the static power prior) cap the maximum Type 1 error over the residual-bias range considered. Our two-step model achieved comparable results for power, type 1 error, and effective number of external events borrowed compared to other borrowing methodologies.

Conclusion: By pairing high-quality external data with rigorous simulations, researchers have the potential to design hybrid controlled trials that better meet the needs of patients and drug development.

1. Background

Randomized controlled trials (RCTs) remain a gold standard for general clinical research and as regulatory approval support, but their conduct may become increasingly challenging in oncology [1]. While accelerated regulatory approvals facilitate patients' timely access to effective cancer therapies [2,3], real-world data (RWD) could foster further research efficiency. Technology has boosted capabilities for data availability and analyses, unlocking the use of sources such as electronic health records (EHRs) [4–7], and spurring interest in RWD use for drug development and regulatory decisions [8–13].

RWD can be applied to construct fully external comparator cohorts without randomization [14–23]. Alternatively, hybrid controlled trial designs that augment RCT control arms with external cohorts (Fig. 1)

* Corresponding author. Flatiron Health, 233 Spring St, New York, NY, 10013.
E-mail address: ktan@flatiron.com (W.K. Tan).
[1] Current affiliation: Indigo Ag; Boston, MA 02129.
[2] Current affiliation: EQRx; Cambridge, MA 02139.

https://doi.org/10.1016/j.conctc.2022.101000
Received 24 November 2021; Received in revised form 13 July 2022; Accepted 8 September 2022
Available online 20 September 2022

can capitalize on well-developed RWD and still retain the benefits of some randomization. Patients in the external cohort must be the closest possible approximation to the trial control arm, in terms of eligibility, clinical history, and treatment received [17,24,25]. In the case of EHR-derived data, patients in the external cohort can be contemporaneous to the trial. The external cohort is typically downweighted relative to the randomized control arm at the interim and final analyses based on an *a priori* decision rule to protect against potential biases.

The FDA has discussed hybrid control arms in the rare-disease guidance [26,27], and these are examples where hybrid controlled trials could be beneficial in oncology (Table 1):

- Phase III programs facing challenging timelines (due to long enrollment or follow-up periods) or with secondary interest in low prevalence subgroups [28], where hybrid controlled designs might mitigate the risk for premature terminations.
- Single-arm phase II trials using response rate as primary endpoint, which may lead to high type I error rates [29], where a hybrid controlled trials with progression-free (PFS) or overall survival (OS) as endpoints might provide more reliable evidence.
- Randomized phase II trials, oftentimes underpowered to support binary decisions [30,31], can lead to high sign and magnitude error rates [32]. Hybrid controlled designs could increase statistical power and reliability, although the balance between power and bias must be assessed on a case-by-case basis.

For instance, IMpassion130 was a phase III RCT studying the addition of atezolizumab to nab-paclitaxel to treat metastatic triple-negative breast cancer (mTNBC) [33]. This is a high unmet-need setting, but patients may be averse to randomization to the current standard of care (SOC) of single-agent anthracycline- or taxane-based treatments; whereas immunotherapies such as atezolizumab, are more tolerable and have shown early promise [34].

Conversely, hybrid controlled trials would be hard to justify where adequate RCTs are possible. Additionally, there may be cases where it is impossible to construct adequate hybrid control arms without unacceptable bias (the external data may not be fit for purpose).

This article discusses methods and considerations for hybrid controlled trials. In particular, we evaluate a few commonly used dynamic borrowing methods, and propose a new frequentist method that, despite its simplicity, performs equally to more complex methods. While these methods can help to protect against unmeasured confounders and other biases, they cannot overcome fundamental differences in patient populations, patterns of care, or endpoint measurements. Any valid RWD application must carefully assess whether or not the data source is fit for purpose, and prospective validation of the data source and trial design may be necessary [10,12,16,18,19,22,23,35,36].

2. Materials and methods

From an analytical perspective, of the four main steps in a hybrid controlled trial beyond typical RCT procedures (external cohort selection; baseline covariate balance; endpoint, index dates, and follow-up time definition; implementation of a borrowing method), the first three are common to fully external controls, and have been described before [17,19,22,23].

In the implementation of a borrowing method, there are a number of approaches described below, all of which effectively downweight the external data. Whereas the aim of the first three steps is to account for observed patient and trial characteristics, the aim of this step is to protect against sources of bias that are unknown or cannot otherwise be accounted for.

2.1. Existing methods

There are a few commonly used types of borrowing methods [37,38], including Bayesian approaches such as power prior models, commensurate prior models, meta-analytic predictive (MAP) models, robust MAP models [39], and hierarchical models, frequentist approaches such as simple test-then-pool procedures, as well as variations of these approaches.

Recent proposals call for integrating propensity scores into power prior models [40]. However, when patient-level data are available, patient-level matching and weighting (e.g. inverse propensity score weights) may be preferable to strata-level weights, both to retain a clear estimand and possibly to achieve more precise estimates [40]. It would be important to compare these new methods to established patient-level matching and weighting methods before adopting them. For this reason, we currently recommend that patient-level information be used for matching or weighting prior to dynamic borrowing (please see Web Appendix A for details).

Commonly used borrowing methods (Table 2) can be categorized as static (the downweighting factor is fixed *a priori*) or dynamic (the downweighting factor is a function of observed outcomes) [37]. Each method has a tuning parameter that allows a study team to pre-specify how much they are willing to borrow from the external data.

Some methods can determine how much to borrow from external data without accessing data on the experimental patients, whereas other methods do require accessing data on experimental patients. If experimental patient data are used in deciding how much to downweight

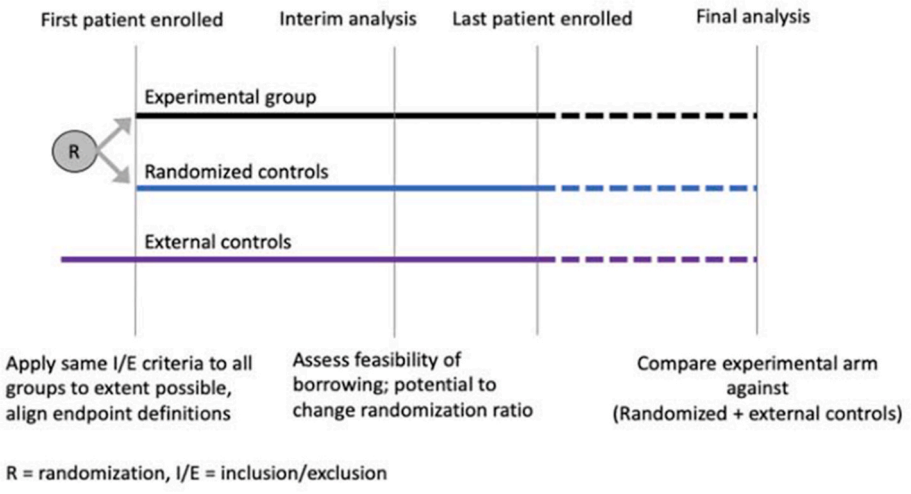


Fig. 1. Example schema for a hybrid controlled trial using external RWD.

Table 1
Example disease settings and trials for which a hybrid controlled trial may be appropriate to consider. In addition to the considerations outlined in this table, it is critical to weigh the considerations in Section 3.2 to determine whether a hybrid controlled trial is appropriate and whether the external data are fit for purpose.

Disease setting and representative trials	Low prevalence disease	Long time to events	SOC with low clinical benefit and/or toxic	Comments
Metastatic triple negative breast cancer (mTNBC) ● IMpassion130 (phase III for atezolizumab) (Schmid et al., 2018)			✔	● Median OS < 18 months ● Lack of targeted therapies ● SOC can be difficult to tolerate (e.g. anthracycline- and taxane-based chemotherapy)
Chronic myeloid leukemia (CML) ● Phase II for imatinib mesylate (Kantarjian et al., 2002)			✔	● Five-year survival for patients diagnosed in 1996–2002, 44.7% (Ries et al., 2006) ● SOC at the time (interferon alfa) had limited efficacy and serious side effects
Progressive Medullary Thyroid Cancer ● EXAM (phase III for cabozantinib) (Eisei et al., 2013)			✔	● 10 year survival percentage of 95.6% for local cancers and 40% for metastatic cancers (Roman et al., 2006) ● SOC is ineffective, so placebo was used for control therapy in EXAM. This raises issues as to whether randomization was ethical.
Notch activating Adenoid Cystic Carcinoma (ACC) ● ACURRACY (clinicaltrials.gov NC T03691207, phase II single-arm) ● A future phase III trial	✔		✔	● Median OS of ~14 months in general population for ACC (Sharma et al., 2008) (not subset to patients with an activating notch mutation) ● Lack of targeted therapy ● No established SOC, and common treatments are ineffective and have serious side effects (chemotherapy, surgery, radiation)
Adjuvant therapy for early breast cancer ● NATALEE (clinicaltrials.gov NC T03701334,		✔		● NATALEE is expected to take 7 years to complete. ● APHINITY enrolled 4800 patients to

Table 1 (*continued*)

Disease setting and representative trials	Low prevalence disease	Long time to events	SOC with low clinical benefit and/or toxic	Comments
phase III for Ribociclib in HR+/HER2-) ● APHINITY (phase III for Perjeta + Herceptin in HER2+) (Von Minckwitz et al., 2017)				observe 381 invasive disease-free survival events.
Pan-tumor NTRK gene fusions ● NAVIGATE (clinicaltrials.gov NC T02576431, phase II basket study for larotrectinib) ● STARTRK-2 (clinicaltrials.gov NC T02568267, phase II basket study for entrectinib) ● A future phase III basket study	✔		May depend on tumor type	● Cohort selection in EHR-derived data may be challenging for basket trials, but might be possible after first gaining experience with each individual tumor type.
First line Diffuse Large B-Cell Lymphoma (DLBCL) ● GOYA (phase III for Obinutuzumab + CHOP vs Rituximab-CHOP) (Vitolo et al., 2017)		✔		● 5 year survival percentage of 62% (Crump et al., 2017) ● Rituximab-CHOP has been an established SOC for many years ● Approximately one third of patients relapse or are refractory to 1 L treatment (Friedberg 2011)
Relapsed/ Refractory DLBCL ● ARGO (NC T03422523, phase II for Atezolizumab, Rituximab, Gemcitabine and Oxaliplatin ● Potential future studies comparing CAR-NK to CAR-T therapies. This may also be relevant in other disease areas (Liu et al., 2020).			✔	● Median OS of 6.3 months for patients whose disease is refractory (best response of progression or stable disease during chemotherapy) or relapses (within 12 months of autologous stem cell transplantation) (Crump et al., 2017)

external patients, the decision could potentially be influenced by the estimated treatment effect, which could raise concerns over the validity of the trial or the need to adjust for multiple hypothesis tests. However, if experimental patient data are never accessed when making this decision, these concerns would not be applicable. See Web Appendix B for details on measuring the amount of information borrowed, as well as Chen et al. [41]for an overview of effective sample size.

Table 2
Common classes of borrowing methods.

Statistical method	Description	Tuning parameter	Pros/cons
Static			
Power prior with fixed power parameter (Chen et al., 2000; Ibrahim et al., 2000)	The contribution of each external patient to the likelihood is weighted by a common "power parameter" between 0 and 1. Typically implemented as a Bayesian model.	Power parameter: Setting it to 1 is equivalent to pooling, and setting it to 0 is equivalent to ignoring external data	**Pro:** Simple and interpretable downweighting factor **Con:** Does not cap type I error inflation or decreases in power
Dynamic			
Test-then-pool (Viele et al., 2014)	A hypothesis test is done to compare the outcomes of external and trial controls after steps 1–3. ● For point null hypotheses, the data are pooled[a] if the null hypothesis of no difference is not rejected, and is ignored otherwise. ● For non-equivalency null hypotheses, the external data are pooled if the null is rejected, and is ignored otherwise	For point null hypotheses: ● The significance level of the test (smaller alpha makes it more difficult to reject the null, and thus more likely to pool) For non-equivalency null hypotheses: ● The significance level of the test (smaller alpha results in wider confidence intervals, making it harder to reject the null and thus less likely to pool) ● The equivalency bounds (larger bounds are more likely to contain the confidence interval, thus making it more likely to reject the null and pool)	**Pro:** ● Simple ● Does not require outcome data for experimental group to determine downweighting factor **Con:** All or nothing approach, resulting in greater variability and uncertainty about how much information will be borrowed
Adaptive/modified power prior model (Duan et al., 2006; Neuenschwander et al., 2009)	Similar to the (static) power prior, but the power parameter is given a prior distribution and allowed to be selected based on the data. The power parameter is estimated simultaneously with all other parameters in the model, including the treatment effect.	Hyperpriors on the power parameter	**Pro:** Retains some of the interpretability of the fixed power prior method **Con:** ● Can be difficult to implement in standard software and can be computationally intensive ● Requires outcome data on experimental group to estimate the downweighting factor
Frequentist version of modified power prior (See two-step approach in Web Appendix A)	Step 1: A regression model is fit to the external and trial controls to estimate the HR between these two arms. The estimated HR is mapped to a downweighting factor, such that HRs near 1 give a downweighting factor close to 1 and HRs far from 1 give a downweighting factor close to 0. Step 2: A second regression model is fit to the pooled external and trial data, giving all external patients the common downweighting factor determined in step 1 and giving all trial patients a weight of 1.	The rate at which the common weights decay to 0 as the HR moves away from 1. For example, the downweighting factor could be defined by the function $w = exp(c*\|log(HR)\|)$ for a tuning parameter $c > 0$. Larger values of c result in a faster decay to 0 as the HR moves away from 1.	**Pro:** ● Simple and interpretable downweighting factor that is chosen dynamically ● Does not require outcome data from experimental group to determine downweighting factor, as the downweighting factor is determined in step 1 and outcome data for the experimental group is not required until step 2 **Con:** Still pending a full evaluation of performance in different settings
Commensurate prior model (Hobbs et al., 2011, 2012)	The outcomes in the randomized controls are centered around the outcomes in the external controls. For example, the log hazard rate of the trial controls might be given a normal prior, centered around the log hazard in the external controls and with hyperprior on the precision of the normal prior.	The hyperpriors on the precision of the normal distribution that shrinks the hazard rate in the randomized controls toward the hazard rate in the external controls. The more this precision is pushed toward zero, the less the hazard in the trial controls is shrunk toward the hazard in the external controls and the more the external controls are effectively downweighted.	**Pro:** Dynamic Bayesian borrowing method that is straightforward to implement in standard software **Con:** ● Downweighting is implicit, so can be more difficult to interpret the amount of borrowed information. ● Requires outcome data on experimental group to estimate the downweighting factor

[a] In this context, pooling refers to combining RWD and trial control data into a single dataset that is then analyzed as though the data were collected together.

Table 2 is not exhaustive, and excludes some notable classes of models, such as MAP models [42]. Note that while MAP models are most appropriate for settings where there are multiple external data sources, our context considers an alternative scenario where borrowing is only from one external data source.

2.2. Proposed "two-step" borrowing method

2.2.1. Overview

We propose a simple "two-step" dynamic borrowing procedure as follows:

1. Fit a regression model to the randomized control and external control cohort and estimate the hazard ratio (HR) between the groups, HR_{RWD} to estimate residual bias between the two cohorts. Note that this step does not involve data for the experimental group. Then, estimate the cohort-level downweighting amount, analogous to the power prior parameter as a function of the HR. The weight function $w = f(HR_{RWD})$ can be any function that fulfills the following criteria: 1) bounded between 0 and 1 (to allow downweighting anywhere from all to none of the external cohort), and 2) monotonically increases with increasing HR_{RWD} (to allow for more downweighting with higher bias between trial and RWD control groups). In our illustrations, we used one example of the weight function $w = exp(-c|log(HR_{RWD})|)$, where $c > 0$ is a constant decay factor selected via simulations that optimize type 1 error and power (see Fig. 2a.) Another example of a weight function is a step function where $w = 1$ when $|\log(HR_{RWD})| = 0$ and then drops to $w = 0$ at an appropriately chosen value of $|\log(HR_{RWD})|$, equivalent to a test-then-pool procedure for a point null hypothesis. Other weight functions may also be considered.
2. Fit a second regression model to a dataset containing both the trial and external patients, giving a weight of 1 to all trial patients and a weight of w to all external patients. The second model is used to estimate the treatment effect of the experimental therapy versus the control therapy.

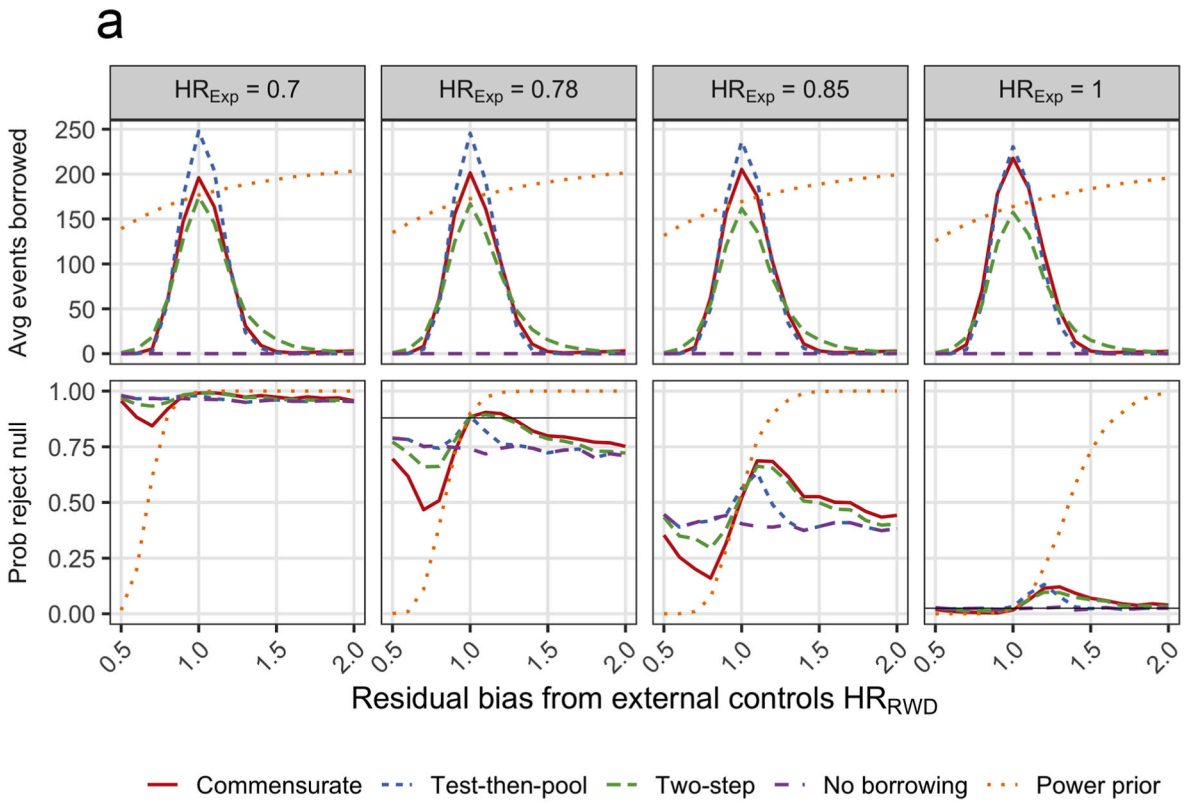


Fig. 2a. Simulation results. X-axis values smaller than 1 indicate that external controls have longer median time-to-event than randomized controls after steps 1–3, and x-axis values larger than 1 indicate that external controls have shorter median time-to-event than randomized controls after steps 1–3. In practice, the full range of residual bias shown on the x-axis may not be relevant (see Section 4.4).

While we used an exponential model in steps 1 and 2 in the simulations of Section 3, it would be straightforward to use a different type of model instead, such as a Weibull or Cox model. Regardless, the same type of model should be used in both steps so that the weights determined in step 1 accurately reflect the amount of residual bias for the model used in step 2.

2.2.2. Details

In more detail, let $z_{i,exp}$ and $z_{i,ext}$ be indicators for whether patient i is in the experimental arm or external cohort, respectively. Also, let D_{exp}, D_{ctr}, and D_{ext} be tabular datasets (all with the same columns and one row per patient) containing data on patients in the experimental arm, control arm, and external cohort, respectively. Suppose that a proportional hazards model is prespecified for the trial analysis.

In step 1 or our proposed approach, the analyst would fit the model $\lambda(t; z_{i,ext}, \beta_{ext}) = \lambda_0(t)exp(\beta_{ext} z_{i,ext})$ using the concatenated row bound pooled dataset (D_{ctr}, D_{ext}), where λ(t) is the hazard at time t and $\lambda_0(t)$ is the baseline hazard at time t. Note that D_{exp} is not involved in step 1, as data on patients in the experimental arm are not required. The dynamic borrowing weight would then be calculated as $w = exp(-c|\widehat{\beta_{ext}}|)$ where the decay factor c is determined prior to analyzing the data through a simulation similar to that in Section 3, in which many values of c are tried in a grid search and one value is selected to achieve the desired operating characteristics.

In step 2 of our proposed approach, the analyst would fit the model $\lambda(t; z_{i,exp}, \beta_{exp}) = \lambda_0(t)exp(\beta_{exp} z_{i,exp})$ using the concatenated row bound dataset (D_{ctr}, D_{ext}, D_{exp}), providing the weight w to all external patients and a weight of 1 to all trial patients. The estimate of the *log HR* β_{exp} and its associated confidence intervals would then be used to determine the effect of the experimental therapy.

As shown in Fig. 2a., the proposed formula for the weight is equal to 1 if HR_{RWD} between the randomized and external controls is equal to 1, and decays to 0 as the HR_{RWD} moves away from 1. Values of c represent a tradeoff between type 1 error and power for the trial: larger values of c result in quicker weight decays, less borrowing, and correspondingly a lower type 1 error at the expense of lower power. Though the weight w is selected dynamically as a function of the data, the procedure for determining the weight (setting the value of c) is specified prior to inspecting and analyzing the data. This is a frequentist analog to the modified power prior, where the weight w is comparable to the power parameter. However, unlike the modified power prior, calculation of w is straightforward and not computationally intensive.

2.3. Data considerations

2.3.1. Treatment time period and data collection methods

Fully contemporaneous external data including only patients who start therapy after the first patient enrolled in the trial and prior to the last patient enrolled would provide the strongest evidence [26]. However, if SOC and diagnostic practices have remained stable in the setting of interest, and there is no evidence of outcome drift prior to the start of the trial, it may be possible to include historical real-world patients. This could increase the size of the external cohort, which could be particularly relevant for rare diseases. If historical patients are included in the real-world cohort, comparability of follow-up times between external and trial patients may be assessed with methodologies such as the reverse Kaplan Meier method [43]. To account for potential differential follow-up times, outcomes may be censored to a pre-specified maximum duration (e.g. censor all events happening after the trial follow-up duration), as long as the censoring algorithm is applied non-differentially across all cohorts.

In addition to the time period of data collection (historical vs contemporaneous, or a mixture of the two), data on the external cohort can be collected either prospectively or retrospectively. Most EHR data

is collected retrospectively without the express purpose of supporting research. However, it is also possible to select patients in a prospectively designed real-world study and follow them through the EHR [44]. While retrospective data capture is less burdensome, prospective intentional data capture may allow for better alignment between the randomized and real-world cohorts.

2.3.2. *Assessment of potential benefits*

The assessment and magnitude of potential benefits of a hybrid controlled trial approach is specific to the trial at hand and depends on several factors and assumptions.

Web Appendix C provides a framework for making these assessments with an illustration for a trial similar to IMpassion130 [33], where a potential reduction of the patients randomized to the control arm by half (by effectively accruing enough control patients in the external data source after accounting for downweighting) might have made it possible to reduce the number of new patients enrolled to the trial by 225 patients yet maintaining required power, read out the study 4 months early, and enroll patients 2:1 (experimental:control) as opposed to 1:1.

3. Simulation study design

To demonstrate how borrowing methods perform, we simulated data resembling a modified IMpassion130 trial [33] if the trial had used a 2:1 instead of a 1:1 randomization ratio, and had been able to effectively borrow half of the control patients from an external data source. Specifically, each simulated dataset had N = 450 trial experimental, N = 225 trial control, and N = 225 expected external RWD events available to borrow. The simulation setup maintained the overall N = 900 sample size of the IMPassion130 trial at its design of 88% power, but with some events coming from the external data source as a hybrid control arm.

To illustrate the performance of statistical borrowing methods across a variety of scenarios plausible in practice, we considered a range of experimental treatment effects, HR_{Exp}, that were more effective or less effective compared to that hypothesized in the IMPassion130 trial ($HR_{Exp} = 0.78$). We also considered values of residual bias between the external real-world (RWD) and randomized controls, HR_{RWD}, ranging from no bias ($HR_{RWD} = 1$) to extreme bias scenarios where the RWD patients were expected to have worse ($HR_{RWD} > 1$) or better outcomes ($HR_{RWD} < 1$) compared to the trial controls. Additional details of the simulation study parameter and values are shown in Table 3 and details of the data generating process, model specifications, and metrics are in Web Appendix B.

For each parameter combination, we simulated 1000 datasets and illustrated performance of five different statistical borrowing methods: 1) commensurate prior model, 2) test-then-pool procedure with a point null hypothesis, 3) our proposed two-step procedure with an exponential model, 4) power prior model with a fixed power parameter ("static power prior"), and 5) an exponential model to the trial data only (no borrowing) for reference. While methods 1–3 are dynamic borrowing approaches, method 4 was a static borrowing method. We averaged the results over the 1000 simulated datasets to compute the average number of effectively borrowed external events, the type I error rate and power for a one-sided hypothesis test at a 0.025 significance level, the mean squared error and bias of the log hazard ratio comparing experimental and control arms, and the standard deviation of the number of events effectively borrowed (See Web Appendix B for details).

These simulations are intended to reflect the type of assessments that might be done at the design stage of a hybrid controlled trial. They emulate a study design in which the external control arm is fully concurrent, and the final analysis is triggered by the total number of events that have occurred across the trial and external arms (downweighting the events in the external arm based on *a priori* assumptions regarding how much information will be effectively borrowed).

As noted above and detailed in Web Appendix B, each borrowing method has a tuning parameter. For each dynamic borrowing method, we used grid search to select tuning parameters that would result in the lowest type I error while maintaining 88% power for the target HR_{Exp} under no residual bias. The approach of minimizing type 1 error at a fixed power (instead of maximizing power at a fixed type 1 error) was used to select c to allow for comparisons across static and dynamic borrowing methods. While dynamic borrowing methods cap the maximum type I error rate over a range of residual biases (of which the actual value is unknown in practice), static borrowing methods do not cap the type I error rate (as they are data agnostic). To enable the side-by-side comparisons of static and dynamic borrowing methods in a practical scenario, we therefore fixed the power under no residual bias and examined potential type I error inflation across a range of residual biases.

Table 3
Simulation setup based on IMpassion130[33].

Parameter	Values
Experimental treatment effect: Hazard ratio between experimental and control arms of trial (HR_{Exp})	0.70 (More effective than expected) 0.78 (Target HR, i.e. alternative hypothesis) 0.85 (Less effective than expected) 1.00 (No treatment effect)
Residual bias: Hazard ratio between real-world controls and randomized controls after careful alignment on I/E criteria, covariate balancing, and alignment of endpoints, index dates, and follow-up time (HR_{RWD}) (composite bias)	Range from 0.5 to 2 by 0.1 (i.e. 0.5, 0.6, …, 1.9, 2.0): 0.5 (Extreme): External patients have longer median time-to-event than randomized controls 1 (No bias) 2 (Extreme): External patients have shorter median time-to-event than randomized controls
Expected downweighting factor for external controls[a]	0.6
Total number of patients in RCT (control + experimental)	675 (out of 900 planned in IMpassion130)
Number of external patients potentially available to borrow	375 (resulting in an expected 375 * 0.6 = 225 effectively borrowed external patients)
Randomization ratio in trial	2:1 (experimental:control)
Target number of events (control + experimental + downweighted external control)	655
Percent lost to follow-up in both the trial and external data source	5%
Accrual rate in trial	34 patients per month
Significance level for hypothesis test of experimental treatment effect	0.025 one-sided

[a] At the time of study design, the downweighting factor is known with certainty if using a power prior model with fixed power parameter, and is predicted if using a dynamic borrowing method.

4. Simulation results

Fig. 2b shows the simulation results for the average number of effectively borrowed external events, power (probability of rejecting the null when $HR_{Exp} < 1$), and type I error (probability of rejecting the null when $HR_{Exp} = 1$). Scenarios to the left of the x-axis represent longer median survival in the external controls than randomized controls. With the tuning parameters selected in these simulations and for $HR_{Exp} = 0.78$ and $HR_{RWD} = 1$ (the target scenario under no residual bias), the commensurate prior model has 88.5% power, the test-then-pool procedure has 88.6% power, the two-step approach has 88.5% power, the power prior model with power parameter fixed at 0.6 has 90.2% power, and the reference model that does not borrow any information from external data has 74.1% power.

As seen in Fig. 2b., the effective number of external events is greatest for the dynamic borrowing methods (commensurate, test-then-pool, and two-step) when the external patients introduce no bias ($HR_{RWD} = 1$), and tapers off as the magnitude of the bias increases. For the static power prior model in this example, the effective number of events is always

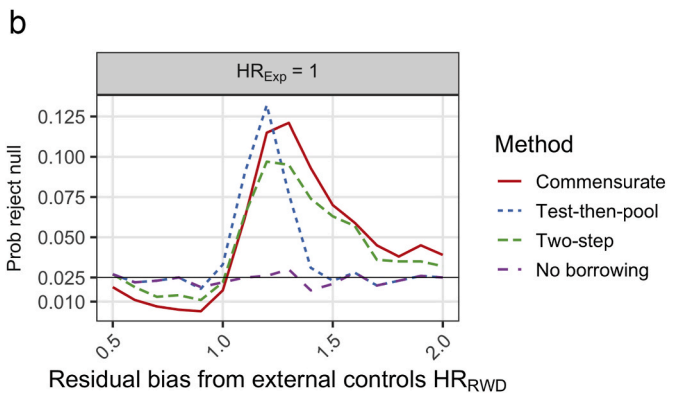


Fig. 2b. Same results for type I error, excluding power prior model and with a different y-axis scale. In practice, the full range of residual bias shown on the x-axis may not be relevant (see Section 4.4).

60% of the total number of external events. As noted above, there tends to be a greater number of external events as HR_{RWD} increases due to the assumption of equal follow-up time for all groups. Also as noted above, there is a decrease in the number of external events as HR_{Exp} increases to 1 (moving from left column to right of Fig. 2b.) because the hazard in the experimental group becomes similar to that in the control arm, and thus more of the total events occur in the experimental group. This can be seen in the results for the test-then-pool, two-step, and power prior methods. Interestingly, the same trend does not occur for the commensurate prior model.

Regarding power, the left-most panel ($HR_{Exp} = 0.7$) represents an overpowered study, so all methods tend to have near 100% power regardless of residual bias and the number of external events borrowed, except for the static power prior model for which power can be dramatically impacted if there is large residual bias. The second column from the left ($HR_{Exp} = 0.78$) represents the expected experimental treatment effect. The horizontal line is at 88%, which corresponds to the designed power of the IMpassion130 trial. All borrowing methods achieve 88% power when there is no residual bias ($HR_{RWD} = 1$) even though the target number of events was not reached with only trial patients. For all methods, power decreases as fewer events are borrowed. This decrease is more pronounced when the median survival in the external controls is longer than in the randomized controls ($HR_{RWD} < 1$), because the few effectively borrowed events reflect a longer median OS, suggesting that the experimental treatment effect is smaller than it is in truth. The third column from the left (HR_{Exp} 0.85) represents a scenario in which the experimental treatment effect is not as strong as anticipated. Hence the power is shifted downward, but the trends are otherwise similar to the scenario in which $HR_{Exp} = 0.78$. The fourth column from the left ($HR_{Exp} = 1$) represents a scenario in which the experiment treatment has no effect, which is required to assess type I error, as discussed below.

The type I error can be dramatically inflated for the power prior method under large residual bias (HR_{RWD} near 2 when $HR_{Exp} = 1$). However, the dynamic borrowing methods all cap the type I error (max type I error rate of 0.13 for test-then-pool, 0.12 for the commensurate prior model, and 0.097 for the two-step regression), which is shown in Fig. 3; these are the same results shown in Fig. 2b., but excluding the static power prior model and with a different y-axis scale. For these simulations and choice of turning parameters, type I error increases for moderate residual bias (HR_{RWD} near 1.2). However, as the residual bias continues to move away from 1, the models stop borrowing, in turn decreasing type I error. This is a key property of dynamic borrowing methods [37]. We also note that type I error decreases below the nominal rate when the median survival in the external controls is longer than in the randomized controls ($HR_{RWD} < 1$); this is for the same reason that power also decreases in this setting. As the residual bias becomes larger ($HR_{RWD} > 1.2$–1.3), and less information is borrowed from the external data, type I error decreases.

By carefully selecting the tuning parameters of the dynamic borrowing methods, we were able to achieve fairly similar performance with the commensurate prior, test-then-pool, and two-step methods. However, in order to obtain 88% in the target scenario with no bias ($HR_{Exp} = 0.78$ and $HR_{RWD} = 1$), the test-then-pool approach incurred the largest max type I error inflation, followed by the commensurate prior model and two-step approach (see Fig. 3). The commensurate prior model is more sensitive to residual bias than the two-step approach in these simulations, as seen by the greater type I error inflation and decrease in power, though it might be possible to improve the performance of the commensurate prior model by using a spike-and-slab prior [41] instead of the Half-Cauchy prior we used in these simulations (see Web Appendix B). However, the spike-and-slab prior is also more difficult to tune.

Web Appendix B also shows results for the mean squared error (MSE) and bias of $\log(HR_{Exp})$, as well as the standard deviation of the number of external events effectively borrowed.

4.1. Assessment of risk and benefits

As noted above, all of these methods have a tuning parameter that can adjust how much is borrowed, which results in different risk/benefit trade-offs [37]. Risk refers to potential inflation of type I error or decrease in power, and benefits refers to potential increases in power, timeline savings, or randomization ratios that allocate more patients to the experimental arm. Fig. 3 shows the simulation results for the two-step method with three different tuning parameters, as well as with a model that is fit to the trial data only (no borrowing). As c increases, the weight decays to 0 faster and borrowing is less likely. This results in lower type I error rates and less of a power decrease when there is

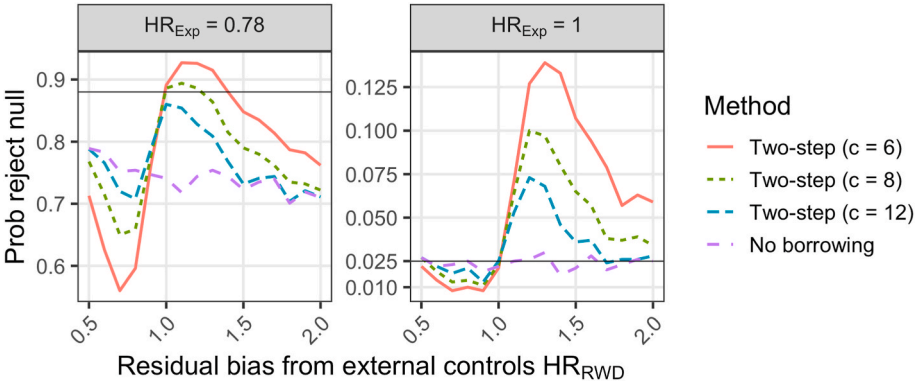


Fig. 3. Two-step procedure with different risk/benefit profiles. In practice, the full range of residual bias shown on the x-axis may not be relevant.

residual bias, but also lower power when there is no residual bias. Similar trends are observed with the other methods [37].

By itself, the results shown in Fig. 2a and b., and 3 may not provide adequate information to support a study team's decision on which risk/benefit profile they prefer. This decision may depend on the amount of residual bias expected in that setting, and the purpose of the trial. To make an informed decision, a study team would need to assess how much bias might be introduced by the external controls after careful cohort selection, covariate balancing, and endpoint, index date, and follow-up alignment (steps 1–3). While it is impossible to know exactly how much residual bias there will be in a particular study, it may be possible to build a body of evidence to suggest likely/plausible scenarios. In particular, by replicating the control arms of recently completed studies (ie, following steps 1–3 and then comparing outcomes with the randomized control) in the same disease area and trial setting, and with the same external data source, it may be possible to develop empirical evidence for how much residual bias might be expected. The operating characteristics of the trial (type I error and power) could then be evaluated accordingly.

For example, Carrigan et al. [19] applied steps 1–3 to Flatiron Health's nationwide EHR-derived de-identified database to emulate the control arms of eleven trials in advanced non-small cell lung cancer (aNSCLC), and found that nine trials had a residual bias HR_{RWD} (obtained by exponentiating the 'Difference in ln(HR)' column of Table 1 in that report) between 0.96 and 1.10 for the Overall Survival (OS) endpoint [19]. These authors speculated that this large residual bias was in part due to the enrichment in the trial population of mesenchymal-to-epithelial transition (MET) positive patients, not accounted for in steps 1–3 [19].

Similarly, Tan et al. (2021) studied 15 trials across multiple tumor types and found that the majority of trials had HR_{RWD} ranging from 0.66 to 1.09 for the OS endpoint [23]. Such evaluations provide a sense of how much residual bias may be plausible and relevant when selecting the value of tuning parameters and assessing the overall suitability of a hybrid design for a future cancer trials with similar I/E criteria. An evaluation in mTNBC, either based on clinical judgment or an analysis similar to Carrigan et al. or Tan et al. [19,23], could help to select the value of tuning parameters and assess overall suitability of a hybrid design for a future mTNBC trial.

5. Discussion

Hybrid controlled trials with external RWD have the potential to improve the efficiency of cancer drug development, which could be particularly beneficial in disease settings with low prevalence or long times to event, or for which the SOC has low clinical benefit and/or is very toxic. While we have primarily focused on two-arm designs in this paper, hybrid control arms could also be extended to multi-arm designs, including platform trials, where several experimental arms are evaluated against a single, shared control arm [45]. Hybrid control arms can be constructed by assessing the borrowing of external data to the shared control arm, and then evaluating treatment effects of multiple experimental arms separately.

Prior to borrowing information from an external source, it is critical to assess whether the external data are fit for purpose. This evaluation involves many factors related to the ability to apply the trial's eligibility criteria to the external dataset (including biomarker and/or genomic information if required), to achieve covariate balance on clinically prognostic characteristics, and to align endpoint definitions, index dates, and follow-up time [17,22,46].

If the data are deemed fit for purpose, then statistical borrowing methods provide a principled way to protect against unknown or unobservable sources of residual bias that persists after alignment with clinical trial information. Our evaluation of borrowing methods varied across the dimensions of static versus dynamic, and Frequentist versus Bayesian. We found that dynamic borrowing methods such as the commensurate prior and two-step regression model tended to protect against type 1 error inflation over a range of residual bias; however, the exact amount of borrowing cannot be pre-specified and is dependent on data similarity. Frequentist dynamic borrowing methods such as the two-step regression model may have the additional advantages of ease of explaining the intuition (i.e. a weighted regression model) and ease of implementing in practice with existing software packages. Yet, there is no one-size-fits-all for every scenario, and therefore for a specific situation, simulations are critical to assess performance of several borrowing methods, as well as for selecting tuning parameters that result in the desired operating characteristics.

While the analytical methods described herein help to address potential discrepancies between the trial and external data source, it is always preferable to minimize these discrepancies at the beginning of the study to the extent possible. To this end, we note that treatment patterns in the real world typically follow standard guidelines, such as the National Comprehensive Cancer Network (NCCN) guidelines, and alignment of the trial protocol with these guidelines could reduce the need to rely on analytical methods later in the study to account for differences. Investigations into treatment patterns and patient characteristics in the real-world can also help to inform trial protocols.

As noted above, there is a history of conducting hybrid controlled trials in cancer, though typically with historical trial data as opposed to external RWD [38,45,47]. When bridging historic and current trials, patient populations, endpoint definitions, and assessment timings may be more similar between trials, as compared to RWD. However, RWD may be more recent or collected concurrently to the trial. Using historical data can be problematic when SOCs (including supportive care) or diagnostic methods have evolved over time, or if I/E criteria have become more inclusive [26,48,49].

For registrational hybrid controlled trials, it could be important for the assessment of comparability between randomized and external controls to be conducted by an independent data monitoring committee in a pre-defined manner. Similar to the two-stage design [50–53], at the interim we recommend that an independent statistician implement the borrowing approach in addition to the weighting or matching. It is typically preferable to have early discussions with regulatory authorities; in the case of the US FDA, we recommend considering study design, operating characteristics of the borrowing methods, and format for submitting RWD [26,54], possibly through the Complex Innovative Trial Designs Pilot Program [49]. There are also operational features of hybrid control designs that require consideration. In particular, the time at which a sufficient number of events have occurred to make an interim assessment on how much information to borrow from the external cohort may occur when the trial is already or nearly fully enrolled. Timing of assessment may be an even more crucial issue when using hybrid controlled designs for platform trials, where multiple experimental treatments are compared against a single control arm using a master protocol, especially if the multiple experimental treatment arms start enrolling at different time points. Furthermore, if the study team had been planning to borrow information from the external data source but the interim assessment shows that it will not be possible, then the trial could potentially be underpowered. In order to mitigate these risks, additional research is needed to develop and assess decision criteria that can be applied early in a trial's enrollment.

In addition, there are many methodological areas for future research to adapt and evaluate borrowing methods for use with external RWD. In particular, it will be important to develop methods for incorporating covariate balancing weights into borrowing methods, including weights to balance post-baseline characteristics and treatment patterns such as differences in treatment duration and subsequent therapies [17,55]. There has already been initial work done in this area [40], but as described above and in Web Appendix A, we think there may be opportunities for simpler solutions that retain a clear causal estimand. There is also a need to evaluate borrowing methods with simulations that reflect the many nuances of RWD, such as missing data and

differential treatment duration, assessment timing, and loss-to-follow-up.

6. Conclusion

This methodological work is done against the backdrop of a large medical need. Nearly two million new cancer cases in the United States are projected for 2022 [56] but only a small fraction will enroll in a clinical trial [57]. Hybrid controlled trials leverage the overlap between clinical trial protocols and routine care, using valuable patient resources more efficiently to better meet the high unmet needs of patients with cancer. As with any use of RWD, the data sources need to be carefully assessed on a case-by-case basis to ensure the data are fit for purpose, and the operating characteristics of the statistical methods need to be assessed through simulations that mimic the trial at hand. By pairing high-quality external data with rigorous simulations, researchers have the potential to design hybrid controlled trials that better meet the needs of drug development and patients.

Funding statement

This study was sponsored by Flatiron Health, Inc., which is an independent subsidiary of the Roche group.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: At the time of the study, BDS, MDC, SSB, WKT, SS, MS report employment in Flatiron Health, Inc., and stock ownership in Roche. BPH reports research fundings from Amgen, scientific advisor role and stock ownership in Presagia. RAH reports grant funding from Pfizer. JZ and WBC report employment in Roche/Genentech and stock ownership in Roche. DSH reports research/grant funding from AbbVie, Adaptimmune, Aldi-Norte, Amgen, Astra-Zeneca, Bayer, BMS, Daiichi-Sankyo, Eisai, Fate Therapeutics, Genentech, Genmab, Ignyta, Infinity, Kite, Kyowa, Lilly, LOXO, Merck, MedImmune, Mirati, miRNA, Molecular Templates, Mologen, NCI-CTEP, Novartis, Pfizer, Seattle Genetics, Takeda, and Turning Point Therapeutics; travel and accommodation expenses from Bayer, LOXO, miRNA, Genmab, AACR, ASCO, SITC; consulting or advisory roles with Alpha Insights, Acuta, Amgen, Axiom, Adaptimmune, Baxter, Bayer, COG, Ecor1, Genentech, GLG, Group H, Guidepoint, Infinity, Janssen, Merrimack, Medscape, Numab, Pfizer, Prime Oncology, Seattle Genetics, Takeda, Trieza Therapeutics, and WebMD; and other ownership interests in Molecular Match, OncoResponse, and Presagia Inc. Other authors: nothing to disclose.

Data availability

No data was used for the research described in the article.

Acknowledgements

We would like to thank Nicole Mahoney, Sheila Nemeth, Khaled Sarsour, and Ashwini Shewade for helpful discussions, and Julia Saiz-Shimosato and Hannah Gilham for editorial assistance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2022.101000.

References

[1] C.S. Bennette, S.D. Ramsey, C.L. McDermott, J.J. Carlson, A. Basu, D.L. Veenstra, Predicting low accrual in the national cancer institute's cooperative group clinical trials, J. Natl. Cancer Inst. 108 (2) (2016) djv324, https://doi.org/10.1093/jnci/djv324. Online publication.
[2] J.A. Beaver, L.J. Howie, L. Pelosof, et al., A 25-year experience of US food and drug administration accelerated approval of malignant hematology and oncology drugs and biologics: a review, JAMA Oncol. 4 (6) (2018) 849–856.
[3] E. Baumfeld Andre, R. Reynolds, P. Caubel, L. Azoulay, N.A. Dreyer, Trial designs using real-world data: the changing landscape of the regulatory approval process, Pharmacoepidemiol. Drug Saf. 29 (10) (2020) 1201–1212, https://doi.org/10.1002/pds.4932.
[4] M.D. Curtis, S.D. Griffith, M. Tucker, et al., Development and validation of a high-quality composite real-world mortality endpoint, Health Serv. Res. 53 (6) (2018) 4460–4476, https://doi.org/10.1111/1475-6773.12872. (Accessed 13 April 2020).
[5] M. Stewart, A.D. Norden, N. Dreyer, et al., An exploratory analysis of real-world end points for assessing outcomes among immunotherapy-treated patients with advanced non-small-cell lung cancer, JCO Clin Cancer Inform 3 (2019) 1–15, https://doi.org/10.1200/CCI.18.00155. (Accessed 24 June 2020).
[6] S.D. Griffith, R.A. Miksad, G. Calkins, et al., (a) Characterizing the feasibility and performance of real-world tumor progression end points and their association with overall survival in a large advanced non-small-cell lung cancer data set, JCO clinical cancer informatics 3 (2019) 1–13.
[7] S.D. Griffith, M. Tucker, B. Bowser, et al., (b) Generating real-world tumor burden endpoints from electronic health record data: comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer, Adv. Ther. 36 (8) (2019) 2122–2136.
[8] S. Khozin, G.M. Blumenthal, R. Pazdur, Real-world data for clinical evidence generation in oncology, JNCI: J. Natl. Cancer Inst. 109 (11) (2017) djx187.
[9] M. Fralick, A.S. Kesselheim, J. Avorn, S. Schneeweiss, Use of health care databases to support supplemental indications of approved medications, JAMA Intern. Med. 178 (1) (2018) 55–63.
[10] H. Eichler, F. Sweeney, The evolution of clinical trials: can we address the challenges of the future? Clin. Pharmacol. Ther. 15 (S1) (2018) 27–32.
[11] US Food and Drug Administration, Framework for FDA'S real-world evidence program (a), https://www.fda.gov/media/120060/download, 2018. (Accessed 11 August 2020).
[12] J.M. Franklin, R.J. Glynn, D. Martin, S. Schneeweiss, Evaluating the use of nonrandomized real-world data analyses for regulatory decision making, Clin. Pharmacol. Ther. 105 (4) (2019) 867–877.
[13] S.V. Ramagopalan, A. Simpson, C. Sammon, Can real-world data really replace randomised clinical trials? BMC Med. 18 (1) (2020) 1–2.
[14] N. Gökbuget, M. Kelsh, V. Chia, et al., Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia, Blood Cancer J. 6 (9) (2016) e473.
[15] Center for Drug Evaluation And Research, Application number: 125557Orig1s000. Statistical review and evaluation. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/125557Orig1s000StatR.pdf, 2014. (Accessed 11 August 2020).
[16] H. Schmidli, D.A. Häring, M. Thomas, A. Cassidy, S. Weber, F. Bretz, Beyond randomized clinical trials: use of external controls, Clin. Pharmacol. Ther. 107 (4) (2020) 806–816.
[17] M. Burcu, N.A. Dreyer, J.M. Franklin, et al., Real-world evidence to support regulatory decision-making for medicines: considerations for external control arms, Pharmacoepidemiol. Drug Saf. 29 (10) (2020) 1228–1235, https://doi.org/10.1002/pds.4975.
[18] I. Chau, D.T. Le, P.A. Ott, et al., Developing real-world comparators for clinical trials in chemotherapy-refractory patients with gastric cancer or gastroesophageal junction cancer, Gastric Cancer 23 (1) (2020) 133–141.
[19] G. Carrigan, S. Whipple, W.B. Capra, et al., 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. 107 (2) (2020) 369–377.
[20] M.A. Hernán, J.M. Robins, Using big data to emulate a target trial when a randomized trial is not available, Am. J. Epidemiol. 183 (8) (2016) 758–764.
[21] Center for Drug Evaluation and Research, NDA/BLA multi-discipline review and evaluation (NDA [NME] 212018) BALVERSATM (erdafitinib). https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212018Orig1s000MultidisciplineR.pdf, 2019. (Accessed 11 August 2020).
[22] Friends of Cancer Research, Characterizing the use of external controls for augmenting randomized control arms and confirming benefits. https://www.focr.org/sites/default/files/Panel-1_External_Control_Arms2019AM.pdf, 2019. (Accessed August 2020).
[23] K. Tan, J. Bryan, B. Segal, et al., Emulating control arms for cancer clinical trials using external cohorts created from electronic health record-derived real-world data [published online ahead of print, 2021 jul 1], Clin. Pharmacol. Ther. (2021), https://doi.org/10.1002/cpt.2351, 10.1002/cpt.2351.
[24] S.J. Pocock, The combination of randomized and historical controls in clinical trials, J. Chron. Dis. 29 (3) (1976) 175–188.
[25] Z.M. Thomas, A hybrid design incorporating real-world evidence for control outcomes, in: ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop. Washington DC, US. September 24, 2019. Paper presented at the 2019.
[26] US Food and Drug Administration, Rare diseases: natural history studies for drug development: guidance for industry: draft guidance. https://www.fda.gov/media/122425/download, 2020. (Accessed 11 August 2020).
[27] J. Wu, C. Wang, S. Toh, F.E. Pisa, L. Bauer, Use of real-world evidence in regulatory decisions for rare diseases in the United States—current status and future directions, Pharmacoepidemiol. Drug Saf. 29 (10) (2020) 1213–1218, https://doi.org/10.1002/pds.4962.

[28] B.P. Hobbs, P.C. Barata, Y. Kanjanapan, et al., Seamless designs: current practice and considerations for early-phase drug development in oncology, JNCI: J. Natl. Cancer Inst. 111 (2) (2019) 118–128.
[29] H. Tang, N.R. Foster, A. Grothey, S.M. Ansell, R.M. Goldberg, D.J. Sargent, Comparison of error rates in single-arm versus randomized phase II cancer clinical trials, J. Clin. Oncol. 28 (11) (2010) 1936–1941.
[30] W.D. Tap, R.L. Jones, B.A. Van Tine, et al., Olaratumab and doxorubicin versus doxorubicin alone for treatment of soft-tissue sarcoma: an open-label phase 1b and randomised phase 2 trial, Lancet 388 (10043) (2016) 488–497.
[31] W.D. Tap, A.J. Wagner, P. Schöffski, et al., Effect of doxorubicin plus olaratumab vs doxorubicin plus placebo on survival in patients with advanced soft tissue sarcomas: the ANNOUNCE randomized clinical trial, JAMA 323 (13) (2020) 1266–1276.
[32] A. Gelman, J. Carlin, Beyond power calculations: assessing type S (sign) and type M (magnitude) errors, Perspect. Psychol. Sci. 9 (6) (2014) 641–651.
[33] P. Schmid, S. Adams, H.S. Rugo, et al., Atezolizumab and nab-paclitaxel in advanced triple-negative breast cancer, N. Engl. J. Med. 379 (22) (2018) 2108–2121.
[34] S. Adams, J.R. Diamond, E. Hamilton, et al., Atezolizumab plus nab-paclitaxel in the treatment of metastatic triple-negative breast cancer with 2-year survival follow-up: a phase 1b clinical trial, JAMA Oncol. 5 (3) (2019) 334–342.
[35] X. Ma, N.C. Nussbaum, K. Magee, et al., Comparison of real-world response rate (rwRR) to RECIST-based response rate in patients with advanced non-small cell lung cancer (aNSCLC), Ann. Oncol. 30 (2019) v651.
[36] S. Ventz, A. Lai, T.F. Cloughesy, P.Y. Wen, L. Trippa, B.M. Alexander, Design and evaluation of an external control arm using prior clinical trials and real-world data, Clin. Cancer Res. 25 (16) (2019) 4993–5001, https://doi.org/10.1158/1078-0432.CCR-19-0820. (Accessed 22 June 2020).
[37] K. Viele, S. Berry, B. Neuenschwander, et al., Use of historical control data for assessing treatment effects in clinical trials, Pharmaceut. Stat. 13 (1) (2014) 41–54.
[38] J. van Rosmalen, D. Dejardin, Y. van Norden, B. Löwenberg, E. Lesaffre, Including historical data in the analysis of clinical trials: is it worth the effort? Stat. Methods Med. Res. 27 (10) (2018) 3167–3182.
[39] H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, B. Neuenschwander, Robust meta-analytic-predictive priors in clinical trials with historical control information, Biometrics 70 (4) (2014) 1023–1032, https://doi.org/10.1111/biom.12242. https://api.istex.fr/ark:/67375/WNG-S5B1VXL8-2/fulltext.pdf.
[40] C. Wang, H. Li, W. Chen, et al., Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies, J. Biopharm. Stat. 29 (5) (2019) 731–748.
[41] N. Chen, B.P. Carlin, B.P. Hobbs, Web-based statistical tools for the analysis and design of clinical trials that incorporate historical controls, Comput. Stat. Data Anal. 127 (2018) 50–68.
[42] H. Schmidli, B. Neuenschwander, T. Friede, Meta-analytic-predictive use of historical variance data for the design and analysis of clinical trials, Comput. Stat. Data Anal. 113 (2017) 100–110.
[43] J.J. Shuster, Median follow-up in clinical trials, J. Clin. Orthod. 9 (1991) 191–192.
[44] M.W. Lu, G. Walia, K. Schulze, et al., A multi-stakeholder platform to prospectively link longitudinal real-world clinico-genomic, imaging, and outcomes data for patients with metastatic lung cancer, J. Clin. Oncol. 38 (suppl) (2020).
[45] J. Normington, J. Zhu, F. Mattiello, S. Sarkar, B. Carlin, An efficient bayesian platform trial design for borrowing adaptively from historical control data in lymphoma, Contemp. Clin. Trials 89 (2020), 105890.
[46] V. Agarwala, S. Khozin, G. Singal, et al., Real-world evidence in support of precision medicine: clinico-genomic cancer data as a case study, Health Aff. 37 (5) (2018) 765–772.
[47] C.J. Lewis, S. Sarkar, J. Zhu, B.P. Carlin, Borrowing from historical control data in cancer drug development: a cautionary tale and practical guidelines, Stat. Biopharm. Res. 11 (1) (2019) 67–78.
[48] E.S. Kim, D. Bernstein, S.G. Hilsenbeck, et al., Modernizing eligibility criteria for molecularly driven trials, J. Clin. Oncol. 33 (25) (2015) 2815–2820.
[49] US Food and Drug Administration, (b) Complex innovative trial designs pilot program, in: https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/ucm617212.htm, 2018. (Accessed 11 August 2020).
[50] N. Lu, Y. Xu, L.Q. Yue, Good statistical practice in utilizing real-world data in a comparative study for premarket evaluation of medical devices, J. Biopharm. Stat. 29 (4) (2019) 580–591.
[51] Y. Xu, N. Lu, L. Yue, R. Tiwari, A study design for augmenting the control group in a randomized controlled trial: a quality process for interaction among stakeholders, Therapeutic innovation & regulatory science 54 (2) (2020) 269–274.
[52] L.Q. Yue, G. Campbell, N. Lu, Y. Xu, B. Zuckerman, Utilizing national and international registries to enhance pre-market medical device regulatory evaluation, J. Biopharm. Stat. 26 (6) (2016) 1136–1145.
[53] L.Q. Yue, N. Lu, Y. Xu, Designing premarket observational comparative studies using existing data as controls: challenges and opportunities, J. Biopharm. Stat. 24 (5) (2014) 994–1010.
[54] US Food and Drug Administration, Submitting documents using real-world data and real-world evidence to FDA for drugs and biologics: guidance for industry: draft guidance. https://www.fda.gov/media/124795/download, 2019.
[55] S.R. Cole, M.A. Hernán, Constructing inverse probability weights for marginal structural models, Am. J. Epidemiol. 168 (6) (2008) 656–664.
[56] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer statistics, 2022, CA A Cancer J. Clin. 72 (2022) 7–33.
[57] J.M. Unger, R. Vaidya, D.L. Hershman, L.M. Minasian, M.E. Fleury, Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation, JNCI: J. Natl. Cancer Inst. 111 (3) (2019) 245–255.