Published in final edited form as:
Pharm Stat. 2014 ; 13(1): 41–54. doi:10.1002/pst.1589.

Use of historical control data for assessing treatment effects in clinical trials

Kert Viele[a,*], **Scott Berry**[a], **Beat Neuenschwander**[b], **Billy Amzal**[c], **Fang Chen**[d], **Nathan Enas**[e], **Brian Hobbs**[f], **Joseph G. Ibrahim**[g], **Nelson Kinnersley**[h], **Stacy Lindborg**[i], **Sandrine Micallef**[j], **Satrajit Roychoudhury**[k], and **Laura Thompson**[l]

[a]Berry Consultants, Austin, TX, USA [b]Novartis Pharma, CIS, Basel, Switzerland [c]LA-SER Analytica, London, UK [d]SAS, Cary, NC, USA [e]Eli Lilly & Company, Indianapolis, IN, USA [f]MD Anderson, Houston, TX, USA [g]University of North Carolina, Chapel Hill, NC, USA [h]F. Hoffman La Roche, Welwyn Garden City, Hertfordshire, UK [i]Biogen IDEC, Cambridge, MA, USA [j]Sanofi-Aventis R&D, Paris, France [k]Novartis, East Hanover, NJ, USA [l]US Food and Drug Administration, Rockville, MD, USA

Abstract

Clinical trials rarely, if ever, occur in a vacuum. Generally, large amounts of clinical data are available prior to the start of a study, particularly on the current study's control arm. There is obvious appeal in using (i.e., 'borrowing') this information. With historical data providing information on the control arm, more trial resources can be devoted to the novel treatment while retaining accurate estimates of the current control arm parameters. This can result in more accurate point estimates, increased power, and reduced type I error in clinical trials, provided the historical information is sufficiently similar to the current control data. If this assumption of similarity is not satisfied, however, one can acquire increased mean square error of point estimates due to bias and either reduced power or increased type I error depending on the direction of the bias. In this manuscript, we review several methods for historical borrowing, illustrating how key parameters in each method affect borrowing behavior, and then, we compare these methods on the basis of mean square error, power and type I error. We emphasize two main themes. First, we discuss the idea of 'dynamic' (versus 'static') borrowing. Second, we emphasize the decision process involved in determining whether or not to include historical borrowing in terms of the perceived likelihood that the current control arm is sufficiently similar to the historical data. Our goal is to provide a clear review of the key issues involved in historical borrowing and provide a comparison of several methods useful for practitioners.

Keywords

priors; borrowing; historical data; Bayesian

1. INTRODUCTION

A large proportion of clinical trials involves the comparison of a novel treatment to an existing control arm, either a placebo or a standard of care. While often the control arm stands on its own within a trial, with parameter estimates for the control group depending

[*]Correspondence to: Kert Viele, Berry Consultants, Austin, TX, USA. kert@berryconsultants.net.

only on the data within the current trial, interest has been growing over the past few decades in leveraging historical clinical trial data on the control arm [1–6]. Often, one or more clinical trials have been conducted involving the control arm (perhaps the current control arm was the novel treatment in the historical trial). In theory, bringing this existing information into the current trial holds the promise of more efficient trial design. Such trials may be smaller, and/or unequal randomization may be used to place proportionately more subjects on the experimental treatment arm in a study, potentially increasing the relative amount of information both on the efficacy and safety of the current novel treatment, as well as on secondary endpoints. In clinical practice, expected results are based on the current set of historical studies, and it makes statistical sense to capitalize on this historical data whenever possible.

In practice, methods for borrowing historical information, and the ramifications of these methods, are less well understood in terms of benefits, effects, and regulatory ramifications. Potentially, the incorporation of quality external information allows for reduced mean square error (MSE), increased power, and reduced type I error within the current trial. In contrast, should the historical data be inconsistent with current trial control arm data, there is a potential for bias and inflated type I error. The relative weights of these risks depend on the phase of development. For example, smaller sample sizes in early phase studies combined with less rigorous control of type I error make the possibility of reduced MSE and increased power very appealing, while in a phase III trial, any possibility of inflated type I error may be controversial. In early phase, development point and interval estimates may carry more weight, but power and type I error remain important as decisions must constantly be made whether or not to continue a development program. Thus, it is important to understand type I error and power in terms of 'how many phase II trials would result in correct go/no-go decisions for phase III'.

Authors of this article are members of the DIA Bayesian Scientific Working Group (BSWG), which was formed in 2011 and includes representatives from industry, regulatory agencies, and academia, with the vision to ensure that Bayesian methods are well understood, accepted more broadly, and appropriately utilized to improve decision making and enhance patient outcomes. Our goal in this article is to illustrate and compare several methods (a test-then-pool approach, power priors, single arm trials, and hierarchical modeling) in a concrete example, showing the amount of weight each method places on the historical data, and the potential MSE, power, and type I error implications.

We specifically emphasize the idea of 'dynamic borrowing' in the approaches considered. It is important that any method for historical borrowing recognizes when the current data appear to be inconsistent with the historical data. We expect variation in the actual parameters from study to study. These may be due to slightly differing patient populations, site locations, improvements in secondary aspects of treatment in the time between the historical and control data, and so forth. A method that incorporates dynamic borrowing borrows most when the current data are consistent with historical data and borrows least when the current data are inconsistent.

To see these issues, we begin with an extreme analogy. Suppose your friend is watching a basketball game and wants to estimate the current (today) free throw shooting percentage of his favorite player (for those unfamiliar with basketball, the key point here is that the player takes a series of 'shots', and each one is either successful or not). Suppose we know that going into the game this season the player has made 130 of 200 free shots (65%). Typically, professional players are fairly consistent over the course of a season, so you argue that this historical data indicates his current true free throw percentage (the parameter) is probably around 65%. There might be some discrepancy today (sampling variability in the historical data, issues today with the particular arena the game is played in, etc.), but you argue you would be surprised if his true shooting percentage is much different than 65%. If you see him shoot five times and hit all five, for example, you might believe his current true shooting percentage is slightly higher than 65%, but you are unlikely to believe he is suddenly a near perfect free throw shooter. Your friend argues 'No! you are going to take the observed results from today and then estimate today's shooting rate as somewhere between the observed data and 65%. That is biased! Suppose my favorite player has corrected his form and now has a true shooting rate of 90%. You will likely reduce the observed rate closer to 65% and thus underestimate my favorite player's true shooting rate for today'. This argument is correct, point estimates constructed in this way are biased if you use the historical data. Your counterargument here is that the data collected in the past has value and that it is quite unlikely for a player to correct their form to this degree, particularly midseason. So do we use what seems like very valuable historical information, or should we be concerned about the possible biases that will result from using it?

While the basketball analogy is not serious, there are several parallels in clinical trials. Typically, an agent is explored in many clinical trials over the course of several years, in situations analogous to the study we want to undertake. We expect there to be some variation in the response rates for our drug across these studies. In the basketball analogy, issues like where today's game is played, and others, may be similar to differing inclusion/ exclusion rules and so forth in the clinical trial. We want to estimate the parameter for our drug for the current study ('today' in the basketball analogy) and need to know how much to incorporate the available historical data. Statistically, incorporating the historical study will produce biases in the presence of 'drift' (if the current study parameters differ from the observed historical rate, we will see biases). For later phase trials involving hypothesis tests, these biases result in inflated type I error depending on the direction and magnitude of the drift. However, if the historical data is on point, we can acquire dramatically better estimates incorporating the historical data, in terms of MSE (we see a variance reduction that more than compensates for the bias) and simultaneous improvements in type I error and power.

Thus, fundamentally the historical data can either help or hurt depending on the relationship between the past data and the current parameter. Our goal in this manuscript is to illustrate these trade-offs in a practical simple analysis. Some methods are more robust to drift than others, and we try to illustrate which methods are the most robust. After assessing the possible benefits and risks, the user must assess whether the benefits exceed the risks, an assessment that should include the likelihood of their occurrence. Returning to the basketball analogy, it may be clear that if the player has corrected their form and now shoots 90%, then borrowing the historical information is detrimental. However, this assumes a change to 90% that may not be plausible. If such changes are unlikely, borrowing from historical data may produce substantial gains over utilizing the limited amount of information in the current day (basketball) or current study (clinical trials).

We describe our example trial in Section 2 as well as the methods we consider for historical borrowing. For each method, we identify parameters the user may control and show how they affect the borrowing behavior, MSE, type I error, and power. While our intent is illustrative rather than a comprehensive review article, we do provide a minimal amount of detail particular to the example and references for more technical details behind the methods. In Section 3, we compare the methods in terms of their borrowing behavior as well as operating characteristics such as MSE, type I error, and power. In Section 4, we provide a 'where to go from here' review of extensions from the current literature to complement the simpler structure of the example, and finally in Section 5, we provide a discussion.

2. METHODS

Suppose we are about to conduct a trial with a dichotomous endpoint where higher rates are preferred. We will enroll 400 subjects. Generally, we will consider designs with equal randomization (200 to control and 200 to treatment).

Looking at the available research on the control arm (this deserves a paper of its own, generally one must be careful in any literature review to identify studies that are 'on point' with similar patient populations, dosing, and so forth to the currently envisioned control arm), we have a historical study that observed 65 responses in 100 subjects on the current control arm. Our goal is to incorporate this information into the current trial. See Section 4 for a description of more complicated scenarios (multiple historical studies, covariates, etc.)

Here, our primary goal will be to maintain our current sample size, using the historical information to increase the power of the trial. Alternatively, we could consider using the historical information and changing to unbalanced randomization (e.g., 2:1 randomization preferential to the treatment arm). In the extreme, single arm trials might be conducted using the historical 0.65 rate as a performance criteria, where the primary analysis indicates that one must beat 0.65 to achieve a trial success. Our goal here is, as much as possible, to perform an 'apples to apples' comparisons of the methods, particularly with respect to a trial that does not borrow any information.

2.1. Methods of borrowing

We consider six methods for incorporating the historical data, the first two acting as 'fenceposts' for understanding our three main historical borrowing methods. We also consider single arm trials, as these are also a form of historical borrowing in that typically the threshold for success (e.g., a null hypothesis response probability) is determined after looking at historical data.

In all examples except for single arm trials, our primary analysis is a hypothesis test of H_0 : $p_0 = p_T$ against H_1 : $p_0 < p_T$, where p_0 is the true rate for the current control arm and p_T is the true rate for the treatment arm. The six methods are as follows:

1. **1.** Separate—we ignore the historical data. This would be viewed as a 'standard analysis'. Here, we would continue with equal randomization on the current treatment and control, with no incorporation of the historical information. We perform a Fisher exact test.

2. **2.** Pooling—suppose we perform equal randomization in the current trial (n = 200 in each arm), but we pool the historical subjects with the current control subjects (thus, if we observe 140/200 = 0.70 in the control arm of the current study, with our 65/100 historical dataset, our actual control estimate would be (140 + 65)/(200 + 100) = 0.683). One could combine pooling with unequal randomization, but we are attempting to maintain an equal number of treatment subjects for all methods. We perform a Fisher exact test but here pool the historical information as if they had been control observations in the current trial.

3. **3.** Single arm trial—while somewhat unusual for these sample sizes, many single arm trials are conducted that look at historical data (often with sample sizes less than our 100 historical subjects) to create a performance criterion that must be beaten in the current study. This performance criterion may be either a point estimate or some upper quantile of a CI based on historical data. Single arm trials may be used in situations where accrual is particularly difficult (thus the goal is to obtain reasonable power from smaller sample sizes) or where it is viewed as unethical to

include a control arm. In our example, suppose we eliminated the control arm and placed 200 subjects on the treatment arm, with a primary analysis testing $H_0 : p = 0.65$ against $H_1 : p > 0.65$, where the 0.65 is acquired from the observed historical rate. We perform an exact binomial test.

4. Test-then-pool—pooling presents an obvious difficulty in that a priori we may not be sure our historical data are sufficiently similar to our current control arm (our efforts in reviewing the literature notwithstanding). We would like a way to avoid pooling in situations where the current control arm appears to be different from the historical data. In 'test-then-pool', we make a choice between the 'separate' and 'pooling' options by first performing a test of $H_0 : p_0 = p_H$ against $H_1 : p_0 \neq p_H$, where p_0 is the current control response rate and p_H is the historical control response rate. If the null hypothesis of equality is not rejected, one uses the pooling approach. If the hypothesis of equality is rejected, then one completely ignores the historical data and performs the separate analysis. This is a basic form of dynamic borrowing, as the amount of weight assigned to the historical data depends on the data in the current trial.

5. Power priors—the power prior ([4], described in more detail in Section 2.4) assigns a 'weight' to the historical data some-where in between the pooling (weight=1) and separate analyses (weight=0). Thus, the historical data are incorporated to a degree into the current analysis.

6. Hierarchical modeling—in a hierarchical model [1, 2, 5, 6], described in more detail in Section 2.5), we assume a distribution across studies (here the current and historical controls) with an explicit parameter τ measuring the variation across studies. A prior distribution is placed on τ that is then updated using the current data. A discrepancy between the historical and current data would put more weight toward larger τ values in the posterior distribution than would an agreement between the current and historical data. As with power priors, the borrowing depends on the parameter τ and incorporates its uncertainty, producing dynamic borrowing.

Generally, these methods move from the simplest to implement to more complicated. Separate, pooling, or single arm trials can be quickly implemented from scratch or have standard implementation in statistical software packages. Test-then-pool requires some basic coding to connect the two hypothesis tests (one for whether to pool, the other to perform the final analysis). Power priors, depending on the likelihood, may be performed in a statistical package or may require MCMC, while hierarchical modeling almost always requires some MCMC implementation, although some commercially available clinical trial simulation software will perform these calculations automatically. In general, none of these methods require excessive computation that would be an obstacle to implementation.

2.2. Comparison of pooling, separate, and single arm trials

We tend to think of the separate and 'pooled' analyses as fence-posts in that they represent the extremes of borrowing. Intriguingly, a single arm trial represents a further extreme of borrowing in that we typically use the historical data to construct a performance criterion. Thus, given that our historical study has an observed rate of 0.65, we might choose a single arm trial where we place 200 subjects on treatment (no control arm) and use a primary analysis of H_0: $p_T = 0.65$ against H_1: $p_T > 0.65$. In effect, in the single arm trial, we choose not to observe the control data (typically this is performed with smaller sample sizes, but the principles described here remain).

With $n = 200$ subjects on the treatment arm and using $\alpha = 0.025$, we find that the null hypothesis H_0: $p_T = 0.65$ is rejected if we observe $Y_T = 144$ responses or more using an exact binomial test.

Figure 1 compares what the trial will conclude for separate, pooling, and single arm trials. The X-axis in Figure 1 shows the observed number of control responses in the current trial, while the Y-axis shows the observed number of treatment responses. The three curves represent the decision boundaries for separate (orange), pooled (red), and single arm (purple) designs. Trials with a (control and treatment) result above the curve are successful, while trials with a (control and treatment) result below the curve are not successful. Note that the red (pooled) curve is more horizontal than the separate (orange) curve. The extra information from the pooled observations results in the test being less sensitive to the current control data. Note the purple (single arm) curve is flat. The current control arm is not observed in a single arm trial and thus has no effect on the results. Generally, the pooled analysis would be more and more horizontal as the sample size of the historical data increases. $N = 200$ is shown, but $N = 0$ would correspond to the separate analysis (nothing to pool). In the extreme, with an infinite sample size in the historical data, we would acquire the flat purple curve. Essentially, the single arm trial completely ignores the uncertainty in the historical data and assumes that the control parameter is known, thus borrowing to a degree beyond pooling (the effective sample size borrowed is infinite).

Once a method has been established, we must consider its operating characteristics. We proceed here similarly to much of the literature, fixing the historical data and computing MSE of point estimates, type I error, and power as a function of the true rate p_0 in the current control arm. Before continuing, it is important to also note that an alternative framework exists, where one may prospectively (prior to the historical study) plan on using the historical data in the current trial. This would occur, for example, in an inferentially seamless phase II/III trial, where one prospectively decides that the phase II data will be combined with the data from the phase III portion of the study. In our current framework with the historical data fixed but the current control parameter varying freely, we always consider situations where the observed historical rate is far from the current control rate, without consideration of how likely this is to occur. To take this to the extreme, suppose we had several large, clearly applicable historical trials all with rates near 0.65. The argument 'if the true current control rate is 0.90, then you will acquire a large bias to your estimates and inflate type I error' is certainly true. However, the historical data themselves (several large trials, all with rates near 0.65 and thus far from 0.90) make the premise of the argument 'if the true current control is 0.90…' questionable at best. In the second framework (borrowing considered prospectively, prior to the historical data), these differences between the current and historical data are part of the likelihood and hence directly computable (of course, issues of drift in the parameter, etc., are still relevant).

Figure 2 shows the MSE, type I error, and power (for detecting a 12% improvement on the treatment arm) conditional on the current control rate (X-axis) and fixing the historical data.

Mean square error is generally flat for the separate analyses (with the usual maximization of the variance for $p_0 = 0.5$). The pooled analysis (red) reduces the MSE when the true control rate is anywhere between 0.58 to 0.73, while the single arm trial (purple, which simply assumes the control rate is 0.65) is perfect when 0.65 is the correct answer but quickly has bias with any change in the true control rate.

Note that the orange curves show that the separate analysis controls type I error under 0.025 (dashed horizontal line) by design while achieving a variable amount of power depending on the control rate (power is around 70% if the true current control proportion is 0.65).

Generally, for all borrowing methods, we can divide the current control rate axis into three regions. In a region near the observed historical rate (0.65), methods that borrow generally simultaneously have lower type I error rates and higher power than a separate analysis, largely combined with reduced MSE. The borrowed data are an accurate estimate of p_0 and thus have all the benefits of simply adding extra data without any added cost. We call this region 'the sweet spot' as here borrowing dominates a separate analysis. For the pooled analysis, the sweet spot extends from a true control rate around 0.61 (where the power for pooling exceeds the separate analysis) to around 0.67 (after which the type I error inflates, although the type I error for pooling is actually still controlled a little farther than that).

For true current control rates below the sweet spot, the pooled analysis controls type I error very well but has reduced power. For observed control rates much lower than 0.65, borrowing increases the estimated control rate, and thus, the bar for the treatment arm is raised (it requires more treatment responses for success than a separate analysis). This makes it harder to declare trial success that simultaneously lowers type I error and power. Similarly, if the true control rate is much higher than 0.65 (how much higher depends on the borrowing method), the bar for success tends to be lowered, making it easier to declare success. This has great benefits for power but comes at the cost of type I error inflation. Generally, these areas also show increased MSE.

Thus, in assessing a borrowing method in terms of MSE, type I error, and power, we have several questions. First, how broad is the sweet spot and how much is the benefit (how much power increase)? The larger a region where borrowing dominates the separate analysis, the more appealing the method will be. Second (and most important for possible confirmatory trials), where is the type I error and how much type I error inflation occurs? While type I error will almost always inflate to some degree, often user selected parameters can reduce the inflation if a particular 'cap' is desired. Third, how much of a power loss do we have when the true control rate is much lower than the historical rate? Again, user controlled parameters can affect this amount.

These three regions (reduced power, sweet spot, and inflated type I error) create a decision problem in deciding whether to adopt borrowing. As noted earlier, a key issue is the likelihood of relative values in the sweet spot and the magnitude of the type I error inflation and power loss. The historical data itself should provide some insight into the likely range for the current control rate (that is the point of using it), but this should also be combined with substantive knowledge regarding possible drift in the true parameters in the time between observing the historical and current data. If the historical data have a sufficiently large sample size and drift may be assumed to be minimal, the sweet spot may be quite likely to occur. Alternatively, in situations where much drift may be expected (or even observed if historical data are available at various time points), we may be quite worried about the extreme regions and question borrowing.

In our example, we noted that the sweet spot is around 0.61–0.67. A negative aspect of pooling is that the inflation of type I error thereafter occurs rapidly and without bound. Should the true control rate be 0.80, the type I error rate approaches 20%. The following alternative methods attempt to improve upon this aspect.

It is important to note that the single arm trial suffers from substantially larger type I error inflation than pooling. The type I error rate is inflated for almost all true control rates above 0.65, and the type I error rate exceeds 20% when the true control rate is 0.70. The central statistical advantage of the single arm trial is its dramatically increased power when the true control rate is exactly 0.65, but its performance quickly degrades should there be any drift in the control population. Generally, type I error is considered 'controlled' under the nominal null of H_0 : $p = 0.65$, but given that the actual trial conclusion may be interpreted as testing a null of H_0: $p_0 = p_T$, we note that generally the performance of the single arm trial is worse than any common method for historical borrowing. Note that this problem is not removed simply by choosing a higher performance criteria (such as running a single arm trial with a null hypothesis of H0 : $p = 0.70$). This will shift the power and type I error curves downward and thus may afford type I error control up to a true current parameter of 0.7, but will not change the fact that the type I error rate inflates greatly beyond whatever null value is chosen.

2.3. Test-then-pool

A difficulty in the pooled (and single arm) examples is the dramatic type I error inflation. This is a result of ‘static borrowing’ in that the pooled analysis always borrows the complete historical dataset. In contrast, we would prefer dynamic borrowing where the amount of historical data borrowed is related to the agreement between the current control and historical data.

In test-then-pool (described in Section 2.1), we first perform a hypothesis test of equal rates between the current and historical control subjects. If this hypothesis is not rejected, we use the pooled analysis. If the null hypothesis of equality is rejected, we perform the separate analysis. This results in an ‘all or nothing’ approach, using one of the extremes of borrowing.

The user may choose the size of the test (α) of equality between the current and historical controls (one might choose $\alpha = 0.20$, 0.10, 0.05, or 0.01 independently of the size of the test comparing treatment and control for the primary analysis, which remains 0.025 regardless of the borrowing behavior). By changing the size of this test, one can produce rejection regions anywhere in between the pooled and separate analyses (large sizes always reject the null hypothesis and thus are nearly always separate analyses, while small sizes almost never reject and thus almost always pool).

Figure 3 illustrates the decisions made by the test-then-pool approach for tests of equality of the current and historical controls with sizes 0.20, 0.10, 0.05, and 0.01. combined with a 0.025 size test comparing treatment to control. Similar to Figure 1, the orange and red curves provide the decision boundaries for separate (orange) and pooled (red) trials. Trials above the decision boundary are successful. Note that the blue curves represent the test-then-pool decisions. These overlap the pooled analysis when the test of equality is not rejected and overlap the separate analysis when the test of equality is rejected. The size of the test determines the degree of overlap with separate and pooled. Small sizes require extreme differences between the current and historical data to reject. Thus, small sizes emulate pooled analyses over a broader range of control data than large sizes. Here, the test of equality at size 0.10 (solid blue line) between the current and historical controls would not reject (and hence pool) if the current control arm has between 109 and 149 responses (observed rates between 0.545 and 0.745). Thus, we acquire ‘cliffs’ in the borrowing structure. If we observe 108 responses, we completely ignore the historical data, while for 109 responses, we pool. Changing the size of the test of equality simply widens or narrows the region in which pooling occurs.

One important advancement here over pooling is that there is a constraint on the borrowing. If the current control arm differs sufficiently from the historical data, the model will cease borrowing heavily (here it will not borrow at all). This allows for dynamic borrowing and thus caps the amount of type I error inflation possible.

Figure 4 shows the borrowing behavior, MSE, type I error, and power characteristics for test-then-pool (blue curves) compared to separate (orange) and pooled (red) analyses, similar to Figure 2. The four blue curves in each plot correspond to different sizes ($\alpha = 0.20$, 0.10, 0.05, and 0.01) for the test of equality of the historical and current control data. Recall that a size of 0.10 (solid blue line) is shown in Figure 3, and changing the size of the test of equality (e.g., the criteria for pooling) can expand or contract the region where pooling takes place. Thus, the pooled and separate analyses can be viewed as extremes of test-then-pool.

The top left panel in Figure 4 shows the expected number of borrowed subjects. Given test-then-pool either borrows all (100 subjects) or none (0 subjects), the expected number of borrowed subjects simply is 100 multiplied by the probability one will observe a dataset where pooling occurs. Thus, for example, we see that when the true control rate is 0.75, the solid blue line (size = 0.10 for test of equality) shows we expect around 40 borrowed subjects, which indicates around a 40% chance of pooling (a true 0.75 current control parameter has a reasonable likelihood of generating data close to the historical control rate of 0.65). Note that if the true control rate is near 0.65, the model is extremely likely to emulate pooling, but as the true control rate differs from the historical 0.65, we are less likely to borrow. This is dynamic borrowing in that the weight given to the historical data depends on the current control data.

The upper right panel shows the MSE of test-then-pool. For any chosen threshold, test-then-pool performs similarly to pooling when the true control rate is near 0.65. In contrast to pooling, as we move away from $p_0 = 0.65$, we borrow less, and hence, the MSE does not continue to increase but rather decreases to approach the separate analyses.

The bottom panels of Figure 4 are similar to Figure 2, again showing four blue curves. As test-then-pool tends to emulate pooling for control rates near 0.65, we see a similar sweet spot of improved MSE, increased power, and reduced type I error for test-then-pool. In addition, as an advantage over pooling, we see the dynamic borrowing behavior. As the true control rate moves away from 0.65, the model borrows less, thus capping the amount of possible type I error inflation or potential power loss outside of the sweet spot. If one is given a goal, for example, we want a type I error rate of 0.025 around $p_0 = 0.65$, but are willing to tolerate 'X' amount of inflation, then one can find an appropriate test-then-pool parameter that achieves that goal, while preserving most of the power gains near $p_0 = 0.65$. Also note that the maximal type I error inflation occurs for a current control rate around 0.75 or higher, so this may or may not be a concern for the current trial, depending on the anticipated current control rate.

There are several possible variants of test-then-pool. For example, the point where null hypothesis could be replaced by an equivalence test. One could also consider a Bayesian model averaging [7] approach that considers the posterior probabilities of pooling and separate approaches. This would create a smoother version of test-then-pool. The remaining methods (power priors and hierarchical models) allow a continuum of borrowing (e.g., one can weight the historical data somewhere between pooling and separate analyses).

2.4. Power priors

The power prior (see, e.g., [4]) is a useful class of informative priors for historical borrowing. Here, we focus on the particulars of applying the power prior to the binomial dataset; a more complete description may be found in [2, 6] and [8].

Intuitively, the goal behind a power prior is to 'downweight' the historical data to some degree. Thus, in our example, we have 100 historical observations available. Pooling would borrow these at full weight, adding 100 observations to the 200 we will have on the current control arm, and would result in an effective sample size of 300 for the current control arm. Given that we may have differences between the historical studies and the current control arm beyond simple sampling error (if that were the only discrepancy, pooling would be appropriate), one option is to downweight the historical data, treating the historical data as if it had the same observed rate but with a smaller sample size.

Thus, in our example, suppose prior to the historical study we began with a vague Beta(0.001, 0.001) prior on the response rate. After observing 65 of 100 responses in the historical study, we would update this prior to a Beta(0.001 + 65, 0.001 + 35) posterior distribution. With downweighting, we treat the data as if the observed proportion was 0.65 but with a smaller sample size. Suppose our weight was $a = 0.4$ (the power parameter), indicating we want to count the 100 observations as if they held the information of 40 observations. Our posterior distribution after the historical data is observed would be Beta(0.001 + (65 * 0.4), 0.001 + (35 * 0.4)).

We would then use this posterior (Beta(26.001, 14.001)) as an informative prior for the current control study. This weights the historical data less than pooling but still incorporates the point estimate from the historical study to some degree. A weight of 0 corresponds to ignoring the historical data and performing a separate analysis, while a weight of 1 corresponds to pooling.

After observing Y_0 responses on the current control arm and Y_T responses on the treatment arm, combined with our power prior p_0 ~ Beta(26.001, 14.001) and a non-informative p_T ~ Beta(0.001, 0.001) (we may want to be more careful about these priors if we expected very few or very many responses), we would acquire the posterior distribution $p_0|\text{Data}$ ~ Beta(26.001 + Y_0, 14.001 + 200 − Y_0) and $\mathrm{p_T}|\text{Data}$~Beta(0.001 + Y_T, 0.001 + 200 − Y_T). As this is a Bayesian analysis, our primary analysis is conducted by declaring trial success if $\Pr(p_0 < p_T|\text{Data}) > 0.975$. Note that we can change the 0.975 threshold to further tune type I error and power.

Formally, power priors are defined as a generalization of the usual Bayesian updating step, where the prior for the current study is a product of an initial prior (often non-informative) and the likelihood for the historical study. In a power prior, the likelihood for the current study is raised to a power (between 0 and 1), and the resulting 'posterior' is used as an informative prior for the current study. For a Beta prior, this results in the calculation earlier. For more general information, see Chen and Ibrahim [9].

Figure 5 shows how this method would make decisions using 20%, 40%, 60%, or 80% weight. Almost by definition, the decisions are proportionally between pooling and separate analyses. Note this creates static borrowing, in that the amount of weight given to the historical data does not depend on the current control data. If you use 40% weight (solid curve), you always borrow 40 observations. Generalizations to dynamic borrowing are briefly mentioned in the succeeding paragraphs.

Figure 6 compares the borrowing, MSE type I error, and power characteristics using 20%, 40%, 60%, or 80% weight. As noted earlier, the weight assigned to the historical data is fixed regardless of the current control data. Thus, the top panel of Figure 6 shows horizontal blue lines at 20, 40, 60, and 80 observations. As with previous comparisons of type I error and power, we observe a sweet spot region, with lower current control rates resulting in reduced power and higher current control rates resulting in type I error inflation.

Generally, the sweet spot is wider and shifted to the left (lower current control rates) compared to the pooled analysis. The shift is largely a result of the shift from a frequentist to a Bayesian analysis combined with the discreteness of the problem (thus, with different sample sizes, this may not be the case). The larger sweet spot region is thus a function of the downweighting resulting in increased power over a much broader range of control values less than 0.65. Instead of only going down to 0.61 (where the pooled and separate analysis power curves cross), downweighting of 40% (solid blue line) has higher power all the way down to a control rate of 0.58, with a 20% downweighting descending even farther. A similar situation is observed with the MSE, where low weights result in broad regions of slight improvement over separate analyses, while high weights result in MSE curves closer to the pooled analysis.

Also note that, like test-then-pool, the weight parameter can be set to control the amount of type I error inflation. Unlike test-then-pool, there is no cap on the amount of type I error inflation. While the graphs shown reach a maximum, as the true current control rate exceeds 0.80, the type I error continues to increase. This is in contrast to test-then-pool and other dynamic borrowing methods that reach a maximum and then decrease as they borrow less and less for extreme datasets. By choosing a weight of 20% or 40%, one achieves far less type I error inflation than a pooled analysis.

Hobbs *et al.* [3] as well as Ibrahim and Chen [4] and Neuenschwander, Branson, and Spiegelhalter [10] consider dynamic borrowing versions of power priors that consider the weight random and estimate it using the agreement between the current and historical data.

An empirical estimate for the power parameter (or a fixed value) appears to present an easier posterior analysis than a random power parameter. Specifically, even with a properly normalized conditional power prior [11], Hobbs *et al.* [12] point out that the posterior distribution for the power parameter does not involve the current data. Thus, the commensurability of the historical and current data is not easily captured using random power priors. Neelon and O'Malley [13] warn that using random power parameters (in their case, with a beta prior distribution) tends to over-attenuate the influence of the historical data, and so one might have to use highly informative priors on the power parameter.

Chen and Ibrahim [9] discuss relationships between power priors (with a fixed or estimated power) and the hierarchical priors that are covered in the next section. Consult the reference for case examples of the use of power priors in practice.

2.5. Hierarchical models

In a hierarchical model, one places an explicit distribution across the true response rates in the different studies and estimates those across group parameters to facilitate the borrowing of information. Generally, let p_0 be the true control rate in the current study, and let p_1, p_2, …, p_H be the true response rates in the H historical studies (H may be 1). Define γ_0, γ_1, …, γ_H to be the logits of the true control rates ($\text{logit}(p) = \log(p/(1 - p))$) and further assume

$$\gamma_0, \gamma_1, \ldots, \gamma_H \sim N(\mu, \tau^2)$$

Thus, μ and τ represent the between-study mean and standard deviation. The key parameter for borrowing is τ. For $\tau \approx 0$, it is extremely likely that all the γ values will be similar (and thus borrowing extensively would be appropriate), while for large τ, we may acquire quite different true control rates in the different studies (and thus minimal borrowing is appropriate).

Unfortunately, of course, we do not typically know τ (or μ). Thus, we place priors on both μ and τ, adding a second layer to the model and thus creating a hierarchical structure with

$$\mu \sim N(\mu_0, \tau_0) \text{ and } \tau^2 \sim \text{IGamma}(\alpha, \beta)$$

Note that the choice of IGamma can be controversial (see, e.g., [14]). Here, we consider 'informative' choices such as IGamma(1, β) that, by varying β, allow a mild to moderate degree of prior information on the scale of τ^2. We also consider the more controversial IGamma(ε, ε) prior. As noted in Gelman [14] and Lambert *et al.* [15], despite appearances this is not 'uninformative' as the choice of ε drastically changes the properties of the prior. That said, as can be seen in the graphs in Figure 8, the results are competitive with other methods.

We place a $N(0, 100)$ prior on μ (essentially non-informative on the logit scale), IGamma(1, β) priors for τ^2, with $\beta = 1$, 0.1, 0.01, and 0.001, and IGamma(ε, ε) priors on τ^2 with $\varepsilon = 0.1$, 0.01, and 0.001. Note that on the logit scale, standard deviations of 1 are fairly broad, so all mass above $\tau = 1$ represents a small amount of borrowing. Thus, we compare priors with differing scales for smaller τ.

For this analysis, we assume that the control data is $Y_0 \sim \text{Bin}(n = 200, p_0)$, and the treatment data is $Y_T \sim \text{Bin}(n = 200, p_T)$ where $\text{logit}(p_T) = \gamma_0 + \theta$. Thus, θ represents the log odds treatment effect, on which we place a non-informative $\theta \sim N(0, 100)$ prior. With this prior, the posterior distribution of the control arm is minimally dependent on the treatment data.

The posterior distribution can be acquired via standard MCMC techniques.

Our primary analysis of $H_0 : p_0 = p_T$ against $H_1 : p_0 < p_T$ is equivalent to testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Thus, we declare a primary analysis success if, after acquiring the posterior distribution of θ, we observe $\Pr(\theta > 0|\text{Data}) > 0.975$ (note that this threshold may be altered to change the type I error and power trade-off if desired).

The estimation of τ produces dynamic borrowing. Situations where the current and historical data are far apart produce a posterior distribution more heavily weighted toward large τ, while situations where the current and historical controls are similar produce a posterior distribution more heavily weighted toward small τ. Given the limited number of studies typically present in historical borrowing, these posterior distributions are still quite vague (e.g., they reflect the uncertainty in the prior), but the general shift produced by the data is sufficient to result in dynamic borrowing behavior.

Figure 7 illustrates the decisions made by hierarchical modeling for various priors on τ^2. Generally, we observe the behavior to produce an S-shaped interpolation between the separate and pooled analyses, approached the pooled analysis when the observed control data are similar to the historical data, and approaching the separate analysis as the current control data diverge from the observed historical rate. This prior is fairly informative in terms of the prior median being near strong borrowing and also has a long tail that allows larger τ to be considered.

Similar to Figures 1, 3, and 5, the background in Figure 7 indicates the trial decisions for separate analyses and the black outlines indicate the trial decisions from pooling. The blue dots here indicate the trial decision from hierarchical borrowing. As with the other methods, the hierarchical model produces borrowing between separate analyses and pooling. Generally, we acquire behavior similar to pooling when the current control data are close to the observed historical rate, and gradually move back toward separate analyses as the current control rate differs from the historical data.

Figure 8 shows the borrowing behavior, MSE, type I error, and power comparisons for the seven different choices of prior on τ. In the top panel, we see that borrowing with the hierarchical model is dynamic, with maximal borrowing occurring for a true control rate of 0.65 and decreasing as the true control rates vary from this value. Generally, the decrease is gradual, with some borrowing occurring even at control rates fairly distant from the historical 0.65. Note that for each prior, there is a cap on the expected amount of borrowing somewhere less than pooling. Thus, by choosing the prior distribution for τ, one can maximize the possible amount of weight given to the historical data. Note that with binomial likelihoods, effective sample sizes are somewhat approximate, and thus, the expected borrowing can be negative

In terms of type I error and power, one sees the (now) usual pattern of decreased power when the true control rate is much lower than 0.65, followed by a sweet spot of reduced type I error and increased power compared to separate analyses around 0.65, followed by inflated type I error and increased power for values much higher than 0.65. In Figure 8, the blue curves correspond to the IGamma(1, β) priors while the green curves correspond to the IGamma(ε, ε) priors. Note that the green IGamma(ε, ε) priors tend to borrow less and thus generally lack the inflated MSE and inflated type I error rate that occurs with the more informative IGamma(1, β) choices. The solid lines (green and blue) show choices of β = 0.01 and ε = 0.001 where the type I error power trade-off would appear to be heavily in favor of the ε = 0.001 choice. Similar to the curves for power priors, the type I error and power curves are ‘flatter’ than what we see for the other methods such as pooling or test-then-pool, with the sweet spot extending farther below 0.65 and the type I error inflation, while certainly occurring at the lower rate. Note that by changing the threshold for trial success (e.g., by changing $\Pr(\theta > 0|\text{Data}) > 0.975$ to $\Pr(\theta > 0|\text{Data}) > 0.98$), one can move the sweet spot to some degree at the cost of a small amount of power. As with other methods, by changing the prior on τ, one can achieve goals for minimizing the amount of type I error inflation, again potentially at the cost of some of the power gains.

Bayesian hierarchical models have been used extensively in the literature. Some texts to consult for references include [16] and [17].

2.6. Comparison of methods

The previous sections have focused on each method (single arm trials, test-then-pool, power priors, and hierarchical modeling), demonstrating how user controlled parameters can affect the borrowing behavior of each method. Here, we make a brief comparison of the methods against each other.

As noted in the previous sections, by selecting appropriate parameters, one can essentially select the degree of type I error inflation. We have chosen to compare test-then-pool with size 0.10 on the test of equality, downweighting with 40% weight, the hierarchical model with τ^2 ~ IGamma(1, 0.01), and the hierarchical model with τ^2 ~ IGamma(0.001, 0.001). The type I error and power of these three choices are shown in Figure 9 (test-then-pool in purple, power prior in blue, IGamma(1, 0.01) hierarchical model in dashed green, and IGamma(0.001, 0.001) in solid green). These were chosen based on their maximal amount of type I error inflation (note that the IGamma(ε, ε) priors did not reach a particularly high amount of type I error inflation). Test-then-pool reaches its cap on type I error earlier and also tends to have lower power than the power priors/hierarchical models over much of the parameter space shown. Generally, the particular choices shown have common values of type I error and power for true control proportions immediately near 0.65, although they diverge as the true control proportion diverges from 0.65. In particular, test-then-pool has more dramatic type I error inflation followed by an equally dramatic reduction in type I error. The IGamma(0.001, 0.001) tends to have milder type I error inflation over much of the range considered, at some moderate cost to power compared to the other methods (although generally the IGamma(0.001, 0.001) outperforms test-then-pool).

Figure 9 of course does not show the entirety of the current control parameter space. For current control rates above 0.8, the power priors with 40% downweighting have increasing type I error, while the hierarchical model begins to descend because of the dynamic borrowing. Given the historical data, it is unclear whether values above 0.8 are likely, but if they are viewed as plausible, then this range should of course be considered.

3. TYPE I ERROR INFLATION AND CONDITIONAL/UNCONDITIONAL CALCULATIONS

The fundamental concern of borrowing in drug development is the inflation of type I error, which presents a difficulty for using borrowing in phase III (confirmatory) trials. We noted earlier that part of this difficulty involves the definition of type I error inflation in that typically type I error is computed conditional on both the current control rate and the historical data. By viewing the entire space of current control rates, it is always possible to find a current control rate far from the observed historical rate and thus acquire type I error inflation from borrowing.

This places us in the following situation, returning to the basketball analogy from the introduction. Suppose a coach has a new player (novel treatment) and wants to determine if the new player is better than an existing player (current control) in game situations (one could acquire many data at practice, a weakness in the analogy we are ignoring here). The coach has 2 weeks to make a decision (this may involve as few as 25 observations/'free throws'). The coach knows the existing player has a 0.650 free throw percentage for the season to date (assume a large number of observations). If the coach argues that he will use the 0.650 free throw percentage in making a decision, the existing player can legitimately argue 'but coach, suppose all my practice has paid off and I'm now shooting 90%. Using my results earlier this season will bias your estimates and increase your type I error (if the new player is no better than me, I'll be more likely to lose my job by chance to the new player than I should be)'. This is quite true, and a key issue for the coach is to determine how likely it is that the practice has paid off to that degree (players may be known not to change that much, or equivalently very few players have that high a free throw percentage). The simple existence of type I error inflation may not be the issue as much as its likelihood and degree.

Returning to a clinical trials application, a rigid 'no type I error inflation at all is acceptable' approach can create some clearly irrational decisions. Suppose (possibly impractically) that one were able to guarantee that $p_C = p_H$ (the true rate for the current control rate is identical to the true rate for the historical data). Of course, the data are still prone to sampling error, but the parameters are identical. Here, the correct decision must be to pool the historical data.

However, all the arguments about type I error inflation remain. Suppose we observed 65/100 responses in the historical dataset, while also knowing $p_C = p_H$. It is possible that $p_C = p_H = 0.8$ and we by chance observed a low observed historical rate. Conditional on $p_C = 0.8$ and the historical data, pooling inflates type I error (as shown in Figure 2, the type I error inflation is quite large). If we did not allow any possibility of type I error inflation conditional on the historical data, we would not be allowed to pool the historical data. But we know that $p_C = p_H$, and thus, we should pool. Essentially, values outside the sweet spot arise from the possibility that the historical data may lead us in the wrong direction. If we cannot allow any possibility that the historical data are misleading, then we are forbidden from using any data in the past, and essentially 'the only good data are unseen data'.

In practice, we are never granted such a powerful assumption. We are faced with the uncertainty of drift in the control population, the chance that our literature search is not 'on point' or suffers from publication bias, and many other factors. To properly evaluate the possibility of borrowing, one must consider the possible likelihood of such drift combined with the straightforward possibility of sampling error in the historical data and assess whether the benefits of increased power and reduced type I error are more likely to occur than the pitfalls of reduced power for some values of p_0 and type I error inflation for others.

As noted earlier, these calculations change if one prospectively intends to use the data from one study in another study. One must still account for the possibility of drift, but prior to observing the data from the first study, the agreement between the historical and current control data is random and thus part of the type I error calculations. This is not the most common situation, but it also deserves investigation as such prospective designs have the possibility of further limiting the possible type I error inflation.

4. EXTENSIONS FROM THE LITERATURE

There have been many extensions to the models demonstrated thus far. For example, exchangeability between historical studies and current studies in the hierarchical model may be a concern. Neelon and O'Malley [13] propose an 'exchangeable power prior' to limit the impact of historical data in a hierarchical model. Consequently, even if the between-study variance parameter indicates borrowing, the power parameter will temper the borrowing by a fixed amount. The same authors also warn about too much attenuation of historical data when a power prior is used with a random power parameter (unless a very informative prior is used for the power parameter). *Commensurate* power priors [12] were proposed as an alternative. The commensurate prior adjusts the power parameter prior conditionally through a measure of the degree to which the historical and current data are commensurate, analogous to a measure of bias between the historical and current controls [1]. Thus, the borrowing obtained via the power parameter can be adjusted based on commensurability, and attenuation occurs when it is appropriate.

Furthermore, several authors have expanded models for incorporating historical controls. Extensions include the addition of covariates (e.g. [18]), multiple historical controls or multiple clinical sites [3], and adaptive randomization using historical controls [19]. Bayesian design including sample size determination is discussed in De Santis [20] and Chen *et al.* [21]. In this section, we discuss some useful extensions in the literature for incorporating historical controls in Bayesian trials.

4.1. Multiple historical controls

Multiple historical controls entail multiple power parameters in the power prior model (see Ibrahim and Chen [4]) and a more complicated relationship between the commensurability parameter and the between-study variance from the hierarchical model [3]. For the power prior example in Section 2.4, if the historical controls are assumed independent of one another (conditional on the control proportion), multiple historical controls would allow for multiple weights, one for each historical control. For hierarchical modeling, multiple historical controls make the estimation of the between-study precision more reliable (τ in Section 2.6). As such, the extent of borrowing as well as pre-posterior risk (i.e., the expected posterior squared error loss prior to observing data) is not overly sensitive to the forms of hyperpriors (see [15] and [3]).

Chen *et al.* [21] investigate Bayesian design and analysis of a non-inferiority trial and compare hierarchical priors with power priors when both incorporate multiple historical controls. For the power prior, they study both fixed and random power parameters. In their example, they study a 12-month target lesion failure (TLF, a binary endpoint) in a new generation drug-eluting stent compared to a historical control composed of two studies on previous generation drug-eluting stents. TLF was modeled using a binomial likelihood with beta priors on the TLF rates. Chen *et al*. computed operating characteristics when a future study had a control TLF rate that was the pooled average of the two historical rates and when the device TLF rate was either equal to the control (power) or equal to the control plus non-inferiority margin. When the between-study precision for the hierarchical prior had a Gamma(0.001, 0.001) prior, the power weights all had Beta(1, 1) priors or were all fixed at 0.03, and power was relatively similar across priors, even as sample size increased. When the between-study precision had a Gamma(0.01, 0.01) hyperprior, power was somewhat lower, and type I error rate was somewhat higher than for the other priors. The type I error rate for both fixed and random power priors was somewhat lower than for the other priors.

4.2. Adaptive allocation

When historical controls are used within a randomized controlled trial, the realized randomization ratio of treatment to (current) control may be much higher than 1:1 because historical control information can be used in place of current control subjects so that fewer current controls need to be randomized. Such ‘information-balanced’ randomization has been researched by Hobbs *et al.* [19]. Here, the randomization ratio adapts as a function of the relative informativeness of the historical control data for evaluating the endpoints. Their method requires interim assessment of relative informativeness of the historical control(s) using an expression for the interim effective sample size in the control. To the extent that the treatment effect from the already enrolled current control subjects is similar to the effect estimate from the historical control(s), the higher the effective sample size in the control group, and the fewer new control subjects need to be randomized. While Hobbs *et al*. are not the first to suggest adapting the randomization ratio, other authors have mostly focused on using response-adaptive or outcome-adaptive randomization (see [2]). Hobbs *et al*. illustrate their method using data from a colon cancer trial, with prior data coming from a previous trial. The trial begins non-adaptively, with 1:1 randomization until a certain number of events occurs; then, at each of a specified number of interim looks, the allocation ratio adapts to the information obtained from the historical controls.

Several authors show that the effective sample size could be estimated a priori to design a clinical trial (see [5, 20, 22, 23]). Here, if the historical control data are worth n^* prior control subjects, then under a 1:1 randomization ratio to treatment and control with n subjects per group, the trial would need only n–n^* control subjects. Naturally, the lower the between-trial heterogeneity among the current and historical controls, the higher the prior effective sample size and the lower the mean-squared error of the control estimate [5].

4.3. Role of covariates

Study-level and/or patient-level covariates may be important in determining the relevance of historical data and could be accounted for in the modeling. However, to this point, the use of covariates in formal borrowing of historical data has seemingly limited investigation/ discussion in the literature, revealing a potential research gap. O’Malley *et al.* [18] illustrated sample-size methodology for a Bayesian trial that used historical control information with relevant covariates that influence the treatment effect. More recently, Hobbs *et al.* [3] developed their commensurate prior approach using a general linear regression model format, which is also adapted for other types of linear models, such as generalized linear models, mixed effects models, and failure time models. Hobbs *et al.* [3] assume that $p - 1$ identical fixed covariates of interest are measured in both the current and historical trials. Commensurability then depends on similarity in the intercepts as well as the covariate effects (and, in the case of accelerated failure time models, the scales too).

Often, modeling covariates can help the assumption of exchangeability of the current and historical controls. Partial exchangeability [24] is applicable when the control groups are exchangeable after accounting for baseline covariates. Thus, if the control populations differ in a measurable way, exchangeability may still be viable. As a simple illustration, consider a new weight loss device to be compared with a control device, where some control patients may be obtained from one or more historical studies. Suppose the patients from the historical studies had a higher baseline weight on average than the currently enrolled patients. In order to borrow appropriately from the historical controls, patients' initial weight should be accounted for in the model so that outcomes are exchangeable after conditioning on baseline weight. For a numerical illustration, see Pennello and Thompson [25].

In certain situations, calibrating the current and historical controls is imperative for making correct inferences. Pennello and Thompson [25] consider a randomized controlled non-inferiority study of a complication rate. A hierarchical model is employed to borrow strength from historical controls to help estimate the concurrent control rate in order to reduce the sample size of the concurrent control. If the current control subjects are less healthy than the historical control subjects, they are more likely to have complications. Consequently, borrowing strength from the historical controls without proper adjustment for covariates that measure health status will bias downward the concurrent control complication rate, making it more difficult to demonstrate non-inferiority of the new device. An analogous situation could describe an anti-conservative bias where non-inferiority (and superiority) is easier to demonstrate because the historical subjects are less healthy than the current control subjects.

Baseline (and time-varying) covariates could also play an imperative role in borrowing strength across different populations, where one population is limited or scarce. A timely example is where an indication is sought in a pediatric population for a medical device that already has an adult indication. Parents may be hesitant to enroll their children in a study where there is a possibility of acquiring the control treatment. If adult studies are determined to be appropriate for borrowing strength in the control group, then a necessary task to determining the extent of borrowing is to identify covariates that make the adult and pediatric populations different in the effectiveness of the control treatment.

Covariates are also included in power prior models, as the power prior is simply the historical likelihood raised to a power [4]. Coefficients associated with the covariates in the model may or may not be the same for the historical likelihood as the current likelihood.

5. DISCUSSION

Our central goal has been to illustrate the main practical issues in drug development utilizing historical borrowing for the current control arm (presuming little to nothing can be assumed about the test treatment). At its best, borrowing good information from past studies allows for reduced type I error and power in a current study, which may either be used at face value or translated into a smaller sample size for the current trial and/or unequal randomization.

Generally, borrowing is dominant (reduced type I error and higher power) when the current control rate is close to the historical observed rate. This is intuitive as we are borrowing information nearly identical to the true current value. As the true current control rates diverge from the observed historical data, we can acquire reduced power (in one direction) and inflated type I error (in the other direction). Assessing the magnitude and relative likelihood of these costs in comparison to the possible benefits is the key issue in determining whether historical borrowing is appropriate in any given setting.

We have compared several possible borrowing mechanisms. We note that pooling produces good results in the region very near the historical observed rate but can suffer from greatly inflated type I error (while also showing that commonly used single arm trials can actually inflate type I error more than any proposed borrowing mechanism). Alternative methods such as test-then-pool, power priors, and hierarchical models can produce lower type I error inflation, and in fact, each of these methods has user-settable parameters that allow the user to cap the amount of type I error inflation in a specific range of control rates.

Finally, we note that a more prospective viewpoint of historical borrowing, always viewing today's trial as information for future trials, can create a different calculation for computing type I error, as the discrepancy between the historical and control datasets then has an explicit distribution. This is particularly relevant for seamless phase II/III trials, which were not considered here.

With the current environment demanding more efficient clinical trials, proper balancing of the risks and benefits of historical borrowing has the potential to further streamline the development of drugs and medical devices.

REFERENCES

1. Pocock SJ. The combination of randomized and historical controls in clinical trials. Journal of Chronic Diseases. 1976; 29:175–188. [PubMed: 770493]
2. Berry, SJ.; Carlin, BP.; Lee, JJ.; Mueller, P. Bayesian Adaptive Methods for Clinical Trials. Boca Raton, FL: CRC Press; 2011.
3. Hobbs BP, Carlin BP, Sargent DJ. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. Bayesian Analysis. 2012; 7(2):1–36.
4. Ibrahim JG, Chen MH. Power prior distributions for regression models. Statistical Science. 2000; 15:46–60.
5. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter D. Summarizing historical information on controls in clinical trials. Clinical Trials. 2010; 7:5–18. [PubMed: 20156954]
6. Spiegelhalter, DJ.; Abrams, KR.; Myles, JP. Bayesian Approaches to Clinical Trials and Health-care Evaluation. Chichester, UK: John Wiley & Sons Ltd; 2004.
7. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Statistical Science. 1999; 14:382–401.
8. Chen M-H, Ibrahim JG, Shao Q-M. Power prior distributions for generalized linear models. Journal of Statistical Planning and Inference. 2000; 84:121–137.
9. Chen MH, Ibrahim JG. The relationship between the power prior and hierarchical models. Bayesian Analysis. 2006; 1:551–574.
10. Neuenschwander B, Branson M, Spiegelhalter D. A note on the power prior. Statistics in Medicine. 2009; 28:3562–3566. [PubMed: 19735071]
11. Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. Environmetrics. 2006; 17:95–106.
12. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics. 2011; 67(3):1047–1056. [PubMed: 21361892]
13. Neelon B, O'Malley J. Bayesian analysis using power priors with application to pediatric quality of care. Journal of Biometrics and Biostatistics. 2010; 1(1):1–9.
14. Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis. 2006; 1:515–533.
15. Lambert P, Sutton A, Burton P, Abrams K, Jones D. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Statistics in Medicine. 2005; 24:2401–2428. [PubMed: 16015676]
16. Gelman, A.; Carlin, J.; Rubin, D.; Stern, H. Bayesian Data Analysis. Boca Raton, FL: CRC Press; 2003.
17. Carlin, B.; Louis, T. Bayesian methods for data analysis. Boca Raton, FL: CRC Press; 2009.

18. O'Malley J, Normand S, Kuntz R. Sample size calculation for a historically controlled clinical trial with adjustment for covariates. Journal of Biopharmaceutical Statistics. 2002; 12(2):227–247. [PubMed: 12413242]

19. Hobbs BP, Carlin BP, Sargent DJ. Adaptive adjustment of the randomization ratio using historical control data. Clinical Trials. 2013; 10(3):430–440. [PubMed: 23690095]

20. DeSantis F. Using historical data for Bayesian sample size determination. Journal of the Royal Statistical Society Series A. 2007; 170:95–113.

21. Chen M, Ibrahim J, Lam P, Yu A, Zhang Y. Bayesian design of non-inferiority trials for medical devices using historical data. Biometrics. 2011; 67:1163–1170. [PubMed: 21361889]

22. Morita S, Thall P, Muller P. Determining the effective sample size of a parametric prior. Biometrics. 2008; 64:595–602. [PubMed: 17764481]

23. Pennello G, Thompson L. Design considerations for Bayesian medical device studies. Journal of Biopharmaceutical Statistics. 2013 submitted to.

24. Bernando, J.; Smith, A. Bayesian theory. Chichester, UK: John Wiley and Sons; 2000.

25. Pennello G, Thompson L. Experience with reviewing medical device clinical trials. Journal of Biopharmaceutical Statistics. 2008; 18(1):81–115. [PubMed: 18161543]

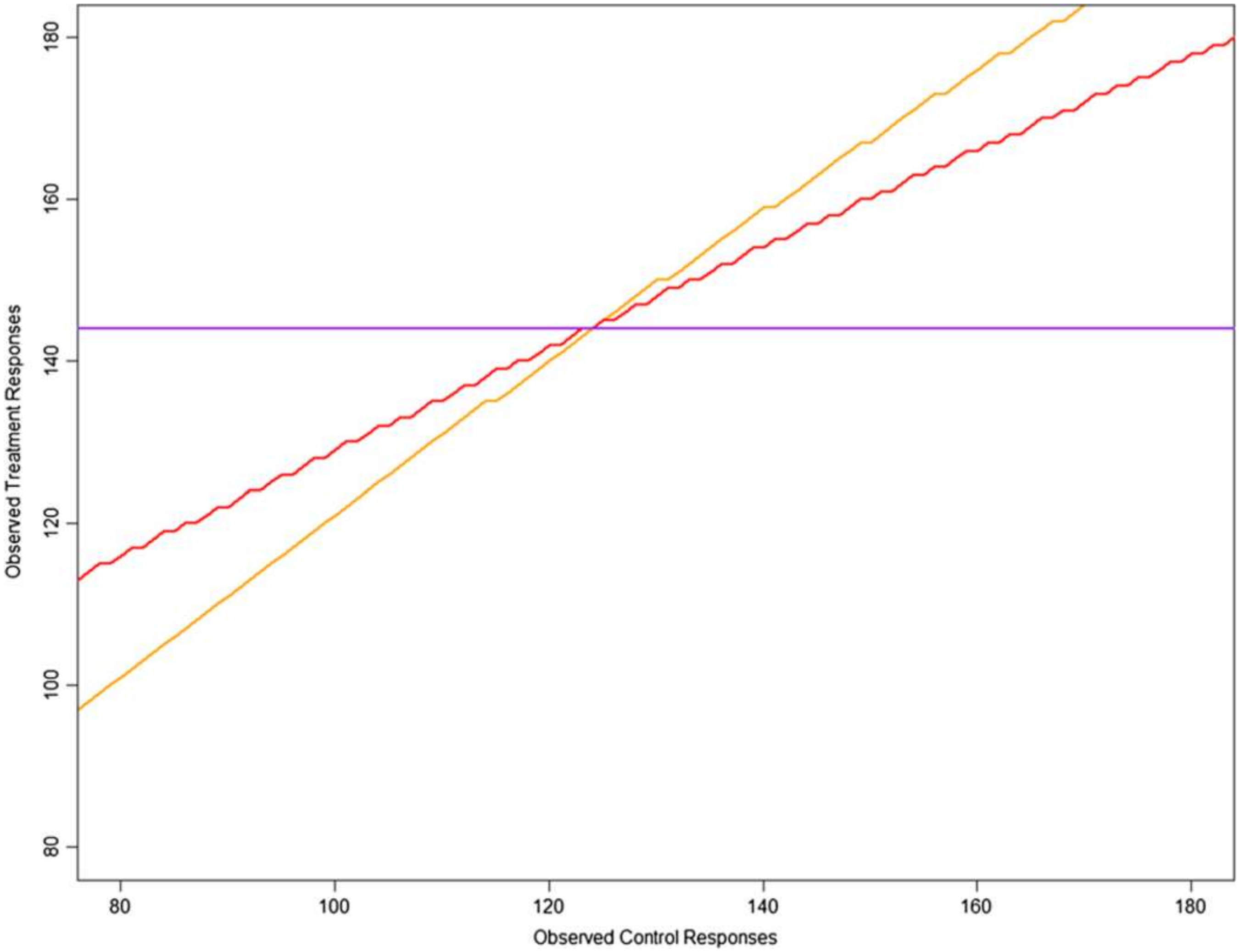

Figure 1.

Conclusions reached by separate, pooled, and single arm trials. The X (control) and Y (treatment) axes show the possible values of the observed data, while the three curves show the decision boundaries for the separate (orange), pooled (red), and single arm (purple) trials. Note that in a single arm trial, control data are not collected, and hence, the decision is based on the treatment data alone.

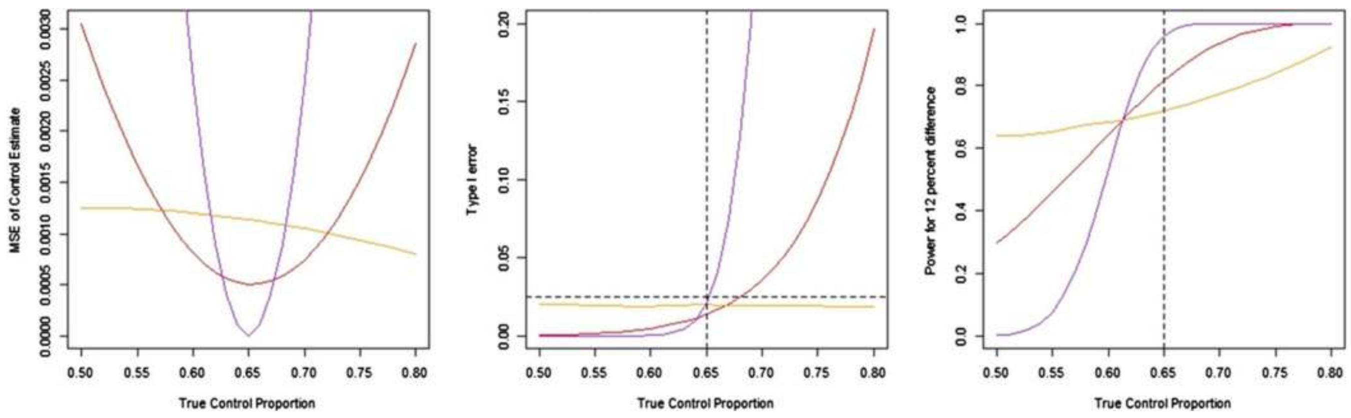

Figure 2.

Comparison of the mean square error (MSE) (left), type I error (middle), and power (right) for separate (orange), pooled (red), and single arm trial (purple) designs. Generally, there is a 'sweet spot' near 0.65 where borrowing simultaneously achieves lower MSE, lower type I error, and higher power compared to the separate analysis. Below the sweet spot, we see diminished power with borrowing, and above the sweet spot, we see inflated type I error. Assessing the relative likelihood of these regions is important to assessing the costs and benefits of borrowing.

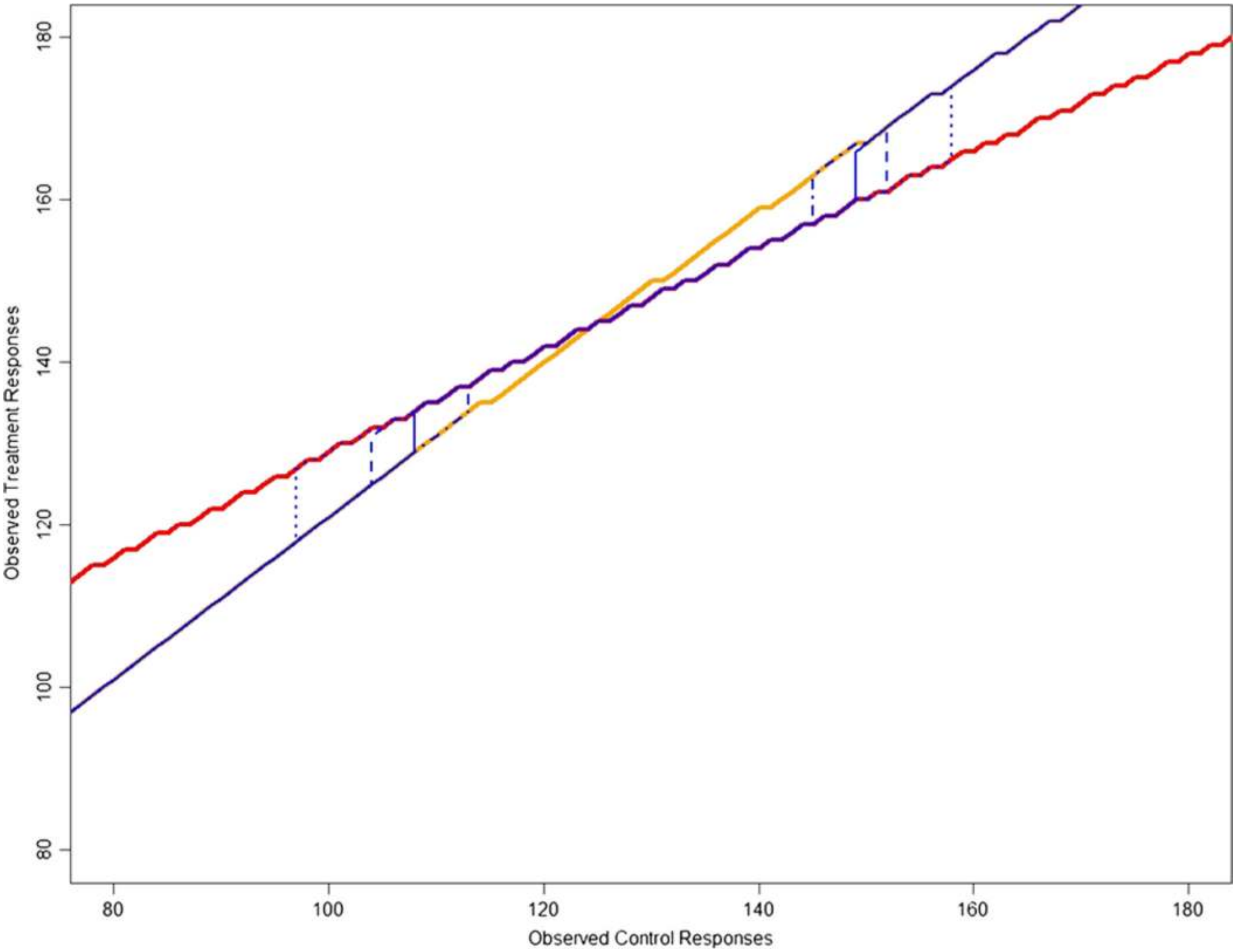

Figure 3.

Conclusions drawn by separate, pooled, and test-then-pool (using sizes 0.20, 0.10, 0.05, and 0.01 for the test of equality between current data and historical data) analyses. The curves indicate the decision boundaries for each design, showing separate (orange), pooled (red), and test-then-pool (blue). Results above the curves are successful trials. Note that small sizes for the test of equality produce the greatest overlap between test-then-pool and the pooled analysis. Thus, the 0.01 size test of equality (the dotted line) has the greatest overlap with pooling. For control values between 109 and 149 responses, test-then-pool (at size 0.10) chooses the pooled analysis, while outside this region, the test-then-pool approach emulates a separate analysis.

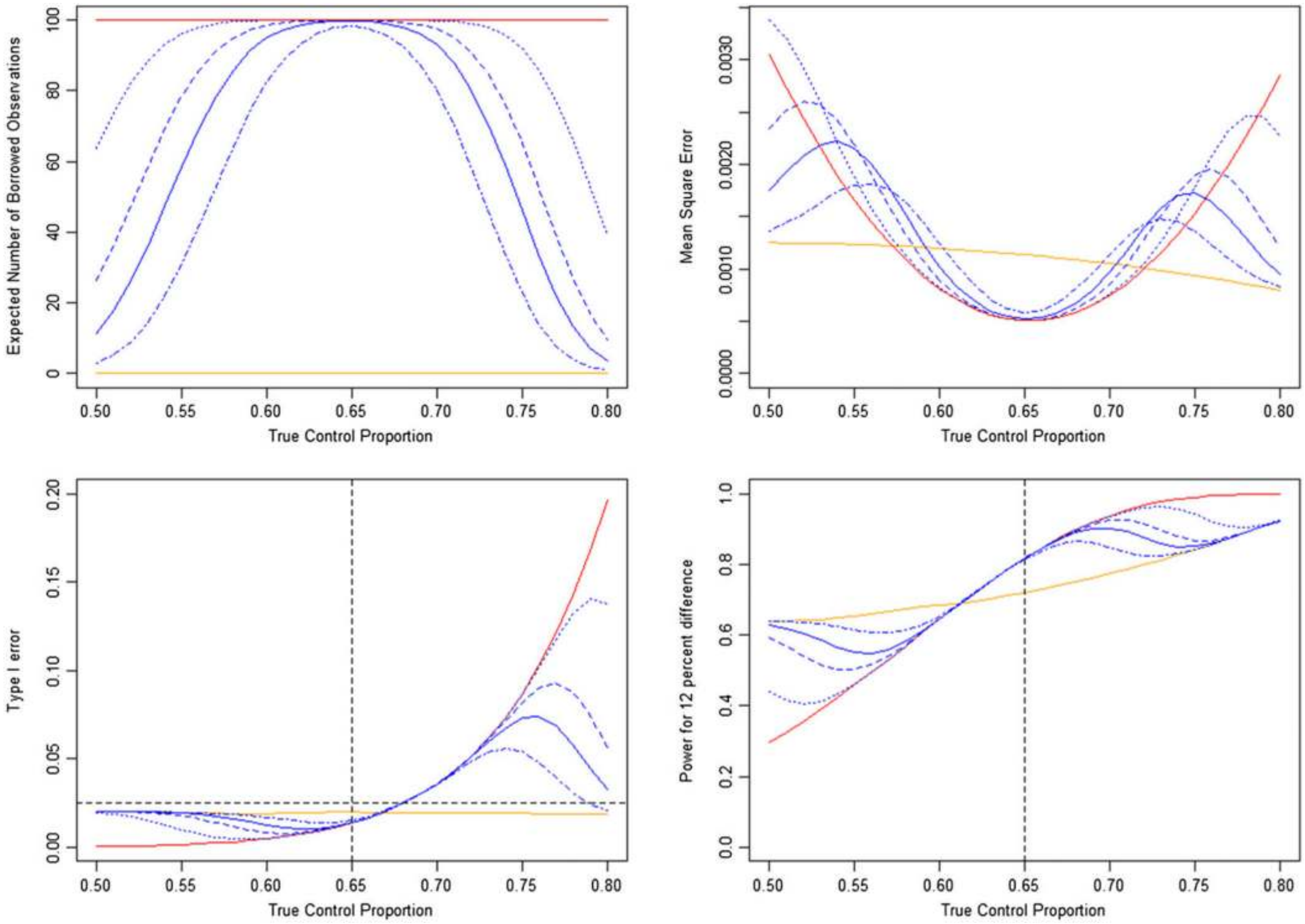

Figure 4.

Comparison of borrowing, mean square error, type I error, and power for test-then-pool. The red curves indicate pooled analyses, the orange curves separate analyses, and the blue curves test-then-pool analyses with sizes of 0.20, 0.10, 0.05, and 0.01 for the test of equality (the primary analysis for testing the novel treatment still uses size 0.025). Test-then-pool incorporates dynamic borrowing (the model borrows less as the historical and current control rates diverge). This caps the amount of type I error inflation. In addition, by changing the size of the test, one can construct a continuum of procedures that can achieve any particular goal for type I error.

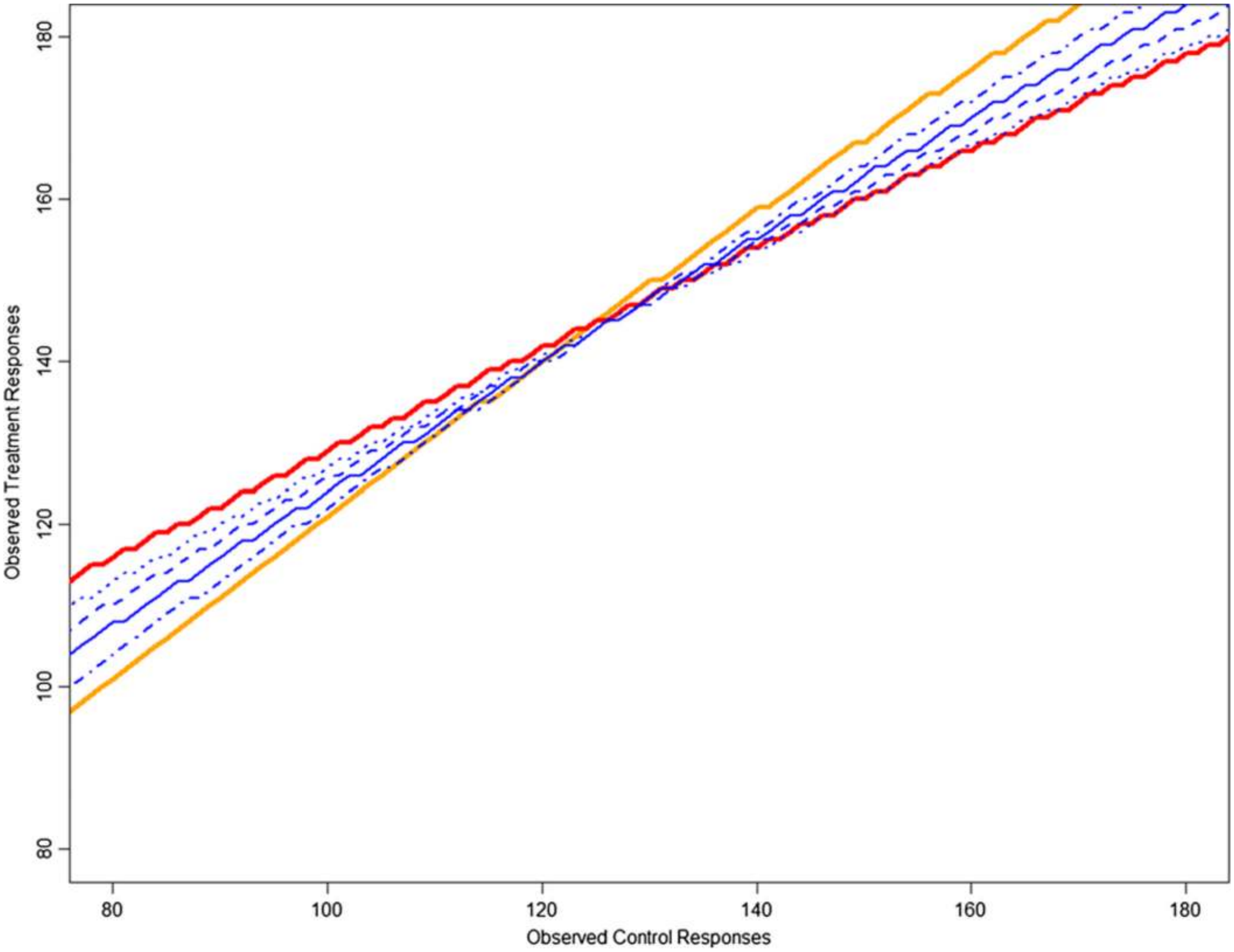

Figure 5.

Decisions made by downweighting using a power prior. Similar to Figures 1 and 3, the curves indicate the decision from a separate analysis (orange), pooled analysis (red), or a 20%, 40%, 60%, or 80% downweighting (dot dashed, solid, dashed, and dotted lines). Data above a curve result in trial success for that design. Downweighting essentially acts proportionally between the separate and pooled analyses.

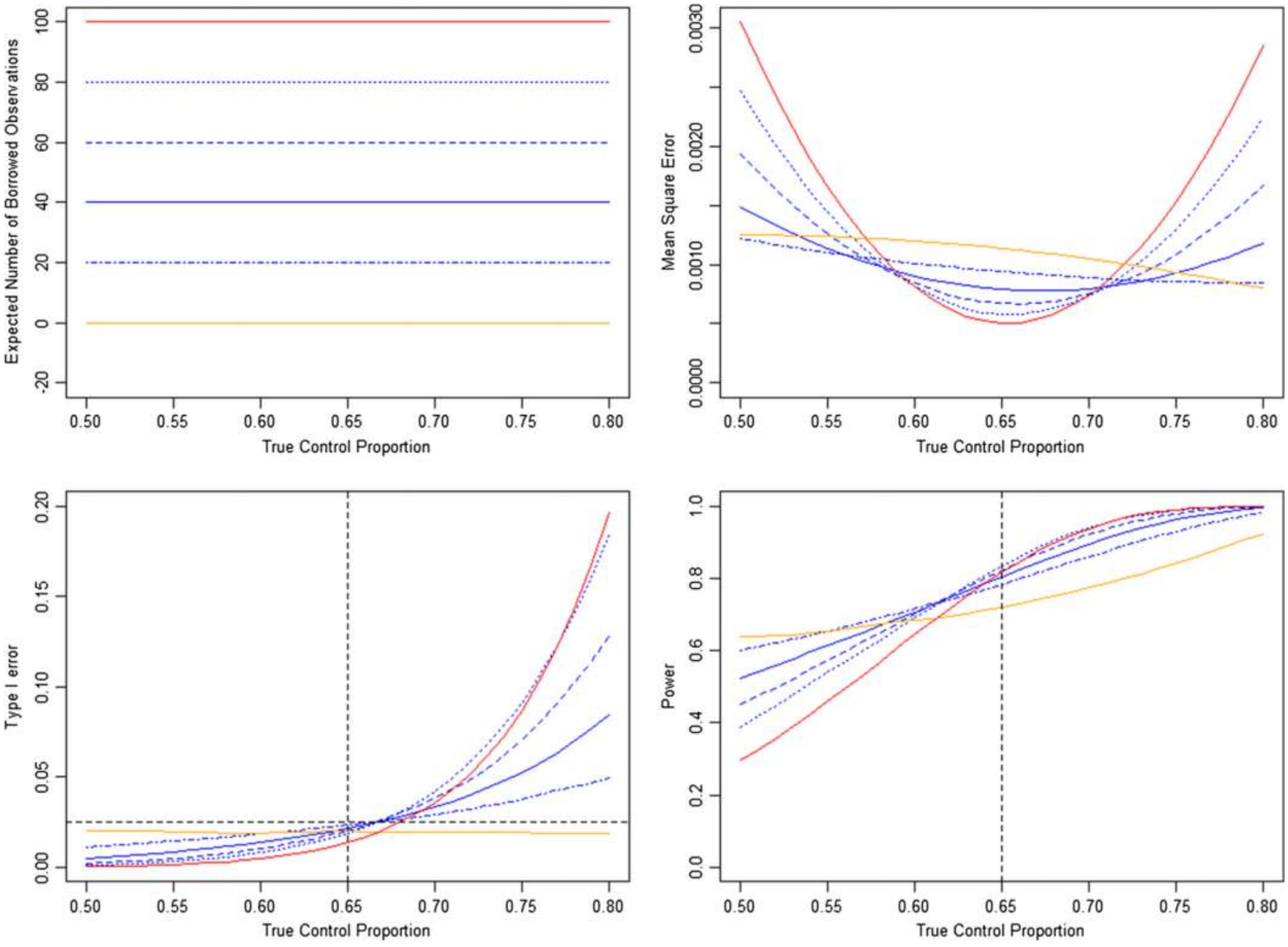

Figure 6.

Borrowing, mean square error type I error, and power comparison for downweighting. The top panel provides the effective number of borrowed observations (here directly set by the weight parameter, one of 20%, 40%, 60%, or 80%). The bottom left panel shows the type I error as a function of the true control proportion, while the bottom right panel shows the power. Generally, the ‘sweet spot’ where borrowing dominates a separate analysis is longer for downweighting than pooling, and downweighting with low weights can limit the amount of type I error inflation for values near the historical rate of 0.65.

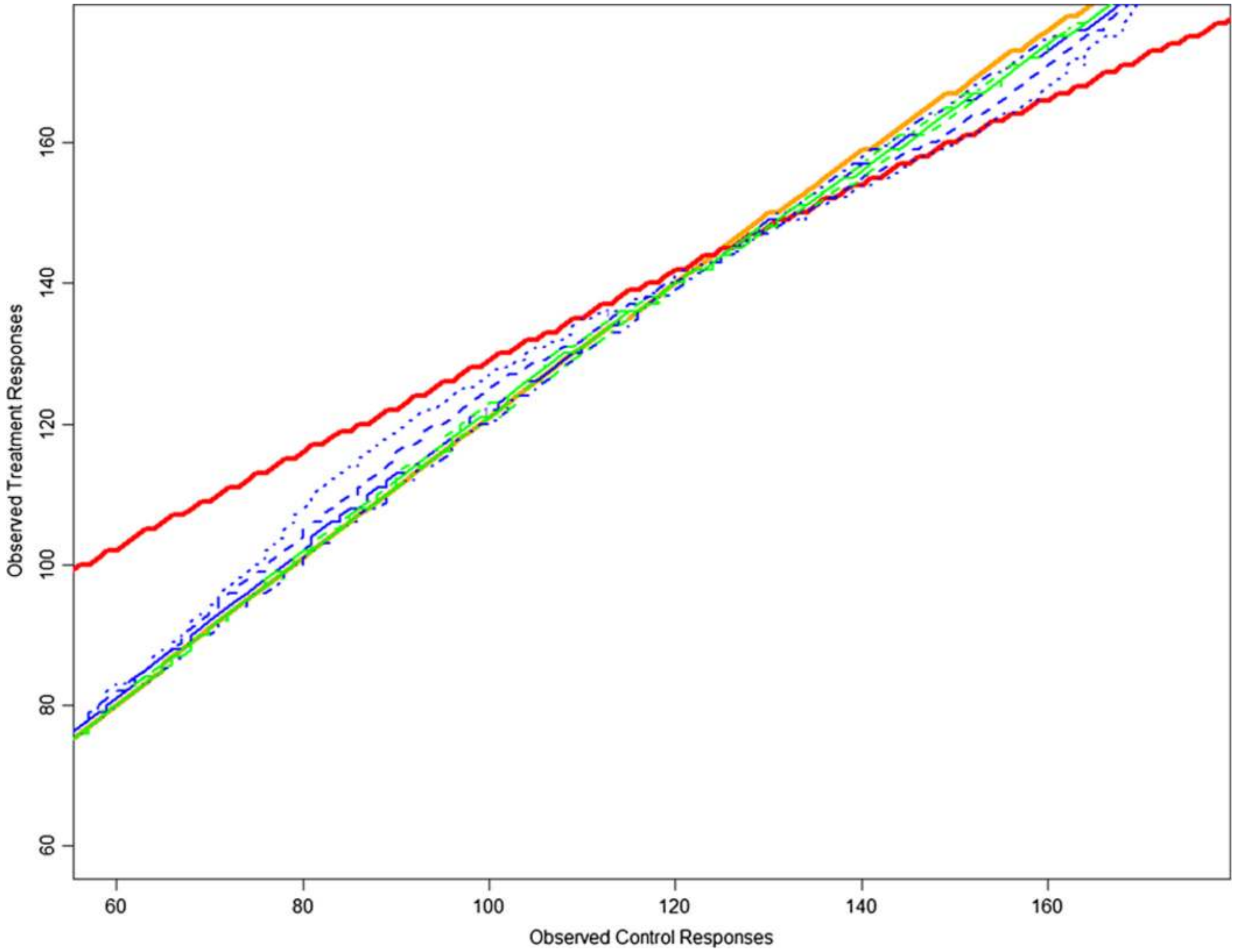

Figure 7.

Decisions made using hierarchical borrowing with different priors on τ^2. Hierarchical borrowing is dynamic, emulating pooling when the current controls agree with historical data and coming closer to separate analyses as the current control data diverge from the historical data, and interpolating between those extreme with an S-shape. The green curves (quite similar to a separate analysis represent the Gamma(ε, ε) priors, while the blue curves represent Gamma(1, ε) priors.

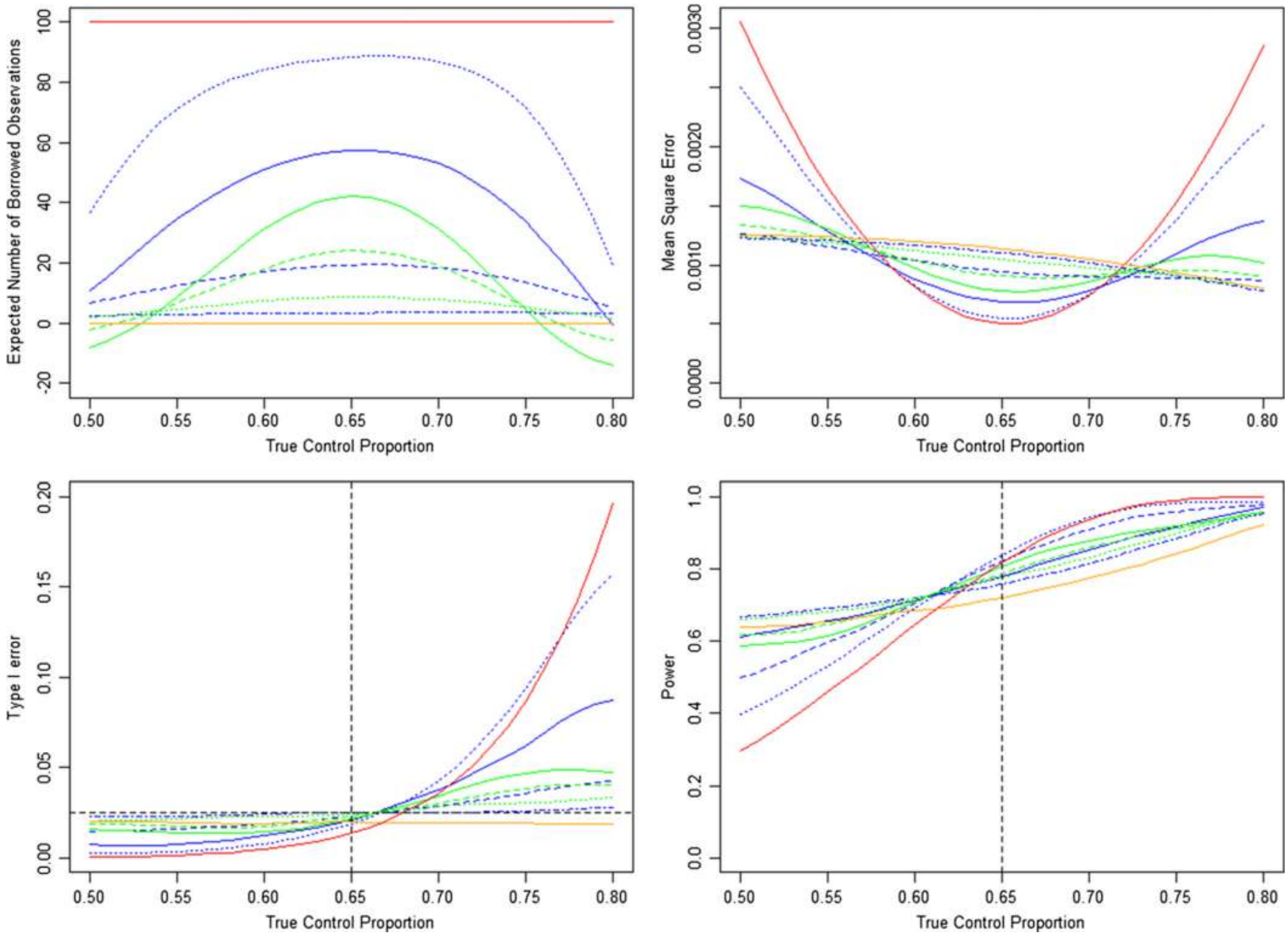

Figure 8.

Comparison of borrowing, mean square error, type I error, and power for hierarchical models. As before, orange line represents the separate analyses while the red represents the pooling analyses. Blue curves correspond to IGamma(1, β) priors on τ^2 while green curves correspond to IGamma(ε, ε) priors on τ^2. Borrowing behavior tends to be 'flatter' for hierarchical models, borrowing moderately over a long range, while still displaying dynamic borrowing (borrowing is reduced when the true control rate is far from the historical data). This moderate, long range borrowing is also reflected in the type I error inflation that has a lower slope than other methods (although it still does reach reasonably high values). Generally, the 'sweet spot' of improved type I error and higher power extends farther down (for values under 0.65) than other methods. Also note that the green IGamma(ε, ε) choices tend not to inflate type I error as much as the more informative IGamma(1, β) choices.

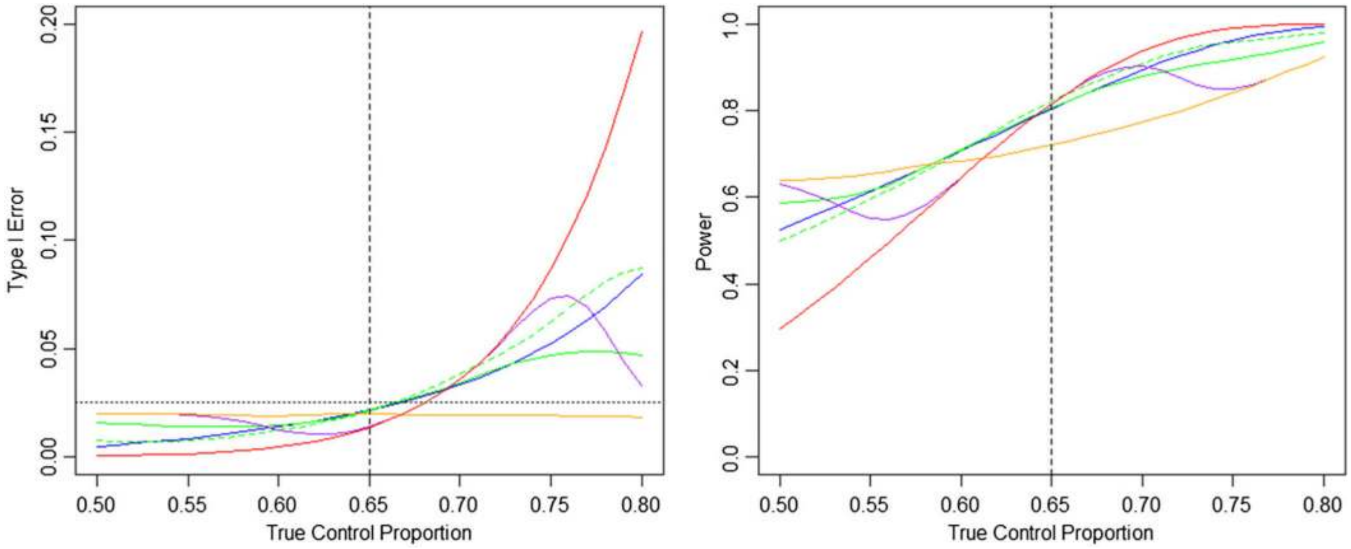

Figure 9.

Type I error and power comparison for separate (orange), pooling (red), selected test-then-pool (size 0.10, purple), downweighted power prior (40% weight, blue), and hierarchical model (IGamma(1, 0.01) in dashed green, and IGamma(0.001, 0.001) in solid green). Generally, the test-then-pool approach has lower type I error and also lower power near a control rate of 0.65, but has reduced power compared to power priors and hierarchical models outside that range. For control rates near 0.65, all methods achieve similar power gains as pooling (red) with much less type I error inflation.