Evaluation of False Discovery Rate in Platform Trials

Diana Chen[a*] , Nicole Li[a], Cong Chen[a] and Linda Sun[a]

[a] Biostatistics and Research Decision Sciences, Merck & Co., Inc., Kenilworth, NJ 07033, USA

[*] Corresponding author: MAILSTOP UG-1CD44, 351 North Sumneytown Pike, North Wales, PA 19454, USA, qiusheng.chen@merck.com, 267-305-3032

Abstract

Recent development of PD-1 immunotherapy in cancer drug development leads to the thinking of using an immune-cancer therapy as a backbone therapy and adding targeted therapies or chemotherapies as combination to boost the efficacy. Due to the many choices of potential combination therapies, platform studies will be utilized. A platform study can be conducted with or without a control arm. One may be interested in how many of the selected regimens would be active regimens and how many are falsely discovered. We will evaluate the false discovery rate (FDR) based on Storey's (2001) positive FDR definition under different parameter settings. We'll use the FDR as one criterion to evaluate the options of with and without a control arm in a platform study. Furthermore, we evaluate the impact on FDR for treatment effect size and the percentage of sample size in the shared control arm.

Keywords: Platform study; False discovery Rate; Share control arm; Immunotherapy

1 Introduction

Traditional clinical drug development typically evaluates one investigational therapy in a single disease in its own clinical trial. Due to the recent success of cancer immunotherapy and its mechanism of action, clinical society undergoes the theory that it is possible to use cancer immunotherapy as a primer to boost the immune system of cancer patients, and at the same time combine the cancer immunotherapy with either chemotherapies, targeted therapies or another type of cancer immunotherapy to boost the efficacy. Scientists identified many potential cancer drug candidates from pre-clinical research and there is a need to know if their combination with cancer immunotherapy would further increase the efficacy. During the efficacy screening proof-of-concept phase (Phase Ib/II), many combinations of interest are to be evaluated in the common tumor types, say, non-small-cell-lung-cancer (NSCLC). Setting up one clinical trial for each combination therapy is less efficient operationally. Due to site variation and other confounding factors, outcomes from different clinical trials are also hard to directly compare without adjustment.

Platform trial, one key type of master protocols, is to study multiple therapies in the context of a single disease in a perpetual manner, with therapies allowed to enter or leave the platform on the basis of a decision algorithm[1,2]. The advantages of conducting platform trials instead of multiple individual clincal trials for each therapy include but not limited to: streamlined and reduce the operational overhead cost; made comparison between therapies, though not part of the primary goals, feasible since now they are conducted in the same study with the same site, under the same set of inclusion/exclusion criteria. For an efficacy screening exploratory platform trial, the goal is to select active investigational therapies for phase III development. An exploratory platform trial can be without a comparator, similar to the single arm study using Simon's two-stage design, or it can include a comparator arm such as active control or placebo, and the efficacy Go/No-Go to phase III development will depend upon the comparison with the control arm. There have been quite some successful exploratory platform trials thus far. One example is I-SPY2[3,4], an exploratory platform trial designed to investigate new treatments for biomarker-identified subtypes of early-stage breast cancer in the neoadjuvant setting. Since a platform study is conducted in a perpetual manner, once it is set up, it is possible that it will evaluate many therapies over the years. For each therapy, a decision will be made whether it will be considered "active" for phase III development. Since these platform trials are exploratory in nature, a nominal type I error is usually used for each investigational therapy when deciding whether it is considered 'active' for further development, similar to what has been done in individual clinical trials. When multiple therapies are tested at a nominal type I error, one may wonder among the ones that are considered 'active' in the platform study, how many of them are truly active in the phase III studies. In this paper we will introduce the metric 'false positive rate' that measures among the declared 'positive/active' treatment arms, # of them are not truly active, in the context of platform trials.

Soric (1989) [5] first proposed the framework of "false discovery rate". A discovery is defined as one hypothesis test which is claimed significant. Instead of focusing on each individual hypothesis, we examine the portion of false discoveries when testing multiple hypotheses. Compared to type I error rate measurement such as family-wise type I error rate, false discovery rate (FDR) is concerned on the portion of true discoveries. Benjamini and Hochberg (1995)[6] formally introduced the concept of FDR and proposed an FDR controlling procedure. The FDR controlling procedure is designed to control the expected proportion of false rejections among all rejected null hypotheses and it is more commonly used when the interest is on the proportion of wrongly rejected

hypotheses among all the rejected hypotheses. Storey and Tibshirani [7] (2001) modified the definition of FDR to positive FDR which only calculate the false discovery rate conditional at least one hypothesis is rejected. Storey (2002)[8] proposed to directly estimate the FDR when following fixed the rejection region for the type I error control in each individual hypothesis testing. Throughout this paper, we adopt Storey's positive FDR definition for the FDR estimate. Furthermore, we will use the FDR as one criterion to compare the options of with or without a control arm in a platform trial.

In this paper, we discuss how to evaluate FDR in platform trials with multiple arms. The paper is organized as follows. In Section 2, we first present the definition of positive FDR given by Storey and Tibshirani (2001) [7], and then introduce the setup of an oncology platform trial. The FDR estimates of tests in a platform trial without a control arm (Option 1) and with a control arm (Option 2) are given in subsections 2.3 and 2.4, respectively. In Section 3, simulation studies are conducted to evaluate the FDR in a platform trial without a control arm (Option 1) and in a platform trial with a shared control (Option 2). At last, we discuss and provide conclusion and future work in Section 4.

2 Estimation of the FDR

2.1 Definition of FDR

Assume there are a total of K experimental arms (with or without a shared control arm) in a platform trial. Table 1 shows the possible outcomes when conducting *K* hypotheses testing, one for each experimental arm. We have the following notations: V is the number of false positive outcomes (i.e., non-active arms but declared 'active' in the platform trial) and R is the total number of tests that are rejected (i.e., declared 'active' in the platform trial).

Table 1: Possible Outcomes from K Hypotheses Tests

	Null hypothesis is true	Alternative hypothesis is true	Total
Rejected	V	S	R
Not rejected	U	T	$K-R$
Total	m_0	$K-m_0$	K

In Table 1, *K* is the known total number hypotheses tested, m_0 is the unknown number of true null hypotheses (i.e., non-active treatment arms), *(K -* m_0) is the number of true alternative hypotheses (i.e., active treatment arms), *R* is the number of hypotheses rejected, an observable random variable. Moreover, *U, V, S*, and *T* are unobservable random variables.

The traditional family-wise type I error (FWER) is $\Pr(V \geq 1)$. We use the positive false discovery rate introduced by Storey and Tibshirani (2001) [6] for our FDR definition in this paper , i.e., only consider FDR when at least one hypothesis is rejected:

$$FDR = E\left(\frac{V}{R}\middle| R > 0\right) \quad (1)$$

When K hypotheses are tested, denote the p-values of the K tests by $\{p_1, p_2, \ldots, p_K\}$. For a given threshold $0 < \alpha$ let
$V(\alpha)$ = # { false positive with $p_i \leq \alpha$ for $i = 1, \cdots,$
K when the Null hypothesis is true}
and
$R(\alpha)$= # $\{p_i \leq \alpha$ for $i = 1, \cdots, K\}$, the total number of hypotheses rejected.

When $R(\alpha) > 0$, Storey and Tibshirani (2001) suggested to estimate FDR by the following approximation:

$$FDR(\alpha) = E(V(\alpha)/R(\alpha)|R(\alpha) > 0) \approx E(V(\alpha))/E(R(\alpha)) \quad (2)$$

Storey and Tibshirani (2001) pointed out that a simple estimate of $E[R(\alpha)]$ is the observed $R(\alpha)$. In a simulation study when Null hypotheses are known, $E(V(\alpha))$ can be estimated by the number of false positives in the tests.

2.2 Setup of an Oncology Platform Trial

In this section, we describe the two options for the oncology platform trial setup. Option 1 does not include a control arm. All subjects would be equally randomized into one of the K investigational regimens without a control arm. Option 2 includes a shared control arm. All subjects would be equally randomized into either one of the *K* investigational regimens or a control arm that is shared among the *K* investigational regimens.

We assume two scenarios that we think to be realistic in conducting clinical trials. The first scenario is to mimic the reality where there is limited budget and resource in development. Suppose the total number of subjects in a platform trial is fixed at N. In option 1, the sample size is the integer part of (N/K) and for option 2 each experimental arm as well as control will have sample size as the integer part of (N/(K+1)). In the second scenario, we assume the sample size per arm is fixed at *n*. Suppose that in the platform trial a total number of *K* investigational regimens is of interest; the total sample size will be K*n in option 1 where a historical control arm is involved for each experimental arm's analysis. The total sample size will be (K+1)*n in option 2 where it *n* additional subjects are for the shared control arm.

For simplicity, we assume each regimen has probability p to be active and probability $(1 - p)$ to be inactive and we utilize equal randomization ratio across all treatment arms. For this efficacy exploratory platform trial, we use objective response rate (ORR) as our primary endpoint to evaluate efficacy. The response rate of an non-active treatment arm would be $ORR = \pi_0$ and the response rate of an active treatment would be $ORR = \pi_1 > \pi_0$. We denote the effect size as $\Delta = \pi_1 - \pi_0$.

We will discuss the FDR estimation and simulations for each Option using scenario 1 in Section 2.3 and 2.4. The calculation of FDR for scenario 2 is similar, with the sample size as the only difference.

2.3 Option 1: A Platform Trial without A Control Arm

In Option 1, we consider a platform trial with *K* arms without a control arm. Historical control from previous studies, published data, meta-data analyses, or real-world data would be used when making a decision on whether an experimental arm in the platform study is active or not. One concern of conducting single arm study and assume we know the null response rate π_0 is that, we may as well shoot off the target. Therefore, instead of assuming the null response rate is know as π_0, for the FDR calculation here, we assume that the true historical response rate is unknown and it follows a Beta distribution with mean π_0 and stand deviation SD. Even though we still perform hypotheses testing based compared with π_0 and a given significant level α and a fixed rejection region with critical value C such that C satisfies $P(X > C|\pi_{hc}) \leq \alpha$, due to the variation from the historical control, the expected type I error α_{E1} (Thomas Jemielita, Archie Tse, and Cong Chen (2018))[9] would be larger than α.

$$\alpha_{E1} = E[P(X > C|\pi_{hc})] = \int_0^1 P(X > C|\pi_{hc}) f(\pi_{hc})\, d\pi_{hc},$$

Where $f(\pi_{hc})$ is a Beta distribution with mean π_0 and standard deviation *SD*.

Naturally, when the variation of the Beta distribution is large, the expected type I error would be large, and vice versa.

The power $P(X > C|\pi_1)$ of each hypothesis testing can be calculated after *C* is attained based on π_{hc}. For example, for $n = 50,\ \pi_{hc} = \pi_0 = 0.2,\ \alpha = 0.03,\ \pi_1 = 0.4$ we have $C = 15$ and the power is $P(X > C|\pi_1) = 0.9045$.

In this study, performance of different platform trials will be compared by FDR values estimated empirically via statistical simulations. Suppose that the total sample size for the *K* arms is *N*. In this study we assume each arm has the same sample size *n*=integer part of (*N*/*K*). For Option 1, the simulation procedure consists of the following steps:

1. For a given mean π_0 and some standard deviation SD, the parameters a and b in the beta distribution Beta (a, b) are calculated by solving the equations where $\pi_0 = \frac{a}{a+b}$ and $SD^2 = \frac{ab}{(a+b)^2(a+b+1)}$;

2. A historical control response rate π_{hc} is chosen for the given beta distribution Beta (a, b);
   For each of the *K* hypotheses, the alternative hypothesis with $\pi_1$is randomly selected with the given probability *p* (=0.3 in our simulation), and the Null hypothesis with π_{hc} is selected with the probability of (*1-p*);

3. For a given sample size *n*, a random sample is generated from the binomial distribution with probability π_1 or π_{hc} for each of the *K* hypotheses.

4. The *K* hypotheses are tested by the one-sample test of proportions (using prop.test in R and $\alpha = 0.05$). Numbers of total hypotheses rejected R(α) and false positive V(α) are recorded;

5. If R(α)>0, $FDR = \mathrm{E}\left[\frac{\mathrm{V}}{\mathrm{R}}|R > 0\right]$ is estimated by V(α)/R(α).

The above five steps are repeated *M* times (we chose *M*=10,000 in our simulation). Average of the FDR values are used as the final estimate.

2.4. Option 2: A Platform Trial with a Shared Control Arm.

In the second option, a control arm would be shared for all *K* experimental arms. Subjects are equally randomized into either *K* experimental arms or the control arm. Each experimental arm would be compared with the shared control and be evaluated separately at the type I error of α using the Miettinen-Nurminen method[9].

Similar to the simulation procedure given for Option 1, FDR values for Option 2 are estimated by the following steps:

1. For each of the *K* hypotheses, the alternative hypothesis with $\pi_1$is randomly selected with the given probability *p* (=0.3 in our simulation), and the Null hypothesis with π_0 is selected with the probability of (*1-p*);

2. For each of the K hypotheses and a given sample size *n*, two random samples are generated from the binomial distributions with probability π_0 for the control group, and π_0 or π_1 for the treatment group, respectively;

3. The *K* hypotheses are tested by the two-sample test of proportions using the two-sample proportion test;
4. If R(α)>0, $FDR = \mathrm{E}\left[\frac{\mathrm{V}}{\mathrm{R}} | R > 0\right]$ is estimated by V(α)/R(α).

The above four steps are repeated *M* times (again we chose *M*=10,000 in our simulation). Average of the FDR values are used as the final estimate.

For scenario 2, all steps are the same except for changing the value of the sample size in each option .

3. Comparison of Option 1 vs Option 2 based on FDR estimates.

In this section we compare the FDR between Option 1 and Option 2 in each of the two scenarios. The two scenarios are introduced in section 2.2. For this purpose, we run the simulation on a grid with π_0 between 0.2 to 0.5 and π_1 between 0.3 and 0.6, with a 0.01increment. Figure 2 compares the estimated FDR in the simulated experimental design of option 1 (without a control arm) and option 2 (with a shared control arm) when the total sample size is 200 and the standard deviation in the beta distribution for the historical control is 0.05. We evaluate four different cases on number of experimental arms, i.e., $K = 2, 3, 4, 5$. Note that in Option 1, sample size in each experimental arm is the integer part of (N/K), and for Option 2 sample size in each experimental arm is the integer part of (N/$(K+1)$). Sample size in each arm is always smaller in Option 2 since some of the sample size has to be allocated to the control arm. By evaluating the ratio of FDR, $\frac{Option1\ FDR}{Option\ 2\ FDR}$, the smaller the FDR the better the option. We use blue color to indicate that the FDR ratio is over 1.1, i.e., Option 1 has larger estimated FDR than that in Option2; Red color indicates the FDR ratio is between 0.9 and 1.1, and green color indicates the FDR ratio is less than 0.9, i.e, Option 1 is preferred compared to Option 2. Results in Figure 2 show that in general, when the target effect size Δ is small (for example, Δ<0.1) , FDR is smaller in Option 1. When the target effect size Δ is large, FDR is smaller in Option 2. When number of arms increases, i.e., number of patients in each arm decreases, FDR may also be larger in Option 2 compared to Option 1.

Figure 3 provides the estimated ratio of FDR for the two options where *N*=200 and standard deviation of Beta distribution is 0.1. Compared to Figure 2, when the uncertainty of historical control increases, the advantage of Option 2 (with a shared control arm) also increases.

In addition to the impact of the total sample size and the uncertainty on the historical control estimates, number of experimental arms also plays a role in the FDR comparison. Since the more the experimental arms, the smaller sample size each arm will receive. The sample size impact is even more detrimental to Option 2. When *K*=4 or 5, Option 1 is almost always preferred in terms of FDR.

Figure 4 shows the results for scenario 2 when sample sizes are all the same in both two options. Similar to comparison results in scenario 1, when the target effect size Δ is small (for example, Δ<0.1) , FDR is smaller in Option 1. If the target effect size Δ is larger, FDR is smaller in Option 2.

Figure 2: Scenario 1 FDR Comparison of without control arm and with a control arm when N=200, SD=0.05

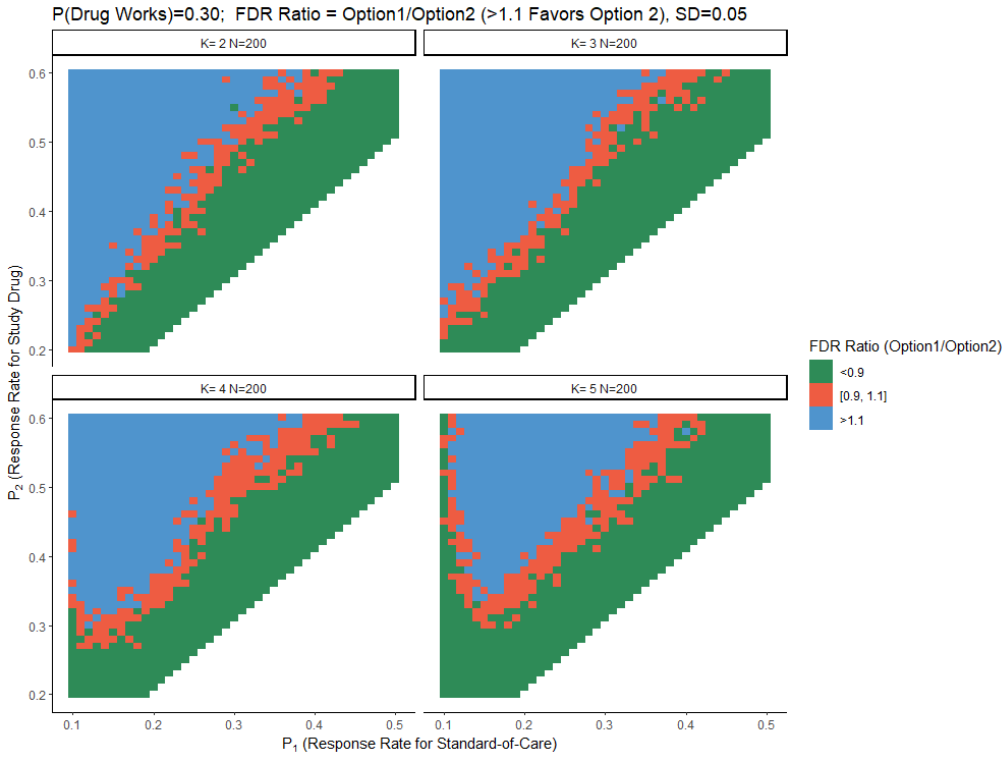


Figure 3: Scenario 1 FDR Comparison of without control arm and with a control arm when N=200, SD=0.1

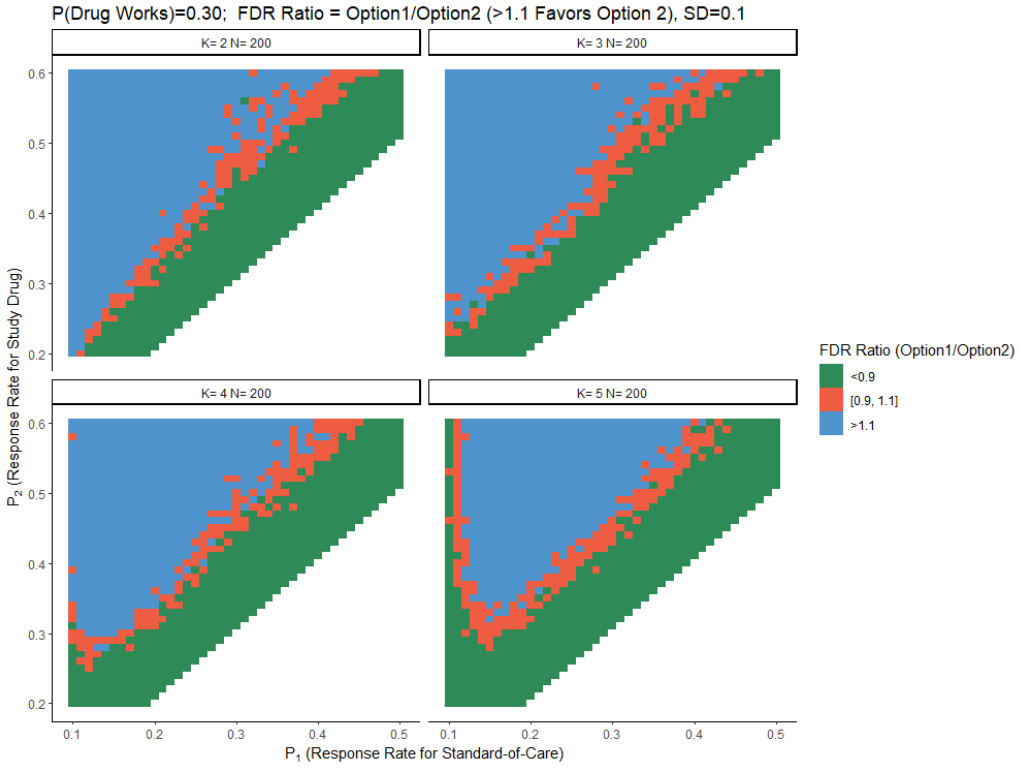
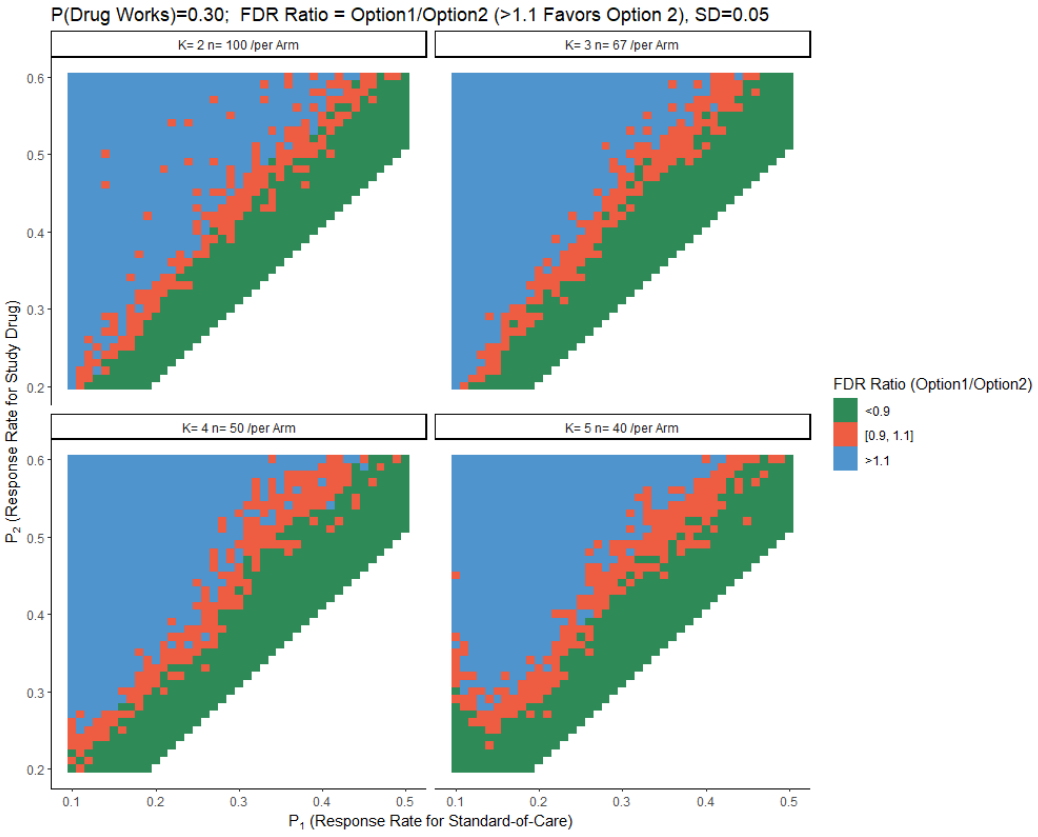


Figure 4: Scenario 2 FDR comparisons of without control arm and with a control arm when sample size n are all the same in both two options, SD=0.05.



4 Discussions

Howard et al. [10] suggested that no type I error control is required in platform studies where each experimental arm is for its own claim. However, even though there is no need for the multiplicity FWER control, one may be interested in the false discovery rate as part of a company's long-term investment and gain in a platform trial. In this paper, we compared the FDR in a platform trial where options of 1) with, 2)without a control arm are considered. Simulation studies were conducted to compare which option would yield to smaller FDR in different parameter settings. Overall, the design with a control arm (Option 2) has smaller FDR when number of experimental arms *K* is small, when effect size is larger, and there is sufficient sample size in each arm. When number of experimental arms K is large, or when the effect size Δ is small, or the sample size is limited, FDR is smaller in the option without a control arm (Option 1). For future work, will extend our evaluation to consider the rolling arm scenario where not all experimental arms are enrolled at the same time and therefore, their portion of shared control is not always 100%.

References

1. Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. New England Journal of Medicine. 2017 Jul 6;377(1):62-70.

2. Benjamin R Saville and Scott M Berry (2016) Efficiencies of platform clinical trials. Clinical Trials, Vol. 13(3) 358-366.

3. Barker, A.D., Sigman, C.C., Kelloff, G.J., Hylton, N.M., Berry, D.A. and Esserman, L., 2009. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clinical Pharmacology & Therapeutics, 86(1), pp.97-100.

4. Berry, S.M., Connor, J.T. and Lewis, R.J., 2015. The platform trial: an efficient strategy for evaluating multiple treatments. Jama, 313(16), pp.1619-1620.

5. Sorić B. Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*. 1989 Jun 1;84(406):608-10.

6. Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society*, Series B. 57 (1): 289300. MR 1325392.

7. John D. Storey and Robert Tibshirani (2001). "Estimating False Discovery Rates under Dependence, with Applications to DNA Microarrays," *Technical Report of Department of Statistics*, Stanford University.

8. Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society,* Series B 64: 479-498.

9. Thomas Jemielita, Archie Tse, and Cong Chen (2018), "Oncology Phase II Proof-of-Concept Studies with Multiple Targets: Randomized Controlled Trial or Single Arm?"

10. Howard, D.R., Brown, J.M., Todd, S. and Gregory, W.M., 2018. Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. Statistical methods in medical research, 27(5), pp.1513-1530.