Investigating the Impact of Prognostic Factor Imbalance in a Randomized Controlled Trial Study

Yuxin Zhang

3/10/2021

Abstract

Background: The situation of prognostic factor imbalance often happens in a randomized controlled trial from pediatric study or oncology study due to a smaller study population. The actual study power in the case of baseline imbalance may be higher or lower than that in a balanced state according to the distribution of imbalance. The objective of this study is to estimate the influence of a prognostic imbalance in terms of power by developing a simulation system.

Methods: A simulation framework which can estimate power was established. The parameters of the simulation program were determined in the initial stage. Patients' data were simulated according to the simulation setting in each simulation round. Study power was estimated by frequency of observing significant treatment effects. Here we utilized two methods to examine the treatment effect (Wald & Fisher). In the stage of investigation, in each simulation study one parameter was relaxed in a study range when adjusting other parameters. The goal was observing whether the prognostic imbalance caused power to rise or decrease and whether the shift or variation in power was related to the relaxed parameter.

Findings: A total number of 14 simulation studies were conducted. When a binary prognostic factor was unbalanced by an imbalance followed $Unif(0.03, 0.07)$, 11 simulation studies showed the moderate or strong degree of variation in power (> 0.001). Degree of shift in power got higher in the same direction when the the prevalence of imbalance between two arms were greater. Our simulation results also indicated that the shift in power changed to the opposite direction after the direction of association between PF and response was reversed.

Conclusions: Results showed that the study power in a perfectly randomized controlled trial would change regularly according to the distribution of the imbalance and the association between PF and treatment response. However, no significant relation between degree of shift in power and conventional RCT characteristics (sample size, treatment effect, etc.) was found due to a number of limitations from the simulation method.

1. Introduction

1.1 Background Information

In the context of randomized clinical trial, usually a large number of baseline factors are collected. Balance in baseline factors is not only a crucial standard to assess the effect of treatment, but also a necessary condition for achieving a perfect randomization. Some prevalent sampling methods [1] such as clustered sampling and multi-stage sampling method somehow effectively reduce the factor imbalance between groups. Nevertheless, these sampling methods do have a limitation which is the balance in baseline covariates will only achieve when sample size is relatively large. In a pediatric or oncology study, it is often unfeasible to perform sampling strategies on available population (patients). Especially for pediatric studies, the sample size [2] is often based on feasibility of recruiting rate and study period given the fact that juveniles have a relatively small population compared to adults. Therefore, expecting a perfect randomization or perfect balance in baseline factors is extremely unlikely in a pediatric RCT study. Compared to many studies that discuss and propose design methods improving randomization such as minimization and stratification [3], this study focuses on how to interpret or assess the statistical result in terms of study power when baseline (prognostic) factor has already occurred.

In an equivalence trial [4], power is a measure of how frequent the null hypothesis can be rejected by statistical tests under the assumption that the effect of proposed treatment is different from a conventional treatment (treatment in control group). The null hypothesis in an equivalence trial is that there is no difference in effectiveness from the proposed treatment and conventional treatment. According to our prior knowledge in statistics, the power of a perfect randomized (i.e. all the baseline covariates are balanced or the unbalanced baseline factor [5] does not influence the probability of an outcome event) controlled trial is related to effect size of treatment, probability of an event (e.g. response rate in oncology study) in control arm, sample size in each arm and type I error. When a prognostic factor that affects the response outcome are unbalanced, the ordinary power calculation will not be precise since the true probability of response for one subject is no longer only depending on the treatment type and it also requires the information regarding the distribution of imbalance.

Simulation [6] is one of the most popular methods for estimation. In case of the predicting statistics or probability is complicated or incapable to be calculated, simulation would be a very effective approach to perform estimation without deducing the true underlying distribution of the statistics. For estimating study power which is a probability measure, this could be done by repeatedly simulating the same scenario (i.e. under same simulation settings) and repeating simulation for 1000 to 10000 times. The more iterations that simulation is conducted, the more precise the estimation is.

With the powerful estimation technique, simulation, this study aims to develop a simulation framework quantifying the magnitude of shift (change) in study power when there is a prognostic imbalance in a randomized controlled trial and investigate the relationship between the magnitude of power (shift in power) and each parameter in RCT.

1.2 Study Objectives and Primary Hypothesis

This is an exploratory study and there is no specific research question to follow or answer. The first objective of this study is proposing a simulation algorithm and methodology that works for estimating power in a RCT scenario with imbalance (between two arms) of a influential binary prognostic factor which is described in section 2. The next goal is to investigate the association between the study power (or the shift in power caused by the prognostic imbalance) measured by the proposed simulation method and each parameter in the RCT such as sample size, treatment effect, the response rate in each arm and so on. The simulation results and evaluation of bias are shown in section 3. Our primary hypothesis is the magnitude and direction of shift in power are also related to the size of treatment effect, sample size and the strength of prognostic factor affecting the response.

2. Methods

In this part, a simulation framework for estimating the power in a randomized controlled trial study with a prognostic imbalance is described. The order of sections follow the actual operations.

2.1 Simulation Settings

Our simulation methodology will be based on a parallel group randomized controlled trial study with binary outcomes. Simulation setting is depending on seven parameters including four parameters for a conventional RCT simulation study: sample size in control group (n_1), sample size in treatment group (n_2), the response rate in control arm (i.e. the prevalence of the outcome event) without considering any prognostic intervention (p_1) and effect size ($\delta = p_2 - p_1$). Three new parameters are now considered when one binary prognostic factor is involved: odds ratio of the effect of the prognostic factor on response outcome (θ) in a logistic regression framework, prevalence of the prognostic factor in control arm and the distribution of the imbalance (for prognostic factor) between two arms. In our study, we assume the distribution of the imbalance follows a uniform distribution with minimum a and maximum b ($Unif(a, b)$). (Note: this is a stochastic parameter and it is randomly generated according to the given distribution in each iteration)

2.2 Data Generation

Variable Information

Data in each simulation round is randomly generated according to the given parameters. To achieve the objective of this study, generated dataset must contain at least three variables, a binary variable refers to group allocation (control or treatment) for each subject, a binary variable indicating the existence of prognostic factor for each subject and a binary variable refers to the occurrence of the outcome event or patient's response in medicine.

Generating Prognostic Factor

Group label of each patient can be created given the sample size in each group. The next step is generating the prognostic factor based on each subject's allocated group. Prognostic factor for patients in control group are generated according to the given prevalence in the simulation settings. Prevalence in treatment group is the one after reducing the imbalance in this round. For example, if the prevalence in control group is 25%, the PF is randomly sampled from Bernoulli (0.25). In contrast, the PF in treatment group is randomly generated from Bernoulli (0.25 - ϵ) assume that the imbalance ϵ is a random sample from $Unif(0.03, 0.07)$.

Computation of Probability

Due to the existence of prognostic factor, the underlying true probability of response for each patient is now different than the usual case without any prognostic factor. Since we assume the effect of the prognostic factor as an exponential of coefficient in a logistic regression (θ), the model framework is now shown as model 2 compared to model 1. Subsequently, the true probability of response after considering prognostic factor are calculated by equation (1). Finally, the outcome data are simulated under the Bernoulli distribution with the updated probabilities.

Model 1: Logistic regression model without considering prognostic factor:

$$log(\frac{P(Response = 1)}{1 - P(Response = 1)}) = \alpha \ + \beta_1 I(treatment)$$

Model 2: Logistic regression model with a binary prognostic factor:

$$log(\frac{P(Response = 1)}{1 - P(Response = 1)}) = \alpha \ + \beta_1 I(treatment) + \beta_2 I(PF)$$

$$P(Response = 1)_i = \frac{e^{\alpha+\beta_1 I(treatment)_i+\beta_2 I(PF)_i}}{1 + e^{\alpha+\beta_1 I(treatment)_i+\beta_2 I(PF)_i}} \tag{1}$$

Dataset Description:

Variable	Description
Response	1 = Yes, 0 = No
Group Label	1 = Treatment, 0 = Control
Prognostic Factor	1 = Yes, 0 = NO

Data Quality Check
Data quality are evaluated through check for NA values (NA value is not expected), check for frequency and each type of response for each variable.

2.3 Statistical Tests

The treatment effect from the simulated data can be assessed through two suggested statistical methods. Given the fact that we are simulating a equivalence trial, the choice of test is two-sided.

Wald Test

Treatment effect is tested through a Wald Test in a logistic regression framework given our null hypothesis is that there is no treatment effect or equivalently the ratio of response odds in treatment group to response odds in control group is 1. ($H_0 : \beta_1 = 0$) The Wald statistic is the square of the regression estimate divided by the standard error of the regression estimate. And it is asymptotically followed a chi-square distribution. The null hypothesis is rejected if the Wald test statistic is greater than $\chi^2_{1-\alpha}$

$$W = \frac{\hat{\beta}_1^2}{SE^2(\hat{\beta}_1)} \sim \chi^2 \tag{2}$$

Fisher Exact Test

The treatment effect is also assessed through a Fisher Exact Test. Suppose X_1 is total number of responses in control arm that we observed and $X_1 = \sum_{i=1}^{n_1} e_i$, where e_i is the observed response outcome for each subject in the control group. Similarly, $X_2 = \sum_{i=1}^{n_2} e_i$. Based on these facts, one may assume that $X_1 \sim Hypergeometric(N = n_1 + n_2, K = x_1 + x_2, n = n_1)$, then the probability of observed total counts in the control group is equivalent to:

$$Pr(X_1 = x_1) = \frac{\binom{K}{x_1} * \binom{N-K}{x_2}}{\binom{N}{n_1}}, where\ K\ = x_1 + x_2 \tag{3}$$

Null hypothesis for the Fisher's exact test is exactly same as Wald Test (the response rate in two groups are identical) but with a different underlying distribution. In other words, we expect the total number of response outcomes are equally distributed between groups under the null hypothesis. Consequently, the p-value of Fisher Exact Test is the summation of probabilities which are more extreme than (or equal to) the observed case.

$$P - value = 2 * \sum_{k=0}^{x_1} P(X_1 = k) \quad for\ a\ two - sided\ test \tag{4}$$

2.4 Power Estimation

According to Monte Carlo method, in case of repeating the simulation steps for a large number of times under a specified simulation settings, Type II error ($P(retain\ H_0|H_0\ is\ false)$) keeps constant in each iteration, one may conclude that frequency of rejecting null hypothesis could be an estimator for the actual power in the light of maximum likelihood estimation. In this scenario, suppose the result from the hypothesis testing in ith iteration, d_i, is recorded as 1 if the null hypothesis is rejected successfully and 0 otherwise.

$(d \sim Bernoulli(p = P(reject\ H_0|H_0\ is\ false)))$ For a total number of simulation rounds, k, the power can be estimated by

$$P(reject\ H_0\hat{|}H_0\ is\ False) = \frac{\#\ of\ simulation\ reject\ H_0}{total\ \#\ of\ simulation} = \frac{\sum_{i=0}^{k} d_i}{k} \tag{5}$$

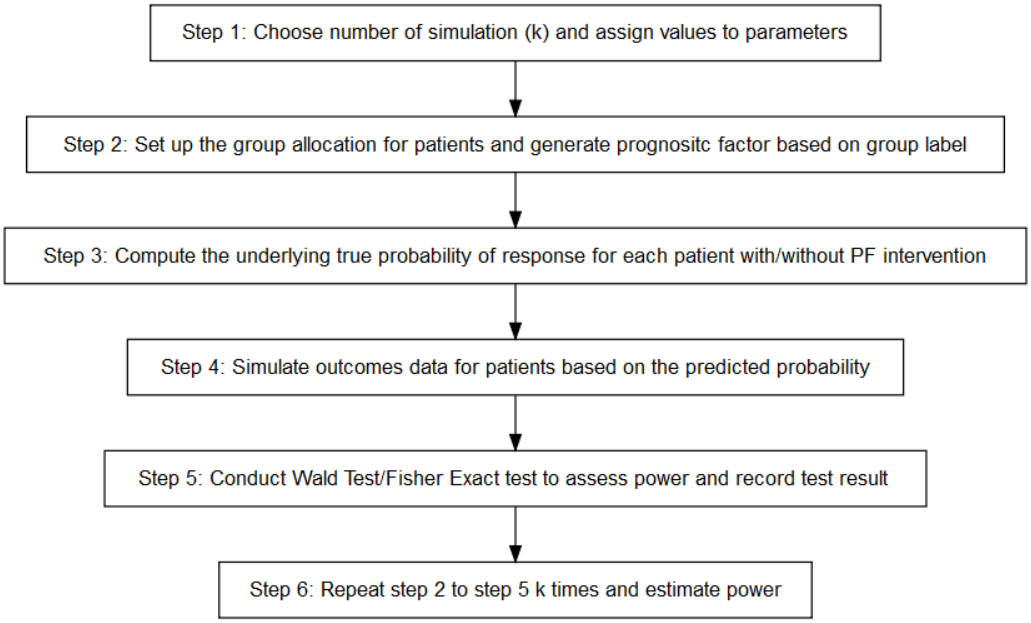

Figure 1: Whole simulation framework for Power Estimation

2.5 Comparative Study and Investigation on Power Shift

A RCT study with a balanced binary prognostic factor was chosen as our comparative study. Prevalence of PF is equal to 0.25 in both arms and imbalance is now ignored. Power of two studies were estimated through the above simulation method and the difference was computed for exploration. The calculation and evaluation of the difference in power are described in detail in the results section.

2.6 Software

Simulation algorithms were implemented through R software. ggplot2 package was utilized for plotting purpose. Sample code for simulation and statistical analysis is provided in the Appendix section.

3. Results and Analysis

Simulation results of RCT with different types of settings are shown in this section. To verify the impact of the prognostic factor imbalance on study power, RCT with a balanced binary prognostic factor (without considering the existence of imbalance) was chosen as our comparative study. Comparative measure was the estimated power from both Wald Test and Fisher Exact Test methods. Given that there were multiple parameters in the proposed simulation program, our research strategy was changing one parameter at a time while adjusting other parameters. Estimated powers were obtained under different scenarios (with balanced PF and with unbalanced PF). Difference between estimated powers was also calculated and recorded as a shift in power. In this analysis we mainly focus on the relationship between the shift in power and parameters such as sample size in each group, response rate in control group, effect size, direction and scale of the effect of the prognostic factor on response outcome in a logistic regression framework as well as distribution of imbalance. Type-I error was set at 0.05 in this study. Meanwhile, the prevalence of prognostic factor in control arm was always equal to 0.25. The number of simulation runs in each study was 1000. In terms of prognostic imbalance, we would only consider two uniform distributions on a range [0.03, 0.07] and [0.07, 0.1] respectively.

Study parameters and study range:

Parameter	Notation	Discussed Section	Range
sample size in control group	n_1	3.1	10-50
sample size in treatment group	n_2	3.1	10-50
treatment effect	δ	3.2	0.1 - 0.3
response rate in control group	p_1	3.3	0.1 - 0.3
odds ratio of PF in logistic regression	θ	3.4	1.5, 2, 2.5

Assessment on shift in power:

$$Magnitude\ of\ Shift = Power_{imbalance} - Power_{balance}$$

where $Power_{imbalance}$ and $Power_{balance}$ are estimated power respectively from an unbalanced PF study and balanced PF study with identical simulation settings.

$$Degree\ of\ Shift = \begin{cases} weak & |Magnitude\ of\ Shift| \leq 0.01 \\ moderate & 0.01 < |Magnitude\ of\ Shift| \leq 0.02 \\ strong & 0.02 < |Magnitude\ of\ Shift| \end{cases}$$

3.1 Exploration on the impact of sample size on study power

Simulations were conducted with different number of total sample size. In order to more easily capture the association between total sample size and study power (bias in power), all other parameters were fixed at a constant level during the whole simulation process. We firstly simulated the situation that there were same number of samples in control group and treatment group (equal allocation), followed by a case when group sample sizes were not equal.

Simulation #1

Simulation Settings:
response rate in control group (p_1): 0.2
effect size (δ): 0.1
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: positive (+)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.03, 0.07)$

total sample size: 30-100
allocation: equal allocation

True underlying probability of response for all combination of covariates:

Treatment Type	Prognostic Factor	Probability
Control (0)	No (0)	0.2
Treatment (1)	No (0)	0.3
Control (0)	Yes (1)	0.333
Treatment (1)	Yes (1)	0.462

Simulation #1 Results:

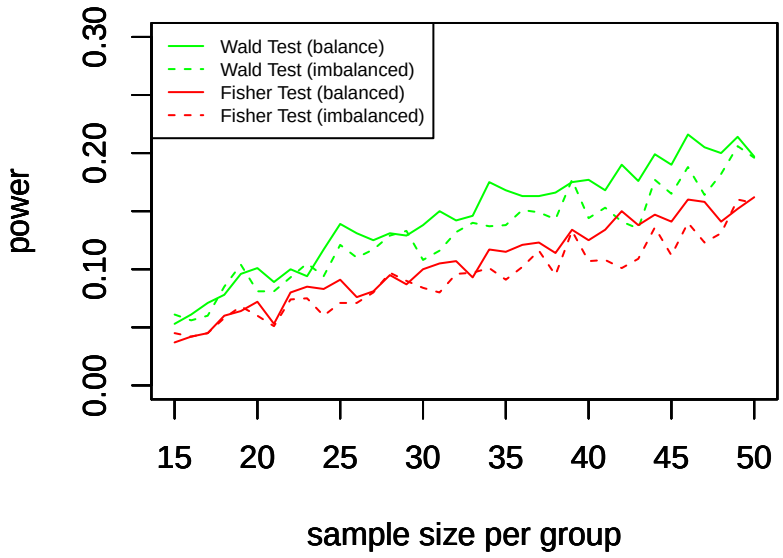

Figure 2: estimated power under simulation setting 1

Estimated Power from RCT with balanced prognostic factor:

sample size in control arm	sample size in treatment arm	power (Wald)	power (Fisher)
15	15	0.053	0.037
25	25	0.139	0.091
35	35	0.168	0.115
50	50	0.197	0.162

Estimated Power from RCT with unbalanced prognostic factor:

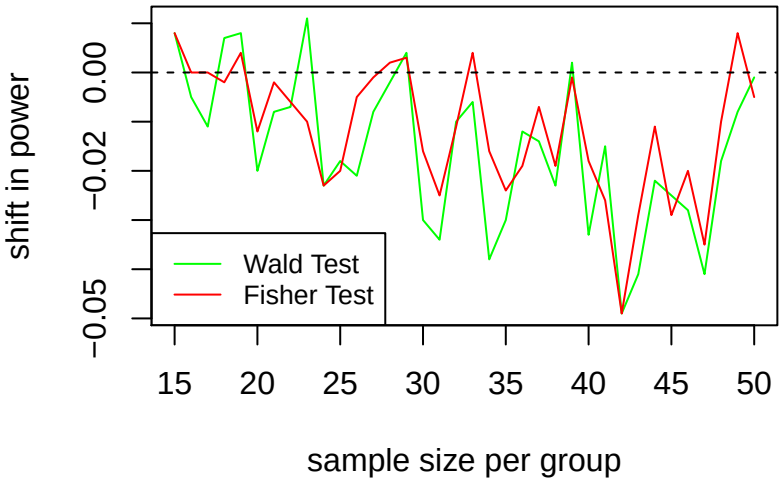

Figure 3: change in power (simulation setting 1)

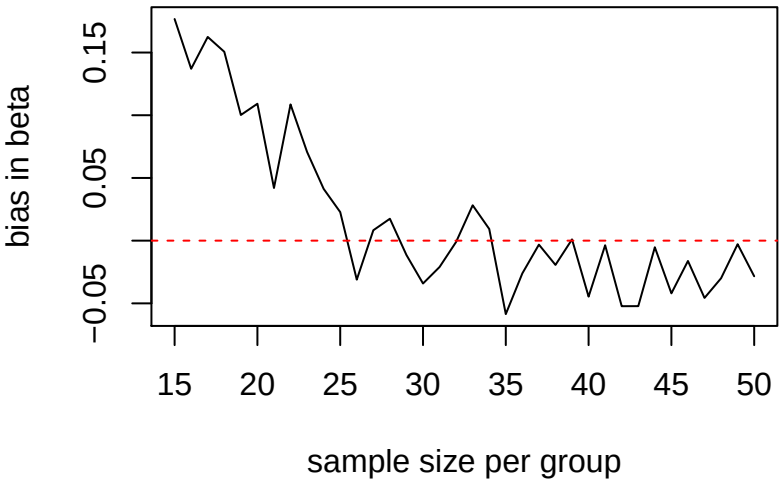

Figure 4: bias in beta (simulation setting 1)

sample size in control arm	sample size in treatment arm	power (Wald)	power (Fisher)
15	15	0.061	0.045
25	25	0.121	0.071
35	35	0.138	0.091
50	50	0.196	0.157

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.016	decrease	moderate
Fisher Exact Test	0.012	decrease	moderate

As shown in the above figures and tables, study power of RCT decreased moderately after making prognostic factor unbalanced. Furthermore, to confirm the validity of the simulation results, an analytical way to solve this problem is computing the *expected value of number of response* (ER) in each arm according to the distribution of prognostic factor. One may derive that the expected value of number of response in control arm is equivalent to:

$$E(number\ of\ response) = n_1 * 0.75 * p_{control,PF=0} + n_1 * 0.25 * p_{control,PF=1}$$

where $p_{control,PF=0}$ and $p_{control,PF=1}$ are the corresponding true underlying probabilities of observing a response. (see table of true underlying probability of response) Similarly, the expected count of response in RCT with unbalanced prognostic factor can also be calculated in this way:

$$E(number\ of\ response) = n_1 * (0.75 + (a+b)/2) * p_{trt,PF=0} + n_1 * (0.25 - (a+b)/2) * p_{trt,PF=1}$$

Expected value of number of response in each arm and for each type of study:

BALANCE STATUS	Group	ER($n_1 = n_2 = 15$)	ER($n_1 = n_2 = 25$)	ER($n_1 = n_2 = 35$)	ER($n_1 = n_2 = 50$)
BALANCED	CONTROL	3.50	5.83	8.16	11.66
BALANCED	TREATMENT	5.11	8.51	11.92	17.03
UNBALANCED	CONTROL	3.50	5.83	8.16	11.66
UNBALANCED	TREATMENT	4.99	8.31	11.63	16.62

Based on the results from imbalance study, The expected number of response in treatment arm was less than that of balanced study in each corresponding case under the condition that the expected value in control arm remained the same. Consequently, the decline of study power was also explained in terms of expected value of response by the reduction in expected difference between two arms.

Simulation #2 (Change the direction of effect compared to simulation 1)

Simulation Settings:
response rate in control group (p_1): 0.2
effect size (δ): 0.1
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: negative (-)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.03, 0.07)$
total sample size: 30-100
allocation type: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability
Control (0)	No (0)	0.2
Treatment (1)	No (0)	0.3
Control (0)	Yes (1)	0.111
Treatment (1)	Yes (1)	0.176

Expected value of number of response in each arm and for each type of study:

BALANCE STATUS	Group	ER($n_1 = n_2 = 15$)	ER($n_1 = n_2 = 25$)	ER($n_1 = n_2 = 35$)	ER($n_1 = n_2 = 50$)
BALANCED	CONTROL	2.67	4.44	6.22	8.89
BALANCED	TREATMENT	4.04	6.73	9.42	13.45
UNBALANCED	CONTROL	2.67	4.44	6.22	8.89
UNBALANCED	TREATMENT	4.13	6.88	9.63	13.76

Simulation #2 Results:

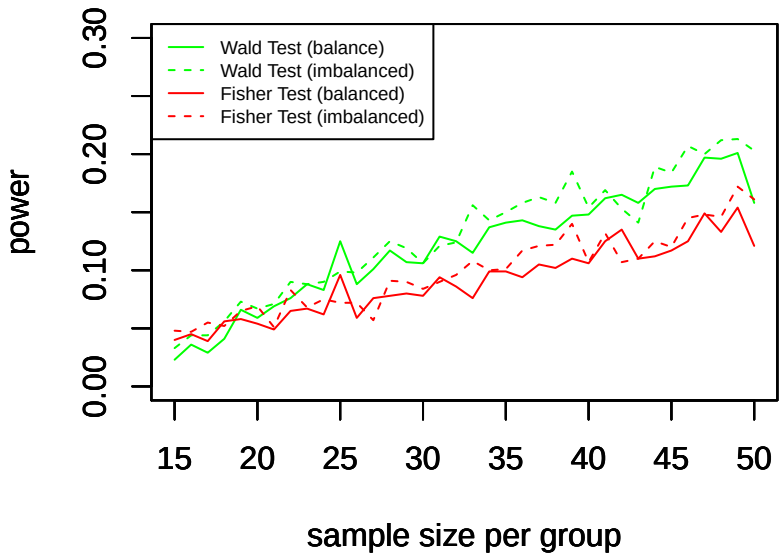

Figure 5: estimated power under simulation setting 2

Estimated Power from RCT with balanced prognostic factor:

sample size in control arm	sample size in treatment arm	power(Wald)	power(Fisher)
15	15	0.023	0.040
25	25	0.125	0.096
35	35	0.141	0.099
50	50	0.158	0.121

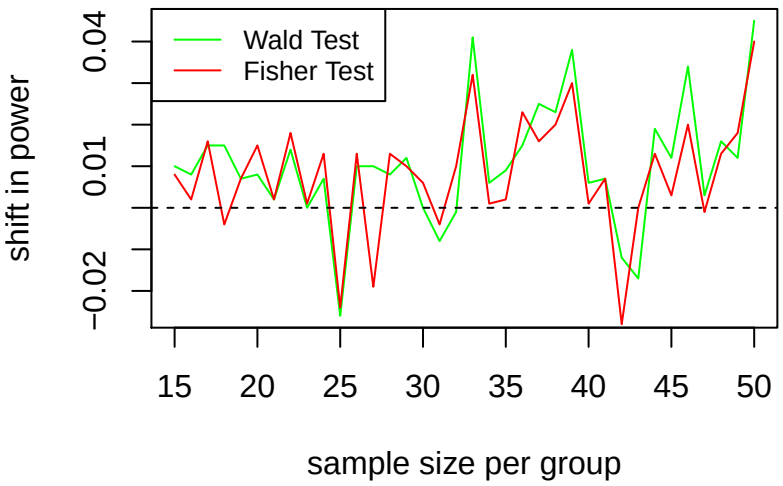

Figure 6: change in power (simulation setting 2)

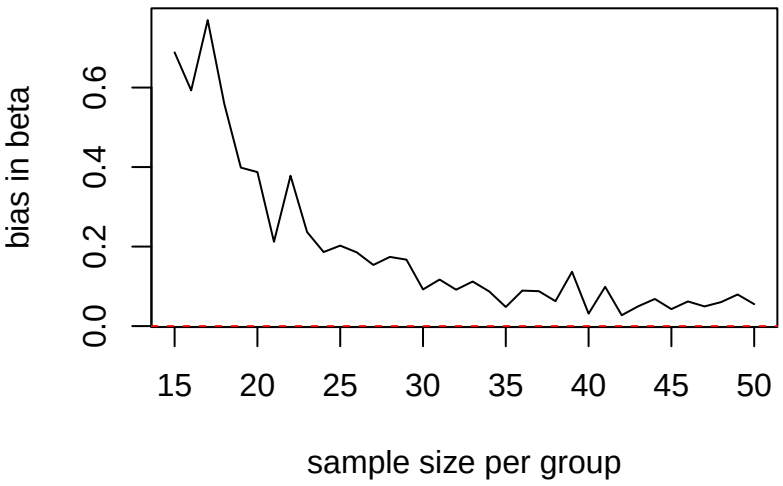

Figure 7: bias plot under simulation setting 2

Estimated Power from RCT with unbalanced prognostic factor:

sample size in control arm	sample size in treatment arm	power(Wald)	power(Fisher)
15	15	0.033	0.048
25	25	0.099	0.072
35	35	0.150	0.101
50	50	0.203	0.161

Evaluation on shift (in power):

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.010	increase	moderate
Fisher Exact Test	0.008	increase	weak

Simulation #3 (Change the distribution of imbalance compared to simulation 1)

Simulation Settings:
response rate in control group (p_1): 0.2
effect size (δ): 0.1
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: positive (+)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.07, 0.1)$
total sample size: 30-100
allocation type: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability
Control (0)	No (0)	0.2
Treatment (1)	No (0)	0.3
Control (0)	Yes (1)	0.333
Treatment (1)	Yes (1)	0.462

Expected value of number of response in each arm and for each type of study:

BALANCE STATUS	Group	ER($n_1 = n_2 = 15$)	ER($n_1 = n_2 = 25$)	ER($n_1 = n_2 = 35$)	ER($n_1 = n_2 = 50$)
BALANCED	CONTROL	3.50	5.83	8.16	11.66
BALANCED	TREATMENT	5.11	8.51	11.92	17.03
UNBALANCED	CONTROL	3.50	5.83	8.16	11.66
UNBALANCED	TREATMENT	4.90	8.17	11.44	16.34

Simulation #3 Results:

Estimated Power from RCT with balanced prognostic factor:

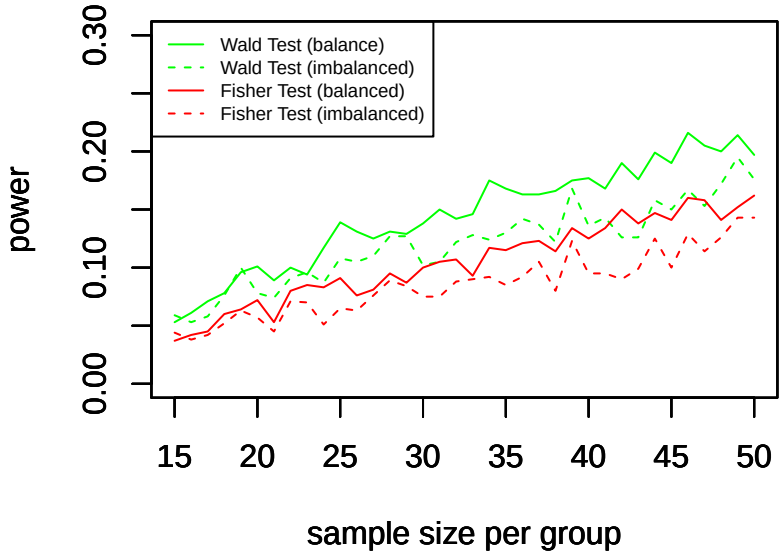

Figure 8: estimated power under simulation setting 3

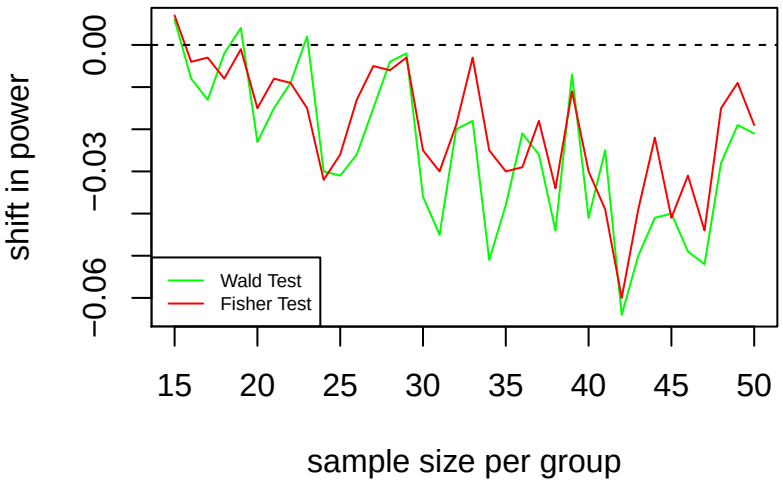

Figure 9: change in power (simulation setting 3)

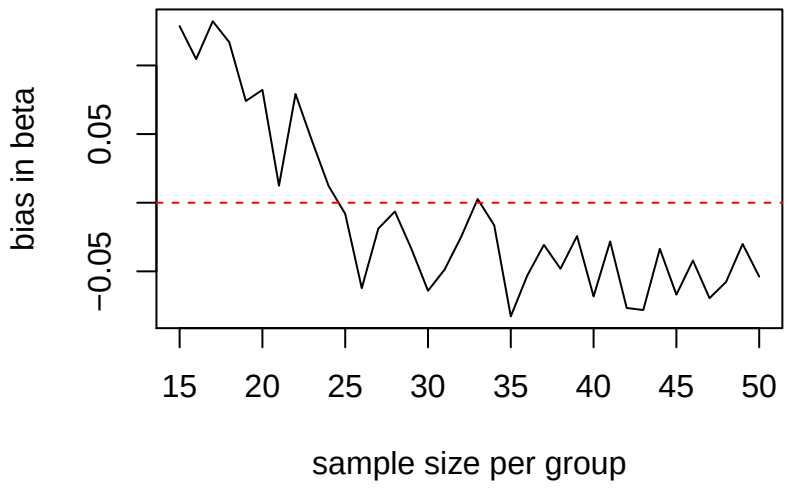

Figure 10: bias plot under simulation setting 3

sample size in control arm	sample size in treatment arm	power (Wald)	power (Fisher)
15	15	0.053	0.037
25	25	0.139	0.091
35	35	0.168	0.115
50	50	0.197	0.162

Estimated Power from RCT with unbalanced prognostic factor:

sample size in control arm	sample size in treatment arm	power(Wald)	power(Fisher)
15	15	0.059	0.044
25	25	0.108	0.065
35	35	0.130	0.085
50	50	0.176	0.143

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.025	decrease	strong
Fisher Exact Test	0.020	decrease	strong

Simulation #4 (change distribution of imbalance compared to simulation 2)

Simulation Settings:
response rate in control group (p_1): 0.2
effect size (δ): 0.1

odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: negative (-)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.07, 0.1)$
total sample size: 30-100
allocation type: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability
Control (0)	No (0)	0.2
Treatment (1)	No (0)	0.3
Control (0)	Yes (1)	0.111
Treatment (1)	Yes (1)	0.176

Expected value of number of response in each arm and for each type of study:

BALANCE STATUS	Group	ER($n_1 = n_2 = 15$)	ER($n_1 = n_2 = 25$)	ER($n_1 = n_2 = 35$)	ER($n_1 = n_2 = 50$)
BALANCED	CONTROL	2.67	4.44	6.22	8.89
BALANCED	TREATMENT	4.04	6.73	9.42	13.45
UNBALANCED	CONTROL	2.67	4.44	6.22	8.89
UNBALANCED	TREATMENT	4.19	6.99	9.78	13.98

Simulation #4 Results:

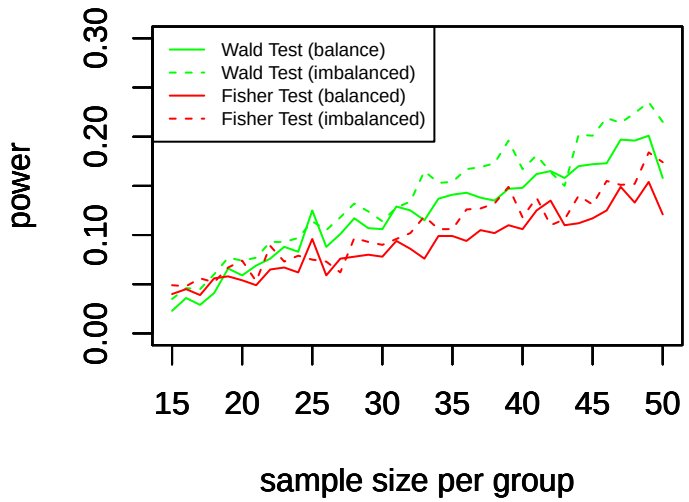

Figure 11: estimated power under simulation setting 4

Estimated Power from RCT with balanced prognostic factor:

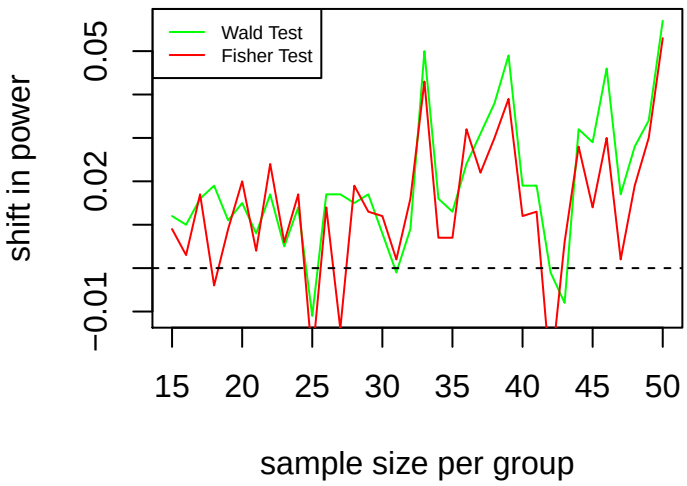

Figure 12: shift in power (simulation setting 4)

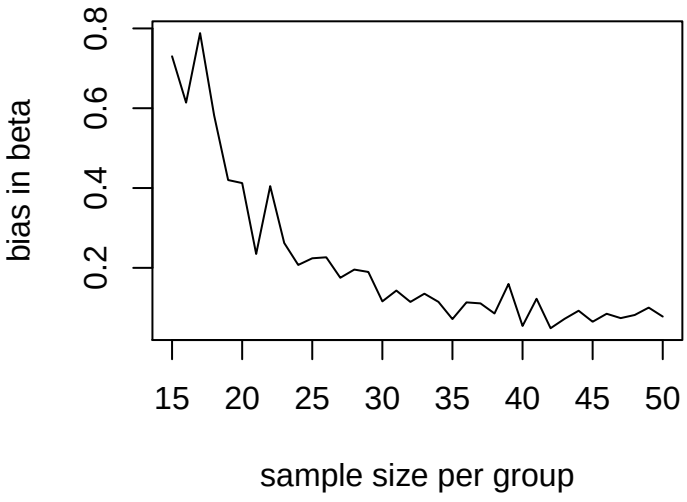

Figure 13: bias plot under simulation setting 4

sample size in control arm	sample size in treatment arm	power(Wald)	power(Fisher)
15	15	0.023	0.040
25	25	0.125	0.096
35	35	0.141	0.099
50	50	0.158	0.121

Estimated Power from RCT with unbalanced prognostic factor:

sample size in control arm	sample size in treatment arm	power(Wald)	power(Fisher)
15	15	0.035	0.049
25	25	0.114	0.075
35	35	0.154	0.106
50	50	0.215	0.174

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.019	increase	moderate
Fisher Exact Test	0.014	increase	moderate

Simulation #5 (Unequal allocation study)

Simulation Settings:
response rate in control group (p_1): 0.2
effect size (δ): 0.1
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: positive (+)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.03, 0.07)$
total sample size: 30-100
allocation type: unequal allocation (Randomization Ratio = 1:2 or $n_2 = 2n_1$)

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability
Control (0)	No (0)	0.2
Treatment (1)	No (0)	0.3
Control (0)	Yes (1)	0.333
Treatment (1)	Yes (1)	0.462

Expected value of number of response in each arm and for each type of study:

BALANCE STATUS	Group	ER (10,20)	ER (17,33)	ER (23,47)	ER (33,67)
BALANCED	CONTROL	2.33	3.97	5.36	7.70
BALANCED	TREATMENT	6.81	11.24	16.00	22.81
UNBALANCED	CONTROL	2.33	3.97	5.36	7.70
UNBALANCED	TREATMENT	6.65	10.97	15.62	22.27

Simulation #5 Results:

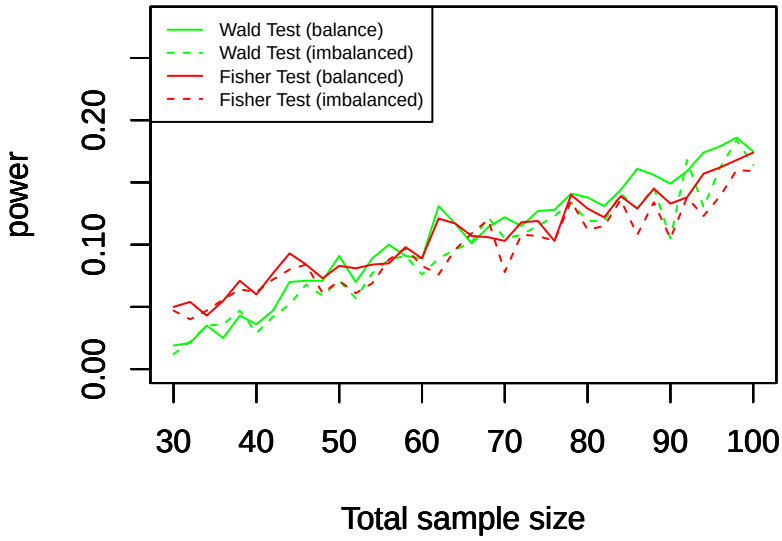

Figure 14: estimated power under simulation setting 5

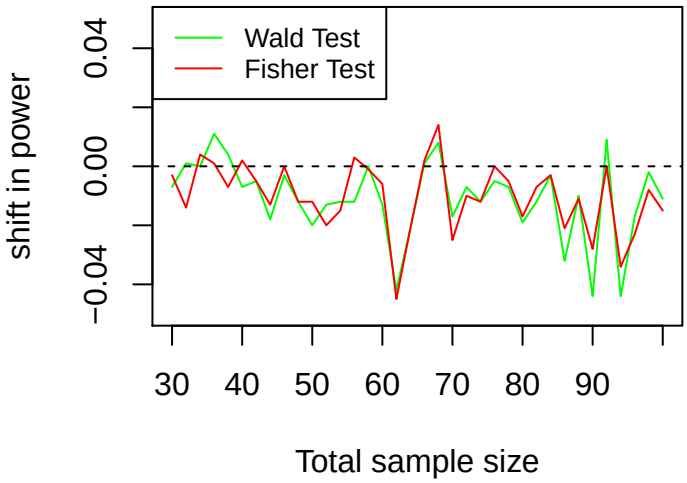

Figure 15: change in power under simulation setting 5

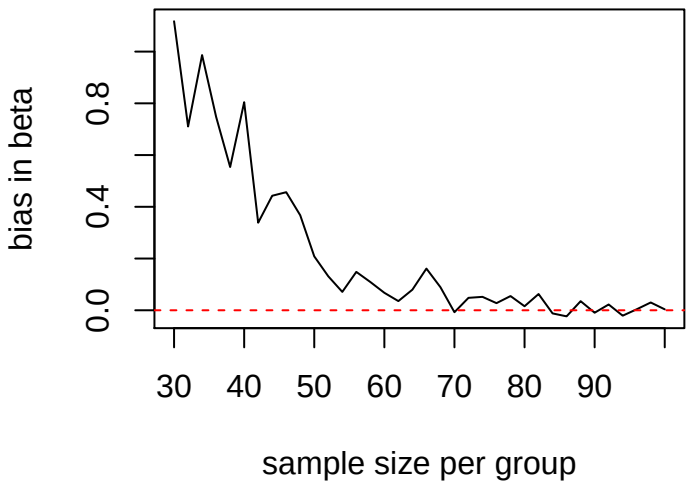

Figure 16: bias plot under simulation setting 5

Estimated Power from RCT with balanced prognostic factor:

sample size in control arm	sample size in treatment arm	total sample size	power(Wald)	power(Fisher)
10	20	30	0.019	0.050
17	33	50	0.091	0.083
23	47	70	0.122	0.103
33	67	100	0.175	0.174

Estimated Power from RCT with unbalanced prognostic factor:

sample size in control arm	sample size in treatment arm	total sample size	power(Wald)	power(Fisher)
10	20	30	0.012	0.047
17	33	50	0.071	0.071
23	47	70	0.105	0.078
33	67	100	0.164	0.159

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.0109	decrease	moderate
Fisher Exact Test	0.0102	decrease	moderate

Summary and Conclusion

Association between power and sample size
Increase in the number of samples resulted in an increase in power when other parameters were helding constant, which was agreed by all the five simulations and also in accordance with statistical common sense.

Association between power shift and sample size
There was a slight trend that the shift in power increases as the sample size increases based on simulation study #1 and simulation #3. On the other hand, no obvious trend could be captured according to the trace plot of shift in study #2, #4 and #5. Overall, no significant relationship was captured between the magnitude of shift and sample size.

Direction of the effect of prognostic factor (Study #1 vs. Study #2 and Study #3 vs. Study #4)
The direction of shift (in power) reversed when the direction of association between the prognostic factor and probability of response (outcome) changed. By comparing study #1 and study #2, all the parameters were consistent expect for the direction of the effect of PF in logistic regression. As a result, the shift in power caused by the prognostic imbalance changed from negative (decrease in power) to positive (increase in power). Same conclusion was also drawn by comparing study #3 and study #4. One reasonable explanation is the emergence of imbalance (on the treatment arm) reduces the prevalence of the subject having PF in the treatment arm when PF has a positive effect on the probability of response, which leads the study power to decrease. Conversely, in case of PF decreases the probability of response, power increases due to there are fewer subjects affected by PF in the treatment arm.

Distribution of imbalance (Study #1 vs. Study #3 and Study #2 vs. Study #4)
The degree of imbalance did not change the direction of shift but magnitude. Two imbalance distributions were selected in our simulation study. $(Unif(0.03, 0.07), Unif(0.07, 0.1))$ The expected value of distribution $Unif(0.03, 0.07)$ is 0.05 while $Unif(0.07, 0.1)$ produces a greater prevalence of imbalance, 0.085. We selected different distribution of imbalance in study #1 and study #3 while holding the value of other parameters. Simulation results showed that increasing imbalance strengthen the magnitude of shift in power without changing the direction. Same conclusion was obtained by comparing study #2 and study #4.

Allocation (Study #1 vs. Study #5)
Compared to study #1, study #5 performed an unequal subject allocation and study power was decreased as expected. Mean shift in power in study #5 obtained from both two methods were also decreased, but this might be a consequence of decreasing in balanced study power.

3.2 Exploration on the impact of treatment effect

Group sample size were fixed at 20 each arm. Treatment effect varied from 0.1 - 0.3 to investigate the impact of effect size on study power (shift in power). Similar to section 3.1, we attempted to compare and contrast the results by changing the direction of association between PF and response as well as the prevalence of imbalance.

Simulation #6

Simulation Settings:
response rate in control group (p_1): 0.2
effect size (δ): 0.1 - 0.3
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: positive (+)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.03, 0.07)$
total sample size: 40

allocation type: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability($\delta = 0.1$)	Probability($\delta = 0.2$)	Probability($\delta = 0.3$)
Control (0)	No (0)	0.2	0.2	0.2
Treatment (1)	No (0)	0.3	0.4	0.5
Control (0)	Yes (1)	0.333	0.333	0.333
Treatment (1)	Yes (1)	0.462	0.571	0.667

Expected value of number of response in each arm and for each type of study:

BALANCE STATUS	Group	ER ($\delta = 0.1$)	ER ($\delta = 0.2$)	ER ($\delta = 0.3$)
BALANCED	CONTROL	4.65	4.65	4.65
BALANCED	TREATMENT	6.81	8.86	10.84
UNBALANCED	CONTROL	4.65	4.65	4.65
UNBALANCED	TREATMENT	6.64	8.68	10.67

Simulation #6 Results:

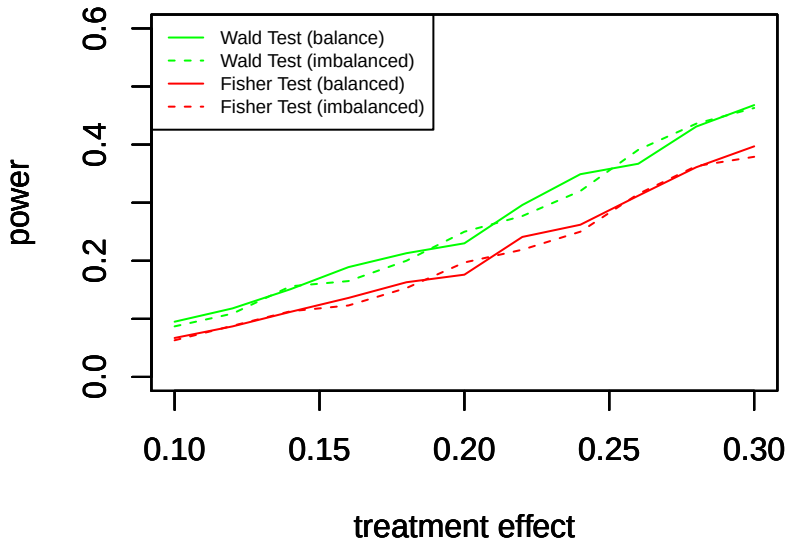

Figure 17: estimated power under simulation setting 6

Estimated Power from RCT with balanced prognostic factor:

n_1	n_2	p_1	d	power (Wald)	power (Fisher)
15	15	0.1	0.1	0.095	0.067
15	15	0.1	0.2	0.230	0.176
15	15	0.1	0.3	0.468	0.397

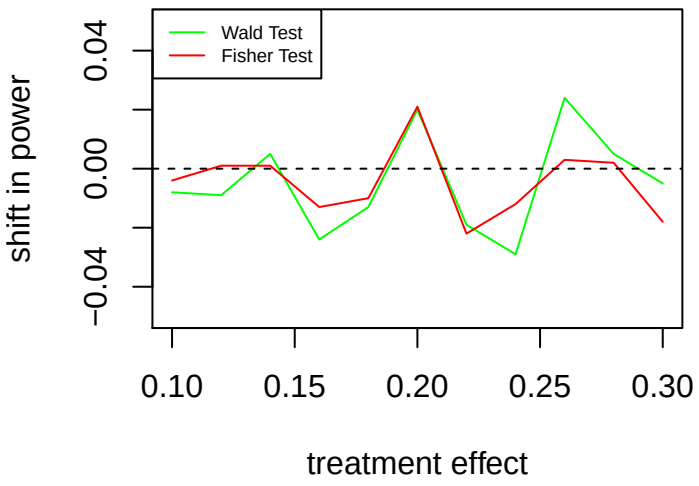

Figure 18: change in power under simulation setting 6

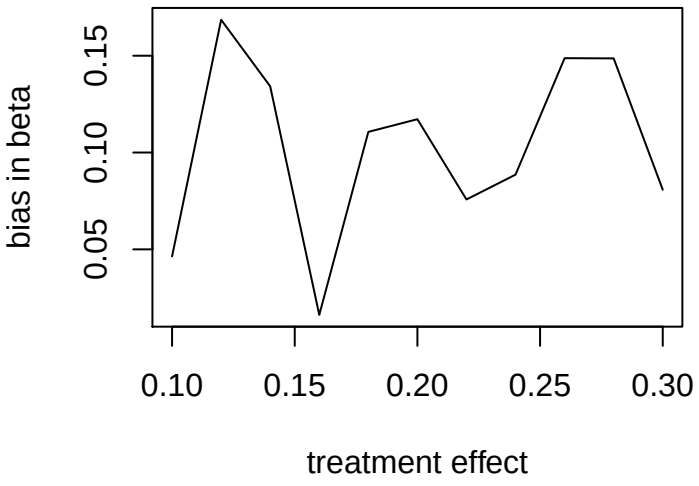

Figure 19: bias plot under simulation setting 6

Estimated Power from RCT with unbalanced prognostic factor:

n1	n2	p1	d	power (Wald)	power (Fisher)
15	15	0.1	0.1	0.087	0.063
15	15	0.1	0.2	0.250	0.197
15	15	0.1	0.3	0.463	0.379

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.0048	decrease	weak
Fisher Exact Test	0.0046	decrease	weak

Simulation #7 (change the direction of effect of PF compared to simulation 6)

Simulation Settings:
response rate in control group (p_1): 0.2
effect size (δ): 0.1 - 0.3
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: positive (-)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.03, 0.07)$
total sample size: 40
allocation type: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability(d = 0.1)	Probability(d = 0.2)	Probability(d = 0.3)
Control (0)	No (0)	0.2	0.2	0.2
Treatment (1)	No (0)	0.3	0.4	0.5
Control (0)	Yes (1)	0.111	0.111	0.111
Treatment (1)	Yes (1)	0.176	0.25	0.333

Expected value of number of response in each arm and for each type of study:

BALANCE STATUS	Group	ER ($\delta = 0.1$)	ER ($\delta = 0.2$)	ER ($\delta = 0.3$)
BALANCED	CONTROL	3.56	3.56	3.56
BALANCED	TREATMENT	5.38	7.25	9.17
UNBALANCED	CONTROL	3.56	3.56	3.56
UNBALANCED	TREATMENT	5.50	7.40	9.33

Simulation #7 Results:

Estimated Power from RCT with balanced prognostic factor:

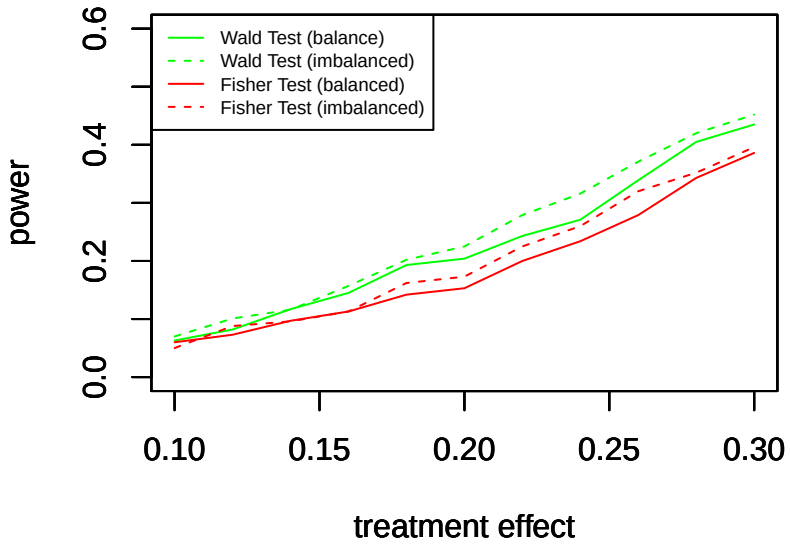

Figure 20: estimated power under simulation setting 7

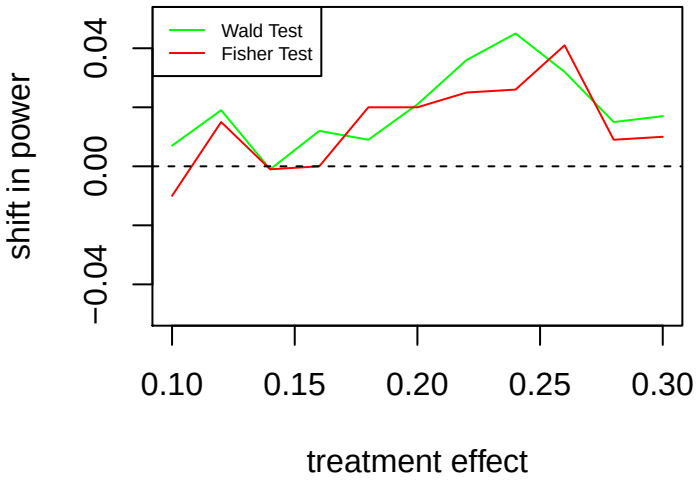

Figure 21: change in power under simulation setting 7

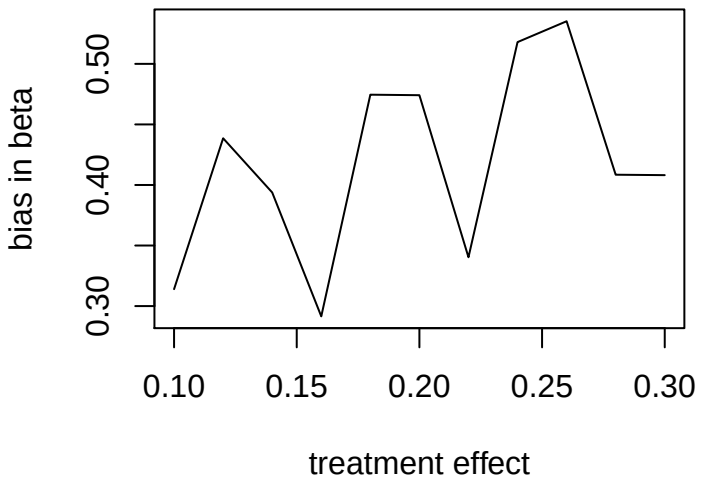

Figure 22: bias plot under simulation setting 7

n1	n2	p1	d	power (Wald)	power (Fisher)
15	15	0.1	0.1	0.063	0.06
15	15	0.1	0.2	0.204	0.153
15	15	0.1	0.3	0.435	0.386

Estimated Power from RCT with unbalanced prognostic factor:

n1	n2	p1	d	power (Wald)	power (Fisher)
15	15	0.1	0.1	0.070	0.050
15	15	0.1	0.2	0.225	0.173
15	15	0.1	0.3	0.452	0.396

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.019	increase	moderate
Fisher Exact Test	0.014	increase	moderate

Simulation #8 (change the distribution of imbalance compared to simulation 6)

Simulation Settings:
response rate in control group (p_1): 0.2
effect size (δ): 0.1 - 0.3
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: positive (+)
prevalence of prognostic factor in control arm: 0.25

distribution of imbalance: $Unif(0.07, 0.1)$
total sample size: 40
allocation: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability($\delta = 0.1$)	Probability($\delta = 0.2$)	Probability($\delta = 0.3$)
Control (0)	No (0)	0.2	0.2	0.2
Treatment (1)	No (0)	0.3	0.4	0.5
Control (0)	Yes (1)	0.333	0.333	0.333
Treatment (1)	Yes (1)	0.462	0.571	0.667

Expected value of number of response in each arm and for each type of study:

BALANCE STATUS	Group	ER ($\delta = 0.1$)	ER ($\delta = 0.2$)	ER ($\delta = 0.3$)
BALANCED	CONTROL	4.65	4.65	4.65
BALANCED	TREATMENT	6.81	8.86	10.84
UNBALANCED	CONTROL	4.65	4.65	4.65
UNBALANCED	TREATMENT	6.53	8.56	10.55

Simulation #8 Results:

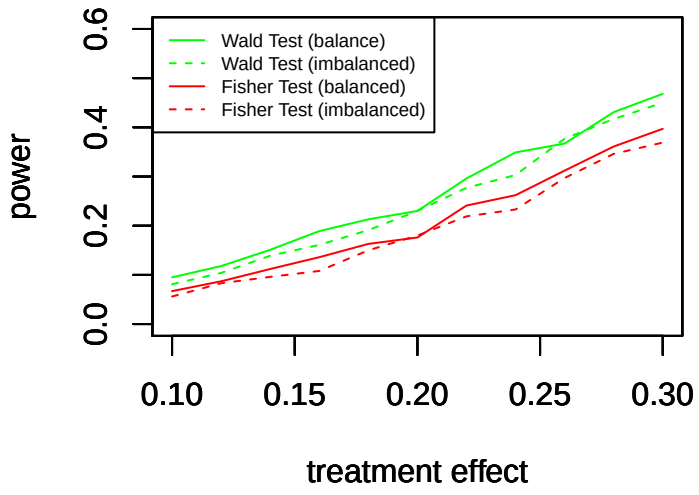

Figure 23: estimated power under simulation setting 8

Estimated Power from RCT with balanced prognostic factor:

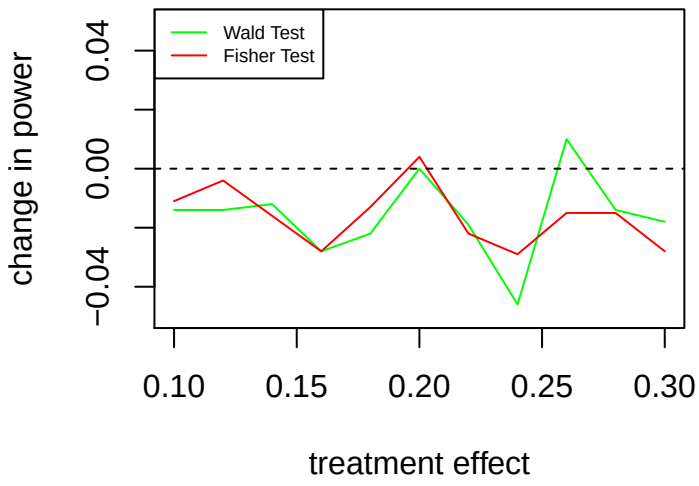

Figure 24: change in power under simulation setting 8

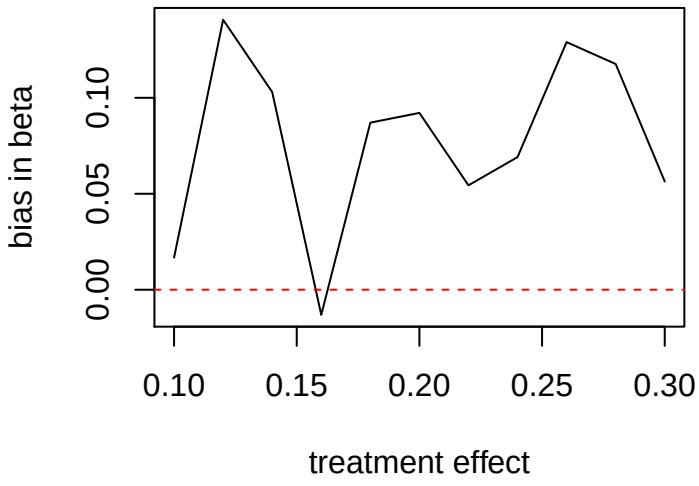

Figure 25: bias plot under simulation setting 8

n1	n2	p1	d	power (Wald)	power (Fisher)
15	15	0.1	0.1	0.095	0.067
15	15	0.1	0.2	0.230	0.176
15	15	0.1	0.3	0.468	0.397

Estimated Power from RCT with unbalanced prognostic factor:

n1	n2	p1	d	power (Wald)	power (Fisher)
15	15	0.1	0.1	0.081	0.056
15	15	0.1	0.2	0.230	0.180
15	15	0.1	0.3	0.450	0.369

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.016	decrease	moderate
Fisher Exact Test	0.016	decrease	moderate

Simulation #9 (change the distribution of imbalance compared to study 7)

Simulation Settings:
response rate in control group (p_1): 0.2
effect size (δ): 0.1 - 0.3
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: negative
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.07, 0.1)$
total sample size: 40
allocation type: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability(d = 0.1)	Probability(d = 0.2)	Probability(d = 0.3)
Control (0)	No (0)	0.2	0.2	0.2
Treatment (1)	No (0)	0.3	0.4	0.5
Control (0)	Yes (1)	0.111	0.111	0.111
Treatment (1)	Yes (1)	0.176	0.25	0.333

Expected value of number of response in each arm and for each type of study:

BALANCE STATUS	Group	ER ($\delta = 0.1$)	ER ($\delta = 0.2$)	ER ($\delta = 0.3$)
BALANCED	CONTROL	3.56	3.56	3.56
BALANCED	TREATMENT	5.38	7.25	9.17
UNBALANCED	CONTROL	3.56	3.56	3.56
UNBALANCED	TREATMENT	5.59	7.51	9.45

Simulation #9 Results:

Estimated Power from RCT with balanced prognostic factor:

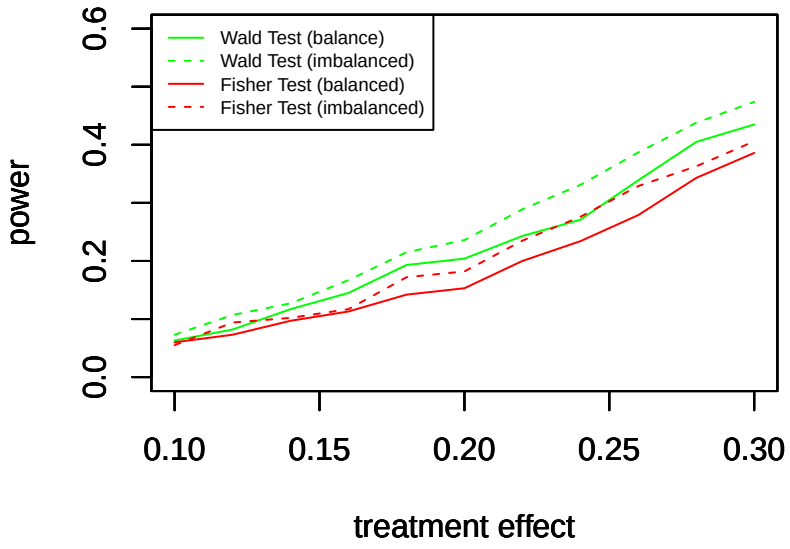

Figure 26: estimated power under simulation setting 9

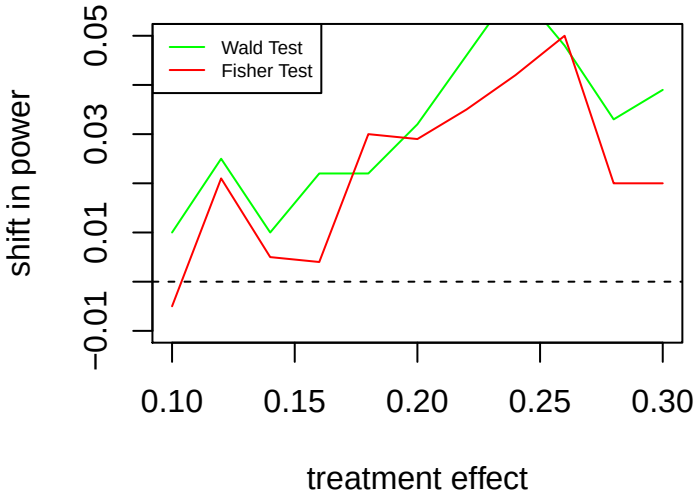

Figure 27: shift in power under simulation setting 9

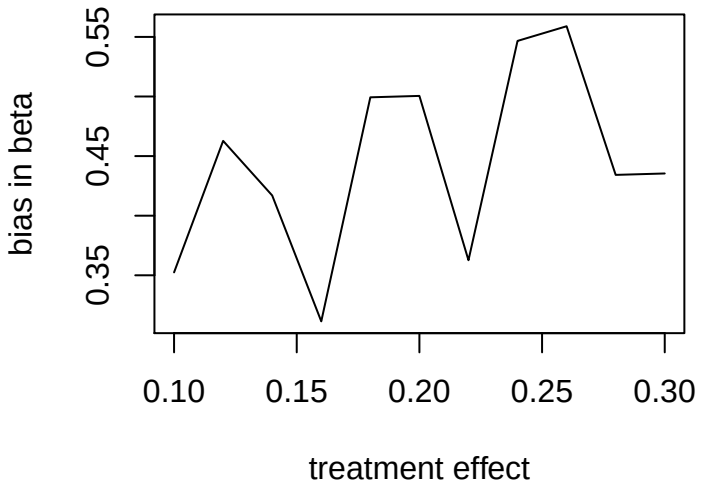

Figure 28: bias plot under simulation setting 9

n1	n2	p1	d	power (Wald)	power (Fisher)
15	15	0.1	0.1	0.063	0.060
15	15	0.1	0.2	0.204	0.153
15	15	0.1	0.3	0.435	0.386

Estimated Power from RCT with unbalanced prognostic factor:

n1	n2	p1	d	power (Wald)	power (Fisher)
15	15	0.1	0.1	0.073	0.055
15	15	0.1	0.2	0.236	0.182
15	15	0.1	0.3	0.474	0.406

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.031	increase	strong
Fisher Exact Test	0.023	decrease	strong

Summary and Conclusion

Association between treatment effect and study power
All simulations demonstrated that larger effect size brings more power for the statistical tests. This fact also complied the principle in statistics.

Association between shift in power and treatment effect
From study #7 and study #9, trace plot of shift in power respect to effect size somehow indicated the

magnitude of shift increased sharply when effect size was between 0.1 and 0.2 and it became more stationary between 0.2 and 0.3. Even though it is difficult to summarize the law of increase in bias, the ER value would have a high reference value for explaning the magnitude of shift. According to the ER table in study #7, under the circumstance that effect size is only 0.1, the emergence of PF imbalance increases the ER in the treatment arm by 0.12. (5.50 - 5.38) As effect size reached to 0.2, the ER gap increased to 0.15! However, in the next 0.1 interval (0.2 - 0.3) where the increase trend of bias stopped, the gap of ER attained 0.16, only increasing by 0.01.

Direction of the effect of prognostic factor (Study #6 vs. Study #7 and Study #8 vs. Study #9)
The conclusion was consistent with 3.1. By changing the direction of the relationship between PF and response, power will change with a same scale but in a different direction.

Distribution of imbalance (Study #6 vs. Study #8 and Study #7 vs. Study #9)
The conclusion was consistent with 3.1. Increasing the size of imbalance would not affect the direction of shift but amplify the size of shift in this direction. Compared to study #6, degree of shift upgraded from weak to moderate. Degree of shift upgraded from moderate to strong from study #7 to study #9.

3.3 Exploration on the impact of response rate in control arm

Simulations are now conducted when group sample size were fixed at 20 each arm with an effect size of 0.15. Relax the response rate in control arm from 0.1 to 0.3. Similar to section 3.1 and 3.2, we attempted to compare and contrast the results by changing the direction of association between PF and response and the distribution of imbalance.

Simulation #10

Simulation Settings:
response rate in control group (p_1): 0.1 - 0.3
effect size (δ): 0.15
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: positive (+)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.03, 0.07)$
total sample size: 40
allocation type: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability($\delta = 0.1$)	Probability($\delta = 0.2$)	Probability($\delta = 0.3$)
Control (0)	No (0)	0.1	0.2	0.3
Treatment (1)	No (0)	0.25	0.35	0.45
Control (0)	Yes (1)	0.182	0.333	0.462
Treatment (1)	Yes (1)	0.4	0.519	0.621

Expected value of number of response in each arm and for each type of study:

PF invention	Group	ER ($p_1 = 0.1$)	ER ($p_1 = 0.2$)	ER ($p_1 = 0.3$)
NO	CONTROL	2.41	4.67	6.81
NO	TREATMENT	5.75	7.85	9.86
YES	CONTROL	2.41	4.67	6.81
YES	TREATMENT	5.6	7.68	9.68

Simulation #10 Results:

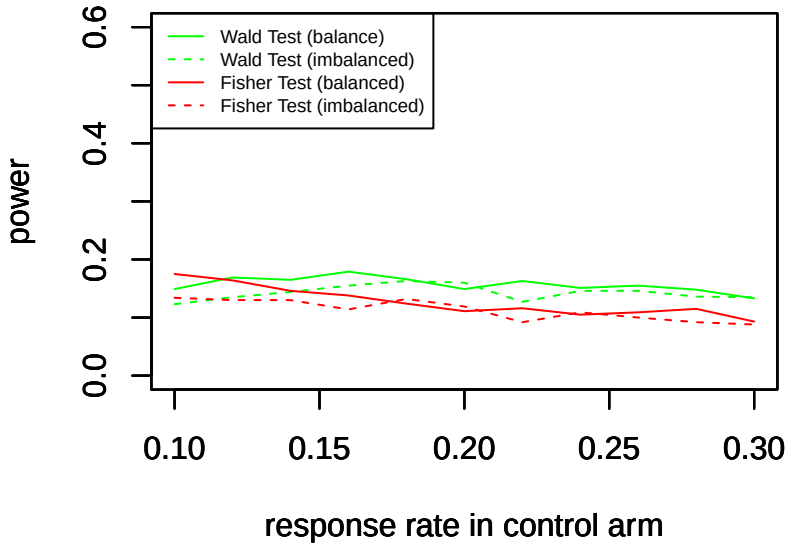

Figure 29: estimated power under simulation setting 10

Estimated Power from RCT with balanced prognostic factor:

p1	power (Wald)	power (Fisher)
0.1	0.149	0.175
0.2	0.149	0.111
0.3	0.133	0.093

Estimated Power from RCT with unbalanced prognostic factor:

p1	power (Wald)	power (Fisher)
0.1	0.123	0.134
0.2	0.160	0.119
0.3	0.135	0.088

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.014	decrease	moderate
Fisher Exact Test	0.014	decrease	moderate

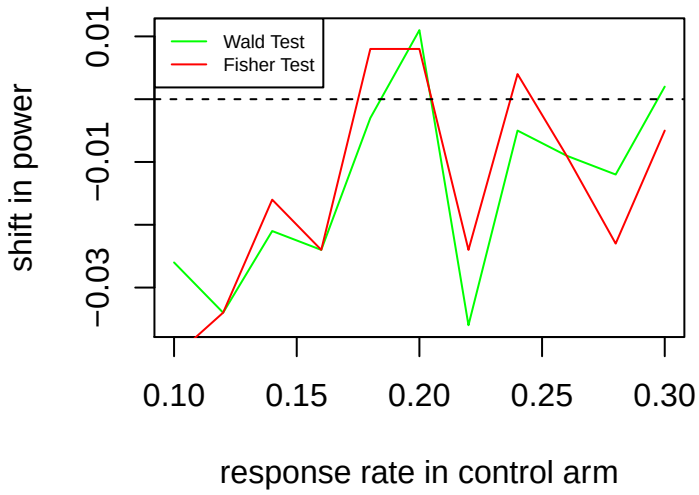

Figure 30: shift in power under simulation setting 10

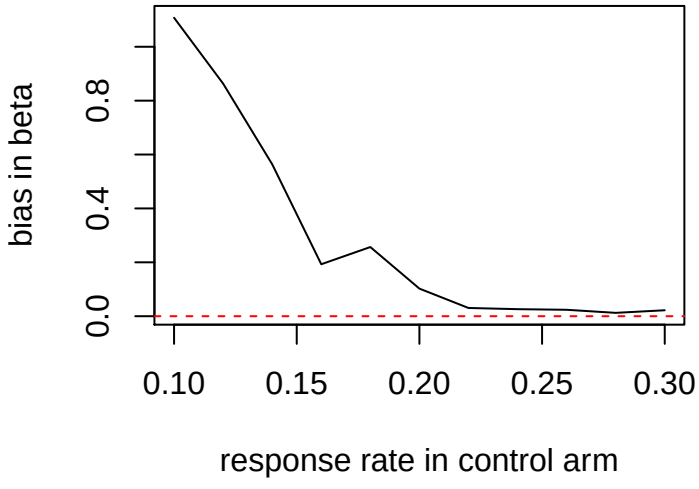

Figure 31: bias plot under simulation setting 10

Simulation #11

Simulation Settings:
response rate in control group (p_1): 0.1 - 0.3
effect size (δ): 0.15
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: negative (-)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.03, 0.07)$
total sample size: 40
allocation: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability($\delta = 0.1$)	Probability($\delta = 0.2$)	Probability($\delta = 0.3$)
Control (0)	No (0)	0.1	0.2	0.3
Treatment (1)	No (0)	0.25	0.35	0.45
Control (0)	Yes (1)	0.053	0.111	0.176
Treatment (1)	Yes (1)	0.142	0.212	0.290

PF invention	Group	ER ($p_1 = 0.1$)	ER ($p_1 = 0.2$)	ER ($p_1 = 0.3$)
NO	CONTROL	1.77	3.56	5.38
NO	TREATMENT	4.46	6.31	8.2
YES	CONTROL	1.77	3.56	5.38
YES	TREATMENT	4.57	6.45	8.36

Simulation #11 Results:

Estimated Power from RCT with balanced prognostic factor:

p1	power (Wald)	power (Fisher)
0.1	0.049	0.129
0.2	0.118	0.090
0.3	0.120	0.092

Estimated Power from RCT with unbalanced prognostic factor:

p1	power (Wald)	power (Fisher)
0.1	0.067	0.123
0.2	0.128	0.099
0.3	0.153	0.113

Evaluation on shift in power:

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.013	increase	moderate
Fisher Exact Test	0.007	increase	weak

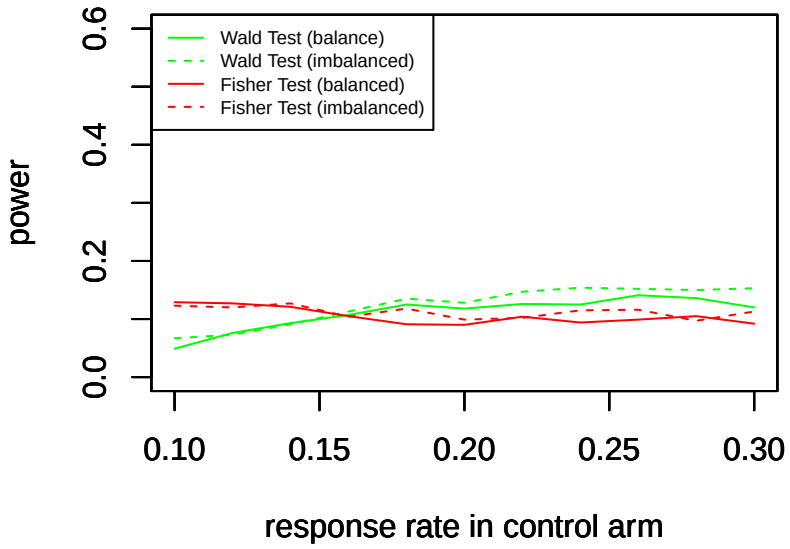

Figure 32: estimated power under simulation setting 11

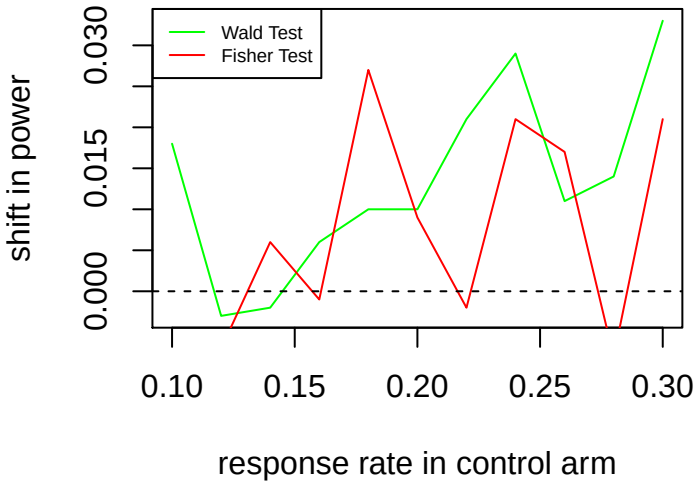

Figure 33: change in power under simulation setting 11

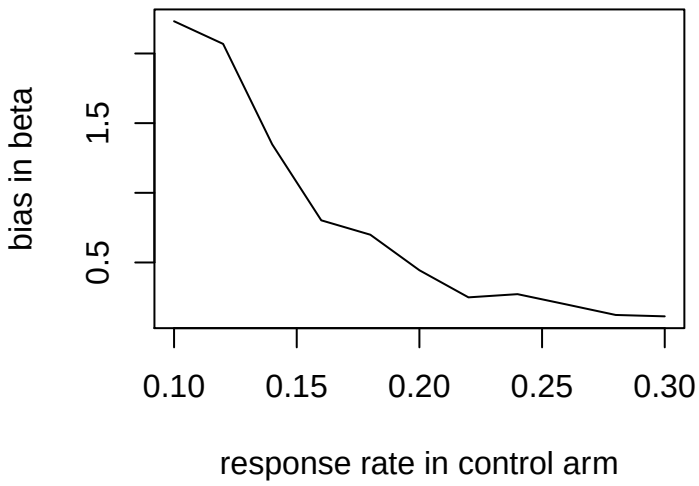

Figure 34: bias plot under simulation setting 11

Simulation #12

Simulation Settings:
response rate in control group (p_1): 0.1 - 0.3
effect size (δ): 0.15
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: positive (+)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.07, 0.1)$
total sample size: 40
allocation type: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability($\delta = 0.1$)	Probability($\delta = 0.2$)	Probability($\delta = 0.3$)
Control (0)	No (0)	0.1	0.2	0.3
Treatment (1)	No (0)	0.25	0.35	0.45
Control (0)	Yes (1)	0.182	0.333	0.462
Treatment (1)	Yes (1)	0.4	0.519	0.621

PF invention	Group	ER ($p_1 = 0.1$)	ER ($p_1 = 0.2$)	ER ($p_1 = 0.3$)
NO	CONTROL	2.41	4.67	6.81
NO	TREATMENT	5.75	7.85	9.86
YES	CONTROL	2.41	4.67	6.81
YES	TREATMENT	5.49	7.56	9.56

Simulation #12 Results

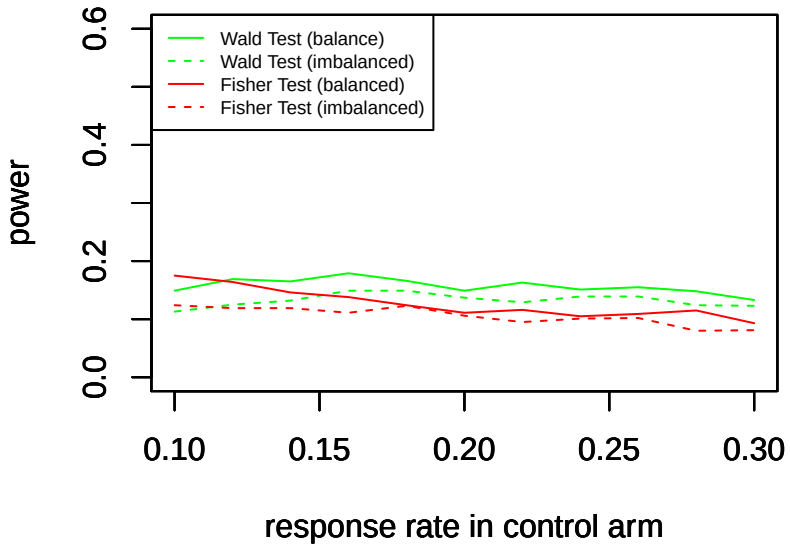

Figure 35: estimated power under simulation setting 12

Estimated Power from RCT with balanced prognostic factor:

p1	power (Wald)	power (Fisher)
0.1	0.149	0.175
0.2	0.149	0.111
0.3	0.133	0.093

Estimated Power from RCT with unbalanced prognostic factor:

p1	power (Wald)	power (Fisher)
0.1	0.113	0.124
0.2	0.137	0.106
0.3	0.123	0.081

Evaluation on shift in power:

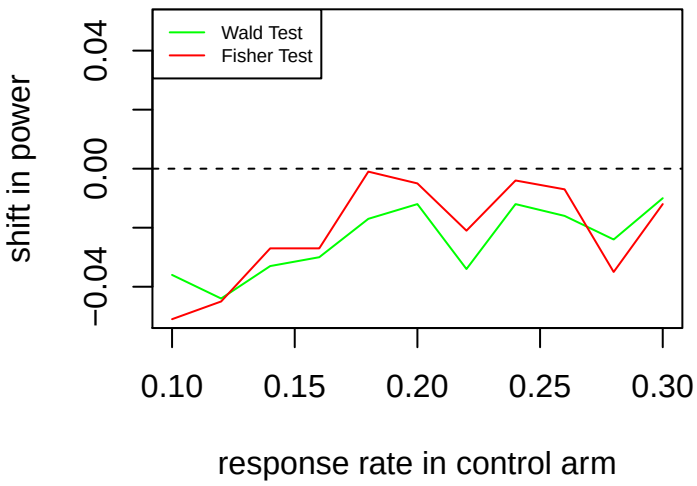

Figure 36: shift in power under simulation setting 12

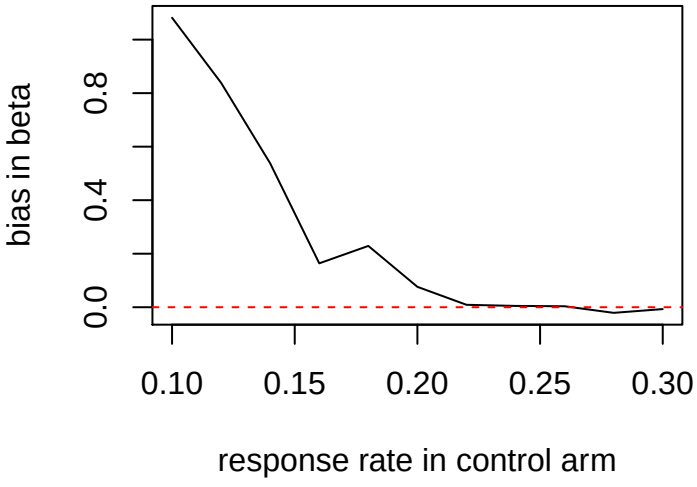

Figure 37: bias plot under simulation setting 12

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.024	decrease	strong
Fisher Exact Test	0.021	decrease	strong

Simulation #13

Simulation Settings:
response rate in control group (p_1): 0.1 - 0.3
effect size (δ): 0.15
odds ratio of effect of prognostic factor (θ): 2
direction of the effect of prognostic factor: negative (-)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.07, 0.1)$
total sample size: 40
allocation: equal allocation

True underlying probability of response for each type of subject:

Treatment Type	Prognostic Factor	Probability($\delta = 0.1$)	Probability($\delta = 0.2$)	Probability($\delta = 0.3$)
Control (0)	No (0)	0.1	0.2	0.3
Treatment (1)	No (0)	0.25	0.35	0.45
Control (0)	Yes (1)	0.053	0.111	0.176
Treatment (1)	Yes (1)	0.142	0.212	0.290

PF invention	Group	ER ($p_1 = 0.1$)	ER ($p_1 = 0.2$)	ER ($p_1 = 0.3$)
NO	CONTROL	1.77	3.56	5.38
NO	TREATMENT	4.46	6.31	8.2
YES	CONTROL	1.77	3.56	5.38
YES	TREATMENT	4.64	6.54	8.47

Simulation #13 Results:

Estimated Power from RCT with balanced prognostic factor:

p1	power (Wald)	power (Fisher)
0.1	0.049	0.129
0.2	0.118	0.090
0.3	0.120	0.092

Estimated Power from RCT with unbalanced prognostic factor:

p1	power (Wald)	power (Fisher)
0.1	0.072	0.127
0.2	0.132	0.104
0.3	0.160	0.117

Evaluation on shift in power:

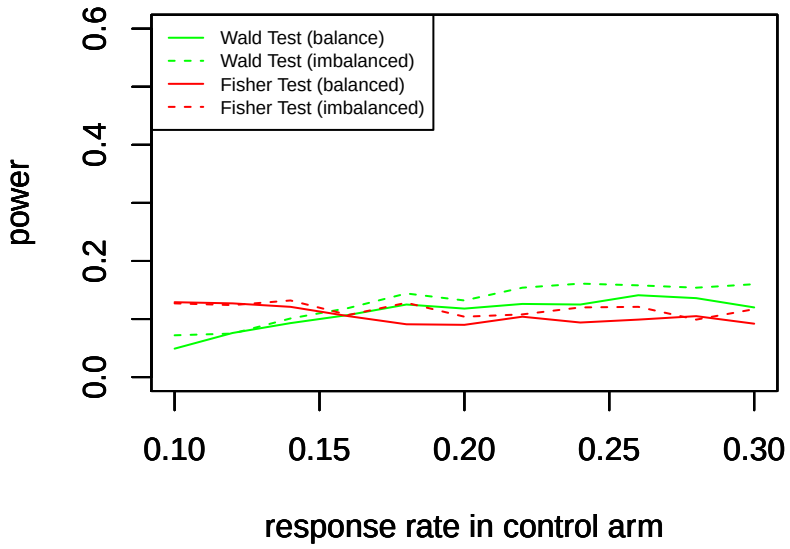

Figure 38: estimated power under simulation setting 13

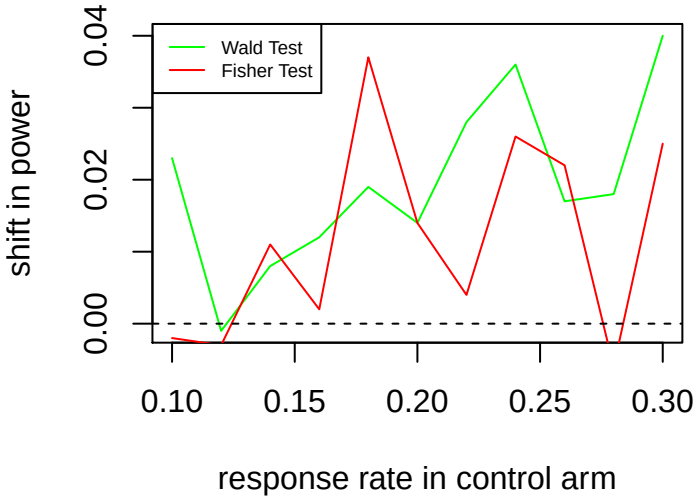

Figure 39: shift in power under simulation setting 13

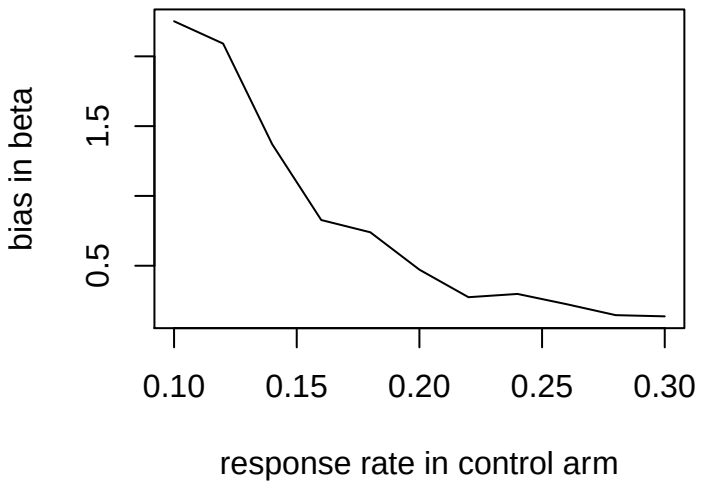

Figure 40: bias plot under simulation setting 13

Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
Wald Test	0.019	increase	moderate
Fisher Exact Test	0.012	increase	moderate

Summary and Conclusion

Association between response rate and study power
Based on statistical theories, a greater the response rate in control arm is more likely to produce a smaller power when holding other parameters are constant. This is because the variance of probabilities at 0.5 are greatest. The trajectory of power (when p_1 goes from 0.1 to 0.3) estimated using Fisher exact test in this section was consistent with our prior knowledge. In contrast, the power measured using Wald test showed a completely different conclusion. It might be the consequence due to the scenario of 20 samples in each group does not seem to comply the normality assumption of Wald test.

Association between shift in power and response rate
The trace plot of shift in power in all four simulation studies demonstrated that no significant relationship between magnitude of bias and response rate could be captured. As the response rate increased, bias in power was randomly distributed around mean magnitude of shift.

Direction of the effect of prognostic factor (Study #10 vs. Study #11 and Study #12 vs. Study #13)
The conclusion was consistent with 3.1 and 3.2. Changing the direction of the association between PF and response reversed the direction of shift.

Distribution of imbalance (Study #10 vs. Study #12 and Study #11 vs. Study #13)
The result was same as 3.1 and 3.2.

3.4 Exploration on the impact of strength effect of PF in logistic regression

Simulation Setting 14

response rate in control group (p_1): 0.2
effect size (δ): 0.1
odds ratio of effect of prognostic factor (θ): 1.5, 2, 2.5
direction of the effect of prognostic factor: positive (+)
prevalence of prognostic factor in control arm: 0.25
distribution of imbalance: $Unif(0.03, 0.07)$
total sample size: 30 - 70
allocation: equal allocation

True underlying probability of response for each type of subject ($\theta = 1.5$):

Treatment Type	Prognostic Factor	Probability
Control (0)	No (0)	0.2
Treatment (1)	No (0)	0.3
Control (0)	Yes (1)	0.272
Treatment (1)	Yes (1)	0.391

True underlying probability of response for each type of subject ($\theta = 2$):

Treatment Type	Prognostic Factor	Probability
Control (0)	No (0)	0.2
Treatment (1)	No (0)	0.3
Control (0)	Yes (1)	0.333
Treatment (1)	Yes (1)	0.462

True underlying probability of response for each type of subject ($\theta = 2.5$):

Treatment Type	Prognostic Factor	Probability
Control (0)	No (0)	0.2
Treatment (1)	No (0)	0.3
Control (0)	Yes (1)	0.385
Treatment (1)	Yes (1)	0.517

Expected value of number of response in each arm and for each type of study ($\theta = 1.5$):

BALANCE STATUS	Group	ER($n_1 = n_2 = 15$)	ER($n_1 = n_2 = 25$)	ER($n_1 = n_2 = 35$)	ER($n_1 = n_2 = 50$)
BALANCED	CONTROL	3.27	5.45	7.63	10.9
BALANCED	TREATMENT	4.84	8.07	11.3	16.14
UNBALANCED	CONTROL	3.27	5.45	7.63	10.9
UNBALANCED	TREATMENT	4.77	7.96	11.14	15.91

Expected value of number of response in each arm and for each type of study ($\theta = 2$):

BALANCE STATUS	Group	ER($n_1 = n_2 = 15$)	ER($n_1 = n_2 = 25$)	ER($n_1 = n_2 = 35$)	ER($n_1 = n_2 = 50$)
BALANCED	CONTROL	3.50	5.83	8.16	11.66
BALANCED	TREATMENT	5.11	8.51	11.92	17.03
UNBALANCED	CONTROL	3.50	5.83	8.16	11.66
UNBALANCED	TREATMENT	4.99	8.31	11.63	16.62

Expected value of number of response in each arm and for each type of study ($\theta = 2.5$):

BALANCE STATUS	Group	ER($n_1 = n_2 = 15$)	ER($n_1 = n_2 = 25$)	ER($n_1 = n_2 = 35$)	ER($n_1 = n_2 = 50$)
BALANCED	CONTROL	3.69	6.16	8.62	12.31
BALANCED	TREATMENT	5.31	8.86	12.40	17.71
UNBALANCED	CONTROL	3.69	6.16	8.62	12.31
UNBALANCED	TREATMENT	5.15	8.59	12.02	17.17

Simulation #14 Results:

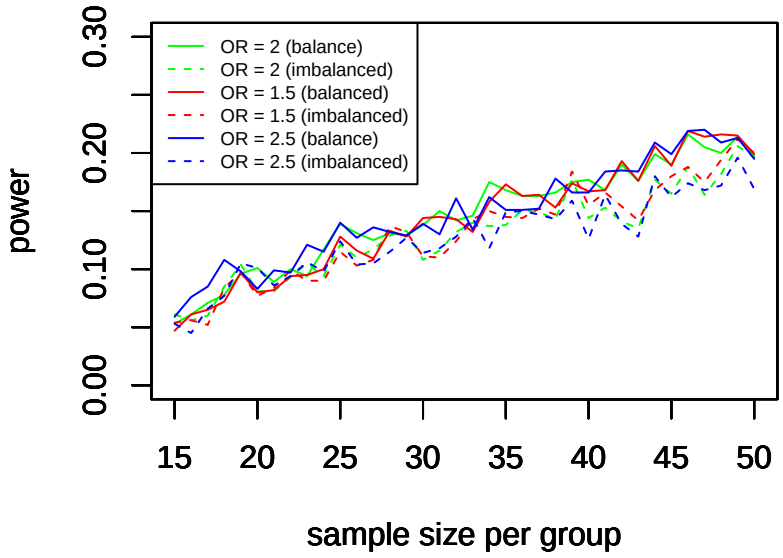

Figure 41: estimated power under simulation setting 14

Estimated Power from RCT with balanced PF and with unbalanced PF when $\theta = 2$:

total sample size	power (W)	power (F)	unbalanced power (W)	unbalanced power (F)
30	0.053	0.037	0.061	0.045
50	0.139	0.091	0.121	0.071
70	0.168	0.115	0.138	0.091
100	0.197	0.162	0.196	0.157

Estimated Power from RCT with balanced PF and with unbalanced PF when $\theta = 1.5$:

total sample size	power (W)	power (F)	unbalanced power (W)	unbalanced power (F)
30	0.047	0.037	0.054	0.049
50	0.128	0.085	0.115	0.069
70	0.173	0.109	0.145	0.101
100	0.199	0.156	0.200	0.165

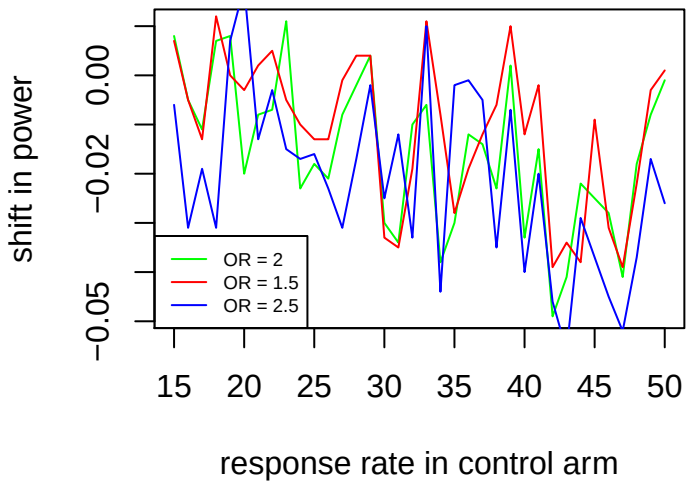

Figure 42: change in power under simulation setting 14

Estimated Power from RCT with balanced PF and with unbalanced PF when $\theta = 2.5$:

total sample size	power (W)	power (F)	unbalanced power (W)	unbalanced power (F)
30	0.059	0.043	0.053	0.031
50	0.140	0.075	0.124	0.075
70	0.151	0.108	0.149	0.097
100	0.195	0.151	0.169	0.134

Evaluation on shift in power:

OR	Method	Mean shift magnitude	Power increase/decrease after imbalance	Degree of shift
2	Wald Test	0.016	decrease	moderate
2	Fisher	0.012	decrease	moderate
1.5	Wald Test	0.011	decrease	moderate
1.5	Fisher	0.009	decrease	weak
2.5	Wald Test	0.021	decrease	strong
2.5	Fisher	0.019	decrease	moderate

Summary and Conclusion

Association between effect of PF and study power
There was no obvious relationship between power and strength of association between PF on the prevalence of response. As odds ratio gets larger or smaller, expected value of response in both arms increases or decreases accordingly.

Association between effect strength of PF and shift in power
There was a strong evidence that size of shift in power associated with θ. The assessment and plot of shift

in power both demonstrated this relationship. The volatility of shift becomes greater as the OR increases. According to the final bias assessment, mean magnitude of power shift was greatest when OR = 2.5 no matter the method for testing treatment effect.

4. Discussion and Limitations

4.1 Summary of Findings

1. A prognostic imbalance does affect the study power compared to a balanced RCT study.

2. Change the direction of association between PF and response reverses direction of the shift in power

3. Increase the prevalence of imbalance (e.g. $Unif(0.03, 0.07)$ -> $Unif(0.07, 0.1)$) changes the magnitude of shift but does not influence the direction of the shift in power

4. Strengthen the association of PF and response does increase the size of the power shift. Conversely, weaken the association decrease the size of shift

5. Unequal sampling allocation between groups results with a smaller power due to a greater variance in test statistics

4.2 Limitations

Mid P-value
In the previous simulation studies,the power was measured and estimated through two kinds of statistical method (Wald and Fisher) to examine the significance of treatment effect. The rationale for that is cross-referencing the results from two different methodologies. Coincidentally, in most of our studies, Fisher test reported a p-value less than that of Wald test. One reasonable explanation is the calculation of p-value for Fisher exact test is equivalent to equation (4) in section 2.3 and that is the default setting in R. However, one limitation is the value of P (x = k) will be very high when the total sample size is small due to a smaller sample space, which will also cause the p-value to be large and difficult to reject the null hypothesis. A recommended adjustment is using mid P-value instead of conventional p-value. Mid P-value [8] reduces the probability of observed count by half. An alternative is applying a one-sided test instead of two-sided (since the treatment effect in our scenario is always positive).

$$Mid\ P-value = 2 * \sum_{k=0}^{x_1-1} P(X_1 = k) + P(X_1 = x_1) \quad for\ a\ two-sided\ test$$

Study Range for Parameters and zero adjustment
In our analysis, there is no significant relationship shown between shift in power and treatment effect, sample size or response rate. This might be the consequence due to a undistinctive study range for parameters. For instance, in Section 3.3, a sample size of 20 was set to each arm which does not comply to Wald test's assessment and cause estimates for power are not robust. The range for response rate was from 0.1 to 0.3. However, it is likely to produce a zero response in control arm and a extremely large value of beta in logistic regression. This also explains why the bias of beta seems to be extremely large when $p_1 = 0.1$. A zero count adjustment might be an effective resolution.

4.3 Future Considerations

In this study, the properties of binary PF imbalance impact on power is discussed. When PF is a continous variable followed normal or exponential distribution. Will the properties discussed comply to a case when PF is in a continous scale? Moreover, in some cases the distribution of PF is unknown, determining the direction and mean of imbalance is still a meaningful question worth to think about.

References

[1] P.C. ELWOOD, RANDOMISED CONTROLLED TRIALS: SAMPLING, *Br. J. clin. Pharmac.*, 13, 631-636 (1982),
[2] Stavros Nikolakopoulos, Johanna H van der Lee, and Ingeborg van der Tweel, Sample size calculations in pediatric clinical trials conducted in an ICU: a systematic review, *Trials*, volume 15, Article number: 274 (2014)
[3] Mahmoud Saghaei, An Overview of Randomization and Minimization Programs for Randomized Clinical Trials, *J Med Signals Sens*, 1(1): 55–61,2011 Jan-Apr
[4] Emmanuel Lesaffre, Superiority, equivalence, and non-inferiority trials, *Bull NYU Hosp Jt Dis*, 66(2):150-4 (2008)
[5] Bolaji E Egbewale, Martyn Lewis and Julius Sim, Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study, *BMC Medical Research Methodology* volume 14, Article number: 49 (2014)
[6] Chenglin Ye, Joseph Beyene, Gina Browne, Lehana Thabane, Estimating treatment effects in randomised controlled trials with non-compliance: a simulation study, *BMJ Open* 4 (2014)
[7] Hae-Young Kim, Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test, *Restor Dent Endod* 2017 May; 42(2): 152–155.
[8J . T. Gene Hwang and Ming-Chung Yang, An optimality theory for mid p-value in $2 \cdot 2$ contingency tables, *Statistica Sinica* 11(2001), 807-826

APPENDIX

R functions for simulation:

```
calculate_p.value <- function(n1, n2, p1, p2, theta, a, b){
  #balanced prognostic factor
  p.pf.control <- 0.25; p.pf.trt <- 0.25
  pf.control <- sample(c(0,1), n1, replace = TRUE,
                       prob = c(1-p.pf.control, p.pf.control))
  pf.trt <- sample(c(0,1), n2, replace = TRUE,
                   prob = c(1-p.pf.trt, p.pf.trt))
  # group allocation
  group <- c(rep(0,n1), rep(1,n2))
  # calculate intercept for logistic regression
  b0 <- log(p1/(1-p1)) # P(y = 1|group = control, pf = 0) = P(y = 1|group = control)
  # coefficient for treatment allocation
  b1 <- log(p2*(1-p1)/(1-p2)/p1)
  # coefficient for prognostic factor
  b2 <- log(theta)
  #create subject data
  data.balance <- data.frame(trt = group, pf = c(pf.control, pf.trt))
  # compute the underlying true probability of response for each patient
  data.balance$p <- exp(b0 + b1*data.balance$trt +
                          b2*data.balance$pf)/(1+exp(b0 +b1*data.balance$trt + b2*data.balance$pf))
  # simulate outcome data
  for (i in 1:nrow(data.balance)){
    data.balance$y[i] <- sample(c(0,1), 1,
                                prob = c(1-data.balance$p[i], data.balance$p[i]))
  }
  # generate imbalance
  e <- runif(1, a, b)
  # set prevalance of pf for control arm
  p.pf.control <- 0.25
  # set prevalance of pf for trt arm
  p.pf.trt <- 0.25 - e
  # generate binary bf data for control arm
  pf.control <- sample(c(0,1), n1, replace = TRUE,
                       prob = c(1-p.pf.control, p.pf.control))
  # generate binary bf data for trt arm
  pf.trt <- sample(c(0,1), n2, replace = TRUE,
                   prob = c(1-p.pf.trt, p.pf.trt))
  # group allocation
  group <- c(rep(0,n1), rep(1,n2))
  # create subject data
  data.imbalance <- data.frame(trt = group, pf = c(pf.control, pf.trt))
  # calculate intercept for logistic regression
  b0 <- log(p1/(1-p1)) # P(y = 1|group = control, pf = 0) = P(y = 1|group = control)
  # coefficient for treatment allocation
  b1 <- log(p2*(1-p1)/(1-p2)/p1)
  # coefficient for prognostic factor
```

```
  b2 <- log(theta)
  # compute the underlying true probability of response for each patient
  data.imbalance$p <- exp(b0 +
                            b1*data.imbalance$trt +
                            b2*data.imbalance$pf)/(1+exp(b0 + b1*data.imbalance$trt + b2*data.imbalance$

  # simulate outcome data
  for (i in 1:nrow(data.imbalance)){
    data.imbalance$y[i] <- sample(c(0,1), 1,
                                  prob = c(1-data.imbalance$p[i], data.imbalance$p[i]))

  }
  #fit a logistic regression
  fit <- glm(y ~ trt, family = "binomial", data = data.balance)
  fit2 <- glm(y ~ trt, family = "binomial", data = data.imbalance)

  # calculate p-value from Wald test
  wald.p.value = coef(summary(fit))[2,4]
  wald.p.value.im = coef(summary(fit2))[2,4]
  beta.bias <- coef(summary(fit2))[2,1] - log(p2*(1-p1)/(1-p2)/p1)
  # calculate p-value from fisher exact test
  x1 <- sum(data.balance[data.balance$trt == 0,]$y == 1)
  x2 <- sum(data.balance[data.balance$trt == 1,]$y == 1)
  x1.im <- sum(data.imbalance[data.imbalance$trt == 0,]$y == 1)
  x2.im <- sum(data.imbalance[data.imbalance$trt == 1,]$y == 1)
  fisher.p.value <- fisher.test(matrix(c(x1, n1-x1, x2, n2-x2), ncol = 2),
                                alternative = "two.sided")$p.value
  fisher.p.value.im <- fisher.test(matrix(c(x1.im, n1-x1.im, x2.im, n2-x2.im), ncol = 2),
                                alternative = "two.sided")$p.value
  return (list(wald = c(wald.p.value, wald.p.value.im),
               fisher = c(fisher.p.value, fisher.p.value.im),
               beta.bias = beta.bias))
}



#simulation program
sim.power.pf <- function(n1,n2,p1,p2,theta,a, b, n.iter = 1000){
  wald.count <- 0
  wald.count.im <- 0
  fisher.count <- 0
  fisher.count.im <- 0
  beta.bias <- 0
  for (i in 1:n.iter){
    res <- calculate_p.value(n1=n1, n2=n2, p1=p1, p2=p2, theta=theta, a = a, b= b)
    if (res$wald[1] <= 0.05){
      wald.count <- wald.count + 1
    }
    if (res$wald[2] <= 0.05){
      wald.count.im <- wald.count.im + 1
    }
    if (res$fisher[1] <= 0.05){
      fisher.count <- fisher.count + 1
```

```
    }
    if (res$fisher[2] <= 0.05){
      fisher.count.im <- fisher.count.im + 1
    }
    beta.bias <- beta.bias + res$beta.bias
  }
  wald.power <- wald.count/n.iter
  wald.power.im <- wald.count.im/n.iter
  fisher.power <- fisher.count/n.iter
  fisher.power.im <- fisher.count.im/n.iter
  wald.bias <- wald.power - wald.power.im
  fisher.bias <- fisher.power - fisher.power.im
  beta.bias <- beta.bias/n.iter
  return (list(wald = c(wald.power, wald.power.im),
               fisher = c(fisher.power, fisher.power.im),
               wald.bias = wald.bias, fisher.bias = fisher.bias,
               beta.bias = beta.bias))
}
```