A Statistical Analysis of Phase II Clinical Trials and Contributing Systematic Bias

Samantha Benedict
Cesar Rene Pabon Bernal
Professor William Williams
Fall 2019: Case Seminar
City University of New York, Hunter College

Abstract

Systemic bias is defined as the inherent tendency of a process to support particular outcomes due to repeatable error often associated with a flawed experimental design. [1] The study of systematic bias is a well documented technique in clinical research, due to its complex layering and repetitive nature.[2] Our objective is to perform statistical analyis of a simulated phase II clinical trial and explore extreme systematic biases, specifically differential non-response. We will examine the encumbering effects which bias can have on the overall advancement of clinical development plans. [3]

1. Introduction

Understanding the structure of clinical research is an essential first step in detecting systematic bias. Clinical trials are often divided into three phases. We define the phases as such:

1. Phase I: the objective is to determine a tolerable dose range of the administered drug pertaining to a clinical trial to evaluate possible side effects – toxicity, pharmacokinetics, bioavailability
2. Phase II: the earliest trials of a drug on a restricted yet homogenous population (test subjects who exhibit similar symptoms)
3. Phase III: (final step) the objective is to determine the efficiency of a drug on a heterogeneous population [3]

Due to experimentation and not just observation present in these phases, clinical trials are fundamentally complicated to dissect. For this reason, we focus only on phase II clinical trials, as they appear to be the most detectable level for differential non-response.[4] Before we examine possible bias, let us first simulate a phase II clinical trial and perform some exploratory analysis on the data.

2. Data

2.1 A Simulated Clinical Trial

We have constructed a hypothetical clinical trial designed to emulate a study published in the American Journal of Respiratory and Critical Care Medicine entitled "Randomized Controlled Trial of Oral Antifungal Treatment For Severe Asthma." [5] The purpose of this clinical trial was to study the effects of antifungal medication on patients with severe asthma.

We randomly simulate a simple two-arm clinical trial, with equal probability and no replacement, to compare a new drug to placebo on reducing the point index, peak flow (PF) of asthma in adults ages 20-69. Peak flow is a breathing measurement used to diagnosis asthma. A peak flow reading below 200 is considered in the asthmatic zone.

The sample size required to detect a specified treatment difference is n = 860 for both treatment groups (drug vs. placebo). For the 860 participants, we record their weight and PF asthma difference just before randomization; weight is an important risk factor linked to higher levels of asthma. [6]

The new drug and placebo are administered and the PF for asthma is measured and recorded periodically. A final PF measurement is recorded at the end of the trial. The change in the PF between the endpoint and the baseline is calculated and used to evaluate the efficiency of the new drug. Appropriate statistical analysis and diagnostics analysis are performed on the above data.

3. Exploratory Analysis

The simulated clinical trial assumes that the PF asthma values for these n = 1720 (n = 860 for each treatment) recruited participants are normally distributed with mean, $\mu = 100$ and standard deviation, $\sigma = 10$. The weight for the participants is assumed to be normally distributed with weight.$\mu = 195$ (lbs) and standard deviation weight.$\sigma = 15$. In addition, we assume the new drug will increase the PF asthma value by increase.$\mu = 200$; the higher the PF value, the easier it becomes for that person to breath. [7]

Tables 1 and 2 provide a snapshot for the first six values of the simulated data for the n placebo and n drug participants, respectively, with weight, baseline PF index (denoted PF.base), endpoint PF index (denoted PF.end), and change in PF index from the baseline to endpoint (denoted PF.diff).

Table 1. SIMULATED PLACEBO INPUT VALUES

Weight	*PF.base*	*PF.end*	*PF.index*
206	*94*	*100*	*7*
187	*109*	*86*	*-23*
170	*103*	*100*	*-3*
209	*111*	*100*	*-11*
176	*104*	*92*	*-12*
206	*85*	*101*	*16*

Table B. SIMULATED DRUG INPUT VALUES

Weight	*PF.base*	*PF.end*	*PF.index*
178	*102*	*-85*	*-188*
209	*97*	*-108*	*-205*
211	*102*	*-119*	*-220*
168	*96*	*-104*	*-201*
177	*118*	*-93*	*-210*
225	*106*	*-99*	*-205*

Figure 1a provides boxplot distributions for data generated placebo. The data appears to be normally distributed with an exception of a few outliers. However, they don't seem to deviate too far away from the mean, and therefore, they will be included in the final analysis. **Figure 1b** provides boxplot distributions for data generated drug. The boxplot for the endpoint is 10 points lower than the PF baseline.

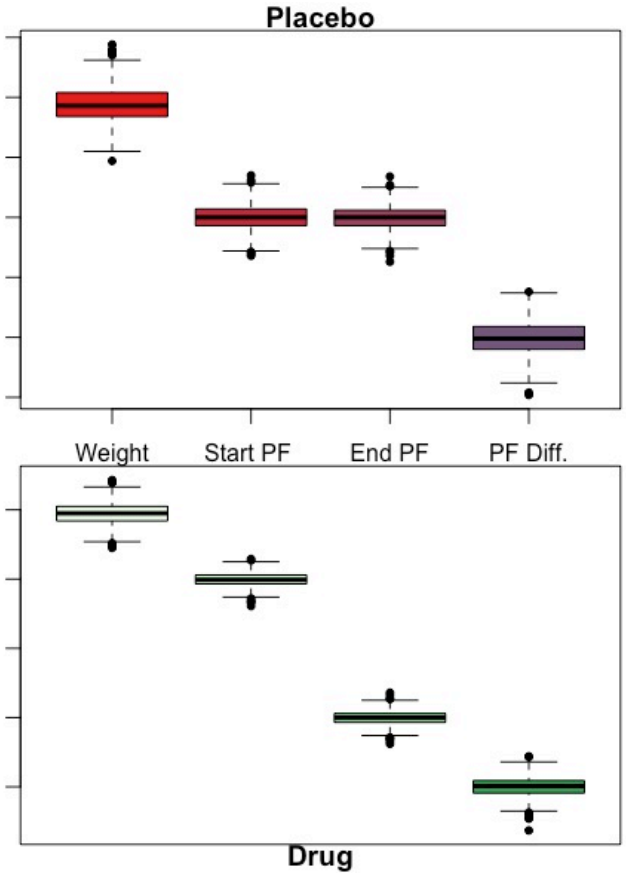

Figure 1a boxplots of placebo (top) and Figure 1b boxplots of drug (bottom)

To explore the relationship between PF of asthma difference as a function of weight for each treatment, we assess whether there exists a statistical significance or not. **Figure 2** concludes that the relationship between PF asthma difference and weight may not be truly significant, however, the new drug did reduce the PF index. The p-value for the difference is 0.879 yet the p-value for the model itself is 2.2e-16 with an adjusted R^2 value of 0.9809.

Figures 3-5 display model diagnostic plots. The Residual vs. Fitted plot, as shown in **Figure 3**, is appropriate for both treatments as they share random evenly spaced distributions. The QQ plot, as shown in **Figure 4,** confirms normality with minimal heavy lower tail action and only a few outliers. **Figure 5**, The Residuals vs. Leverage plot, confirms the above as there are no influential data points that are particularly worth checking for validity, using cook's distance.

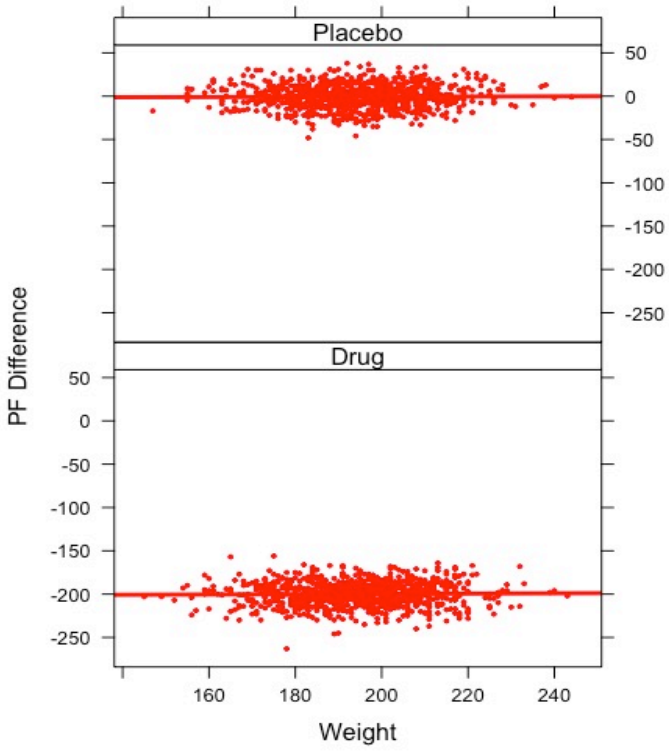

Figure 2 Scatter plots for PF asthma difference as a function of weight for each treatment

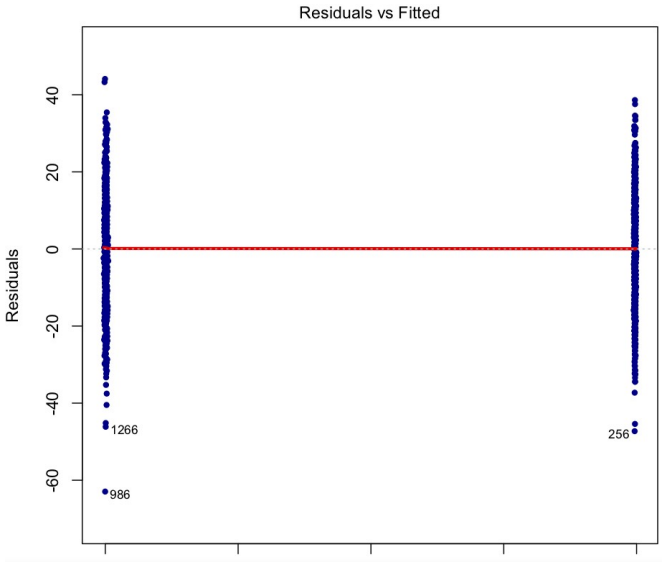

Figure 3 Residual Vs. Fitted plot

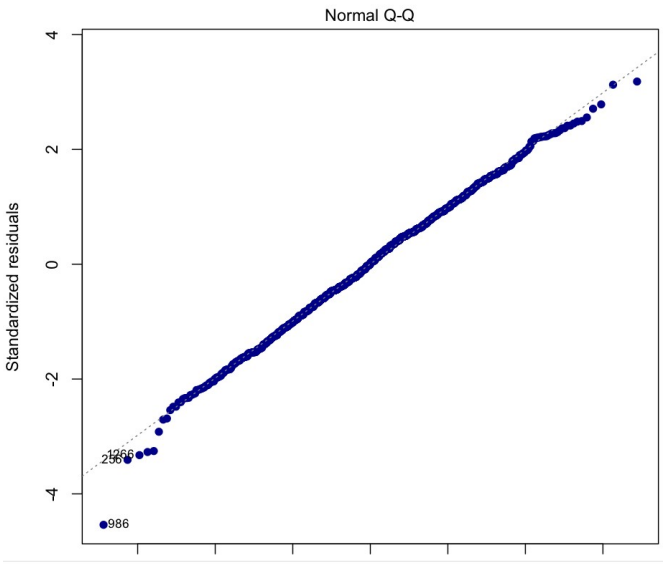

Figure 4 Normal Probability QQ plot

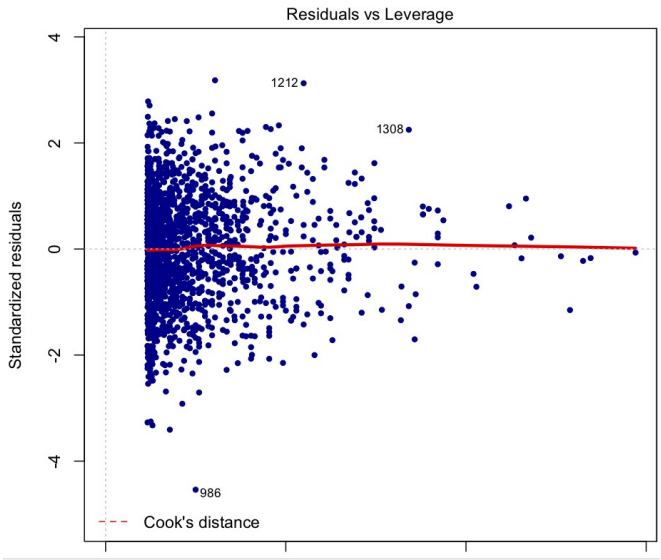

Figure 5 Residuals vs. Leverage plot

Based on the above analysis, further exploration needs to be performed on this model. It appears that an underlying discrepancy exists between placebo, drug, weight and PF of asthma. According to research, evaluation of response and non-response in phase II of clinical trials is an acceptable reason for concern. We explore the latter in the following sections.

3. Modeling

3.1 Statistical Model

Let us now investigate differential non-response and introduce the model in which its detection is made possible. We begin by creating a simple two category statistical model, "Symptoms" and "No-Symptoms," at T_1 and T_2 for phase II clinical trials.

Table 3. NUMBERS OF PATIENTS WITH SYMPTOMS AND NO SYMPTOMS

		Symptom status at time T_2	
		Symptoms	*No Symptoms*
Symptom status at time T_1	*Symptoms*	N_{ss}	N_{sn}
	No Symptoms	N_{ns}	N_{nn}

Using the notation of **Table 3**, the true symptoms rate at T_1, is given by

$$R_1 = (N_{ss} + N_{sn}) / N \qquad \text{Eq. 1}$$

and at T_2 by

$$R_2 = (N_{ss} + N_{ns}) / N \qquad \text{Eq. 2}$$

where

$$N = N_{ss} + N_{ns} + N_{sn} + N_{nn} \qquad \text{Eq. 3}$$

When the actual sampling for the selected trial phase begins, the classification of the data will not be accurate since not all patients designated for the research will take part in the study. Therefore, a more involved model is necessary. In **Table 4**, a simple three category statistical model, "Symptoms," "No-Symptoms," and "Lost to Follow-up" at T_1 and T_2 for phase III clinical trials, is introduced. After T_2 of a phase III trial, nine frequencies in this table will be known. We assume here that patients who do not take part in the study at either T_1 or T_2 can still be counted, so that the frequency F_{oo} is known.

Table 4. OBSERVED NUMBER OF PATIENTS IN VARIOUS CATEGORIES

		Symptom status at time T_2		
		Symptoms	*No Symptoms*	*Lost to Follow-up*
Symptom Status at time T_1	*Symptoms*	F_{ss}	F_{sn}	F_{so}
	No Symptoms	F_{ns}	F_{nn}	F_{no}
	Lost to Follow-up	F_{os}	F_{on}	F_{oo}

We state at this point that due to the complexity and privacy of much of the data collected from clinical trails, Foo either remains unknown or a rough estimate of it may become available. However, its value does not affect the findings of the rest of the eight frequencies and subsequently, its respective estimations of the symptom rates.

We construct an elementary expectation model and extend **Table 4**. Let:

P_n = Probability that a "No-Symptoms" person actually appears at T_1

P_s = Probability that a "Symptoms" person actually appears at T_1

P_{nn} = Probability that a drug was administered and an individual appeared at T_2, given that the drug was administered and that that he/she appeared at T_1 and showed "No-Symptoms" at both T_1 and T_2

P_{sn} = Probability that a drug was administered and an individual appeared at T_2 given that the drug was administered and that that he/she appeared at T_1 and showed "Symptoms" at T_1 and "No-Symptoms" at T_2

P_{ns} = Probability that a drug was administered and an individual appeared at T_2, given that the drug was administered and that he/she appeared at T_1 and showed "No-Symptoms" at T_1 and "Symptoms" at T_2

P_{ss} = Probability that a drug was administered and an individual appeared at T_2, given that the drug was administered and that that he/she appeared at T_1 and showed "Symptoms" at both T_1 and T_2

Finally, let Q_{ss}, Q_{sn}, Q_{ns}, and Q_{nn} represent probabilities similar to P_{ss}, P_{sn}, P_{ns}, and P_{nn}, except that Q's are conditional on the patient not showing symptoms at T_1. Ideally, each of these probabilities would equal its counterpart (ex. $P_{ss}=Q_{ss}$) because all patients in the clinical trial are theoretically included in the sample. However, lost to follow-up problems will almost always ensure that P's and Q's are not always in unity. Therefore, exploration of expected sample numbers is an attractive option given three-by-three classification analysis. These expectations are displayed in **Table 5.**

Table 5. EXPECTED SAMPLE NUMBERS

		Symptom status at time T_2		
		Symptoms	*No Symptoms*	*Lost to Follow-up*
Symptom status at time T_1	*Symptoms*	$N_{ss}P_sP_{ss}$	$N_{sn}P_sP_{sn}$	$N_{ss}P_s(1-P_{ss}) + N_{sn}P_s(1-P_{sn})$
	No Symptoms	$N_{ns}P_nP_{ns}$	$N_{nn}P_nP_{nn}$	$N_{ns}P_n(1-P_{ns}) + N_{nn}P_n(1-P_{nn})$
	Lost to Follow-up	$N_{ss}(1-P_s)Q_{ss} + N_{ns}(1-P_n)Q_{ns}$	$N_{sn}(1-P_s)Q_{sn} + N_{nn}(1-P_n)Q_{nn}$	$N_{ss}(1-P_s)(1-Q_{ss}) + N_{sn}(1-P_s)(1-Q_{sn}) + N_{ns}(1-P_n)(1-Q_{ns}) + N_{nn}(1-P_n)(1-Q_{nn})$

3.2 The Study of Identical Patients

Table 4 can be used to construct an estimator based only on patients who are observed both at T_1 and T_2. It can be seen that the number of patients who have symptoms at T_1 and are found to either be lost to follow-up or not at T_2 is given by $F_{nn} + F_{ns}$. Consequently, the number of patients who have no symptoms at T_1 and either have been lost to follow-up or not at T_2 is given by $F_{sn} + F_{ns}$. We can then state that the symptoms rate at T_1, based only on those patients who appear both at T_1 and T_2, is given by:

$$\hat{R}_1 = (F_{ss} + F_{sn}) / F \quad \text{Eq. 4}$$

where

$$F = F_{ss} + F_{sn} + F_{ns} + F_{nn} \quad \text{Eq. 5}$$

The symptom rate at T_2 for this same group of identical patients is given by:

$$\hat{R}_2 = (F_{ss} + F_{ns}) / F \quad \text{Eq. 6}$$

where

$$F_{sn} = F_{ns} \quad \text{Eq. 7}$$

Since we assume that we are exploring a large clinical trial, under the model of section 2, we neglect sampling variability and **Eq. 7** becomes

$$N_{sn}P_sP_{sn} = N_{ns}P_nP_{ns} \quad \text{Eq. 8}$$

In this case there is no overall change in no-symptoms response and **Eq. 8** can be expressed as

$$P_s/P_n = P_{ns}/P_{sn} \quad \text{Eq. 9}$$

The change in proportions of the observed symptom rate will occur even though there is no change in the true symptom rate. An easier way to interpret **Eq.9** is by,

$$P_{1s}/P_{1n} = P_{2s}/P_{2n} \quad \text{Eq. 10}$$

Where the ratio of the probability of a patient attending a trial as demonstrating symptoms to the probability of a patient attending a trial as showing no-symptoms must be the same at T_1 and T_2. Otherwise, there will be a change in the expected symptom rate from T_1 and T_2.

In addition, we state, that due to the lack of information in determining the theoretical proposals of P_s and Q_s, we find difficulties in benchmarking our results and state summaries as is.

3.3 Comparison of the estimates based on identical, unmatched (single), and all patients

The estimate based on "single" patients and the estimate based on all available patients can be formed by estimating $\hat{R}_{1t}$ and $\hat{R}_{2t}$. Their values are as such:

$$\hat{R}_{1t}/1-\hat{R}_{1t} = (F_{ss} + F_{sn} + F_{so})/(F_{ns} + F_{nn} + F_{no}) \approx [(N_{ss} + N_{sn})/(N_{ns} + N_{nn})]^* (P_{1s}/(P_{1n}) \quad \text{Eq. 11}$$

and

$$\hat{R}_{2t}/1-\hat{R}_{2t} = (F_{ss} + F_{sn} + F_{so})/(F_{ns} + F_{nn} + F_{no}) \approx [(N_{ss} + N_{ns})/(N_{ns} + N_{nn})]^* (P_{2s}/(P_{2n}) \quad \text{Eq. 12}$$

The estimates $\hat{R}_{1m}$, $\hat{R}_{2m}$ based on patients who appear only on a single occasion are given (in the simple case of independence) by,

$$\hat{R}_{1m},/1-\hat{R}_{1m} = (F_{sn})/(F_{ns}) \approx ([N_{ss}(1-P_{2s}) + N_{sn}(1-P_{2n})]/[N_{ns}(1-P_{2s}) + N_{nn}(1-P_{2n})])^* (P_{1s}/(P_{1n}) \quad \text{Eq. 13}$$

$$\hat{R}_{2m},/1-\hat{R}_{2m} = (F_{os})/(F_{on}) \approx ([N_{ss}(1-P_{1s}) + N_{ns}(1-P_{1n})]/[N_{sn}(1-P_{1s}) + N_{nn}(1-P_{1n})])^* (P_{2s}/(P_{2n}) \quad \text{Eq. 14}$$

According to Williams,[2] $\hat{R}_1 - \hat{R}_2$, $\hat{R}_{1t} - \hat{R}_{2t}$, and $\hat{R}_{1m} - \hat{R}_{2m}$ do not have easy algebraic reductions and can be studied numerically. Below are comments about those results.

4. Experimental

We've structured the following examples based on the phase II simulated clinical trial data from our previous exploratory analysis. Sample size is held constant in each example however we've shifted the values and probabilities between examples in effort to draw comparisons between each. We will now explore the models in which we've just defined in application with the simulated data and comment on the results of this experiment as they may pertain to the effects that dropouts may have on clinical trials.

Example A

First Stage Response Probabilities:

$P_n = 0.94$ $P_s = 0.88$

Second Stage Probabilities:

$P_{nn} = 0.97$ $P_{ns} = 0.83$ $P_{sn} = 0.95$ $P_{ss} = 0.88$

$Q_{nn} = 0.97$ $Q_{ns} = 0.83$ $Q_{sn} = 0.95$ $Q_{ss} = 0.88$

True Population Figures:

$N_{nn} = 710$ $N_{ns} = 30$ $N_{sn} = 30$ $N_{ss} = 90$

Table A1. EXPECTED SAMPLE NUMBERS

		Symptom status at time T_2			
		Symptoms	*No Symptoms*	*Lost to Follow-up*	*Total*
Symptom status at time T_1	*Symptoms*	*70*	*25*	*11*	*106*
	No Symptoms	*23*	*647*	*25*	*695*
	Lost to Follow-up	*11*	*45*	*3*	*59*
	Total	*104*	*717*	*39*	*860*

Table A2. Estimates of Symptom Rates (Percent)

R	*TRUE*	*Identicals*	*Total*	*Singles*
R_1	*14*	*11.05*	*15.18*	*38.74*
R_2	*14*	*10.81*	*14.17*	*23.83*
$R_1 - R_2$	*0*	*0.24*	*1.01*	*14.91*

Example B

First Stage Response Probabilities:

$P_n = 0.94$ $P_s = 0.88$

Second Stage Probabilities:

$P_{nn} = 0.95$ $P_{ns} = 0.55$ $P_{sn} = 0.95$ $P_{ss} = 0.55$
$Q_{nn} = 0.95$ $Q_{ns} = 0.35$ $Q_{sn} = 0.95$ $Q_{ss} = 0.35$

True Population Figures:

$N_{nn} = 710$ $N_{ns} = 30$ $N_{sn} = 30$ $N_{ss} = 90$

Table B1. EXPECTED SAMPLE NUMBERS

		Symptom status at time T_2			
		Symptoms	*No Symptoms*	*Lost to Follow-up*	*Total*
Symptom status at time T_1	*Symptoms*	*44*	*25*	*37*	*106*
	No Symptoms	*16*	*634*	*46*	*696*
	Lost to Follow-up	*4*	*44*	*11*	*59*
	Total	*64*	*703*	*94*	*861*

Table B2. Estimates of Symptom Rates (Percent)

R	*TRUE*	*Identicals*	*Total*	*Singles*
R_1	*14*	*8.01*	*15.18*	*80.24*
R_2	*14*	*6.97*	*9.39*	*15.79*
$R_1 - R_2$	*0*	*1.04*	*5.79*	*64.45*

Example C

First Stage Response Probabilities:

$P_n = 0.94$ $P_s = 0.88$

Second Stage Probabilities:

$P_{nn} = 0.97$ $P_{ns} = 0.83$ $P_{sn} = 0.95$ $P_{ss} = 0.88$
$Q_{nn} = 0.97$ $Q_{ns} = 0.83$ $Q_{sn} = 0.95$ $Q_{ss} = 0.88$

True Population Figures:

$N_{nn} = 350$ $N_{ns} = 80$ $N_{sn} = 80$ $N_{ss} = 350$

Table C1. EXPECTED SAMPLE NUMBERS

		Symptom status at time T_2			
		Symptoms	*No Symptoms*	*Lost to Follow-up*	*Total*
Symptom status at time T_1	*Symptoms*	*271*	*67*	*40*	*378*
	No Symptoms	*62*	*319*	*23*	*404*
	Lost to Follow-up	*41*	*29*	*7*	*77*
	Total	*374*	*415*	*70*	*859*

Table C2. Estimates of Symptom Rates (Percent)

R	*TRUE*	*Identicals*	*Total*	*Singles*
R_1	*50*	*39.35*	*93.62*	*191.15*
R_2	*50*	*38.77*	*87.37*	*133.62*
$R_1 - R_2$	*0*	*0.58*	*6.25*	*57.53*

4. Results

As previously stated, in all of the above examples we have held our sample size constant and additionally, used identical first stage probabilities in each in effort to draw comparison. In both examples A and B we have kept our true population figures constant leaving us with a true, constant symptom rate of 14%. In example C, however, we raised our values to obtain a constant symptom rate of 50%, in order to explore what happens when our true symptom rate increases.

In example A, our total symptom rates at T_1 and T_2 are 15.18 and 14.17 respectively, which are both very close to our true symptom rates. The identical symptom rates are also fairly close to the true rates however, the singles rates are significantly larger than true. In this example, we've used identical P and Q values.

In example B, however, we have decreased our probabilities as well as varied the probabilities between P's and Q's. While both the identical and total rates are still fairly close to true, this has a drastic effect on the singles rates, causing them to skyrocket to more than 4 times the true rate.

Example C is the least representative example, but perhaps the most telling. Unlike Examples A and B, Example C represents a true population rate of 50%. The rates for identical are vastly underrepresented while the rates for total and singles grossly overrepresent the true symptom rates. Like Example A, Example C uses equal values for P's and Q's but represents an evidently biased estimate of symptom rates.

5. Discussion

In this report, the preliminary exploratory analysis of a phase II clinical trial for the treatment of asthma was performed. Through statistical analysis, the simulated clinical data showed an underlying bias. On this same set of data, this bias was explored for possible differential non-response and tested with a pre-existing model. [8] Although it is still unclear as to how to eliminate systematic bias in clinical trials due to dropouts, we can definitively state that this bias does in fact exist and greatly impacts the analysis of clinical data. Some known factors contribute to its presence.

Estimating rates with a low true symptom rate results in seemingly unbiased estimated rates. However, as the true population of patients exhibiting symptoms increases, our estimates of symptom rates become increasingly more biased. While population figures appear to play a role in the presence of bias, conditional probabilities play a seemingly small role in the estimation of symptom rates in clinical trials.

In this paper, we only explored the bias existing in phase II clinical trials. In future work, we will consider both phase I and phase III of clinical trials. Theses phases are prone to selection bias, a phenomenon that occurs when the study population does not reflect a representative sample of the target population; classification bias, resulting from improper recording of individual factors; and confounding bias, a false association made between the outcome and a factor that is not itself causally related to the outcome. [9]

All of these biases can produce extreme discrepancies in overall results and ultimately deliver insufficient or potentially harmful yields. Statistical analysis on data from these other phases should be executed in order to develop a model in an effort to correct some of the systematic bias that is so prevalent in clinical trials.

References

1. Stephanie. “Systematic Error / Random Error: Definition and Examples.” *Statistics How To*, 29 Mar. 2019, https://www.statisticshowto.datasciencecentral.com/systematic-error-random-error/.

2. Williams, W. H., and C. L. Mallows. “Systematic Biases in Panel Surveys Due to Differential Nonresponse.” Journal of the American Statistical Association, vol. 65, no. 331, 1970, p. 1338., doi:10.2307/2284300.

3. Chen, Ding-Geng, and Karl E. Peace. *Clinical Trial Data Analysis Using R*. CRC Press, 2011.

4. Guthery, Stephen. “Faculty of 1000 Evaluation for Why Most Clinical Research Is Not Useful.” F1000 - Post-Publication Peer Review of the Biomedical Literature, 2016, doi:c/f.726438120.793520174.

5. Denning, D. W., Odriscoll, B. R., Powell, G., Chew, F., Atherton, G. T., Vyas, A., . . . Niven, R. M. (2009). Randomized Controlled Trial of Oral Antifungal Treatment for Severe Asthma with Fungal Sensitization. American Journal of Respiratory and Critical Care Medicine, 179(1), 11-18. doi:10.1164/rccm.200805-737oc

6. Gordon, Stephen B., et al. “Respiratory Problems in the Tropics.” *Manson's Tropical Infectious Diseases (Twenty-Third Edition)*, W.B. Saunders, 21 Oct. 2013, https://www.sciencedirect.com/science/article/pii/B9780702051012000716.

7. Millar, W J. “Distribution of Body Weight and Height: Comparison of Estimates Based on Self Reported and Observed Measures.” *Journal of Epidemiology & Community Health*, vol. 40, no. 4, Jan. 1986, pp. 319–323., doi:10.1136/jech.40.4.319.

8. Williams, W.h. “Selection Biases in Fixed Panel Surveys.” Contributions to Survey Sampling and Applied Statistics, 1978, pp. 89–112., doi:10.1016/b978-0-12-204750-3.50014-9.

9. Lambert, Jerome. “Statistics in Brief: How to Assess Bias in Clinical Studies?” *Clinical Orthopaedics and Related Research®*, vol. 469, no. 6, 2010, pp. 1794–1796., doi:10.1007/s11999 010-1538-7.