1-2021

Observational studies in group testing and potential applications.

Alexander Christopher Noll
University of Louisville

OBSERVATIONAL STUDIES IN GROUP TESTING AND POTENTIAL APPLICATIONS

By
Alexander Christopher Noll
B.S. in Mathematics, 2017

A Thesis
Submitted to the Faculty of the
School of Public Health and Information Sciences of the
University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

May 2021

OBSERVATIONAL STUDIES IN GROUP TESTING AND POTENTIAL APPLICATIONS

By

Alexander Christopher Noll
B.S. in Mathematics, 2017

Thesis approved on

April 27, 2021

by the following Thesis Committee:

Thesis Director
Dr. Qi Zheng

Dr. Karunarathna Kulasekera

Dr. Maiying Kong

Dr. Aruni Bhatnagar

DEDICATION

For Amy

λ

ACKNOWLEDGMENTS

Thank you to Dr. Qi Zheng of the University of Louisville for guiding me through the process of this project. The guidance and theoretical framework that he provided for the conduction of the simulations described in this thesis were indispensable.

ABSTRACT

OBSERVATIONAL STUDIES IN GROUP TESTING AND POTENTIAL APPLICATIONS

Alexander Christopher Noll

April 27, 2021

The use of group testing to identify individuals with targeted outcomes in a population can greatly improve the efficiency, speed, and cost effectiveness of testing a population for an outcome, or at least for identifying the prevalence of an outcome in a population. The implementation of causal inference techniques can provide the basis for an observational study that would allow an investigator to gather estimates for treatment effectiveness if group testing was conducted on the population in a certain way.

This thesis examines a simulation of the above outlined principles in order to demonstrate a potential application for determining treatment efficacy from observational data obtained via testing for disease outcome in a partially treated population. It is made evident that it is reasonable to make conclusions about treatment conditional incidence of an outcome in a sample of tested individuals based on outcome tests conducted on groups of individuals. Examining group study observations in the manner described in this thesis will allow researchers to estimate treatment effectiveness from partial data in situations where outcome testing may have been limited or where quick results are required from limited data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1 Introduction

The problem of determining the effectiveness of treatments has long been studied by academics. The principles of causal inference allow the investigator to make an unbiased estimate of the average treatment effect while accounting for the confounder that at risk individuals are priortized to receive treatment. However, the effort associated with gathering necessary observations to obtain reliable estimates can result in tremendous expenditure in terms of both time and cost. This problem can be intensified when the incidence rate of the outcome in question is very low, or if the treatment is particularly effective as these conditions would result in the necessity of collecting a much larger sample size in order to achieve a similar estimate as that resulting from a situation with a larger incidence rate in the treated or untreated populations. It is therefore necessary to consider alternative methods for gathering the samples.

Group testing proposed by Dorfman (1943) is a method of disease testing that has been utilized in the past in order to reduce the total number of tests that need to be used in order to test an entire population for the incidence of a disease. It works by dividing a population into groups and performing the test for the disease simultaneously on combined samples from each group. Assuming perfect sensitivity and specificity, if a group tests negative for the disease, it is obvious that to no one in the group has the disease and they can be excluded from further testing. Groups whose combined sample tested positive are therefore implied to contain at least one individual who has the disease. These groups can then be either split into smaller groups for more group tests with the process repeated, or instead then tested at the individual level. There are more complex methods than this that improve the efficiency of the process and can even account for non-perfect sensitivity and specificity (e.g., Lin et al., 2019). However, these methods are beyond the scope being discussed here.

2 Group Testing Applied to Causal Inference

When considering a situation where the effectiveness of a treatment needs to be rapidly assessed it is reasonable to look towards the concepts of group testing for a possible solution. One may consider a situation where an emergent disease has become pandemic and the process for testing the disease is very resource-intensive, very time consuming, or reliant on difficult to obtain materials or special equipment. In this situation it would certainly be important for public health authorities to reduce the

required number of tests that need to be run in order to isolate diseased individuals. By testing individuals in a group setting using a method with sufficient accuracy, the number of tests that are utilized can be reduced. By making use of data from a group testing scenario and implementing causal inference concepts, a consistent estimate of treatment efficacy can be achieved.

The mathematics behind this theory are not the focus of this thesis. Instead, we have examined the use of these principles in a simulated environment in order to examine their practical applications in a hypothetical situation. This process will create a groundwork for future situations where this methodology is necessary. Of particular concern is the impact that the group testing will have on estimation accuracy. While the traditional approach to causal inference is unbiased, we seek to determine whether a practically unbiased result can be achieved with a reasonable sample size in a group testing environment. Additionally, we seek to determine guidelines for how large groups can be in order to further minimize testing.

CHAPTER II

METHODOLOGY

1 Definitions

Let $\mathbf{X}$ denote the vector of q pre-treatment covariates for a subject in the study, A and Y denote, respectively, the treatment received (e.g., getting COVID-19 vaccinated) and the observed outcome for the subject. In this project, we only consider the case with 2 treatments: control ($A = 0$) and treated ($A = 1$) and the binary outcome: $Y = 0$ or 1 (whether the subject is infected; 0 denotes no infection and 1 denotes infection). Each subject would then have had 2 potential outcomes $Y^{(0)}$ and $Y^{(1)}$, where $Y^{(a)}$ would be the outcome if the subject receives the treatment a, $a \in \{0, 1\}$. However, as each subject can only receive one treatment, the observed outcome is the potential outcome corresponding to the treatment assigned.

Let $p_a^* := E[Y^{(a)}]$, that is the average infectious rate of the whole population receiving treatment a. We can use the average treatment effect (ATE), defined as $p_0^* - p_1^* = E[Y^{(0)}] - E[Y^{(1)}]$ to assess the effectiveness of the treatment. For example, in the evaluation of the effectiveness of COVID-19 vaccines, the ATE measures the reduction of the average infectious rate led by the vaccine. We would like to accurately estimate ATE for group testing data in observational studies.

Suppose there are M groups in the group testing data and there are N_i subjects in the ith group. We denote the covariates, treatment received, and the outcome of the jth subject by $\mathbf{X}_{ij}, A_{ij}$, and Y_{ij}, respectively. In group testing data, we do not directly observe the outcome of each subject (i.e., Y_{ij}). Instead, we observe $\tilde{Y}_i$, the outcome of the group, that is, whether there is any subject in the group being infected. It is easy to see that $\tilde{Y}_i = \max_{1 \le j \le N_i} Y_{ij}$.

Let $D_i = \{(\tilde{Y}_i, \mathbf{X}_{ij}, A_{ij}), j = 1, \cdots, N_i\}$ denote the data observed for the ith group. The whole data we observe is $D = \{D_i, i = 1, \cdots, M\}$.

2 Randomized Controlled Trials

We first consider the randomized controlled trials (RCT), that are considered as the gold standard to determine the treatment effect between different treatment groups. In an RCT, the subjects are randomly assigned to different treatment groups and all confounding baseline covariates, either measured or unmeasured, are assumed to be balanced. Therefore, p_a^* and ATE can be directly estimated (Friedman et al., 2015).

In RCT, we observe the following

$$\begin{aligned} &P(Y=1|A=a) = E\left[Y|A=a\right] = E\left[E[Y|A=a,\mathbf{X}]|A=a\right] \\ &= E\left[E[Y^{(a)}|A=a,\mathbf{X}]|A=a\right] = E\left[E[Y^{(a)}|\mathbf{X}]|A=a\right] \\ &= E\left[E[Y^{(a)}|\mathbf{X}]\right] = E[Y^{(a)}] = p_a^*, \end{aligned}$$

where the third equality follows from the consistency (Hernán and Robins, 2010), the fourth equality follows from the conditional exchangeability (Hernán and Robins, 2010), and the fifth equality follows from the fact that the assignment is independent of $\mathbf{X}$ in RCT. Thus,

$$\begin{aligned} &P(\tilde{Y}_i = 0|A_{ij}=a_{ij}, j=1,\cdots,N_i) = P(Y_{ij}=0|A_{ij}=a_{ij}, j=1,\cdots,N_i) \\ &= \prod_{j=1}^{N_i} P(Y_{ij}=0|A_{ij}=a_{ij}) = \prod_{j=1}^{N_i}(1-P(Y_{ij}=1|A_{ij}=a_{ij})) = \prod_{j=1}^{N_i}(1-p^*_{a_{ij}}), \end{aligned}$$

and then $P(\tilde{Y}_i = 1|A_{ij}=a_{ij}, j=1,\cdots,N_i) = 1-\prod_{j=1}^{N_i}(1-p^*_{a_{ij}})$. And the likelihood function using the data $D^{(0)} := \{(\tilde{Y}_i, A_{ij}), j=1,\cdots,N_i, i=1,\cdots,M\}$ is

$$L(\mathbf{p}|D^{(0)}) = \prod_{i=1}^{M}\left[\prod_{j=1}^{N_i}(1-p_{A_{ij}})\right]^{1-\tilde{Y}_i}\left[1-\prod_{j=1}^{N_i}(1-p_{A_{ij}})\right]^{\tilde{Y}_i},$$

where $\mathbf{p} = (p_0, p_1)^{\mathrm{T}}$. Consequently, the log-likelihood function is

$$\ell(\mathbf{p}|D^{(0)}) = \sum_{i=1}^{M}\left\{(1-\tilde{Y}_i)\left(\sum_{j=1}^{N_i}\log\left(1-p_{A_{ij}}\right)\right) + \tilde{Y}_i\log\left[1-\prod_{j=1}^{N_i}(1-p_{A_{ij}})\right]\right\}. \quad (1)$$

If the testing is homogeneous, that is, for each sample j in the group i, all values A_{ij}'s are the same, and $N_i = J$ are the same for $i=1,\cdots,M$, then

$$\begin{aligned} &L(\mathbf{p}|D^{(0)}) \\ &= \prod_{i=1}^{M}\left[1-p_1\right]^{J(1-\tilde{Y}_i)A_i}\left[1-p_0\right]^{J(1-\tilde{Y}_i)(1-A_i)}\left[1-(1-p_1)^J\right]^{\tilde{Y}_iA_i}\left[1-(1-p_0)^J\right]^{\tilde{Y}_i(1-A_i)}, \end{aligned}$$

and the log-likelihood function is

$$\begin{aligned} \ell(\mathbf{p}|D^{(0)}) \quad &= \sum_{i=1}^{M} J(1-\tilde{Y}_i)A_i\log(1-p_1) + \sum_{i=1}^{M} J(1-\tilde{Y}_i)(1-A_i)\log(1-p_0) \\ &\quad + \sum_{i=1}^{M}\tilde{Y}_iA_i\log\left(1-(1-p_1)^J\right) + \sum_{i=1}^{M}\tilde{Y}_i(1-A_i)\log\left(1-(1-p_0)^J\right). \end{aligned}$$

Simple algebra yields the following maximum likelihood estimators:

$$\hat{p}_0 = 1-\left(\frac{\sum_{i=1}^{M}(1-\tilde{Y}_i)(1-A_i)}{\sum_{i=1}^{N}(1-A_i)}\right)^{1/J} \quad \text{and} \quad \hat{p}_1 = 1-\left(\frac{\sum_{i=1}^{M}(1-\tilde{Y}_i)A_i}{\sum_{i=1}^{M}A_i}\right)^{1/J},$$

which coincide with the estimators of p_a^* for the individual testing data.

Therefore, we can easily estimate the ATE for group testing data in RCT.

3 Observational Studies

In practice, conducting an RCT is not always feasible due to many restrictive factors, including ethics, logistical constraints, and patient preferences (Horwitz, 1987). On the other hand, the rapid advance of technology in collecting and storing data makes the observational studies widely available in natural health care settings, which urges us to develop new methods to estimate the ATE.

In observational studies, the treatment received by a subject may be determined by the subject's characteristics, which may also have a significant impact on the outcome. Thus, the covariates between different treatment groups may be unbalanced, and the difference in the outcomes between the two treatment groups is not only attributed to the treatment received (Rubin, 2004). For example, in an observational study evaluating the efficacy of COVID-19 treatments, patients with severe preexisting respiratory diseases were more likely to receive new treatments. Therefore, patients in the treatment group are more likely to have severe respiratory diseases than those without respiratory diseases. Meanwhile, patients with more severe respiratory diseases tend to have a shorter survival time. If one merely compares the simple average survival time between the treatment group and the no-treatment group without controlling for respiratory disease, he or she may reach an erroneous conclusion about the effectiveness of treatments. Covariates that affect both the choices of treatment and outcomes, such as the severity of respiratory diseases in COVID-19 studies, are named confounders in statistics. Thus, we have to control for the confounding factors in order to correctly estimate p_a^* for observational studies(Rosenbaum and Rubin, 1983). In our numerical analysis, we also showed that severe bias would appear if the confounding factors were not adjusted.

Although estimating ATE has been widely studied for individual data in observational studies, the related development for group testing data is relatively sparse. In this project, we develop a new method that can consistently estimate ATE for group testing data in observational studies.

Since the seminal work by Rosenbaum and Rubin (1983), propensity-score-based inverse probability weighting (IPW) method and generalized propensity score (GPS) (Imbens, 2000) have become two popular approaches to control confounding factors and estimating ATE. The term GPS refers to the probability of receiving some treatment assignments conditional on the observed baseline covariates. The alignment of GPS across different treatment groups balances confounders and thus approximates the conditional RCT under the exchangeability condition (Hernán and Robins, 2010). The GPS is often estimated parametrically by logistic models, or nonparametrically by random forest and generalized boosting methods (McCaffrey et al., 2013).

Let $\pi_a(\mathbf{x}) = P(A = a|\mathbf{X} = \mathbf{x}), a = 0, 1$, which are the propensity scores for a subject having characteristics $\mathbf{x}$. We define the propensity scores for the ith group

as follows:

$$W_i = \prod_{j=1}^{N_i} \left(\sum_{a=0}^{1} 1\{A_{ij} = a\} \pi_a^{-1}(\mathbf{X}_{ij}) \right)$$
$$= \prod_{j=1}^{N_i} \left(1\{A_{ij} = 1\} \pi_1^{-1}(\mathbf{X}_{ij}) + 1\{A_{ij} = 0\} \pi_0^{-1}(\mathbf{X}_{ij}) \right) = h^*(D_i).$$

Then we incorporate the proposed group propensity scores into the log-likelihood function for the RCT (1). Consider the following likelihood function.

$$\ell(\mathbf{p}|D) = \sum_{i=1}^{M} \left\{ (1 - \tilde{Y}_i) \left(\sum_{j=1}^{N_i} \left(\sum_{a=0}^{1} 1\{A_{ij} = a\} \log(1 - p_a) \right) \right) \right.$$
$$\left. + \tilde{Y}_i \log \left[1 - \prod_{j=1}^{N_i} \left(\sum_{a=0}^{1} 1\{A_{ij} = a\}(1 - p_a) \right) \right] \right\} W_i, \qquad (2)$$

Take the derivative with respect to p_1 and we obtain that

$$\sum_{i=1}^{M} W_i \left\{ \sum_{j=1}^{N_i} 1\{A_{ij} = 1\} \left[\frac{-(1 - \tilde{Y}_i)}{1 - p_1} + \tilde{Y}_i \frac{\prod_{k \neq j} \left(\sum_{a=0}^{1} 1\{A_{ik} = a\}(1 - p_a) \right)}{1 - \prod_{k=1}^{N_i} \left(\sum_{a=0}^{1} 1\{A_{ij} = a\}(1 - p_a) \right)} \right] \right\},$$

which motivates us to consider the estimating equations $\mathbf{S}(\mathbf{p}) = (S_0(\mathbf{p}), S_1(\mathbf{p}))^{\mathrm{T}} = 0$, where

$$S_0(\mathbf{p}) := M^{-1} \sum_{i=1}^{M} \left[\sum_{j=1}^{N_i} 1\{A_{ij} = 0\} W_i \left\{ -(1 - \tilde{Y}_i) + \prod_{k=1}^{N_i} \left(\sum_{a=0}^{1} 1\{A_{ik} = a\}(1 - p_a) \right) \right\} \right] \qquad (3)$$

and

$$S_1(\mathbf{p}) := M^{-1} \sum_{i=1}^{M} \left[\sum_{j=1}^{N_i} 1\{A_{ij} = 1\} W_i \left\{ -(1 - \tilde{Y}_i) + \prod_{k=1}^{N_i} \left(\sum_{a=0}^{1} 1\{A_{ik} = a\}(1 - p_a) \right) \right\} \right]. \qquad (4)$$

In the Appendix, we show that $\mathbf{S}(\mathbf{p}) = 0$ are valid estimating equations.

In practice, W_i's are unknown but can be estimated either parametrically by logistic models or nonparametrically by random forest and generalized boosting methods (McCaffrey et al., 2013). Let $\hat{h}$ be some estimate of h^*. Then W_i can be estimated by $\hat{W}_i = \hat{h}(D_i)$. Moreover, replace W_i in $\mathbf{S}(\mathbf{p})$ by $\hat{W}_i$ and we obtain the empirical estimating equations $\hat{\mathbf{S}}(\mathbf{p})$.

4 Theoretical results

Given a random sample $Z_1, \cdots, Z_M$, we adopt the following empirical process notations. Let $\mathbb{G}_M(f) = \mathbb{G}_M(f(Z_t)) := M^{-1/2} \sum_{t=1}^{M} (f(Z_t) - E[f(Z_t)])$ and $\mathbb{E}_M f(Z_t) := \sum_{t=1}^{M} M^{-1} f(Z_t)$. We use $\tilde{\mathbf{p}}$ and $\hat{\mathbf{p}}$ to denote some solutions of $\mathbf{S}(\mathbf{p}) = 0$ and $\hat{S}(\mathbf{p}) = 0$, respectively.

Regularity conditions

(A1) (a) Conditional exchangeability: $(Y^{(0)}, Y^{(1)}) \perp A|\mathbf{X}$; (b) Consistency: if $A = a$, $Y^{(A)} = Y^{(a)} = Y$; (c) Positivity: $\pi_1(\mathbf{x}) = P(A = a|\mathbf{X} = \mathbf{x}) > \nu > 0$ for all values of $\mathbf{x}$ that $f(\mathbf{x}) > 0$, where $f(\cdot)$ is the density of $\mathbf{X}$ and ν is some constant less that 1/2.

(A2) Boundedness of $\mathbf{p}^*$: $0 < \nu \leq \min\{p_0^*, p_1^*\} \leq \max\{p_0^*, p_1^*\} \leq \frac{1}{2} - \nu$.

(A3) Group size: $2 \leq N_i \leq N$, where N satisfies (a) $N < 1/\nu$ and (b) $\forall p_0, p_1 \in [\nu, 1/2-\nu]$, $4((N+1)-Np_0-p_1)((N+1)-p_0-Np_1)-(N-1)^2(2-p_0-p_1)^2 > \nu$.

Remark 4.1 *Condition (A1) has been widely assumed in observational studies (e.g., Hernán and Robins, 2010; Yan et al., 2019). The group size conditions (A3) are needed to ensure that the monotonicity of the estimation equations* $\boldsymbol{S}(\boldsymbol{p}) = 0$ *(Fygenson and Ritov, 1994). Simple algebra shows that Condition (A3(b)) is satisfied when* $N = 2, 3, 4, 5, 6$ *under Condition (A2).*

Let $\dot{S}_{bc}(\mathbf{p}) = \partial S_b(\mathbf{p})/\partial p_c, b = 0, 1; c = 0, 1$ and

$$\dot{\mathbf{S}}(\mathbf{p}) = \begin{pmatrix} \dot{S}_{00}(\mathbf{p}) & \dot{S}_{01}(\mathbf{p}) \\ \dot{S}_{10}(\mathbf{p}) & \dot{S}_{11}(\mathbf{p}) \end{pmatrix},$$

Theorem 4.1 *Under Conditions (A1) – (A3), (i) there exists solutions* $\tilde{\boldsymbol{p}}$ *of* $\boldsymbol{S}(\boldsymbol{p}) = 0$ *satisfying* $\tilde{\boldsymbol{p}} \to_p \boldsymbol{p}^*$*; (ii)* $\sqrt{M}\,(\hat{\boldsymbol{p}} - \boldsymbol{p}^*) \to_d N(0, V\boldsymbol{\Sigma}V^{\mathrm{T}})$*, where* $V = E[\dot{\boldsymbol{S}}(\boldsymbol{p}^*)]^{-1}$.

Corollary 4.1 *If* $\|\hat{h} - h^*\|_\infty = O_p(M^{-1/2})$*, under Conditions (A1) – (A3), there exists solutions* $\hat{\boldsymbol{p}}$ *of* $\hat{\boldsymbol{S}}(\boldsymbol{p}) = 0$ *satisfying* $\hat{\boldsymbol{p}} \to_p \boldsymbol{p}^*$.

Corollary 4.1 indicates that our proposed estimator is consistent.

5 Simulated Observational Study

We conduct simulated studies to examine the performance of the proposed method.

The purpose of the simulated observational study is to replicate the conditions a case where a vaccine is being deployed to prevent incidence of a pandemic disease. A simulated population will be assigned treatment according to risk factors. Then, the population will be assigned incidence of the event according to risk factors, with risk negated by administration of the treatment. Following this process, a simulated group testing model will be applied and results of the estimated ATE will be compared to the actual ATE. This will allow us to assess the accuracy of our proposed method.

6 Population Parameters

To simplify the process of simulating the population, only parameters that impact the incidence of treatment and outcome. While there are many covariates that could go into a situation such as the one being examined in this thesis, we will only consider two confounders, and the treatment assignment impacted by these factors. The outcome variable will also be present and will be impacted by the two confounders and the treatment assignment.

Variable	Distribution Type	Variable Parameters	Notes
Age (A)	Normal	$\mu = 39; \sigma = 12$	Negative ages are possible in this model but should not impact outcomes of simulation.
Preexisting Condition (PC)	Bernoulli	$p = .2$	A "Success" indicates that the sample has a preexisting condition.
Treatment (T)	Bernoulli	$p = \frac{1}{1+e^{-(-6.5+.1*A+2*PC)}}$	p is modeled after a logistic regression model and indicates the probability of receiving treatment.
Outcome (O)	Bernoulli	$p = \frac{1}{1+e^{-(-7.2+.1*A+3*PC-5*T)}}$	p is modeled after a logistic regression model and indicates the probability of an outcome disease).

Table 1. Relevant Population Parameters

The selection of the parameters that define the above factors was intended to create a realistic population that was facing a treatment shortage and did not have a perfect treatment. In the creation of the treatment and outcome variables, the intention was for a 65 year old sample with no preexisting condition to have a 50 percent chance of receiving the treatment. It can also be seen that an individual having a preexisting condition is equivalent to that same individual being 20 years older with no preexisting condition. In the design of the outcome feature, the intent was for a typical untreated sample to have a roughly 10 percent chance of experiencing

the outcome. Additionally, it was important that the treatment not be completely effective, and for there still to be at least a 1 percent chance of a typical treated individual to experience the outcome.

7 Group Selection

We consider homogeneous grouping here, that is, all groups have the same group size and groups consist entirely of either treated or of untreated samples. For a generated population of any given size, the population is split up into equal sized groups (with excess samples that could not be assigned a group discarded). For the purpose of this simulated observational study, group assignment is entirely random, saved for the consideration of treatment.

8 Heterogeneous versus Homogeneous Groups

In this study, all groups consist of either treated or untreated samples. In a different situation where groups do not consist of all of the same treatment statuses, the estimates of p_0^* and p_1^* do not have a closed form but can still be obtained by solving (3) and (4) using a searching algorithm. This is beyond the scope of this project. Instead, we will be examining the case with homogeneous groups. The calculation for probability of infection given treatment based on these cases does have a closed and is calculable using trivial arithmetic.

9 Probability Estimation

In a real observational study it would be very unlikely for the investigators to have information on the true probabilities of the outcome. However, as this study is simulated and we have set the parameters ourselves it is trivial to obtain true ATE. To do this, we calculate the individual probabilities of getting $Y = 1$ and 0 for each subject as if they had either received or had not received the treatment.

We restate the notations here. Let Y be the random variable indicating the outcome, such that $Y_{ij} = 1$ indicates that sample j in group i has the disease (e.g. tests positive for COVID-19) and $Y_{ij} = 0$ indicates that sample j in group i does not have the disease. Let A_{ij} be the random variable indicating treatment, such that $A_{ij} = 1$ indicates that sample j in group i has received treatment (e.g. has received the COVID-19 vaccine) and $A_{ij} = 0$ indicates that sample j in group i has not received treatment. With these definitions, let $p_a(\mathbf{X}_{ij})$ indicate the probability that an individual with the conditions for sample $\mathbf{X}_{ij}$ tests positive for the disease given either that they have received treatment if $a = 1$ or that they have not received treatment if $a = 0$. So, for each individual we calculate

$$p_0(\mathbf{X}_{ij}) = \frac{1}{1 + e^{-(-6.5 + .1 * A + 2 * PC)}} \tag{5}$$

and

$$p_1(\mathbf{X}_{ij}) = \frac{1}{1 + e^{-(-7.2 + .1 * A + 3 * PC - 5)}} \tag{6}$$

Using these values, we can then calculate

$$p_a^* \approx \frac{1}{N} \sum_{k=1}^{N} p_a(\mathbf{X}_{ij}) \tag{7}$$

which is the probability of infection for the sample population of size N.

Recall that $\pi_a(\mathbf{X}_{ij}) = P(A_{ij} = a|\mathbf{X}_{ij})$ is the probability of a subject with characteristics of $\mathbf{X}_{ij}$ getting treatment a. We use a logistic regression model: Treat $\sim$ Age + Preexisting Condition to estimate $\pi_a(\mathbf{X}_{ij})$ and obtain the estimates $\hat{\pi}_a$. Then $\hat{p}_a, a = 0, 1$ obtained by solving (3) and (4) are

$$\hat{p}_0 = 1 - \left(\frac{\sum_{i=1}^{M} (1 - \tilde{Y}_i) \prod_{j=1}^{J} 1\{A_{ij} = 0\} \hat{\pi}_0^{-1}(\mathbf{X}_{ij})}{\sum_{i=1}^{M} \prod_{j=1}^{J} 1\{A_{ij} = 0\} \hat{\pi}_0^{-1}(\mathbf{X}_{ij})} \right)^{1/J} \tag{8}$$

and

$$\hat{p}_1 = 1 - \left(\frac{\sum_{i=1}^{M} (1 - \tilde{Y}_i) \prod_{j=1}^{J} 1\{A_{ij} = 1\} \hat{\pi}_1^{-1}(\mathbf{X}_{ij})}{\sum_{i=1}^{M} \prod_{j=1}^{J} 1\{A_{ij} = 1\} \hat{\pi}_1^{-1}(\mathbf{X}_{ij})} \right)^{1/J}. \tag{9}$$

Derivations of these results can be found in Appendix A.

10 Simulation Process

Simulations were run 1,000 times each for reach combination of three sample sizes and twenty group sizes for a total of 60,000 simulations. Represented samples sizes were 2,000, 5,000, and 10,000. Group sizes ranged between 1 and 20. For each simulation the true and estimated probabilities were collected.

CHAPTER III

RESULTS AND ANALYSIS

1 Process for Examination of Results

Following the generation of each sample, the ATE and its estimates were calculated as $d^* = p_0^* - p_1^*$ and $\hat{p} = \hat{p}_0 - \hat{p}_1$. We examine the performance of the proposed method by calculating the average of the absolute bias (Bias), standard error (SE), and means square error (MSE) over one thousands samples.

2 Results

The results of the simulations can be found in Table 2. These results are also presented in Figures 1,2, and 3.

	Sample Size								
	2,000			5,000			10,000		
Group Size	MSE	SE	Bias	MSE	SE	Bias	MSE	SE	Bias
1	0.0001	0.01014	0.00105	0.00004	0.00668	0.00053	0.00003	0.00522	0.00008
2	0.00023	0.01479	0.00285	0.00006	0.00769	0.00077	0.00004	0.00632	0.00018
3	0.00027	0.01651	0.00106	0.00016	0.01284	0.00127	0.00009	0.00925	0.0002
4	0.00043	0.02035	0.00376	0.00042	0.02012	0.00426	0.00016	0.01273	0.00146
5	0.001	0.03146	0.00302	0.00041	0.02005	0.00243	0.00024	0.01527	0.00284
6	0.00194	0.04302	0.00937	0.00102	0.03161	0.00441	0.00034	0.01807	0.00342
7	0.0036	0.05627	0.0208	0.00195	0.04231	0.01245	0.00066	0.02488	0.00623
8	0.00103	0.03186	0.00399	0.0015	0.03644	0.01323	0.00099	0.03034	0.00844
9	0.00384	0.05717	0.02381	0.00156	0.03744	0.01262	0.00089	0.02876	0.00808
10	0.00487	0.06434	0.02709	0.00408	0.05868	0.02526	0.0014	0.03539	0.0121
11	0.00489	0.06228	0.03174	0.00299	0.05013	0.02177	0.00185	0.0402	0.01525
12	0.00502	0.06636	0.02492	0.00306	0.05006	0.02351	0.00197	0.04156	0.01549
13	0.00342	0.05443	0.02136	0.00367	0.05787	0.01793	0.00207	0.04148	0.01866
14	0.00704	0.07654	0.03438	0.00404	0.0552	0.0315	0.00277	0.04799	0.02156
15	0.00727	0.07926	0.03141	0.00336	0.05403	0.02094	0.00241	0.04471	0.02023
16	0.0064	0.07084	0.03723	0.00349	0.04978	0.03185	0.00333	0.0513	0.02638
17	0.00637	0.06618	0.04462	0.0066	0.07197	0.03762	0.00377	0.05487	0.02758
18	0.0114	0.09545	0.04783	0.00438	0.0594	0.02917	0.00415	0.05664	0.03068
19	0.02393	0.13395	0.07739	0.00708	0.07041	0.04605	0.00454	0.05897	0.03267
20	0.01306	0.10168	0.05218	0.00555	0.06265	0.04036	0.00499	0.06073	0.03612

Table 2. Simulation Results

Figure 1. Error Estimate Results for 1,000 Sample Size 2,000 Simulations

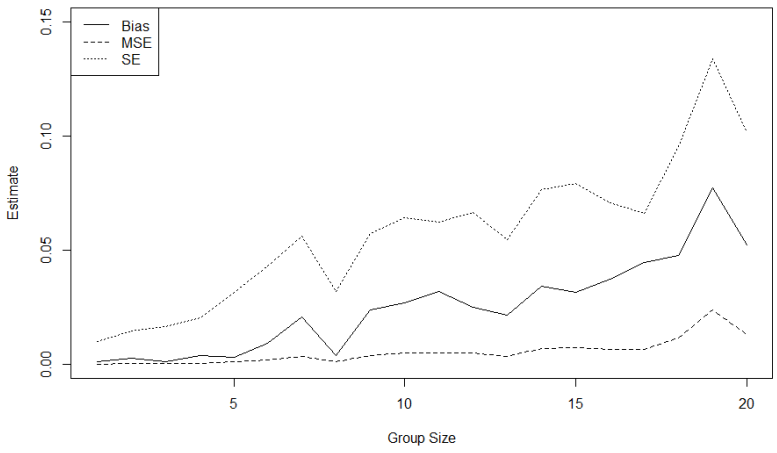
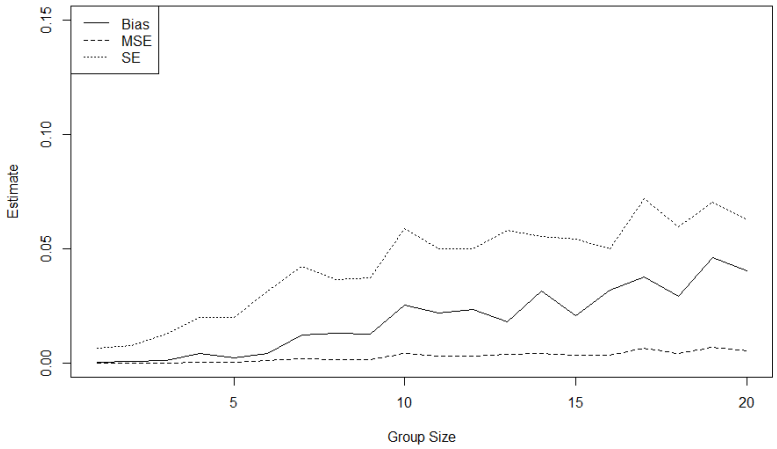


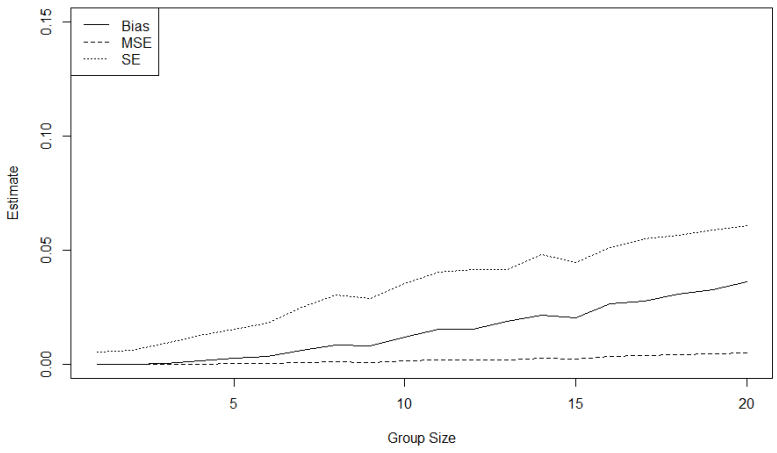
Figure 2. Error Estimate Results for 1,000 Sample Size 5,000 Simulations



3 Naive Estimation

Performing the simulation process again while recalculating infection probabilities using a naive methodology allows us to see the importance of accounting for probability of receiving treatment. For the naive estimation, we estimate p_0^* and p_1^* using the same method but with $\hat{W}_i = 1$, i.e, without considering the confounders. The results are presented in Table 3. It can be seen that the bias of the estimate rises to an unacceptable level, no matter how large the sample size is. It can therefore be seen that the controlling confounders are definitely needed in observational studies.

Figure 3. Error Estimate Results for 1,000 Sample Size 10,000 Simulations



	Sample Size								
	2,000			5,000			10,000		
Group Size	MSE	SE	Bias	MSE	SE	Bias	MSE	SE	Bias
1	0.00784	0.00929	0.08807	0.00796	0.00453	0.08912	0.00788	0.00371	0.08867
2	0.00759	0.00872	0.08666	0.00789	0.00511	0.08869	0.00781	0.00412	0.0883
3	0.00799	0.01084	0.08871	0.00786	0.00541	0.08847	0.00788	0.00423	0.08869
4	0.00794	0.00902	0.08866	0.00788	0.00577	0.08856	0.00788	0.00405	0.08866
5	0.0077	0.00939	0.08727	0.00795	0.00645	0.08891	0.0079	0.00423	0.08876
6	0.0078	0.00906	0.08785	0.00794	0.00619	0.0889	0.00782	0.00445	0.08831
7	0.00811	0.01021	0.08946	0.0077	0.00551	0.08755	0.00785	0.00424	0.08849
8	0.00805	0.01078	0.08906	0.00785	0.00699	0.0883	0.00776	0.00426	0.08797
9	0.00815	0.01022	0.08969	0.00803	0.00691	0.08933	0.00781	0.00477	0.08823
10	0.00802	0.01125	0.08886	0.00793	0.00618	0.08883	0.00785	0.00498	0.08847
11	0.00833	0.0114	0.09057	0.00781	0.0075	0.08806	0.00765	0.00499	0.0873
12	0.00818	0.01174	0.0897	0.00758	0.00705	0.08676	0.00788	0.00559	0.08859
13	0.00778	0.01176	0.08742	0.00788	0.00662	0.08854	0.00788	0.00499	0.08863
14	0.00818	0.01292	0.08953	0.00805	0.00703	0.08947	0.00792	0.00522	0.08885
15	0.00798	0.00982	0.08882	0.00811	0.00816	0.0897	0.0079	0.00516	0.08874
16	0.00853	0.01332	0.09141	0.00798	0.0068	0.08907	0.00781	0.00552	0.0882
17	0.00799	0.01225	0.08855	0.00827	0.00893	0.09048	0.00766	0.00574	0.08731
18	0.00778	0.01507	0.08688	0.00804	0.00804	0.08929	0.00783	0.00489	0.08837
19	0.00811	0.01283	0.08915	0.00804	0.00801	0.08933	0.00776	0.00542	0.08793
20	0.00834	0.01542	0.09	0.00816	0.00864	0.08993	0.00786	0.00627	0.08842

Table 3. Naive Estimation Results

4 Analysis

It is clear upon examining the results that Bias of the estimator tends to increase as group size increases, and tends to decrease as sample size increases. While there are some inconsistencies in the data, the initial relationship appears to be generally monotonic accompanied by a steady increase. There is no apparent sign of increase or decrease in the rate of change of bias as group size increases, and there are not enough sample sizes to make that observation for that dimension.

The difference between biases for simulations of the same group size but different sample sizes appears to be sufficient such that a researcher may want to take sample size into account when determine group size. A group size of 10 with a sample size of 10,000 shows a bias of .012 in our simulations, while the same group size for a sample of 2,000 shows a bias estimated at .027. Therefore, for at least small sample sizes researchers may want to retain relatively small group sizes.

It is notable that for all sample sizes, a group size of 5 appears to be around or below a bias of .003. This would likely be considered acceptable bias for the situation in which this solution would be deployed, and represents a significant decrease in tests administered.

Upon examining the figures, it can be seen that the curves represented by the 10,000 sample size simulations are much smoother than the curves represented by the other populations. We suspect that the large sample size would reduce the variation in estimations.

Upon examination of the results of these simulations, it is apparent that the estimator has negligible bias for small group sizes.

CHAPTER IV

CONCLUSIONS

1 Next Steps

There are several opportunities to build upon the research conducted as part of this project. One opportunity is to build upon the utilization of the model for heterogeneous groups. It may be impractical, either due to time constraints, or due to geographic constraints, to test all samples in groups with like treatment status. There is grounding for making estimations for heterogeneous groups. However, the solutions for estimating p_0^* and p_1^* do not have a closed form. An initial attempt at designing a program that estimates these values can be found in Appendix B.

Another opportunity for improvement is investigating this method for populations with different constraints. It would be important to tests the effectiveness of this method for situations where the treatment or outcome is less frequent, or where the treatment is far less effective. Also, it may be important to investigate populations with more complex criteria for treatment and outcome probabilities. Initial investigations during the course of research indicated that adding additional or more complex criteria greatly increased variance of the prediction, so it may be important to investigate this issue further.

2 Conclusions

We have shown that the proposed estimator for ATE is consistent and our numerical analysis confirmed that when group sizes are low, the bias and mean squared errors are so low that the technique can be used without extreme concern for significant random error.

In a typical situation it would seem that a representative sample of 2,000 with a group size of 5 would be sufficient to accurately estimate outcome probabilities in a situation with factors similar to the ones outlined in this thesis. The implementation of group size constraints in such a study would drastically reduce the cost and time requirements of conducting the study and even greatly increase feasibility of carrying out the study in some situations.

Of course, with greater sample sizes comes greater degrees of certainty with the results provided by the prediction. However, the exact relationship with sample and group size is not yet apparent. For example, it seems that doubling the sample size would not allow you to also double the group size, as the 5,000 population samples with group size 5 yielded a bias of .002 and the 10,000 population samples of group size 10 yielded a bias of .012. Essentially, one cannot increase the sample size and

expect to be able to use the same number of tests. However, keeping the group size constant while increasing the sample size does appear to reduce bias.

Even if investigators are not able to design and implement studies for treatment efficacy on a large scale, the concepts outlined here will allow them to utilize data from group testing efforts made by other researchers or medical professionals for the purposes of estimating disease prevalence. In situations like these, the investigator will be able to determine if it is reasonable to apply these concepts. Ultimately, it is up to the investigator to decide what degree of bias is acceptable, what sample size is available to them, and what resources are available to them in order to make a decision on using the group testing model to generate predictions.

REFERENCES

Crowder, M. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory*, 305–330.

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics 14*(4), 436–440.

Friedman, L. M., C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger (2015). *Fundamentals of clinical trials*. Springer.

Furi, M. and M. Martelli (1991). On the mean value theorem, inequality, and inclusion. *The American Mathematical Monthly 98*(9), 840–846.

Fygenson, M. and Y. Ritov (1994). Monotone estimating equations for censored data. *The Annals of Statistics*, 732–746.

Hernán, M. A. and J. M. Robins (2010). Causal inference.

Horwitz, R. I. (1987). The experimental paradigm and observational studies of cause-effect relationships in clinical medicine. *Journal of chronic diseases 40*(1), 91–99.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710.

Lin, J., D. Wang, and Q. Zheng (2019). Regression analysis and variable selection for two-stage multiple-infection group testing data. *Statistics in medicine 38*(23), 4519–4533.

McCaffrey, D. F., B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine 32*(19), 3388–3414.

McLeod, R. M. (1965). Mean value theorems for vector valued functions. *Proceedings of the Edinburgh Mathematical Society 14*(3), 197–209.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics 29*(3), 343–367.

Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer.

Williams, D. (1991). *Probability with Martingales.* Cambridge mathematical textbooks. Cambridge University Press.

Yan, X., Y. Abdia, S. Datta, K. Kulasekera, B. Ugiliweneza, M. Boakye, and M. Kong (2019). Estimation of average treatment effects among multiple treatment groups by using an ensemble approach. *Statistics in medicine 38*(15), 2828–2846.

APPENDIX A: THEORETICAL FRAMEWORK

The validity of the estimation equations $\mathbf{S}(\mathbf{p}) = 0$: We need to show that $E[\mathbf{S}(\mathbf{p}^*)] = 0$. Consider $E\left[S_1(\mathbf{p}^*)\right]$. Given N_i,

$$
\begin{aligned}
& E\left[\sum_{j=1}^{N_i} 1\{A_{ij}=1\} W_i \left\{-(1-\tilde{Y}_i) + \prod_{k=1}^{N_i}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a^*)\right)\right\}\right] \\
&= E\left\{\sum_{j=1}^{N_i} 1\{A_{ij}=1\} \times \prod_{k=1}^{N_i}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}\pi_a^{-1}(\mathbf{X}_{ik})\right) \times \right. \\
&\qquad \left. \left\{-(1-\tilde{Y}_i) + \prod_{k=1}^{N_i}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a^*)\right)\right\}\right\} \\
&= E\left[\sum_{j=1}^{N_i} E\left\{1\{A_{ij}=1\}\pi_1^{-1}(\mathbf{X}_{ij}) \prod_{k\neq j}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}\pi_a^{-1}(\mathbf{X}_{ik})\right) \times \right.\right. \\
&\qquad \left.\left. \left[-(1-\tilde{Y}_i) + (1-p_1)\prod_{k\neq j}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a^*)\right)\right]\right\}\right] = E\left[\sum_{j=1}^{N_i} T_{ij1} + T_{ij2}\right]
\end{aligned}
$$

where

$$
\begin{aligned}
T_{ij1} &:= E\left[-(1-\tilde{Y}_i)1\{A_{ij}=1\}\pi_1^{-1}(\mathbf{X}_{ij}) \prod_{k\neq j}^{N_i}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}\pi_a^{-1}(\mathbf{X}_{ik})\right)\right], \\
T_{ij2} &:= (1-p_1^*)E\left[\{1\{A_{ij}=1\}\pi_1^{-1}(\mathbf{X}_{ij}) \prod_{k\neq j}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a^*)\pi_a^{-1}(\mathbf{X}_{ik})\right)\right].
\end{aligned}
$$

We deal with T_{ij1} and T_{ij2} separately. Noting that $(A_{ij}, \mathbf{X}_{ij})$'s are independent $j = 1, \cdots, N_i$, and $E\left[1\{A_{ik}=a\}|\mathbf{X}_{ik}\right] = \pi_a(\mathbf{X}_{ik})$, it is easy to see that

$$
\begin{aligned}
T_{ij2} &= (1-p_1^*)E\Big[1\{A_{ij}=1\}\pi_1^{-1}(\mathbf{X}_{ij})\Big] \prod_{k\neq j} E\left[\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a^*)\pi_a^{-1}(\mathbf{X}_{ik})\right] \\
&= (1-p_1^*)E\Big[E\Big[1\{A_{ij}=1\}\Big|\mathbf{X}_{ij}\Big]\pi_1^{-1}(\mathbf{X}_{ij})\Big] \prod_{k\neq j} E\left[E\left[\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a^*)\pi_a^{-1}(\mathbf{X}_{ik})\Big|\mathbf{X}_{ik}\right]\right] \\
&= (1-p_1^*)(2-p_1^*-p_0^*)^{N_i-1}.
\end{aligned}
$$

Next, we evaluate T_{ij1}. Noting that $\tilde{Y}_i = 1 - \prod_{k=1}^{N_i}(1 - Y_{ik})$,

$$
\begin{aligned}
T_{ij1} &= -E\Bigg\{ E\Bigg\{ \prod_{k=1}^{N_i}(1 - Y_{ik}) \times 1\{A_{ij} = 1\}\pi_1^{-1}(\mathbf{X}_{ij}) \\
&\qquad \times \left[\prod_{k\neq j}\left(\sum_{a=0}^{1} 1\{A_{ik} = a\}\pi_a^{-1}(\mathbf{X}_{ik})\right)\right] \Bigg| \{\mathbf{X}_{il}\}_{l=1}^{N_i} \Bigg\}\Bigg\} \\
&= -E\Bigg\{ E\Bigg\{ \prod_{k=1}^{N_i}\left[1 - \left(1\{A_{ik} = 1\}Y_{ik}^{(1)} + 1\{A_{ik} = 0\}Y_{ik}^{(0)}\right)\right] \times 1\{A_{ij} = 1\}\pi_1^{-1}(\mathbf{X}_{ij}) \\
&\qquad \times \left[\prod_{k\neq j}\left(\sum_{a=0}^{1} 1\{A_{ik} = a\}\pi_a^{-1}(\mathbf{X}_{ik})\right)\right] \Bigg| \{\mathbf{X}_{il}\}_{l=1}^{N_i} \Bigg\}\Bigg\} \\
&= -E\Bigg\{ E\left[(1 - Y_{ij}^{(1)})1\{A_{ij} = 1\}\pi_1^{-1}(\mathbf{X}_{ij})\Big|\mathbf{X}_{ij}\right]\Bigg\} \times \\
&\qquad \prod_{k\neq j} E\left\{ E\left[\sum_{a=0}^{1} 1\{A_{ik} = a\}\pi_a^{-1}(\mathbf{X}_{ik})\left(1 - Y_{ij}^{(a)}\right)\Big|\mathbf{X}_{ik}\right]\right\} \\
&= -(1 - p_1^*)\left(2 - p_1^* - p_0^*\right)^{N_i - 1}
\end{aligned}
$$

where the last equality follows from

$$
\begin{aligned}
&E\Bigg\{ E\left[(1 - Y_{ij}^{(1)})1\{A_{ij} = 1\}\pi_1^{-1}(\mathbf{X}_{ij})\Big|\mathbf{X}_{ij}\right]\Bigg\} \\
&= E\Bigg\{ E\left[1 - Y_{ij}^{(1)}\Big|\mathbf{X}_{ij}\right] E\left[1\{A_{ij} = 1\}\pi_1^{-1}(\mathbf{X}_{ij})\Big|\mathbf{X}_{ij}\right]\Bigg\} \\
&= E\Bigg\{ E\left[1 - Y_{ij}^{(1)}\Big|\mathbf{X}_{ij}\right]\Bigg\} = E\left[1 - Y_{ij}^{(1)}\right] = 1 - p_1^*,
\end{aligned}
$$

and $E\left\{E\left[\sum_{a=0}^{1} 1\{A_{ik} = a\}\pi_a^{-1}(\mathbf{X}_{ik})\left(1 - Y_{ij}^{(a)}\right)\Big|\mathbf{X}_{ik}\right]\right\} = 2 - p_1^* - p_0^*$. Thus, $\sum_{j=1}^{N} T_{ij1} + T_{ij2} = 0$ and it holds for all N_i's. Thus $E[S_1(\mathbf{p}^*)] = 0$. And similarly, we can show the same for $E[S_0(\mathbf{p}^*)] = 0$. Thus, $\mathbf{S}(\mathbf{p}^*) = 0$ are valid estimating equations.

If A_{ij}'s are the same for $1 \leq j \leq N_i$ and $N_i = J$ for $1 \leq i \leq N$, $\mathbf{S}(\mathbf{p}) = 0$ become

$$
\begin{aligned}
M^{-1}\sum_{i=1}^{M} J\prod_{k=1}^{J} 1\{A_{ij} = 0\}\pi_0^{-1}(\mathbf{X}_{ik}) \times \left\{-(1 - \tilde{Y}_i) + (1 - p_0)^J\right\} = 0 \\
M^{-1}\sum_{i=1}^{M} J\prod_{k=1}^{J} 1\{A_{ij} = 1\}\pi_1^{-1}(\mathbf{X}_{ik}) \times \left\{-(1 - \tilde{Y}_i) + (1 - p_1)^J\right\} = 0.
\end{aligned}
$$

And we obtain the solutions as

$$\hat{p}_0 = 1 - \left(\frac{\sum_{i=1}^{M} (1-\tilde{Y}_i) \prod_{j=1}^{J} 1\{A_{ij}=0\} \pi_0^{-1}(\mathbf{X}_{ij})}{\sum_{i=1}^{M} \prod_{j=1}^{J} 1\{A_{ij}=0\} \pi_0^{-1}(\mathbf{X}_{ij})} \right)^{1/J} \quad \text{and}$$

$$\hat{p}_1 = 1 - \left(\frac{\sum_{i=1}^{M} (1-\tilde{Y}_i) \prod_{j=1}^{J} 1\{A_{ij}=1\} \pi_1^{-1}(\mathbf{X}_{ij})}{\sum_{i=1}^{M} \prod_{j=1}^{J} 1\{A_{ij}=1\} \pi_1^{-1}(\mathbf{X}_{ij})} \right)^{1/J}.$$

Noting that

$$\partial S_b(\mathbf{p})/\partial p_c = -\mathbb{E}_M \left[\sum_{j=1}^{N_i} 1\{A_{ij}=b\} W_i \left(\sum_{l=1}^{N_i} 1\{A_{il}=c\} \prod_{k \neq l} \left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a) \right) \right) \right],$$

we define classes of functions

$$\mathcal{F}_b := \left\{ f_{b,\mathbf{p}} := \sum_{j=1}^{N_i} 1\{A_{ij}=b\} W_i \left\{ -(1-\tilde{Y}_i) + \prod_{k=1}^{N_i} \left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a) \right) \right\}, \right.$$
$$\left. \mathbf{p} \in [\nu, 1/2-\nu] \times [\nu, 1/2-\nu] \right\}, \quad b = 0, 1.$$

$$\mathcal{G}_{bc} := \left\{ g_{bc,\mathbf{p}} := \sum_{j=1}^{N_i} 1\{A_{ij}=b\} W_i \left\{ \sum_{l=1}^{N_i} 1\{A_{il}=c\} \prod_{k \neq l} \left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a) \right) \right\}, \right.$$
$$\left. \mathbf{p} \in [\nu, 1/2-\nu] \times [\nu, 1/2-\nu] \right\}, \quad b, c = 0, 1.$$

Proof of Theorem 4.1: (i) We prove the result by invoking Theorem 3.2 in Crowder (1986). Thus, we need to check the conditions (i) and (ii) of Theorem 3.2 in Crowder (1986).

By Lemma 2.1, $\mathcal{F}_b, b = 0, 1$'s and $\mathcal{G}_{bc}, b, c = 0, 1$'s are Donsker, and hence Glivenko-Cantelli. Thus,

$$\sup_{\mathbf{p} \in [\nu, 1/2-\nu] \times [\nu, 1/2-\nu]} \max \left\{ \left\| \mathbf{S}(\mathbf{p}) \to E[\mathbf{S}(\mathbf{p})] \right\|, \left\| \dot{\mathbf{S}}(\mathbf{p}) - E\left[\dot{\mathbf{S}}(\mathbf{p}) \right] \right\| \right\} \to_p 0. \tag{10}$$

Thus, condition (ii) of Theorem 3.2 in Crowder (1986) is satisfied.

Let $\partial B_r := \{\mathbf{p} : \|\mathbf{p} - \mathbf{p}^*\| = r\}$ be the boundary of the circle $B_r := \{\mathbf{p} : \|\mathbf{p} - \mathbf{p}^*\| \leq r\}$. Given a $\mathbf{p} \in \partial B_r$, consider $(\mathbf{p} - \mathbf{p}^*)^{\mathrm{T}} E[S(\mathbf{p})]$. Since $E[S(\mathbf{p}^*)] = 0$, by the vector valued mean value theorem (McLeod, 1965; Furi and Martelli, 1991),

$$\begin{aligned}
&(\mathbf{p} - \mathbf{p}^*)^{\mathrm{T}} E[\mathbf{S}(\mathbf{p})] = (\mathbf{p} - \mathbf{p}^*)^{\mathrm{T}} (E[\mathbf{S}(\mathbf{p})] - E[\mathbf{S}(\mathbf{p}^*)]) \\
&= (\mathbf{p} - \mathbf{p}^*)^{\mathrm{T}} \left(\lambda_1 E\left[\dot{\mathbf{S}}(\tilde{\mathbf{p}}_1) \right] + \lambda_2 E\left[\dot{\mathbf{S}}(\tilde{\mathbf{p}}_2) \right] \right) (\mathbf{p} - \mathbf{p}^*) \\
&= (\mathbf{p} - \mathbf{p}^*)^{\mathrm{T}} \left(\lambda_1 \frac{E\left[\dot{\mathbf{S}}(\tilde{\mathbf{p}}_1) \right] + E\left[\dot{\mathbf{S}}(\tilde{\mathbf{p}}_1) \right]^{\mathrm{T}}}{2} + \lambda_2 \frac{E\left[\dot{\mathbf{S}}(\tilde{\mathbf{p}}_2) \right] + E\left[\dot{\mathbf{S}}(\tilde{\mathbf{p}}_2) \right]^{\mathrm{T}}}{2} \right) (\mathbf{p} - \mathbf{p}^*)
\end{aligned} \tag{11}$$

where $\lambda_1, \lambda_2 > 0, \lambda_1 + \lambda_2 = 1$, $\tilde{\mathbf{p}}_1 = t_1\mathbf{p} + (1-t_1)\mathbf{p}^*$, and $\tilde{\mathbf{p}}_2 = t_2\mathbf{p} + (1-t_2)\mathbf{p}^*$, for some $0 < t_1, t_2 < 1$. When $b = c$,

$$
\begin{aligned}
&E\left[\dot{S}_{bb}(\mathbf{p})\right] = -E\left[\sum_{j=1}^{N_i} 1\{A_{ij}=b\} W_i \left(\sum_{l=1}^{N_i} 1\{A_{il}=b\} \prod_{k\neq l}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a)\right)\right)\right] \\
&= -E\left[W_i \sum_{j=1}^{N_i} 1\{A_{ij}=b\} \prod_{k\neq j}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a)\right)\right] \\
&\quad - E\left[W_i \sum_{j=1}^{N_i}\sum_{l\neq j} 1\{A_{ij}=b\}1\{A_{il}=b\}(1-p_b) \prod_{k\neq j,l}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a)\right)\right] \\
&= -E\left[N_i(2-p_0-p_1)^{N_i-1} + N_i(N_i-1)(1-p_b)(2-p_0-p_1)^{N_i-2}\right]
\end{aligned}
$$

where the last equation follows from the same arguments used for T_{ij1} and T_{ij2}. Likewise, when $b \neq c$

$$
\begin{aligned}
&E\left[\dot{S}_{bc}(\mathbf{p})\right] = -E\left[W_i \left(\sum_{l=1}^{N_i} 1\{A_{il}=c\} \prod_{k\neq l}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a)\right)\right)\sum_{j=1}^{N_i} 1\{A_{ij}=b\}\right] \\
&= -E\left[W_i \sum_{j=1}^{N_i}\sum_{l\neq j} 1\{A_{ij}=b\}1\{A_{il}=c\}(1-p_b) \prod_{k\neq j,l}\left(\sum_{a=0}^{1} 1\{A_{ik}=a\}(1-p_a)\right)\right] \\
&= -E\left[N_i(N_i-1)(1-p_b)(2-p_0-p_1)^{N_i-2}\right].
\end{aligned}
$$

Thus, $E[\dot{\mathbf{S}}(\mathbf{p})] = -E\left[N_i(2-p_0-p_1)^{N_i-2}\mathbf{V}_{N_i}(\mathbf{p})\right]$, where

$$
\mathbf{V}_{N_i}(\mathbf{p}) = \begin{pmatrix} N_i + 1 - N_i p_0 - p_1 & (N_i-1)(1-p_0) \\ (N_i-1)(1-p_1) & N_i + 1 - p_0 - N_i p_1 \end{pmatrix},
$$

and consequently,

$$
\begin{aligned}
&\frac{E[\dot{\mathbf{S}}(\mathbf{p})] + E[\dot{\mathbf{S}}(\mathbf{p})]^{\mathrm{T}}}{2} \\
&= -\frac{1}{2}E\left[N_i(2-p_0-p_1)^{N_i-2}\begin{pmatrix} 2N_i + 2 - 2N_i p_0 - 2p_1 & (N_i-1)(2-p_0-p_1) \\ (N_i-1)(1-p_0-p_1) & 2N_i + 2 - 2p_0 - 2N_i p_1 \end{pmatrix}\right].
\end{aligned}
$$

Under Condition (A3), the eigenvalues of $-(E[\dot{\mathbf{S}}(\mathbf{p})] + E[\dot{\mathbf{S}}(\mathbf{p})]^{\mathrm{T}})/2$ are greater than Λ for some $\Lambda > 0$, uniformly over $\mathbf{p} \in [\nu, 1/2-\nu] \times [\nu, 1/2-\nu]$. Therefore, combining (11) and the fact above yield $(\mathbf{p}-\mathbf{p}^*)^{\mathrm{T}}(-E[S(\mathbf{p})]) \geq \Lambda\|\mathbf{p}-\mathbf{p}^*\|^2 = \Lambda r^2$. Thus, condition (ii) of Theorem 3.2 in Crowder (1986) is satisfied as well.

By Theorem 3.2 in Crowder (1986), there exists a sequence $\tilde{\mathbf{p}}$ such that $\tilde{\mathbf{p}} \to_p \mathbf{p}^*$. (ii) By the vector valued mean value theorem (McLeod, 1965; Furi and Martelli, 1991) again,

$$
0 = \mathbf{S}(\tilde{\mathbf{p}}) = \mathbf{S}(\mathbf{p}^*) + \left(\lambda_1 \dot{\mathbf{S}}(\tilde{\mathbf{p}}_1) + \lambda_2 \dot{\mathbf{S}}(\tilde{\mathbf{p}}_2)\right)(\mathbf{p}-\mathbf{p}^*),
$$

where $\lambda_1, \lambda_2 > 0, \lambda_1 + \lambda_2 = 1$, $\tilde{\mathbf{p}}_1 = t_1\tilde{\mathbf{p}} + (1-t_1)\mathbf{p}^*$, and $\tilde{\mathbf{p}}_2 = t_2\tilde{\mathbf{p}} + (1-t_2)\mathbf{p}^*$, for some $0 < t_1, t_2 < 1$. Thus,

$$\sqrt{n}(\tilde{\mathbf{p}} - \mathbf{p}^*) = -\left(\lambda_1\dot{\mathbf{S}}(\tilde{\mathbf{p}}_1) + \lambda_2\dot{\mathbf{S}}(\tilde{\mathbf{p}}_2)\right)^{-1}\sqrt{n}\mathbf{S}(\mathbf{p}^*).$$

By central limit theorem,

$$\sqrt{n}\mathbf{S}(\mathbf{p}^*) \to_d N(0, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma} = E\left[\mathbf{S}(\mathbf{p}^*)\mathbf{S}(\mathbf{p}^*)^{\mathrm{T}}\right]$. Because $\tilde{\mathbf{p}} \to_p \mathbf{p}^*$ by Theorem 4.1 and the continuous mapping theorem, $\left(\lambda_1\dot{\mathbf{S}}(\tilde{\mathbf{p}}_1) + \lambda_2\dot{\mathbf{S}}(\tilde{\mathbf{p}}_2)\right)^{-1} \to_p E[\dot{\mathbf{S}}(\mathbf{p}^*)]^{-1}$. Then by Slutsky's theorem, $\sqrt{n}(\tilde{\mathbf{p}} - \mathbf{p}^*) \to_p N(0, E[\dot{\mathbf{S}}(\mathbf{p}^*)]^{-1}\mathbf{\Sigma}(E[\dot{\mathbf{S}}(\mathbf{p}^*)]^{-1})^{\mathrm{T}}$.

This complete the proof of Theorem 4.1. □

Proof of Corollary 4.1: Since $\|\hat{h} - h^*\|_\infty = O_p(M^{-1/2})$, then $\max_{1\le i\le M}\|\hat{W}_i - W_i\| = O_p(M^{-1/2})$. Noting that $N_i \le N$ by Condition (A3) and

$$\begin{aligned}
&\hat{S}_0(\mathbf{p}) - S_0(\mathbf{p}) \\
&= \mathbb{E}_M\left[\sum_{j=1}^{N_i} 1\{A_{ij} = 0\}(\hat{W}_i - W_i)\left\{-(1-\tilde{Y}_i) + \prod_{k=1}^{N_i}\left(\sum_{a=0}^{1} 1\{A_{ik} = a\}(1-p_a)\right)\right\}\right],
\end{aligned}$$

we obtain

$$\sup_{\mathbf{p}\in[\nu,1/2-\nu]\times[\nu,1/2-\nu]} |\hat{S}_0(\mathbf{p}) - S_0(\mathbf{p})| = O_p(M^{-1/2}). \tag{12}$$

Since both $S_0(\mathbf{p})$ and $\hat{S}_0(\mathbf{p})$ are bounded, by dominated convergence theorem (see, e.g., Williams, 1991), $\sup_{\mathbf{p}\in[\nu,1/2-\nu]\times[\nu,1/2-\nu]} E[|\hat{S}_0(\mathbf{p}) - S_0(\mathbf{p})|] \to 0$.

Similarly, $\sup_{\mathbf{p}\in[\nu,1/2-\nu]\times[\nu,1/2-\nu]} E[|\hat{S}_1(\mathbf{p}) - S_1(\mathbf{p})|] \to 0$. Consequently, we obtain that

$$\sup_{\mathbf{p}\in[\nu,1/2-\nu]\times[\nu,1/2-\nu]} E[\|\hat{\mathbf{S}}(\mathbf{p}) - \mathbf{S}(\mathbf{p})\|] \to 0. \tag{13}$$

Thus,

$$\begin{aligned}
&\inf_{\mathbf{p}\in\partial B(r)}(\mathbf{p} - \mathbf{p}^*)^{\mathrm{T}}(-E[\hat{\mathbf{S}}(\mathbf{p})]) \\
&\ge \inf_{\mathbf{p}\in\partial B(r)}(\mathbf{p} - \mathbf{p}^*)^{\mathrm{T}}(-E[\mathbf{S}(\mathbf{p})]) - \sup_{\mathbf{p}\in\partial B(r)}\|\mathbf{p} - \mathbf{p}^*\|\,\|E[\mathbf{S}(\mathbf{p})] - E[\hat{\mathbf{S}}(\mathbf{p})]\| \\
&\ge \Lambda\|\mathbf{p} - \mathbf{p}^*\|^2.
\end{aligned}$$

Thus, condition (i) of Theorem 3.2 in Crowder (1986) is satisfied.

Moreover, by (10), (12), and (13),

$$\begin{aligned}
&\sup_{\mathbf{p}\in\partial B(r)}\left|\hat{\mathbf{S}}(\mathbf{p}) - E\left[\hat{\mathbf{S}}(\mathbf{p})\right]\right| \\
&\le \sup_{\mathbf{p}\in\partial B(r)}\left|\hat{\mathbf{S}}(\mathbf{p}) - \mathbf{S}(\mathbf{p})\right| + \sup_{\mathbf{p}\in\partial B(r)}|\mathbf{S}(\mathbf{p}) - E\left[\mathbf{S}(\mathbf{p})\right]| + \sup_{\mathbf{p}\in\partial B(r)}\left|E\left[\hat{\mathbf{S}}(\mathbf{p})\right] - E\left[\hat{\mathbf{S}}(\mathbf{p})\right]\right| \\
&\to_p 0.
\end{aligned}$$

Thus, condition (ii) of Theorem 3.2 in Crowder (1986) is satisfied. Therefore, there exists a sequence $\hat{\mathbf{p}}$ such that $\hat{\mathbf{p}} \to_p \mathbf{p}^*$. This complete the proof of Corollary 4.1. □

Lemmas

Lemma 2.1 *Under Conditions (A1) – (A3), $\mathcal{F}_b, b = 0, 1$'s and $\mathcal{G}_{bc}, b, c = 0, 1$'s are Donsker classes.*

Proof: It is easy to see that $\mathcal{F}_b$ is of dimension 2. Then by Lemma 2.6.15 of Vaart and Wellner (1996), $\mathcal{F}_b$ is VC-subgraph of index smaller than or equal to 4, where we refer the definitions of VC-subgraph and VC-subgraph index to Chapter 2 of Vaart and Wellner (1996).

By Conditions (A1) and (A2), $|f_{b,\mathbf{p}}| \leq 2N^2\nu^{-N}$. Thus, $2N^2\nu^{-N}$ is an envelop function for the class $\mathcal{F}_b$. Then by Theorem 2.6.7 of Vaart and Wellner (1996), for any probability measure Q,

$$N\left(2N^2\nu^{-N}\epsilon, \mathcal{F}_b, L_2(Q)\right) \leq 4K(16e)^4\epsilon^{-6},$$

for a universal constant K and $0 < \epsilon < 1$. We refer the definition of covering number $N(\epsilon, \mathcal{F}, L_2(Q))$ to Pages 83 and 98 of Vaart and Wellner (1996). Noting that

$$\int_0^\infty \sup_Q \sqrt{\log N\left(2N^2\nu^{-N}\epsilon, \mathcal{F}_b, L_2(Q)\right)}d\epsilon = \int_0^1 \sup_Q \sqrt{\log N\left(2N^2\nu^{-N}\epsilon, \mathcal{F}_b, L_2(Q)\right)}d\epsilon$$
$$\leq \int_0^1 \sqrt{\log 4K(16e)^4 + \log \epsilon^{-6}}d\epsilon \leq \int_0^1 \sqrt{\log 4K(16e)^4}d\epsilon + \int_0^1 \sqrt{-6\log\epsilon}d\epsilon$$
$$= \sqrt{\log 4K(16e)^4} + \int_0^\infty \sqrt{6u}\exp(-u)du < \infty,$$

where $u = -log\epsilon$ in the second equality, by Theorem 2.8.3 of Vaart and Wellner (1996), $\mathcal{F}_b, b = 0, 1$'s are Donsker. Similarly, we can show that $\mathcal{G}_{bc}, b, c = 0, 1$'s are Donsker as well. This completes the proof of Lemma 2.1. □

APPENDIX B: SIMULATION CODE

Estimation Function

The function below creates estimates for probability of treated or untreated infection rate. The d input is the data to be inputted. The t function indicates where the groups in the sample are treated or untreated. The g function describes group size.

```
prob<- function(d,t,g){
s1<-c()
s2<-c()
stg<- d[which(d$treat == t & d$full.group == 1),]
for (i in min(stg$group):max(stg$group)){
  product <- prod((stg$predict[which(stg$group ==i)]^-1))
  #numerator
  s1<- c(s1,((1-max(stg$positive[which(stg$group==i)]))*product))
  #denominator
  s2<-c(s2,product)
}
return(1 - (sum(s1)/sum(s2))^(1/g))
}
```

Population Creation and Estimation

```
MSE=c()
SE = c()
for (gsize in 1:20){
predicted<-c()
actual<-c()
for (i in 1:100){
#set sample size
num<- 2000

#sample age
age<-round(rnorm(num, 39, 12),0)

#condition
con<-rbinom(num,1,.2)
```

```
xb<- -6.5 + .1*age +2*con
p<-1/(1+exp(-xb))
treat<-rbinom(n=num, size =1, prob = p)



#Create distribution of infections
# Not using actual infection rates or estimated vaccination success rates
# due to small sample size. Try to replicate ~10% positivity among untreated
xb<- -7.2 + .1*age +3*con -5*treat
p<-1/(1+exp(-xb))
positive<-rbinom(n=num, size =1, prob = p)


xb<- -7.2 + .1*age +3*con  -5 ### All subjects were treated
p<-1/(1+exp(-xb))
p1 = mean(p)                    ### average infection rate

xb<- -7.2 + .1*age +3*con ### none got treated
p<-1/(1+exp(-xb))
p0 = mean(p)                    ### average infection rate

############################
############################




#create data frame
data<- data.frame(n=1:num, age, con, treat, positive)


#randomize and group treated individuals
grp<- gsize
treated<- data[data$treat==1,]
rows<- sample(nrow(treated))
treated2<-treated[rows,]
group<-ceiling((1:nrow(treated2))/grp)
treated3<-cbind(treated2, group)
treated3$full.group<-0
treated3$full.group[(1:(floor(nrow(treated3)/grp)*grp))]<-1
```

```
#randomize and group untreated individuals
untreated<-data[data$treat==0,]
rows<- sample(nrow(untreated))
untreated2<-untreated[rows,]
group2<-ceiling((1:nrow(untreated2))/grp)+max(group)
untreated3<-cbind(untreated2, group2)
untreated3$full.group<-0
untreated3$full.group[(1:(floor(nrow(untreated3)/grp)*grp))]<-1
colnames(untreated3)[6]<-'group'

#recombine groups
data.final<-rbind(treated3,untreated3)

data.final.sorted<-data.final[order(data.final$n),]
data.final.sorted$pt<-pt

#
#
#
#
# Calculations
#
#
#
#
#




trtProbModel <- glm(treat~age+con,family = 'binomial', data=data.final.sorted)
data.final.sorted$trtPredictYes<-predict(trtProbModel,
                    newdata = data.final.sorted[,c(2,3)], type = 'response')
data.final.sorted$trtPredictNo<-1-data.final.sorted$trtPredictYes



predicted<-c(predicted,prob(data.final.sorted, 0, grp) -
            prob(data.final.sorted, 1, grp))
actual<-c(actual,p0-p1)
}
MSE<-c(MSE, mean((predicted-actual)^2))
SE<- c(SE,sd(predicted-actual))
}
Bias<- sqrt(abs(MSE - SE^2))
```

Heterogeneous Estimation

Below is the code proposed to estimate conditional infection rates for heterogeneously composed groups. S0 and S1 are functions that make up part of the function that searches for the conditional probability values. P.solve is the only function that needs to be run, and can replace the prob function described in the previous section.

```
s0<- function(d,g,p0,p1){
  s<c()

  stg<- d[which(d$full.group == 1),]
  for (i in min(stg$group):max(stg$group)){
    rowz<-which(stg$group ==i)
    w<-prod(stg$treat[rowz]*stg$p_assign[rowz]^-1 +
            (1-stg$treat[rowz])*(1-stg$p_assign[rowz])^-1)
    s<-c(s,nrow(stg[which(stg$group==i & stg$treat == 0),]) *
        w*(-(1-max(stg$positive[rowz])) +
   prod(stg$treat[rowz]*(1-p1) + (1-stg$treat[rowz])*(1-p0))))

  }
  return(mean(s))
}

s1<- function(d,g,p0,p1){
  s<c()

  stg<- d[which(d$full.group == 1),]
  for (i in min(stg$group):max(stg$group)){
    rowz<-which(stg$group ==i)
    w<-prod(stg$treat[rowz]*stg$p_assign[rowz]^-1 +
            (1-stg$treat[rowz])*(1-stg$p_assign[rowz])^-1)
    s<-c(s,nrow(stg[which(stg$group==i & stg$treat == 1),]) *
            w*(-(1-max(stg$positive[rowz])) +
       prod(stg$treat[rowz]*(1-p1) + (1-stg$treat[rowz])*(1-p0))))

  }
  return(mean(s))
}



p.solve<- function(d){
```

```
  jbot<-0
  jtop<-.1
  j<-mean(c(jtop,jbot))
  while(jtop-jbot>.001){
    top<-.3
    bot<-0
    i<-mean(c(top,bot))
    while (top-bot>.001){
      x<-mean(c(i,top))
      if (ssq(d,5,i,j) < ssq(d,5,x,j)){
        top<-x
      }  else {bot<-i}
      i<-mean(c(top,bot))
    }

    xj<-mean(c(jtop,j))
    top<-.3
    bot<-0
    ij<-mean(c(top,bot))
    while (top-bot>.001){
      x<-mean(c(ij,top))
      if (ssq(d,5,ij,xj) < ssq(d,5,x,xj)){
        top<-x
      }  else {bot<-ij}
      ij<-mean(c(top,bot))
    }

    if (ssq(d,5,i,j) < ssq(d,5,ij,xj)){
      jtop<-xj
    }  else {jbot<-j}

    j<-mean(c(jtop,jbot))
  }
  return(data.frame(p0=i,p1=j, ssq= ssq(d,5,i,j),
                    s0 = s0(d,5,i,j), s1 = s1(d,5,i,j)))
}

ssq<- function(d,g,p0,p1){
  return(sqrt(s1(d,g,p0,p1)^2+s0(d,g,p0,p1)^2))
}
```

CURRICULUM VITAE

Alexander Christopher Noll

Education

M.S. Biostatistics, University of Louisville, Louisville, KY, May 2021 ...

B.S. Mathematics Bellarmine University, Louisville, KY, May 2017 ...

B.A. History Bellarmine University, Louisville, KY, May 2017 ...

B.A. Political Science Bellarmine University, Louisville, KY, May 2017 ...