2012

The Analysis of Acute Stroke Clinical Trials with Responder Analysis Outcomes

Kyra Michelle Garofolo
Medical University of South Carolina

Follow this and additional works at: https://medica-musc.researchcommons.org/theses

The Analysis of Acute Stroke Clinical Trials with

Responder Analysis Outcomes

By

Kyra Michelle Garofolo

A thesis submitted to the faculty of the Medical University of South Carolina in partial fulfillment of the requirement for the degree of Masters of Science in the College of Graduate Studies.

Division of Biostatistics and Epidemiology

2012

Approved By:

_____
Valerie L Durkalski, PhD, MPH
Chairman, Advisory Committee

_____
Sharon D Yeatts, PhD

_____
Viswanathan Ramakrishnan, PhD

_____
Edward C Jauch, MD, MS

_____
Karen C Johnston, MD, MSc

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to acknowledge my research advisory committee for their continued support throughout the development of this thesis. Their advice and critiques have been invaluable during the research process. Dr. Valerie Durkalski has been an excellent committee chairperson, helping me through roadblocks and keeping me on track. Along with Dr. Durkalski, Dr. Sharon Yeatts and Dr. Viswanathan Ramakrishnan have offered unending insight into the statistical and programming aspects of this study. The statistical aspects of this study have been complemented by the clinical insights offered by Dr. Edward Jauch and Dr. Karen Johnston, both of whom have helped make this research relevant in the medical community.

I would also like to thank the members of the Division of Biostatistics and Epidemiology faculty, who have shown me support both inside and outside of the classroom. In addition, I thank my husband Michael for his unending patience and support throughout the research process.

ABSTRACT

KYRA GAROFOLO. The Analysis of Acute Stroke Clinical Trials with Responder Analysis Outcomes. (Under the direction of VALERIE DURKALSKI).

Traditionally in acute stroke clinical trials, the primary outcome has been a dichotomized modified Rankin Scale (mRS). The mRS is a 7-point ordinal scale indicating a patient's level of disability following a stroke. Traditional analyses have used a fixed dichotomization scheme, which dichotomizes 'success' as an mRS of 0-1 or 0-2. This method fails to address the concern that stroke severity may impact the likelihood of a successful outcome; subjects with mild strokes may achieve the defined threshold for success more easily than subjects with severe strokes. Consequently, subjects are unable to contribute equally to the estimation of treatment effect. Stroke studies are increasingly turning to new statistical methods that make more efficient use of available data, including responder analysis.

Responder analysis, also known as the sliding dichotomy, allows the definition of success to vary according to baseline severity. This method puts patients on a more level playing field, producing a more clinically relevant insight into the actual effect of investigational stroke treatments. It is unclear whether or not statistical analyses should adjust for baseline severity when responder analysis is used, as the outcome already takes into account baseline severity. Through the use of simulations, this research compares the operating characteristics of unadjusted and adjusted analyses in the responder analysis scheme. We also compare the treatment effect estimates and their standard errors between methods.

Under various treatment effect settings, the operating characteristics of the unadjusted and adjusted analyses do not appear to differ substantially. Power and type I error were preserved for both the unadjusted and adjusted analyses. Our results suggest that, under the given treatment effect scenarios, the decision whether or not to adjust for baseline severity should be guided by the needs of the study rather than a strict guideline, as type I error rates and power do not appear to vary largely between the methods.

CHAPTER 1: INTRODUCTION AND BACKGROUND

Stroke is a potentially debilitating medical event that affects approximately 800,000 people in the United States each year, leaving as many as 30% of its victims permanently disabled[1]. Given this level of impact, there is great demand for treatments which significantly improve functional outcome following a stroke. To date, there have been few successful trials for clinical treatment of acute stroke. In fact, only 3 treatments in over 125 stroke trials have demonstrated positive results. These treatments include intravenous tissue-type plasminogen activator (tPA) within 3 to 4.5 hours from stroke onset, hemicraniectomy for malignant infarction, and coiling for aneurysmal subarachnoid hemorrhage[2,3].

One of the possible reasons for the excessive number of neutral or unsuccessful stroke trials is the definition of successful outcome[4]. In clinical trials, stroke outcome is most commonly measured by the modified Rankin Scale (mRS) of global disability, usually measured at 90 days following stroke occurrence. The mRS is a 7-point ordinal scale that ranges from 0 (no disability) to 6 (death), and has been shown to be a valid and reliable measure of functional outcome following a stroke[5]. Past trials have dichotomized mRS scores into "success" and "failure", where mRS scores of 0-1 (or 0-2) were considered to be "successes" while scores greater than 1 (or 2) were considered to be "failures," regardless of baseline stroke severity [6,7,8,9]. This method fails to take into

account the understanding that baseline severity is highly correlated with outcome; a subject with a mild baseline severity may make only a slight improvement but be considered a success, while a subject who suffered a very severe stroke may make vast improvement but still be considered a failure due to scoring above a 1 (or 2) on the mRS at the study's end. In addition, traditional dichotomization does not capture small shifts along the outcome scale, which may be the goal of some treatment trials, such as neuroprotective agents and hemostatic agents in intracerebral hemorrhage[10]. Due to its drawbacks, strict dichotomization does not make efficient use of the data collected. New methods are evolving to make better use of the outcome data in stroke trials with the hopes of providing higher sensitivity to detect true treatment effects. These new methods include the global statistic, shift analysis, permutation testing, and responder analysis.

1.1. New Endpoints and Analysis Methods

1.1.1. The global statistic.

In addition to the mRS, there are many other ordinal scales that assess stroke outcome. These other scales include the National Institutes of Health Stroke Scale (NIHSS), the Barthel Index (BI), the Glasgow Outcome Scale (GOS), and the Stroke Outcome Scale (SOS); each of these is related to the others, but may capture a different aspect of stroke functional outcome or disability. The global statistic consolidates several of these measures into one outcome for analyses[6,11]. Though the tPA trials have been reanalyzed and efficacy confirmed using additional statistical techniques, the original outcome was defined by a global measure based on the dichotomization of the NIH Stroke Scale, the Barthel Index, and the Glasgow Outcome Scale[12,13]. Each of these component scales was dichotomized regardless of baseline severity.

Analysis with a global statistic is particularly useful when no single outcome sufficiently captures the desired endpoint[14]. One of the advantages of the global statistic approach is that it allows for the simultaneous analysis of several outcome characteristics. At the end of the analysis, rejection of the null hypothesis provides evidence for overall treatment efficacy rather than efficacy in just one aspect of stroke outcome[11]. In addition, there is often a power advantage when using the global statistic. By combining several correlated measures of stroke outcome on the same subject, some of the noise observed in the single scales alone is eliminated, thus increasing the ability to detect the true treatment effect[6].

While the global statistic undoubtedly has its advantages, it has several disadvantages as well. One major drawback of global statistic analysis is that its results are difficult to translate into meaningful interpretations for physicians and patients[11]. This aspect of the global statistic makes it less appealing, as any clinical study aims to have clinically pertinent results at the study's end. In addition to questionable interpretation, appropriate statistical methods for the analysis of the global statistic are still evolving and often rely on dichotomization[6,11]. Related methods such as principal components analysis and multiple correspondences analysis have also been proposed[15], but are highly statistically intensive and share the same interpretability disadvantages.

1.1.2. Shift analysis and the proportional odds model.

While the global statistic aims to combine several outcome measures into one comprehensive measure, other methods aim to improve the analysis on a single ordinal scale, such as the mRS. Where the traditional dichotomization method focuses only on the distribution of scores on either side of one predetermined cutpoint, shift analysis

3

focuses on the distribution of study subjects across the entire ordinal scale. Rather than asking whether the treatment makes more patients achieve a good outcome, defined as better than threshold X, as in traditional dichotomized analyses, shift analysis asks the question, "does the treatment make the patient better to some degree?"[6]

Shift analyses are advantageous when prior knowledge of an appropriate cutpoint is not available, as the entire scale is examined. In trials where the treatment effect tends to be basically uniform across a large portion of the ordinal outcome scale or where the treatment effect is clustered at an unexpected cutpoint, shift analysis has been shown to be more powerful and more efficient than traditional dichotomization methods[6,10,4,9]. Another advantage of shift analysis is that it does not let baseline stroke severity limit the ability of subjects to contribute to the estimation of treatment effect[6].

Disadvantages of shift analysis include the assumptions of its statistical methods, which may not be met in real studies[2,4]. These assumptions include the "proportional odds assumption" when proportional odds modeling is used, which assumes that the odds ratio for a better outcome is the same at every cutpoint on the scale, and the assumption that the treatment effect occurs only in one direction along the outcome scale[4,16]. Interpretations from shift analyses can also be difficult and are usually stated in terms of number needed to treat (NNT) or combined odds ratios[2,6,4]. The NNT in a shift analysis setting is not as straightforward as in the dichotomized endpoint setting, as derivation of a NNT across an entire ordinal scale must take into account within-patient correlation and the fact that some transitions (such as transitioning from death to a vegetative state) are considered unfavorable by patients[6]. Joint outcome tables can be used to derive a composite measure of the number of patients needed to treat for a single additional

4

patient to benefit, which can be used as a measure of treatment efficacy in shift analysis studies.

1.1.3. Permutation testing.

One of the more recently proposed approaches to handle ordinal data in stroke trials is that of permutation testing[17]. This method, proposed in 2012 by Howard et al, is similar to the Mann-Whitney U test in that both methods nonparametrically investigate the idea of whether a randomly chosen individual from one treatment group has a better outcome than another randomly chosen person from another treatment group. However, unlike the Mann-Whitney U test, the permutation method considers only untied pairs of study subjects, primarily for the sake of interpretation. The permutation method, like most statistical tests, bases its results on a test statistic calculated from the observed data. This test statistic is compared to the estimated distribution of test statistics under the null, which is derived under an iterative process which randomly assigns treatments to individuals assuming no association between the test statistic and treatment.

The primary advantage of this approach is its ease of interpretability. When using the permutation testing approach, the efficacy results can be explained simply in terms of the proportion of people who will do better on the experimental treatment, compared to the proportion who will do better on placebo and the proportion who will do the same on either treatment[17]. Another primary advantage is that this method is nonparametric, and makes no distributional assumptions that must be met for validity. The permutation testing method takes into account changes across the entire spectrum of an ordinal scale such as the mRS, addressing the issues of traditional dichotomization of such scales. In addition, the permutation method easily allows the incorporation of stratification of

important covariates. One potential disadvantage of this method is its relatively more complex computing requirements when compared to similar methods such as the Mann-Whitney U test, which produce similar results.

1.1.4. Responder analysis.

As previously mentioned, a major problem with the traditional dichotomization technique is that it fails to account for baseline severity. Responder analysis, also known as the sliding dichotomy, still dichotomizes the outcomes into "success" and "failure," but addresses this issue by allowing cutpoints to vary. The definition of successful outcome differs by prognosis group; those study subjects in a less severe prognosis group at baseline must achieve a better outcome to be considered a trial "success," whereas study subjects in a more severe baseline prognosis category must achieve a less stringent criterion for success.

Within the responder analysis framework, there are different ways to determine prognosis groups and success cutpoints. Prognosis groups may be determined by as few as one baseline measure, such as the baseline NIHSS score, or the combination of many baseline measures into one prognosis score by means of an algorithm[16,13,18,19,20]. Often times, cutpoints for success will be predetermined as in the AbESTT-II trial and reanalysis of the NINDS-tPA trials[20,13]. Each of these trials classified subjects into mild, moderate, and severe baseline prognosis groups based on NIHSS scores, and defined success as having a 3-month mRS=0 for subjects in the mild, mRS$\leq$1 for subjects in the moderate, and mRS$\leq$2 for subjects in the severe prognosis groups. Alternatively, some studies determine these thresholds based on the empirical distribution of the data and the

probabilities for success within each of the prognosis groups, as in the simulation study based on the IMPACT Project results[16].

One advantage of responder analysis when compared to the traditional dichotomization method is that it allows each subject to have an achievable goal based on their baseline severity. By allowing the definition of success to vary with severity, the ability to detect a true treatment effect is increased without increasing external noise and variability[6]. Responder analysis is a relatively computationally easy method to employ and has been argued to be more powerful than traditional dichotomization methods[8,16,18]. However, a direct comparison of the traditional dichotomization method with responder analysis in the same data must be interpreted with caution, as the control rates in the two cases are not the same by definition, and thus the power comparison may not be appropriate.

While responder analysis has its advantages, there are potential disadvantages to consider as well. Like traditional dichotomization, the ordinal outcome is still collapsed into a dichotomous outcome, thus discarding information about specific mRS categories. In addition, for responder analysis to be most effective, investigators must carefully define appropriate prognosis groups and their respective cutpoints for success[6].

The global statistic, shift analysis, the permutation method, and responder analysis are several of the new methods being employed in stroke clinical trials to overcome the drawbacks of traditional dichotomization. Each of these methods has its own advantages and disadvantages as discussed above. The most appropriate method will depend on the aims of the clinical trial; investigators should consider the unique aspects of their clinical trial when determining the primary analysis technique.

1.2. Covariate Adjustment

Statistical analyses often adjust for prognostic factors, or covariates that may be predictive of the primary outcome. One philosophy that motivates covariate adjustment is that it can confirm that the observed treatment effect is "independent" of these prognostic factors, rather than artificially created by confounding prognostic factors[21]. Another motivation for covariate adjustment is possible covariate imbalance. While covariate adjustment can help temper the effects of a covariate imbalance, it should not be used as a means to address imbalance in baseline characteristics between treatment groups[22]. Instead, the study design should strive to prevent covariate imbalance whenever possible. Adjustment for important covariates also accounts for additional variation in the data. Accounting for this additional variation can lead to increased statistical efficiency, which is a primary reason for covariate adjustment during analysis. In the case of a continuous outcome and classical linear regression, covariate adjustment may increase the precision of a treatment effect estimate, as it may decrease the standard error of the estimate due to a reduction in residual variance[23].

An interesting phenomenon occurs when logistic regression is used in the case of a binary outcome, as in the traditional or sliding dichotomy settings. Covariate adjustment in the case of logistic regression results in a loss of precision of the treatment effect estimate, as the standard error on the treatment effect estimate is increased. However, this increase in standard error is balanced by a movement of the treatment effect estimate away from the null hypothesis. This phenomenon was described by Robinson and Jewell, who concluded that "it is always as or more efficient to adjust for the covariate when logistic regression is used"[23].

1.3. Our Motivation

We wanted to investigate the effect of covariate adjustment in the responder analysis framework, particularly when the covariate is involved in the definition of successful outcome. This problem was first posed by the data and safety monitoring board for the SHINE Trial (discussed below), which questioned whether adjusting for the prognostic variable twice—in the definition of success and in the analysis—would impact the test of treatment effect. Since the literature does not directly address this issue, we used the SHINE Trial as a basis for a simulation study to explore the consequences of covariate adjustment under the responder analysis framework.

1.3.1. The SHINE Trial

The Stroke Hyperglycemia Insulin Network Effort (SHINE) Trial is a large, multicenter, randomized clinical trial to determine the efficacy and safety of targeted glucose control in hyperglycemic acute ischemic stroke patients. Approximately 1400 subjects will be enrolled and randomized to receive either standard of care or targeted glucose control. To be eligible for the study, subjects must be enrolled within 12 hours of symptom onset and within 3 hours of Emergency Department arrival, as well as have a blood glucose concentration greater than 110 mg/dL on initial evaluation. Baseline stroke severity must fall between 3 and 22 (inclusive) on the NIHSS. Subjects in the standard of care treatment arm receive subcutaneous and basal insulin injections according to a sliding scale with a target blood glucose of $\leq$ 180 mg/dL; subjects in the intervention arm receive up to 72 hours of intravenous insulin infusion with a target blood glucose between 80 and 130 mg/dL.

The primary outcome for efficacy in the SHINE trial is the 90-day mRS score. Outcome is dichotomized as "success" or "failure" according to a sliding dichotomy. Those with a "mild" prognosis, defined by a baseline NIHSS score of 3-7, must achieve a 90-day mRS of 0 to be classified as a "success." Those with a "moderate" prognosis, defined by a baseline NIHSS score of 8-14, must achieve a 90-day mRS of 0-1 to be classified as a "success." Finally, those subjects with a "severe" prognosis, defined by a baseline NIHSS score of 15-22, must achieve a 90-day mRS of 0-2 to be classified as a "success." By using responder analysis, the milder strokes must meet a more stringent threshold to achieve success, while the more severe strokes have more leeway to achieve an attainable definition of success. Two pilot studies, THIS and GRASP, were used to establish the initial safety and efficacy estimates of intensive glucose control in hyperglycemic stroke patients[24,25].

1.4. Specific Aims

The focus of this research is covariate adjustment within the responder analysis framework, as the SHINE trial employs responder analysis in its primary statistical analyses. While the literature provides many resources on the design and implementation of responder analysis as well as examples of trials which used responder analysis, there are no clear resources supporting whether or not statistical analyses should be adjusted for the prognostic variables used to define successful outcome. The cutpoints for the SHINE trial are clinically, rather than statistically defined, and so it is conceivable that adjustment for baseline severity in the statistical analysis may account for additional variation. Through simulations, we aimed to explore the results of covariate adjustment

in the responder analysis setting. Our simulation parameters are based on those specified in the SHINE trial. The goals of this study were:

- To compare the operating characteristics of unadjusted and adjusted analyses under several different treatment effect scenarios in the responder analysis setting when treating baseline severity as a categorical variable with three levels.

- To compare the treatment effect estimates and their standard errors in unadjusted and adjusted analyses under several different treatment effect scenarios.

Since the primary outcome for the SHINE trial is binary, we expected to see an increase of standard error on the treatment effect estimates, consistent with the findings of Robinson and Jewell[23].

1.5. Significance

It is unclear whether or not statistical analyses should further adjust for those covariates involved in the definition of favorable outcome in trials that use responder analysis, such as the SHINE trial discussed above. The results of this study will help demarcate the best way to handle such analyses. These results will not only be applicable in the SHINE and other stroke trials which use the mRS for the primary outcome, but also in other trials which use any ordinal scale as a primary outcome measure and have a baseline prognostic factor.

CHAPTER 2: THE SIMULATION STUDY

We performed several simulation analyses where we examined the performance of logistic regression models that were unadjusted and adjusted by baseline severity category. Baseline severity category was defined as in the SHINE Trial described above: an mRS score of 3-7 was defined as "mild," 8-14 was defined as "moderate," and 15-22 was defined as "severe." As in the SHINE study plan, subjects in the mild group must achieve a 90-day mRS of 0, moderate must achieve a 90-day mRS of 0 or 1, and severe must achieve a 90-day mRS of 0, 1, or 2 to be considered a success. The type I error rate and power were calculated and compared for each method, as were the treatment effect estimates and their standard errors.

2.1. Simulation Methods

The simulation parameters were guided by the SHINE trial design. We simulated 1000 clinical trials at sample sizes 498 to 1958. This sample size range allows us to cover the planned sample size of 1400 while also examining model behavior at smaller and larger sample sizes. Though the SHINE study may begin response-adaptive randomization at some point during enrollment, we have assumed a 1:1 randomization scheme for the purposes of our investigation. All analyses were performed using SAS version 9.2 (SAS Institute, Cary, North Carolina).

The prevalence of each baseline severity category was guided by the THIS and GRASP pilot trial data[24,25]. The pilot trials had similar distributions of baseline severity categories, and thus it is reasonable to assume that the SHINE study population will have a similar distribution. In our simulations, we have assumed that 42% of subjects will be classified as "mild" at baseline, 32% will be classified as "moderate" at baseline, and the remaining 26% will be classified as "severe" at baseline. These classifications were randomly assigned using a uniform [0,1] random variable.

The simulation of study outcome (90-day mRS) differed by treatment group. In order to simulate 90-day mRS scores for the control group, we examined the distribution of 90-day mRS scores for the control groups in the THIS and GRASP pilot trials. Since these trials were both very small, we could not derive a good approximation of the distribution of mRS scores in each of the baseline severity strata. We used the control group from the NINDS tPA trial data to help in the approximation of mRS distributions[12]. The control group distribution of 90-day mRS scores used in this simulation study is shown in Table 1.

Table 1: Distribution of 90-Day mRS Scores for Control Group

Baseline Severity	90-Day mRS						
	0	1	2	3	4	5	6
Mild	0.25	0.30	0.20	0.10	0.08	0.02	0.05
Moderate	0.15	0.20	0.23	0.12	0.16	0.04	0.10
Severe	0.03	0.05	0.07	0.19	0.20	0.21	0.25

Type I error rates for each method were obtained by using the distribution of 90-day mRS scores found in Table 1 for both the control and intervention groups, simulating the null hypothesis of no treatment effect. In order to assess the power of each method, a

treatment effect was simulated in the data by altering the distribution of 90-day mRS scores in Table 1 for the intervention group only. We added a 7% treatment effect, as this was the minimum clinically significant difference defined in the SHINE study plan. For these analyses, we only examined power under several different treatment effect scenarios: (1) a "flat" treatment effect scenario, in which a 7% treatment effect was applied in each baseline severity stratum; (2) a "varying" treatment effect scenario, in which there is still an overall 7% treatment effect, but the magnitude within strata varies and the mild and moderate groups see the most benefit; (3) a "varying" treatment effect scenario, in which there is still an overall 7% treatment effect, but the severe group sees the most benefit; (4) a "mild harm" treatment effect scenario, in which there is still an overall 7% treatment effect, but the mild group sees a harmful treatment effect; and (5) a "severe harm" treatment effect scenario, in which there is still an overall 7% treatment effect, but the severe group sees a harmful treatment effect.

The flat treatment effect was achieved by allowing 7% more prevalence in the defined "success" mRS categories for each stratum. In the first varying treatment effect scenario, we applied an 8.6% treatment effect in the mild category, a 9% treatment effect in the moderate category, and a 2% treatment effect in the severe category; that is, there was an 8.6% increase in prevalence of the 0 mRS for the mild stratum, a 9% increase in the prevalence of the 0-1 range of mRS scores for the moderate stratum, and a 2% increase in the prevalence of the 0-2 range of mRS scores for the severe stratum. This scenario is relevant for the SHINE trial; it is similar to what we may observe if the intensive glucose control intervention is largely beneficial to mild and moderate stroke victims, but only marginally beneficial to victims of severe stroke. Similarly, in the

14

second varying treatment effect scenario, we applied a 2% treatment effect in the mild category, a 9% treatment effect in the moderate category, and a 12.6% treatment effect in the severe category. This scenario could also be observed in the SHINE results if the intensive glucose control intervention is largely beneficial to more severe strokes, but only slightly beneficial to those subjects having mild strokes. Tables 2, 3, and 4 show the distribution of 90-day mRS scores for the treatment groups under these flat and varying treatment effects, respectively.

Table 2: Distribution of 90-Day mRS Scores for Treatment Group:
"Flat" Treatment Effect

Baseline Severity	90-Day mRS						
	0	1	2	3	4	5	6
Mild	0.32	0.27	0.19	0.08	0.07	0.02	0.05
Moderate	0.17	0.25	0.21	0.10	0.15	0.03	0.09
Severe	0.04	0.06	0.12	0.18	0.18	0.19	0.23

Table 3: Distribution of 90-Day mRS Scores for Treatment Group:
First "Varying" Treatment Effect

Baseline Severity	90-Day mRS						
	0	1	2	3	4	5	6
Mild	0.336	0.31	0.19	0.06	0.04	0.02	0.044
Moderate	0.19	0.25	0.25	0.10	0.10	0.02	0.09
Severe	0.03	0.055	0.085	0.20	0.19	0.20	0.24

Table 4: Distribution of 90-Day mRS Scores for Treatment Group:
Second "Varying" Treatment Effect

Baseline Severity	90-Day mRS						
	0	1	2	3	4	5	6
Mild	0.27	0.31	0.20	0.09	0.06	0.02	0.05
Moderate	0.19	0.25	0.25	0.10	0.10	0.02	0.09
Severe	0.04	0.09	0.146	0.18	0.12	0.194	0.23

To achieve the scenarios in which one of the strata experienced harm, we applied similar effects as above, only allowing one of the strata to see a decrease in the prevalence within its defined success categories. In the "mild harm" scenario, we applied a -2% treatment effect in the mild category, with a 15% treatment effect in the moderate category and an 11.7% treatment effect in the severe category. Similarly, in the "severe harm" scenario, we applied an 8% treatment effect in the mild category, a 13% treatment effect in the moderate category, and a -2% treatment effect in the severe category. Either of these scenarios could possibly be observed in SHINE if the intervention if the intensive glucose control interferes with the body's natural recovery processes following a milder or more severe stroke, respectively. Tables 5 and 6 show the distribution of 90-day mRS scores for the mild and severe harm scenarios.

Table 5: Distribution of 90-Day mRS Scores for Treatment Group: "Mild Harm" Treatment Effect

Baseline Severity	90-Day mRS						
	0	1	2	3	4	5	6
Mild	0.23	0.29	0.21	0.12	0.08	0.02	0.05
Moderate	0.20	0.30	0.20	0.08	0.10	0.03	0.09
Severe	0.05	0.09	0.127	0.13	0.16	0.20	0.243

Table 6: Distribution of 90-Day mRS Scores for Treatment Group: "Severe Harm" Treatment Effect

Baseline Severity	90-Day mRS						
	0	1	2	3	4	5	6
Mild	0.33	0.29	0.15	0.09	0.07	0.03	0.04
Moderate	0.20	0.28	0.18	0.11	0.12	0.03	0.08
Severe	0.02	0.04	0.07	0.20	0.21	0.21	0.25

The distributions in Tables 1 through 6 were used to randomly assign 90-day mRS scores to each simulated subject in each simulated trial. The distribution in Table 1 was used for the control group, regardless of the treatment effect scenario being examined. When no treatment effect was applied in order to investigate type I error rates, Table 1 was used for the intervention group as well. The distributions in Tables 2 through 6 were used to assign mRS scores to subjects in the intervention group in order to investigate power under fixed and varying treatment effects, respectively. Given a subject's simulated baseline severity stratum (mild, moderate, or severe), an assignment of "success" or "failure" was made according to the sliding dichotomy definitions.

Logistic regression was used to investigate each of these scenarios. We examined unadjusted and categorically-adjusted analyses for each scenario. The unadjusted case models "success" as a function of only treatment group, while the categorically-adjusted case models "success" as a function of treatment group and severity category. For the power and type I error rate estimation, we created an indicator variable to denote the rejection of the null hypothesis for each of the 1000 simulated trials at a given sample size. The proportion of simulated trials at a given sample size which were rejected is our power/type I error rate estimation at that sample size. We also extracted the treatment effect estimate and its standard error for each trial.

2.2. Results

### 2.2.1. Type I error rates and power.

The type I error rate (significance level) at each sample size for each method is plotted below in Figure 1. The nominal 5% reference line is shown, as well as upper and lower 95% confidence limits on this nominal level. The confidence limits were

calculated using the formula for binomial proportion 95% confidence intervals, yielding the following equation:

$$p \pm z_{0.975}\sqrt{\frac{p(1-p)}{n}} = 0.05 \pm 1.96\sqrt{\frac{0.05(1-0.05)}{1000}} = (0.0365, 0.0635)$$

These confidence limits were then multiplied by 100 to be expressed in terms of percentages. The sample size is 1000, since we have simulated 1000 trials at each sample size in order to get our significance level estimates.

Figure 1: Significance Levels of Unadjusted and Categorically-Adjusted Methods
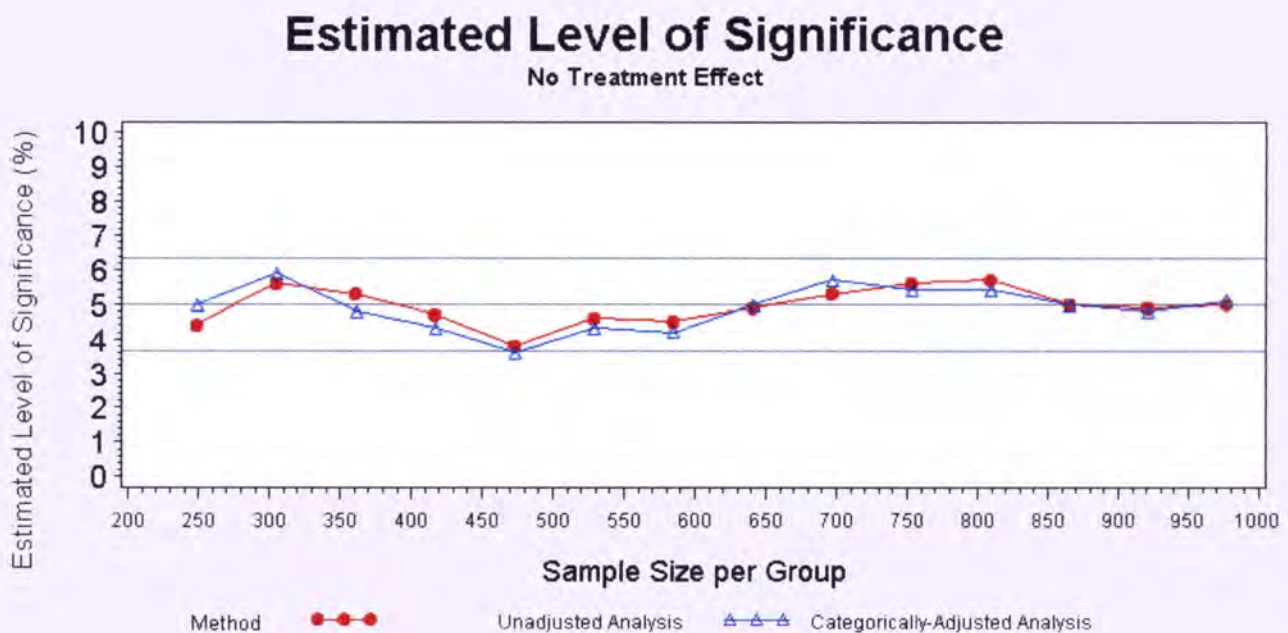
We can observe that both the unadjusted and categorically-adjusted methods have type I error rates within the 95% confidence bounds at each sample size. This is a welcomed result, as an inflated type I error rate results in a test that is too liberal, while a deflated type-I error rate results in a decrease in power and thus a test that is too conservative. The oscillation around the nominal 5% level of significance is due to chance, and is to be expected in experimental or simulated data. Since neither method

18

shows consistently larger type I error rates, we can conclude that there is no meaningful difference between the two methods with respect to significance level under the chosen treatment effect setting.

Our first investigation of power was under a "flat" treatment effect, in which each of the three prognosis groups experienced a simulated 7% treatment effect. The success rates in the control group were 25%, 35%, and 15% in the mild, moderate, and severe prognosis groups, respectively. These control rates are based on the pilot trials for SHINE, and their distributions are further detailed in Table 1. The power estimates for this "flat" treatment effect scenario are plotted in Figure 2 below. Under a true treatment effect of 7%, the SHINE study is designed to have at least 80% power, which is referenced along the plot in Figure 2.

Figure 2: Power of Unadjusted and Categorically-Adjusted Methods Under a Flat 7% Treatment effect
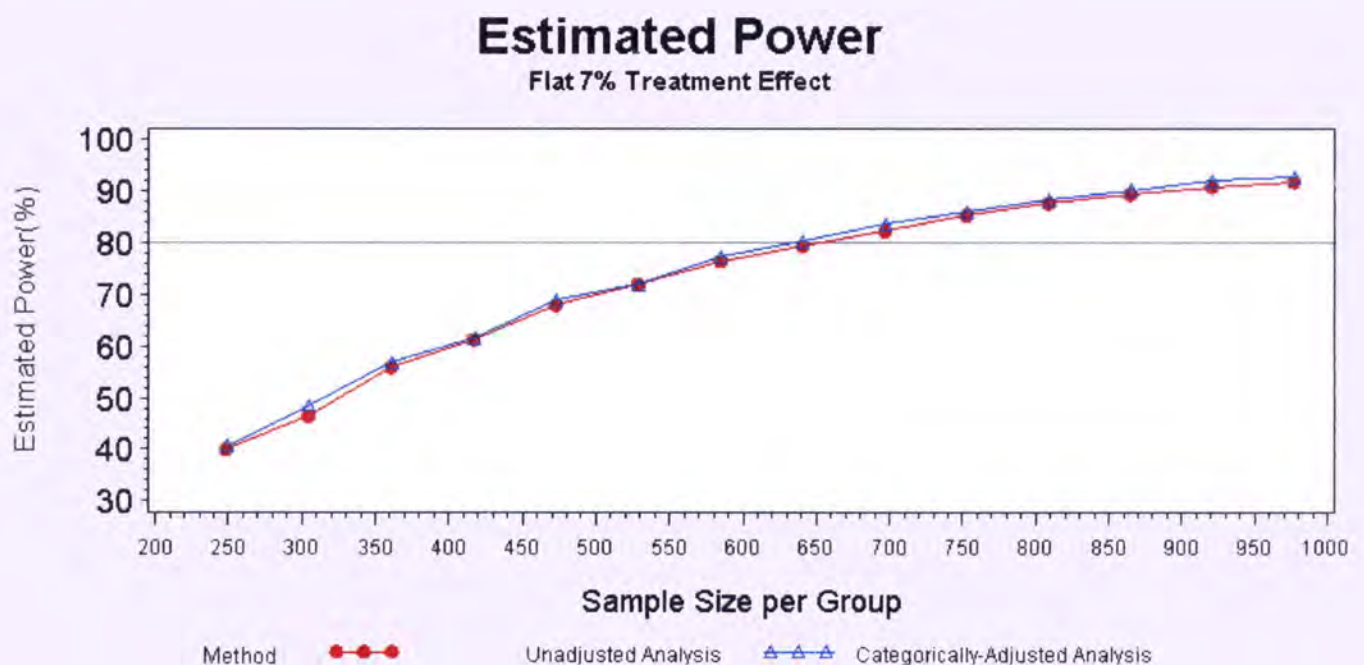
The unadjusted and categorically-adjusted methods do not appear to differ significantly in the plot in Figure 2. The two methods are nearly stacked in most places, with the categorically-adjusted method having very slightly greater power along much of the plot. As planned by the SHINE study investigators, the 80% power threshold is crossed between 650 and 700 subjects per arm (1300-1400 subjects total). The slight appearance of powering beyond the 80% level at the planned 1400 total subjects is expected, as the sample size estimation was slightly inflated for potential non-adherence.

Though our results from the simple flat treatment effect scenario were a good starting point, it is unlikely that we will see a uniform treatment effect across all strata in practice. To continue our investigation under an alternative scenario, we next considered the possibility of a treatment effect that varies across prognosis strata, but maintains the overall 7% treatment effect. We first allowed the mild, moderate, and severe baseline categories to have treatment effects of 8.6%, 9%, and 2%, respectively. Then, we allowed the mild, moderate, and severe baseline categories to have treatment effects of 2%, 9%, and 12.6%, respectively. The power results for these two scenarios are plotted in Figures 3 and 4.

Figure 3: Power of Unadjusted and Categorically-Adjusted Methods Under the First Varying 7% Treatment effect

Estimated Power
Varying Treatment Effect by Prognosis Group
Mild 8.6%, Moderate 9%, Severe 2% Treatment Effect
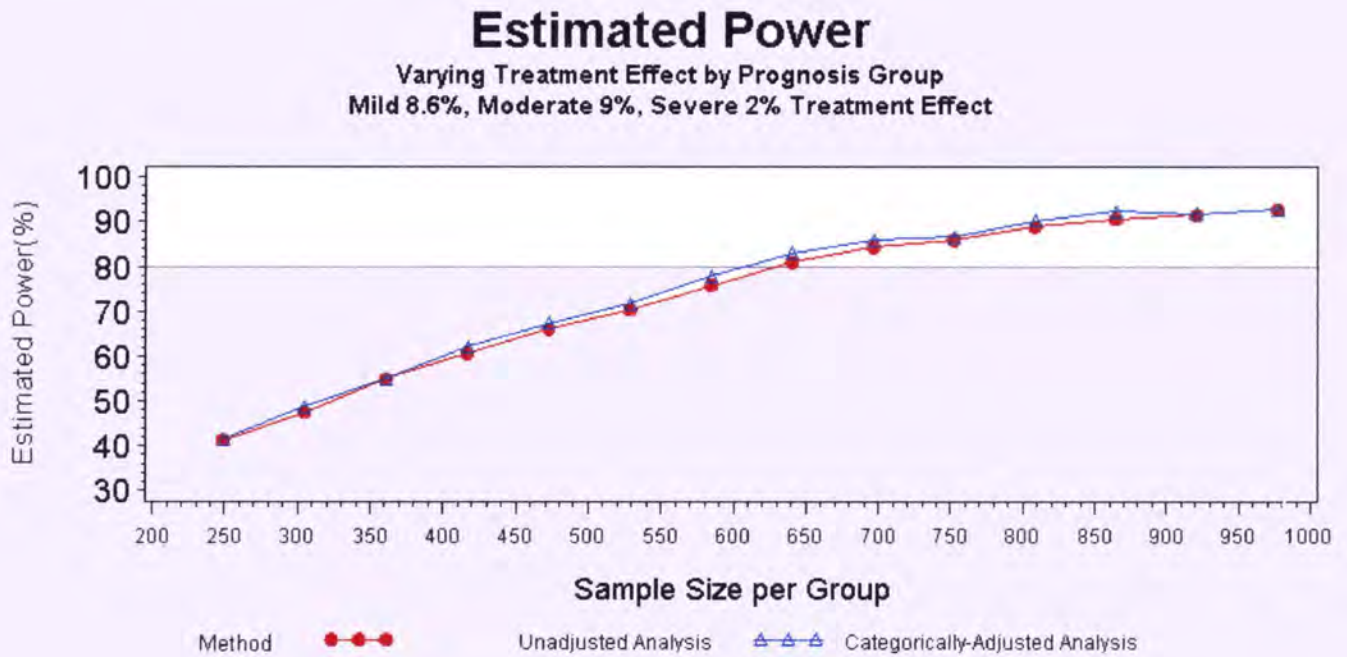
Figure 4: Power of Unadjusted and Categorically-Adjusted Methods Under the Second Varying 7% Treatment effect

Estimated Power
Varying Treatment Effect by Prognosis Group
Mild 2%, Moderate 9%, Severe 12.6% Treatment Effect
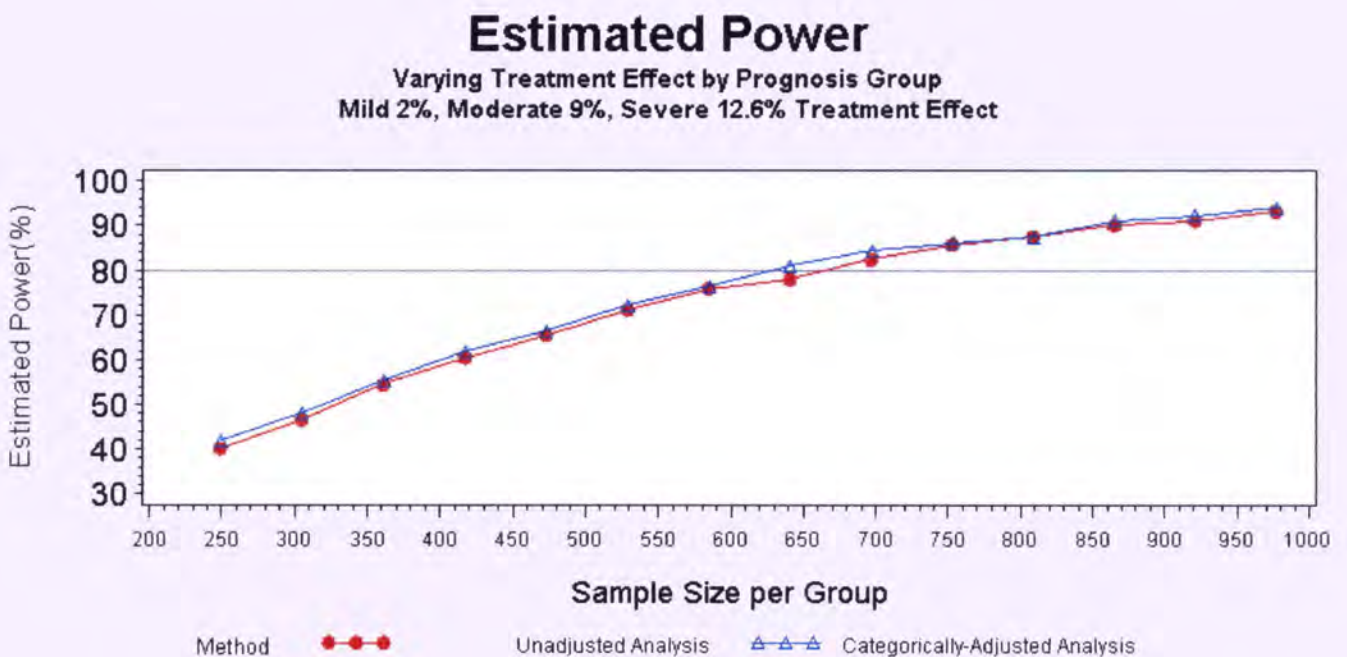
As in the flat treatment effect scenario, we do not see a drastic difference in the

unadjusted and categorically-adjusted methods with respect to power in these varying

21

treatment effect scenarios. There is a slightly larger gap between the categorically-adjusted and unadjusted methods' power curves at points in Figures 3 and 4 when compared to Figure 2, but the difference is not remarkable. It is reassuring, however, that we do not observe a noticeable decrease in power under the varying treatment effects for either method; this will be especially important if the study data reveals that the treatment effect truly does vary by prognosis stratum in either of these manners.

As previously mentioned, it is conceivable that one of the prognosis groups may experience a slightly harmful treatment effect. To investigate the consequences of adjusting for baseline severity in this situation, we examined scenarios in which the mild and severe groups experienced a -2% treatment effect. The results of these simulations are in Figures 5 and 6.

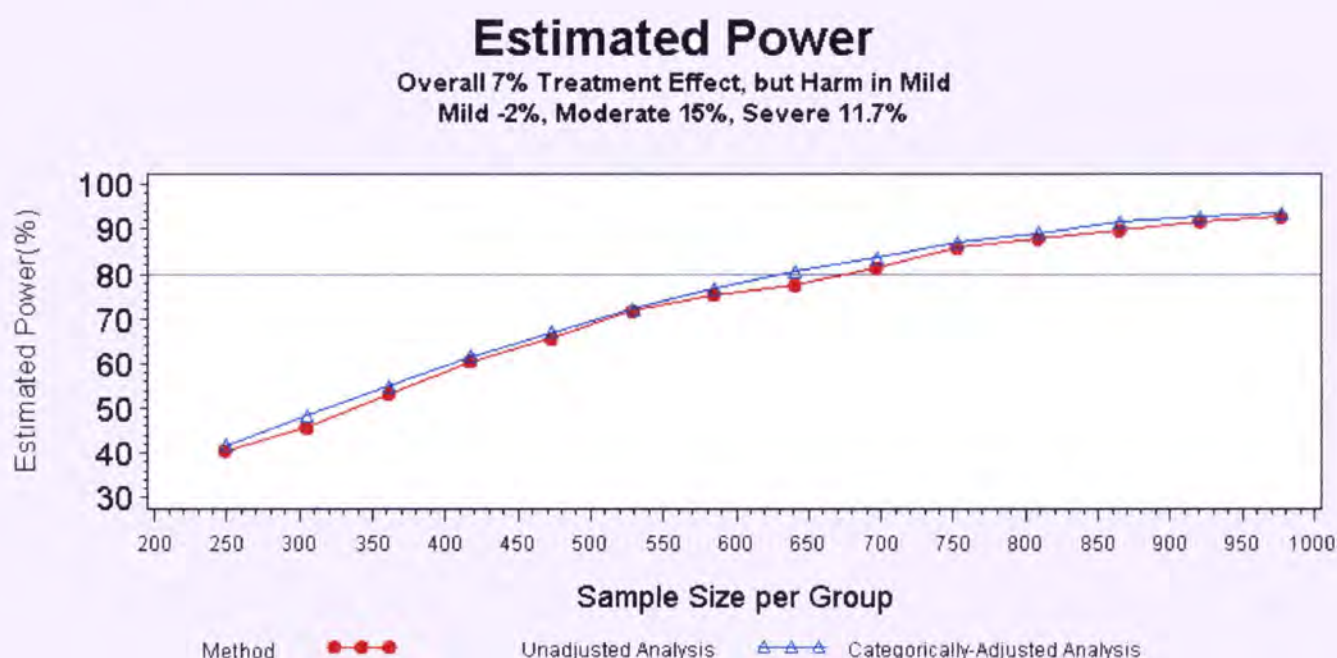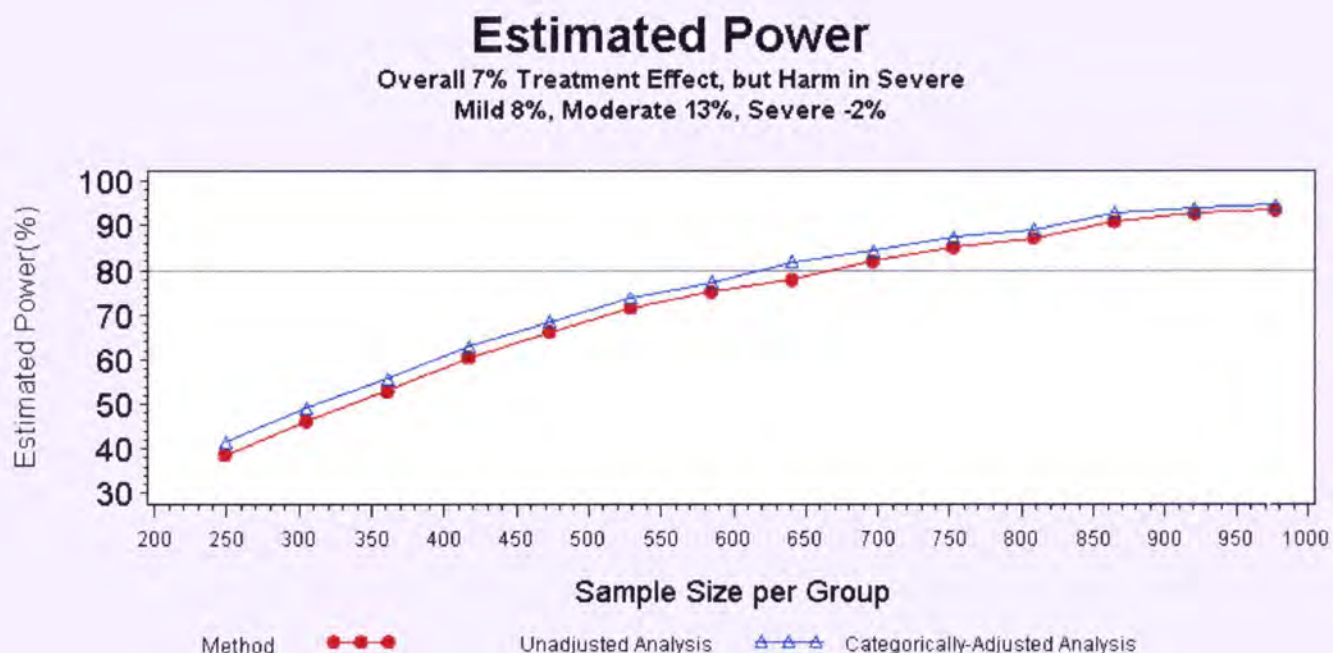*Figure 5: Power of Unadjusted and Categorically-Adjusted Methods Under a Mild Harm Effect*

Estimated Power
Overall 7% Treatment Effect, but Harm in Mild
Mild -2%, Moderate 15%, Severe 11.7%

Figure 6: Power of Unadjusted and Categorically-Adjusted Methods Under a Severe Harm Effect

Estimated Power
Overall 7% Treatment Effect, but Harm in Severe
Mild 8%, Moderate 13%, Severe -2%

Method ●●● Unadjusted Analysis △△△ Categorically-Adjusted Analysis

When 2% harm is experienced in either the mild or the severe baseline prognosis category, the unadjusted and adjusted analyses still appear to perform very similarly. In the mild harm scenario, the unadjusted and adjusted power curves are still nearly stacked upon one another, with the power curve for the adjusted analysis pulling slightly above that of the unadjusted analysis at a few points. A more noticeable difference can be seen in the severe harm scenario, where the adjusted analysis consistently has a slightly higher power than that of the unadjusted analysis. Though the power appears to be slightly higher for the adjusted analysis, this difference is still not very remarkable, and offers little evidence to suggest that adjusting is significantly more powerful under this scenario.

2.2.2. Treatment effect estimates and their standard errors.

In addition to the plots in Figures 2 through 6, we also observed the treatment coefficient estimates and their standard errors for the adjusted and unadjusted models in

the various treatment effect scenarios. The coefficient and standard error estimates are averaged over all simulations and all sample sizes, and are displayed in Table 7.

Table 7: Treatment Coefficient Estimates and Their Standard Errors for Unadjusted and Adjusted Methods Under Different Treatment Effect Scenarios

Scenario	Unadjusted		Adjusted	
	β_{trt} Estimate	SE (β_{trt})	β_{trt} Estimate	SE (β_{trt})
"Flat"	0.3430275	0.1378434	0.3535594	0.1401410
1st Varying	0.3427666	0.1379159	0.3577844	0.1410730
2nd Varying	0.3404175	0.1377643	0.3504053	0.1400125
Mild Harm	0.3409949	0.1377578	0.3568029	0.1411719
Severe Harm	0.3408927	0.1377668	0.3610773	0.1419651

As expected, we observed a slight inflation of the standard errors when adjusting. This phenomenon, first described by Robinson and Jewell[23], is balanced by a slight increase in magnitude of the treatment coefficient estimate away from the null. It is this reciprocating increase that preserves (and may slightly increase) our power in the case of adjustment for categorical baseline severity.

CHAPTER 3: DISCUSSION

Successful stroke treatments are in high demand given stroke's large and detrimental effect on the worldwide population. Consequently, statistical methods that offer higher power to detect a true treatment effect are also in high demand, especially given the large number of unsuccessful stroke trials to date, and the consideration that many of these unsuccessful stroke trials may have failed due to study design. With this simulation study, we sought to determine whether adjustment for baseline severity within the responder analysis setting would be beneficial or harmful in terms of power and type I error rates when compared to an unadjusted analysis.

The results in Chapter 2 show little evidence for or against adjusting for baseline severity in the responder analysis setting. The type I error rates between the two methods did not seem to differ substantially, and the power curves for most treatment effect scenarios examined were practically identical for the adjusted and unadjusted methods. In the case where the interventional treatment caused slight harm to the severe baseline group, the power for the adjusted analysis consistently appeared to be slightly greater than that for the unadjusted analysis; however, this difference was small and not noteworthy. These results suggest that in most treatment effect scenarios, adjustment for baseline severity in the primary analyses may best be guided by individual study needs rather than a blanket guideline for all studies, as neither the adjusted nor unadjusted

method showed notable statistical advantage in our examples. As in any clinical trial setting, sensitivity analyses can be conducted with the alternative approach to provide confirmation of the results found in the primary analyses.

Though we have not shown the results here, we did examine other treatment effect scenarios which also yielded similar results. These scenarios included a flat and varying 15% treatment effect (instead of the 7% specified in the SHINE study plan), as well as a scenario in which the mild group experienced 5% harm. In the 15% treatment effect scenarios, the two methods were practically non-differentiable, as even at the smallest sample sizes examined the study was overpowered for such a large effect and the graphs plotted on top of one another. The 5% mild harm scenario yielded very similar results to those seen in Chapter 2 with the 2% mild harm. Given that none of these results uniquely contributed to our conclusion, we have omitted their results here in interest of space.

It is important to note that these analyses adjust for baseline severity categorically. These categories—mild, moderate, and severe—are defined by the NIHSS score, which is a larger ordinal scale ranging from 0 to 42 (limited to 3 through 22 in SHINE's inclusion criteria). It is possible that adjusting by the actual NIHSS score will provide additional information to the model and increase or maintain power in some treatment effect scenario(s). Though the NIHSS is technically an ordinal scale, it is sometimes used as a continuous measure in the literature[26,27]. However, interpreting the NIHSS as a continuous measure is not necessarily straightforward and should be done with caution; a one-unit increase on the NIHSS at one location on the scale may not have the same implications as a one-unit increase in another location. Development of a

simulation study to investigate adjustment for baseline severity continuously by NIHSS compared with categorically-adjusted and unadjusted analyses has been proposed as future work.

3.1. Future Work

As discussed above, future work will include the investigation into adjustment by NIHSS as a continuous covariate. Though SHINE uses only baseline severity category to define success, some studies—including the GRASP pilot trial for SHINE— use multiple baseline characteristics in a prognostic model to define success thresholds, as discussed in Section 1.1.4[25]. Future work could potentially involve the exploration of adjusting for these baseline characteristics when they are used in such a prognostic model. This study was performed under a perfect one-to-one treatment allocation rate, but we may also use our simulations to investigate how an imbalance in treatment allocation impacts the analyses. Allowing the prevalence within each of the baseline severity categories to vary from what was observed in the pilot trials may also provide interesting results, since baseline severity dictates the definition of successful outcome in responder analysis. When the SHINE Trial concludes, a repeat of these analyses using the actual data may be used to confirm our findings. In addition, future work may include the establishment of the theoretical basis for our findings.

Our immediate next step is to explore the various analysis techniques on a publicly-available clinical trial dataset. We aim to compare not only the unadjusted and categorically-adjusted analysis methods in the responder analysis settings, but then also compare these methods with other analysis techniques available for ordinal data. This comparison will help us verify (or dispute) our findings with respect to adjusting in the

responder analysis setting as well as look at the difference in statistical power between responder analysis and other statistical analysis techniques.

3.2. Conclusion

There does not appear to be an impact in terms of statistical operating characteristics whether the analyses are unadjusted or adjusted by baseline severity category in the treatment effect scenarios examined in this simulation study. While we had hoped to find that one of the two methods had significantly greater power, it should be noted that these results are not negative. Instead, they suggest that adjustment by baseline severity category is a matter of individual study needs and has little effect on the statistical operating characteristics of the analysis. These results are not restricted to use in stroke studies; they are generalizable to any type of study which uses responder analysis to define its primary outcome of interest.

APPENDIX: SIMULATION SAS CODE

Included below is the SAS code for the macro that is used to create each dataset under the various scenarios. In addition, I have included how it would be used in the case of no treatment effect in order to examine type I error rates. Other treatment effect scenarios can be attained similarly by changing the prevalence cutoffs for the mRS distribution, as described in Tables 1 through 6.

```
**Run this macro first.  Then, run section(s) for desired treatment
effects

%macro
simulatetrial(sampsize=,Cmild1=,Cmild2=,Cmild3=,Cmild4=,Cmild5=,Cmild6=
,Cmod1=,Cmod2=,Cmod3=,Cmod4=,Cmod5=,Cmod6=,Csev1=,Csev2=,Csev3=,Csev4=,
Csev5=,Csev6=,Tmild1=,Tmild2=,Tmild3=,Tmild4=,Tmild5=,Tmild6=,Tmod1=,Tm
od2=,Tmod3=,Tmod4=,Tmod5=,Tmod6=,Tsev1=,Tsev2=,Tsev3=,Tsev4=,Tsev5=,Tse
v6=);

data simulatedtrial;
do i=1 to &sampsize;
half=&sampsize/2;                   *sets up treatment randomization;
     if i le half then trt='A';
     else trt='B';
rand_prog=ranuni(80272+&nsim);      *set up prognosis groups;
     if rand_prog le 0.42 then prognosis=1;
     else if rand_prog gt 0.42 and rand_prog le 0.74 then prognosis=2;
     else if rand_prog gt 0.74 and rand_prog le 1.0 then prognosis=3;
     else prognosis=999; *safeguard check;
output;
end;
run;

data simulatedtrial;
set simulatedtrial;
uniform=ranuni(20453+&nsim);
if trt='A' then do;      *TRT A IS CONTROL;

 if prognosis=1 then do;
  if uniform le &Cmild1 then simrankin=0;
  else if uniform > &Cmild1 and uniform le &Cmild2 then simrankin=1;
```

```
     else if uniform > &Cmild2 and uniform le &Cmild3 then simrankin=2;
        else if uniform > &Cmild3 and uniform le &Cmild4 then simrankin=3;
        else if uniform > &Cmild4 and uniform le &Cmild5 then simrankin=4;
        else if uniform > &Cmild5 and uniform le &Cmild6 then simrankin=5;
        else if uniform > &Cmild6 then simrankin=6;
          if simrankin=0 then stratout=1     *Define Success
          else stratout=0;
        end;

    if prognosis=2 then do;
     if uniform le &Cmod1 then simrankin=0;
     else if uniform > &Cmod1 and uniform le &Cmod2 then simrankin=1;
     else if uniform > &Cmod2 and uniform le &Cmod3 then simrankin=2;
     else if uniform > &Cmod3 and uniform le &Cmod4 then simrankin=3;
     else if uniform > &Cmod4 and uniform le &Cmod5 then simrankin=4;
     else if uniform > &Cmod5 and uniform le &Cmod6 then simrankin=5;
     else if uniform > &Cmod6 then simrankin=6;
         if simrankin le 1 then stratout=1;          *Define Success;
         else stratout=0;
    end;

if prognosis=3 then do;
 if uniform le &Csev1 then simrankin=0;
 else if uniform > &Csev1 and uniform le &Csev2 then simrankin=1;
 else if uniform > &Csev2 and uniform le &Csev3 then simrankin=2;
 else if uniform > &Csev3 and uniform le &Csev4 then simrankin=3;
 else if uniform > &Csev4 and uniform le &Csev5 then simrankin=4;
 else if uniform > &Csev5 and uniform le &Csev6 then simrankin=5;
 else if uniform > &Csev6 then simrankin=6;
     if simrankin le 2 then stratout=1;      *Define success;
     else stratout=0;
end;
end;


else if trt='B' then do;

if prognosis=1 then do;
 if uniform le &Tmild1 then simrankin=0;
 else if uniform > &Tmild1 and uniform le &Tmild2 then simrankin=1;
 else if uniform > &Tmild2 and uniform le &Tmild3 then simrankin=2;
 else if uniform > &Tmild3 and uniform le &Tmild4 then simrankin=3;
 else if uniform > &Tmild4 and uniform le &Tmild5 then simrankin=4;
 else if uniform > &Tmild5 and uniform le &Tmild6 then simrankin=5;
 else if uniform > &Tmild6 then simrankin=6;
    if simrankin=0 then stratout=1;    *Define success;
    else stratout=0;
end;

if prognosis=2 then do;
 if uniform le &Tmod1 then simrankin=0;
 else if uniform > &Tmod1 and uniform le &Tmod2 then simrankin=1;
 else if uniform > &Tmod2 and uniform le &Tmod3 then simrankin=2;
 else if uniform > &Tmod3 and uniform le &Tmod4 then simrankin=3;
 else if uniform > &Tmod4 and uniform le &Tmod5 then simrankin=4;
```

30

```
    else if uniform > &Tmod5 and uniform le &Tmod6 then simrankin=5;
   else if uniform > &Tmod6 then simrankin=6;
      if simrankin le 1 then stratout=1;      *Define success;
      else stratout=0;
 end;

 if prognosis=3 then do;
  if uniform le &Tsev1 then simrankin=0;
  else if uniform > &Tsev1 and uniform le &Tsev2 then simrankin=1;
  else if uniform > &Tsev2 and uniform le &Tsev3 then simrankin=2;
  else if uniform > &Tsev3 and uniform le &Tsev4 then simrankin=3;
  else if uniform > &Tsev4 and uniform le &Tsev5 then simrankin=4;
  else if uniform > &Tsev5 and uniform le &Tsev6 then simrankin=5;
  else if uniform > &Tsev6 then simrankin=6;
     if simrankin le 2 then stratout=1;      *Define success;
     else stratout=0;
 end;
 end;
 run;


*Logistic Regression (Adjusted)
proc logistic data=simulatedtrial descending
outest=logisticresults_strat_adj covout noprint ;
      class trt prognosis(ref=last)/  desc param=reference;
      model stratout=trt prognosis;
run;

data estimate_strat_adj;
      set logisticresults_strat_adj;
      where _type_="PARMS";
      or_strat_adj=exp(trtB);
      keep trtB or_strat_adj;
      rename trtB=trtbeta_strat_adj;
run;

data variance_strat_adj;
      set logisticresults_strat_adj;
      where _type_="COV" and _name_="trtB";
      keep trtB;
      rename trtB=var_trtbeta_strat_adj;
run;

*Logistic Regression (Unadjusted);
proc logistic data=simulatedtrial descending
outest=logisticresults_strat_un covout noprint ;
      class trt prognosis(ref=last)/  desc param=reference;
      model stratout=trt;
run;

data estimate_strat_un;
      set logisticresults_strat_un;
      where _type_="PARMS";
      or_strat_un=exp(trtB);
```

31

```
          keep trtB or_strat_un;
          rename trtB=trtbeta_strat_un;
run;

data variance_strat_un;
          set logisticresults_strat_un;
          where _type_="COV" and _name_="trtB";
          keep trtB;
          rename trtB=var_trtbeta_strat_un;
run;

proc iml;
          nsim=&nsim;         *Record the simulation number;

*Determine the significance of the treatment effect based on the
logistic regression model;

use estimate_strat_adj;
read all var {trtbeta_strat_adj} into beta_strat_adj;
read all var {or_strat_adj} into or_strat_adj;
use variance_strat_adj;
read all var {var_trtbeta_strat_adj} into var_beta_strat_adj;
logistic_strat_adj_ts=(beta_strat_adj/sqrt(var_beta_strat_adj))##2;
logistic_strat_adj_reject=logistic_strat_adj_ts>3.84;

use estimate_strat_un;
read all var {trtbeta_strat_un} into beta_strat_un;
read all var {or_strat_un} into or_strat_un;
use variance_strat_un;
read all var {var_trtbeta_strat_un} into var_beta_strat_un;
logistic_strat_un_ts=(beta_strat_un/sqrt(var_beta_strat_un))##2;
logistic_strat_un_reject=logistic_strat_un_ts>3.84;

use simulatedtrial;

read all var {prognosis} into prognosis;
read all var {nihss} into nihss;
read all var {trt} into trt;
read all var {stratout} into stratout;
ntrt=(trt='A')[+]; nplacebo=(trt='B')[+]; *Let a= active trt,
          b=placebo...Create sample sizes for each group;
npergroup=ntrt;

edit trtresults var     {nsim npergroup or_strat_adj
logistic_strat_adj_ts logistic_strat_adj_reject
or_strat_un logistic_strat_un_ts logistic_strat_un_reject
beta_strat_adj var_beta_strat_adj beta_strat_un var_beta_strat_un};

append var              {nsim npergroup or_strat_adj
logistic_strat_adj_ts logistic_strat_adj_reject
or_strat_un logistic_strat_un_ts logistic_strat_un_reject
beta_strat_adj var_beta_strat_adj beta_strat_un var_beta_strat_un};
run;quit;
%mend simulatetrial;
```

```
*Create an empty data set to hold simulation results;
data trtresults;
      input nsim npergroup or_strat_adj logistic_strat_adj_ts
logistic_strat_adj_reject or_strat_un logistic_strat_un_ts
logistic_strat_un_reject beta_strat_adj var_beta_strat_adj
beta_strat_un var_beta_strat_un;
datalines;

;
run;


%macro completesimulation;
      %do samplesize=498 %to 1958 %by 112; *Increases in sample size;
            %do nsim=1 %to 1000; DM 'CLEAR LOG';
      %simulatetrial(sampsize=&samplesize,Cmild1=0.25,Cmild2=.55,Cmild3
      =.75,Cmild4=.85,Cmild5=.93,Cmild6=.95,Cmod1=0.15,Cmod2=.35,Cmod3=
      .58,Cmod4=.7,Cmod5=.86,Cmod6=.9,Csev1=0.03,Csev2=.08,Csev3=.15,Cs
      ev4=.34,Csev5=.54,Csev6=.75,Tmild1=0.25,Tmild2=.55,Tmild3=.75,Tmi
      ld4=.85,Tmild5=.93,Tmild6=.95,Tmod1=0.15,Tmod2=.35,Tmod3=.58,Tmod
      4=.7,Tmod5=.86,Tmod6=.9,Tsev1=0.03,Tsev2=.08,Tsev3=.15,Tsev4=.34,
      Tsev5=.54,Tsev6=.75);
            %end;
      %end;
%mend completesimulation;


%completesimulation;


data sliding.notrt;
      set trtresults;
      run;


proc means data=sliding.notrt mean;
var beta_strat_un beta_strat_adj var_beta_strat_un var_beta_strat_adj ;
run;


proc sort data=sliding.notrt;
by npergroup;
run;
*Compute proportion of rejections;

proc freq data=sliding.notrt;
      tables logistic_strat_un_reject / out=logistic_strat_un_per ;
      by npergroup;
run;


proc freq data=sliding.notrt;
      tables logistic_strat_adj_reject / out=logistic_strat_adj_per ;
      by npergroup;
run;
```

33

REFERENCES

1 Sacco RL, Frieden TR, Blakeman DE, Jauch EC, Mohl S. What the Million Hearts Initiative means for stroke: a presidential advisory from the American Heart Association/American Stroke Association. *Stroke*. 2012;43:924-928.

2 Saver JL. Optimal endpoints for acute stroke therapy trials: best ways to measure treatment effects of drugs and devices. *Stroke*. 2011;42:2356-2362.

3 Hong KS, Lee SJ, Hao Q, Liebeskind DS, Saver JL. Acute stroke trials in the 1st decade of the 21st century. *Stroke*. 2011;42:e314.

4 The Optimising Analysis of Stroke Trials (OAST) Collaboration. Can we improve the statistical analysis of stroke trials?: Statistical reanalysis of functional outcomes in stroke trials. *Stroke*. 2007;38:1911-1915.

5 Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials. *Stroke*. 2007;38:1091-1096.

6 Saver, JL. Novel end point analytic techniques and interpreting shifts across the entire range of outcome scales in acute stroke trials. *Stroke*. 2007;38:3055-3062.

7 Young FB, Lees KR, Weir CJ. Strengthening acute stroke trials through optimal use of disability end points. *Stroke*. 2003;34:2676-2680.

8 Murray GD, Barer D, Choi S, Fernandes H, Gregson B, Lees KR, et al. Design and analysis of phase III trials with ordered outcome scales: the concept of the sliding dichotomy. *Journal of Neurotrauma*. 2005;22:511-517.

9 Savitz SI, Lew R, Bluhmki E, Hacke W, Fisher M. Shift analysis versus dichotomization of the modified Rankin scale outcome scores in the NINDS and ECASS-II trials. *Stroke*. 2007;38:3205-3212.

10 Saver JL, Gornbein J. Treatment effects for which shift or binary analyses are advantageous in acute stroke trials. *Neurology*. 2009;72:1310-1315.

11 Kasner SE. Clinical interpretation and use of stroke scales. *Lancet Neurology*. 2006;5:603-612.

12 The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. *New England Journal of Medicine*. 1995;333:1581-1588.

13 Saver JL, Yafeh B. Confirmation of tPA treatment effect by baseline severity-adjusted end point reanalysis of the NINDS-tPA stroke trials. *Stroke*. 2007;38:414-416.

14 Tilley BC, Marler J, Geller NL, Lu M, Legler J, Brott T, et al. Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial. *Stroke*. 1996;27:2136-2142.

15 Cobo E, Secades JJ, Miras F, Gonzalez JA, Saver JL, Corchero C, et al. Boosting the chances to improve stroke treatment. *Stroke*. 2010;41:e143-e150.

16 McHugh GS, Butcher I, Steyerberg EW, Marmarou A, Lu J, Lingsma HF, et al. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clinical Trials*. 2010;7:44-57.

17 Howard G, Waller JL, Voeks JH, Howard VJ, Jauch EC, Lees KR, et al. A simple, assumption-free, and clinically interpretable approach for analysis of modified Rankin outcomes. *Stroke*. 2012;43:664-669.

18 Young FB, Lees KR, Weir CJ. Improving trial power through use of prognosis-adjusted end points. *Stroke*. 2005;36:597-601.

19 Mendelow AD, Gregson BA, Fernandes HM, Murray GD, Teasdale GM, Hope DT, et al. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas in the International Surgical Trial in Intracerebral Haemorrhage (STICH): a randomised trial. *Lancet*. 2005;365:387-397.

20 Adams HP, Effron MB, Torner J, Davalos A, Frayne J, Teal P, et al. Emergency administration of abciximab for treatment of patients with acute ischemic stroke: results of an international phase III trial. *Stroke*. 2008;39:87-99.

21 Piantadosi S. *Clinical Trials: A Methodological Perspective*. 2nd edition. Hoboken, New Jersey: John Wiley and Sons, 2005, p. 470-473.

22 Harrell FE. The role of covariable adjustment in the analysis of clinical trials. 2010. http://biostat.mc.vanderbilt.edu/twiki/pub/Main/FHHandouts/covadj.pdf. Accessed May 7, 2012.

23 Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*. 1991;58:227-240.

24 Bruno A, Kent TA, Coull BM, Shankar RR, Saha C, Becker KJ, et al. Treatment of hyperglycemia in ischemic stroke (THIS) : a randomized pilot trial. *Stroke*. 2008;39:384-389.

25 Johnston KC, Hall CE, Kissela BM, Bleck TP, Conaway MR. Glucose regulation in acute stroke patients (GRASP) trial: a randomized pilot trial. *Stroke*. 2009;40:3804-3809.

26 Lin HJ, Chang WL, Tseng MC. Readmission after stroke in a hospital based registry: risk, etiologies, and risk factors. *Neurology*. 2011;76:438-443.

27 Dai DF, Thajeb P, Tu CF, Chiang FT, Chen CH, Yang RB, et al. Plasma concentration of SCUBE1, a novel platelet protein, is elevated in patients with acute coronary syndrome and ischemic stroke. *Journal of the American College of Cardiology*. 2008;51:2173-2180.