STATISTICAL ANALYSIS PLAN

Multi-arm Optimization of Stroke Thrombolysis Stroke Trial (MOST)

VERSION 2.0 | OCTOBER 20, 2020

Jordan Elm, Renee H. Martin
Data Coordination Unit
Department of Public Health Sciences
Medical University of South Carolina
Charleston, SC 29425

Scott Berry, Todd Graves
Berry Consultants

Table of Contents

1 Synopsis of the Study

This statistical analysis plan (SAP) is for the MOST trial. The SAP will define all "pre-specified, planned analyses." If there are minor differences between the MOST protocol and the SAP, the SAP will take precedence.

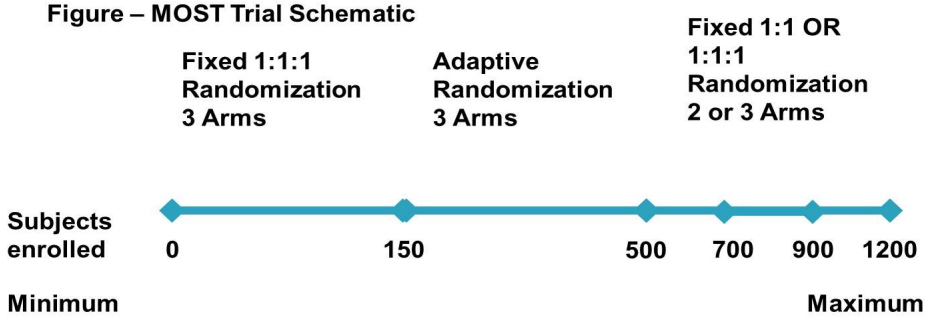
The MOST trial is a three-arm, adaptive, Phase-3, single blinded, randomized controlled clinical trial. The primary efficacy objective of the MOST trial is to determine if argatroban or eptifibatide improves 90-day modified Rankin scores (mRS) as compared with placebo in acute ischemic stroke (AIS) patients treated with IV thrombolysis within three hours of symptom onset.

2 Acronyms

Abbreviation	Description
AIS	Acute Ischemic Stroke
aPTT	activated Partial Thromboplastin Time
DSMB	Data and Safety Monitoring Board
ET	Endovascular thrombectomy
ICA	Internal carotid artery
ITT	Intent to Treat
IV	Intravenous
IQR	Interquartile Range
LVO	Large Vessel Occlusion
MCA	Middle Cerebral Artery
MCMC	Markov Chain Monte Carlo
mRS	modified Rankin Scale
NDMC	National Data Management Center
NIHSS	NIH Stroke Scale
NINDS	National Institute of Neurological Disorders and Stroke
RAR	Response Adaptive Randomization
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Standard Deviation
sICH	symptomatic Intracranial Hemorrhage
UW-mRS	Utility-Weighted modified Rankin Scale

3 Study Design

Figure – MOST Trial Schematic



This trial is designed as a multicenter, prospective, randomized, adaptive phase III clinical trial.

3.1 Treatment Arms

There are three treatment arms defined in the trial.

	Study Arms	Subjects will receive the following for a total of 12 hours of infusion
1	Argatroban	Argatroban (1mg/ml) bolus + Argatroban (1 mg/ml) 2-hour infusion + Argatroban (1mg/ml) 10-hour infusion
2	Eptifibatide	Eptifibatide (1.35 mg/ml) bolus + Eptifibatide (0.25 mg/ml) 2-hour infusion + placebo 10-hour infusion
3	Placebo	Placebo bolus + Placebo 2-hour infusion + placebo 10-hour infusion

NOTE: Commercially available product is assembled into study drug kits. Investigators are unblinded to treatment arms, but subjects and LARs are blinded throughout the duration of the trial.

4 Definition of the Target Population and Study Samples

4.1 Target Population

Acute ischemic stroke (AIS) patients treated with IV thrombolysis within three hours of symptom onset. The specific inclusion/exclusion criteria are listed in the protocol.

4.2 Enrollment Target

We will use 110 recruiting sites. Allowing for study startup and close down activities, we estimate an average of 0.3 patients/site/month or 396 patients/year during the peak enrollment period, consistent with StrokeNet estimates.

Enrollment Milestones and Timelines from MOST Grant Proposal																				
Task	Year 1				Year 2				Year 3				Year 4				Year 5			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Enrollment (goals by quarter)					10	40	65	85	100	100	100	100	100	100	100	100	100	100		

5 General Statistical Considerations

5.1 Subject Accountability

A flowchart will be created to present a summary of participant status. This flowchart will list the number of subjects screened and randomized to each treatment group. Then, within each group, the number of subjects who received at least some treatment, the number of subjects who did not receive any treatment, and the number of missing 90 day visits.

5.2 Randomization

The first 150 subjects will be randomized in a 1:1:1 ratio to either argatroban, eptifibatide, or placebo. From the 150th to the 500th subject enrollment, RAR) will favor the active arm showing the greatest benefit based on accrued data. After 500 subjects, one or both intervention arms may be carried forward for fixed randomization versus placebo.

A web-based central randomization system will be developed by the NDMC and installed on the WebDCU™ MOST study website. Age and NIHSS score will be included in the randomization algorithm to prevent imbalances, thereby enhancing the comparability of treatment groups. The detailed randomization scheme and source codes will be provided in the Randomization Plan document.

Response Adaptive Randomization
RAR begins at 150 subjects and ends at 500 subjects. Allocation probabilities for the two active arms will be updated approximately every month, based on Bayesian predictive probabilities. After RAR updates, the allocation probabilities for the two active arms are proportional to their 1:1 predictive probabilities, and the control arm receives randomization probability proportional to the maximum active arm.

5.3 Blinding

The NIH StrokeNet Central Pharmacy will assemble the commercially available study drug products into kits and ship to Clinical Performing Sites for study use. Investigators are unblinded to treatment arms, but subjects and LARs are blinded throughout the duration of the trial.

Argatroban and placebo subjects will initially receive a bolus followed by a 12-hour infusion. Eptifibatide arm subjects will receive a bolus followed by a 2-hour infusion. A 10-hour saline infusion will be administered after the 2-hour eptifibatide infusion to maintain the single blind.

The two primary concerns of lack of blinding are: 1) knowledge of the treatment arm could bias endpoint assessments of efficacy (mRS at 90 days) or safety (sICH); and, 2) possibility that

knowledge of an active treatment could influence the decision to proceed with ET in some patients. Centralized blinded 90-day mRS assessments and centralized blinded adjudication of sICH will mitigate #1, and requiring a clinical decision regarding a plan to proceed to ET prior to randomization will mitigate #2.

5.4 Missing Data

Under the ITT principle, all subjects who are randomized are included in the analysis. Every attempt will be made to avoid missing outcome data, but missing data is anticipated with any longitudinal study. Reasons for missing data will be examined by comparing means and standard deviations for subjects with and without the relevant data. Missing data on the primary outcome (90-day mRS) will be imputed using a Bayesian multiple imputation method that accounts for baseline NIHSS and the 30-day mRS. If no 30 or 90 day mRS is available, the subject's 30-day mRS is singley imputed using a hot-deck method which randomly selects the outcome from a pool of patients with observed mRS scores matched by treatment group, age and baseline NIHSS, time of symptom onset and EVT.

5.5 Treatment Group Comparability

A description of the baseline characteristics of trial participants will be presented by treatment group. Dichotomous variables will be summarized as number (%). Percentages will be calculated based on the number of participants with available data for that variable. Continuous/ordinal variables will be summarized as either mean (SD) or median (IQR) depending on whether the distribution is skewed. In the case of variables with missing values, the denominator will be stated in the summary table or in a footnote to the summary table.

Baseline Characteristics	Control (N=)	Argatroban (N=)	Eptifibatide (N=)
Age, median IQR			
Female sex, count (%)			
Race, count (%)			
White			
Black			
Asian			
Other			
Hispanic, count (%)			
Type of IV thrombolysis rtPA TNK			
Stroke Onset to thrombolysis bolus (min) , mean (SD)			
Stroke Onset to randomization (min) , mean (SD)			
Baseline NIHSS, median (IQR)			

Baseline aPTT, median (IQR)			
Prior use of , count (%) Aspirin Clopidogrel Warfarin Other Antiplatelet Other Anticoagulant Statins Antihypertensives Diabetic medications			
Past medical history, number (%) Myocardial infarction Atrial fibrillation Hypertension Diabetes Prior Stroke (excluding the qualifying stroke) Hypercholesterolemia			
ASPECTS score on Baseline CT[ŧ], median (IQR)			
Historical mRS, median IQR			
Decision to go to ET at the time of randomization , count (%)			
Presence of LVO at baseline[**ŧ], count (%)			

**Amongst those patients for whom the decision was made to go to ET at the time of randomization

ŧ determined by central neuroimaging readers (unless CTA is not available because of the contrast in which case the LVO determination made at the time of ET initiation will be used.)

5.6 Multiplicity

Inflation of the type I error rate due to the planned interim analysis of the primary outcome is controlled by using $\Pr(\theta_d > 0) \geq 0.985$ rather than 0.975 (details below). This threshold was chosen to control the experiment-wise type I error probability at 0.025.

If both treatments are better than control, then we will compare the two active treatment arms by using $\Pr(|\theta_1 - \theta_2| > 0) \geq 0.975$. All secondary/tertiary analyses will be tested at a significance level of two-sided α=0.05.

For secondary outcomes, we will not account for multiplicity. These analyses are merely supportive and exploratory and will be interpreted as such. The overall trial hinges on the primary analysis and interpretation of secondary results is already tempered by their placement. Accounting for multiple testing would also increase the risk that an interesting finding is not subsequently pursued.

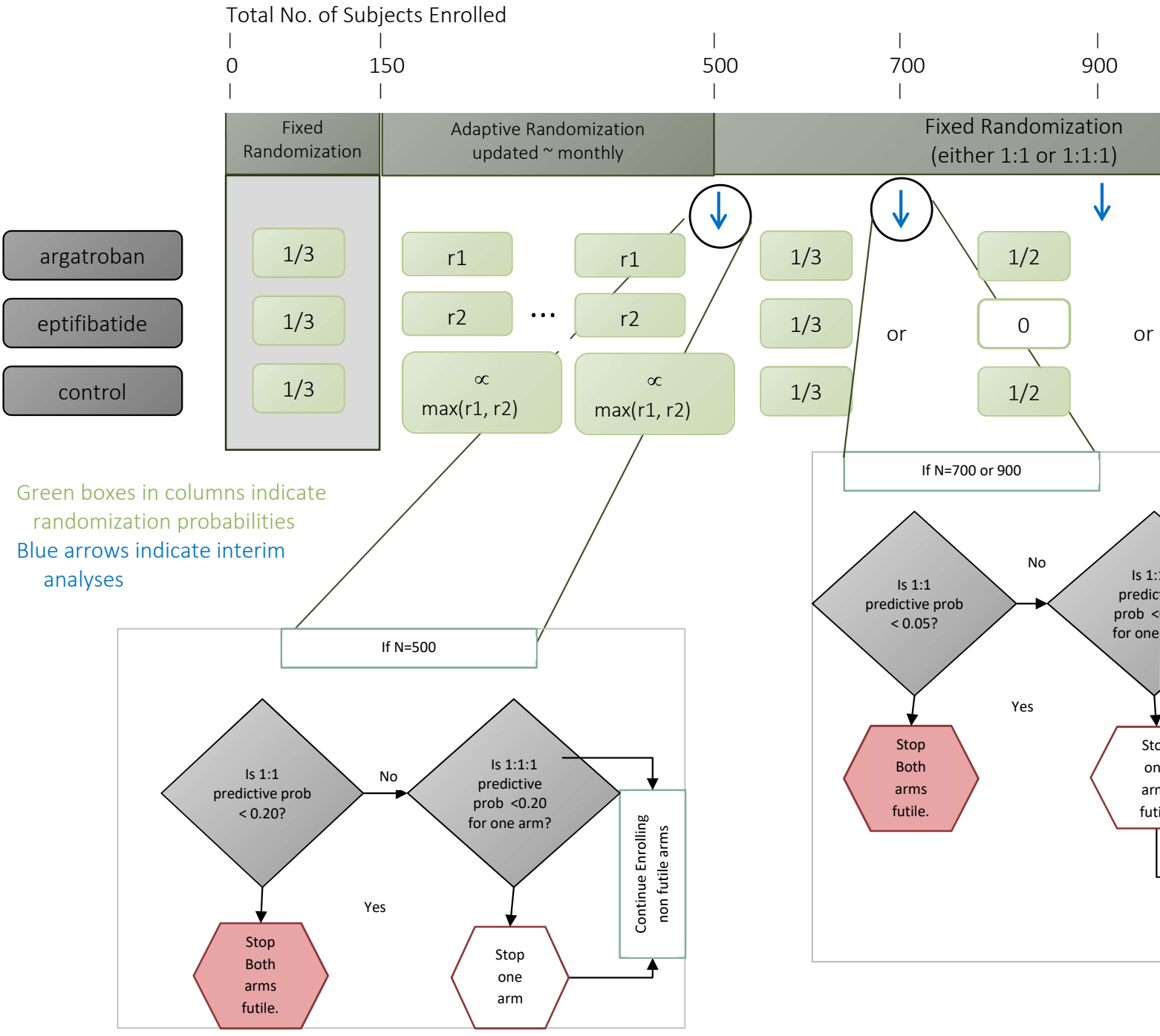
Safety analyses will be tested at a significance level of two-sided α=0.05.

6 Primary Efficacy Analysis

6.1 Overview

The MOST trial will be a single blinded, multi-arm, adaptive, randomized controlled Phase 3 clinical trial at 110 sites to establish whether argatroban and/or eptifibatide is superior to placebo in improving 90-day functional outcomes in AIS patients who are treated with IV thrombolysis within three hours of symptom onset. We will enroll a maximum of 1200 patients. The first 150 subjects will be randomized 1:1:1 to argatroban:eptifibatide:placebo. From 150-500 patients, response adaptive randomization (RAR) will favor the treatment arm(s) showing the greatest benefit based on accrued data. After 500 patients, one or both intervention arms may be carried forward for fixed randomization versus IV thrombolysis. When N=500, the trial may be stopped for futility if there is <20% chance of demonstrating benefit of either intervention (argatroban or eptifibatide) if the trial were to continue. Next, pre-specified thresholds will be used to stop the trial for futility when N=700 or N=900, if there is <5% chance of demonstrating benefit of either intervention if the trial were to continue. After 700 and 900 subjects are enrolled, an arm (or both) may be stopped for expected success. If an arm has an expected success predictive probability ≥99%, the arm will stop for expected success. The final analysis for that arm will occur when all enrolled subjects have been followed up at 90 days. If it is the last remaining active arm, the trial will stop altogether, and all subjects will be followed up to determine the success of the trial. If another arm remains after one arm stops for expected success, the trial will continue with 1:1 randomization between control and the remaining arm. These additional data may demonstrate both regimens have efficacy. Of note, the predictive probability takes into account the patients currently enrolled, but without complete follow-up, and estimates the potential attenuation in any observed treatment effect depending on whether there are many more patients to be enrolled (potential high attenuation) or whether only few patients remain (low likelihood of attenuation). This is different from a *posterior probability* which does not depend on the remaining number of patients to be enrolled.

6.2 Schematic of Trial Design


Total No. of Subjects Enrolled
0
150
500
700
900
Fixed Randomization
Adaptive Randomization updated ~ monthly
Fixed Randomization (either 1:1 or 1:1:1)
argatroban
eptifibatide
control
1/3
1/3
1/3
r1
r2
max(r1, r2)
r1
r2
max(r1, r2)
1/3
1/3
1/3
or
1/2
0
1/2
or
Green boxes in columns indicate randomization probabilities
Blue arrows indicate interim analyses
If N=500
Is 1:1 predictive prob < 0.20?
No
Is 1:1:1 predictive prob <0.20 for one arm?
Yes
Stop Both arms futile.
Stop one arm
Continue Enrolling non futile arms
If N=700 or 900
Is 1:1 predictive prob < 0.05?
No
Yes
Stop Both arms futile.

6.2.1 Primary Efficacy Outcome

Centrally determined 90-day mRS scores re-scaled using the patient-centered utility weighted mRS (UW-mRS) listed in the table below.

Utility Scores for Each mRS Score							
mRS Score	0	1	2	3	4	5	6
MOST Trial utility weighted mRS (UW-mRS)	10	9.1	7.6	6.5	3.3	0	0

6.3 Final Analysis of the Primary Efficacy Outcome

The primary analysis will be conducted on the intent-to-treat (ITT) analysis population. The final analysis is performed with either one or two active arms. If both active arms reach the end of the trial, they will each be compared to control independently.

The parameter of interest is for a given treatment d, θ_d, the difference in the expected utility for the active treatment minus control. For each active arm, the hypothesis test is

$$H_0: \theta_d \leq 0$$
$$H_A: \theta_d > 0$$

Define the primary efficacy outcome Y_j as the 90-day UW-mRS.

Also assume that $Y_j \sim N(\phi_i + \theta_d, \sigma_d^2)$ where θ_d is the difference in the expected utility for the active treatment minus control (with $\theta_0 = 0$ by convention and $d \in \{0,1,2\}$) and ϕ_i for $6 \leq i \leq 42$ is function of the initial NIHSS (i). We model the ϕ_i flexibly, and assume that they come from a second order normal dynamic linear model (NDLM).

6.3.1 Threshold for a Successful Trial

The primary output of the final analysis is the posterior probability that $\theta_d > 0$, for any d's that remain in the trial. The trial is successful if in the final analysis, the posterior probability of a positive benefit is at least 0.985. That is, success is claimed if

$$\Pr(\theta_d > 0) \geq 0.985$$

6.4 Predictive Probabilities

The predictive probability of a successful final analysis will be calculated based on different assumptions about the remaining subjects to be enrolled.

1. First, for each active arm, we assume that the remainder of the trial is 1:1 randomization between that arm and the control arm. We then calculate the predictive probability that the remaining patients generate data that lead to a significant result (high posterior probability of a positive treatment effect). These predictive probabilities will be used in futility decisions and RAR probabilities. For convenience we will refer to these predictive probabilities as 1:1 predictive probabilities.

2. Second, we will assume that the remaining subjects are allocated 1:1:1 to the two active arms and control. We will calculate the predictive probability that those subjects generate data resulting in a significant final analysis for each arm. We refer to these predictive probabilities as 1:1:1 predictive probabilities.

3. Third, we calculate the predictive probability that an arm would have a successful final analysis if enrollment were stopped immediately, based on predictions of results for enrolled subjects without final data. These predictive probabilities are referred to as expected success predictive probabilities.

   Details are in Appendix A.

6.5 Interim Analyses

Interim analyses for safety and futility will take place after 500 subjects are enrolled. Subsequent interim analyses (if the trial is not stopped at 500 patients) take place after 700 and 900 subjects. All interim analysis results will be confidential and only presented to the DSMB in a formal report prepared by the unblinded statistical team at the NDMC. Following review of the interim analysis results and safety data, the DSMB will make recommendations regarding stopping for safety, futility or expected success. These formal interim analyses are distinct from the monthly reports generated for ongoing safety monitoring throughout the trial.

6.6 Stopping for Futility

There are multiple ways in which the trial may be stopped early for futility. When N=500, the trial may be stopped for futility if there is <20% chance of demonstrating benefit of either intervention if the trial were to continue. Next, if the trial proceeds beyond 500 subjects, the trial can be stopped altogether when N=700 or N=900, if there is <5% chance of demonstrating benefit of either intervention if the trial were to continue. Thus, the no-go criterion is predictive probabilities <20% at 500 subjects, or a predictive probability <5% at 700 or 900 patients.

6.7 Stopping for Expected Success

After 700 and 900 subjects are enrolled, an arm (or both) may be stopped for expected success. If an arm has an expected success predictive probability ≥99%, the arm will stop for expected success. The final analysis for that arm will occur when all enrolled subjects have been followed up at 90 days. If it is the last remaining active arm, the trial will stop altogether, and all subjects will be followed up to determine the success of the trial. If another arm remains after one arm stops for expected success, the trial will continue with 1:1 randomization between control and the remaining arm. These additional data may demonstrate both regimens have efficacy, which may yield another treatment option in cases of allergies or contraindications to one of the interventions. Of note, the predictive probability takes into account the patients currently enrolled, but without complete follow-up, and estimates the potential attenuation in any observed treatment effect depending on whether there are many more patients to be enrolled (potential high attenuation) or whether only few patients remain (low likelihood of attenuation). This is different from a posterior probability which does not depend on the remaining number of patients to be enrolled. Indeed, the posterior probability that an intervention is better than control would have to substantially exceed 99% (e.g., a boundary of <0.0001 as used in group sequential designs) for the predictive probability to be 99%.

6.8 If both treatments are better than control

If both arms are effective when compared with control, we will compare both active arms with each other as a secondary analysis using the approach described above and the following criteria for success $\Pr(|\theta_1 - \theta_2| > 0) \geq 0.95$.

7 Secondary and Tertiary Analyses

7.1 Secondary Analysis of the Primary Outcome adjusting for baseline covariates

The primary outcome of UW-mRS values will be reanalyzed in a linear regression adjusting for baseline covariates: enrolling site, NIHSS, age, time to IV thrombolysis, thrombolysis type (TNK/rtPA) location of LVO at the time of start of ET (defined as intracranial ICA, MCA-M1, M2, or basilar occlusion, in which small categories may be combined into a single group) and time from onset to groin puncture in subjects undergoing ET.

Each covariate is evaluated individually first with a model that includes a main effect of the covariates, dummy variables for the treatment groups, and interaction effects between the covariates and the treatment groups. Then a multivariable model that includes all covariates and

interaction terms that contributed significantly ($p<0.05$) individually will be constructed. We will check for collinearity, and re-fit a final model, dropping collinear terms if any are identified.

7.2 Tertiary Analysis of the Primary Outcome: Adjusting for Subgroups

Each subgroup is evaluated individually in a linear model of UW-mRS values that includes a main effect of the subgroup, dummy variables for the treatment groups, and interaction effects between the subgroup levels and treatment group. The p-values for the interaction terms will be reported. The adjusted mean difference in UW-mRS (95% CI) for the treatment effect versus control will be reported for each subgroup (for both active treatment groups).

- Age (< 80 years, ≥80 years)
- Sex (Male, Female)
- Race (Asian, Black, White, Other, etc. Small race categories, such as Asian, Native American, Pacific Islander, may be combined into a single group)
- Ethnicity (Hispanic, Not Hispanic)
- Type of IV thrombolysis (rtPA/TNK)
- Time to IV thrombolysis (<2 hr, >2hr)
- Time to study drug initiation (<2 hr, >2hr)
- ASPECT score (< 6, 6-10)
- Decision to go to ET prior to randomization (Yes/No)
- Presence of LVO (Yes/No)
- Location of LVO (intracranial ICA, MCA-M1, M2, or basilar occlusion. Small categories may be combined into a single other group)
- Received ET (Yes/No)
- Time to reperfusion for ET (cutpoint TBD)

A table/forest plot will be created with shows the following by Subgroup level:
Subgroup/ Number of Patients/ Mean (SD) UW-mRS for all 3 treatment group/ Mean difference from control (95% CI) for each active arm/ p-value for the interaction terms (subgroup x indictor for each active treatment arm).

	Argatroban N=400			Eptifibatide N=400	
Subgroup	No of patients	Mean (95% CI) UW-mRS difference from control	p-value for interaction term	Mean (95% CI) UW-mRS difference from control	p-value for interaction term
Age	1200		*		*
<80 yr	800				
≥80 yr	600				

Sex	*		*
Male			
Female			
Race	*		*
White			
Black			
Asian			
Other			
Ethnicity, Hispanic	*		*
Not Hispanic			
Stroke Onset to thrombolysis bolus	*		*
<2hr			
≥2hr			
Type of IV Thrombolysis			
rtPA			
TNK			
Time to study drug initiation	*		*
<2hr			
≥2hr			
ASPECT score on Baseline CT	*		*
< 6			
6-10			
Decision to go to ET at the time of randomization	*		*
No			
Yes			
Received ET	*		*
No			
Yes			
Presence of LVO (for ET subgroup)	*		*
No			
Yes			
LVO location (for ET subgroup)	*		*
intracranial ICA			
MCA-M1			
M2			
basilar occlusion			
Time to reperfusion (for ET subgroup)**	*		*

< Cutpoint			
≥ Cutpoint			

**A small number of patients may not be able to be reperfused at the end of ET, and so these patients will be excluded. Additionally, some patients will have no occlusion at the start of ET, and so they will be excluded.

7.3 Secondary Clinical Efficacy Outcomes

1. proportion with 90-day mRS 0-1 (or return to their historical mRS)
2. proportion with 90-day mRS 0-2 (or return to their historical mRS)
3. 90-day ordinal analysis of the mRS

7.4 Secondary Radiological Reperfusion Outcomes

1. proportion of participants who are 2b or 3 on the (Baseline) Modified TICI score of 2b or 3 prior to thrombectomy amongst those for whom there was a decision to go to ET
2. proportion of participants who have thrombectomy (Final TICI score of 2b or 3) amongst those for whom there was a decision to go to ET

	Control (N=)	Argatroban (N=)			Eptifibatide (N=)		
Secondary Clinical Efficacy Outcomes—no. (%)	No (%)	No (%)	Odds Ratio (95% CI)	P value	No (%)	Odds Ratio (95% CI)	P value
90-day mRS 0-1‡	Xx (%)	Xx (%)		[a]	Xx (%)		[a]
90-day mRS 0-2‡	Xx (%)	Xx (%)		[a]	Xx (%)		[a]
90-day ordinal analysis of the mRS	median (IQR)	media n (IQR)	----	[b]	median (IQR)	----	[b]
Secondary Radiological Reperfusion Outcomes (for whom there was a decision to go to ET) —no. (%)	N=nn	N=nn	Odds Ratio (95% CI)	P value	N=nn	Odds Ratio (95% CI)	P value
Modified TICI score of 2b or 3 prior to thrombectomy	Xx (%)	Xx (%)		[a]	Xx (%)		[a]
Thrombectomy: Final Modified TICI score of 2b or 3	Xx (%)	Xx (%)		[a]	Xx (%)		[a]

‡(or return to their historical mRS)

[a] Odds ratio will be expressed as an unadjusted risk ratio with its 95% confidence interval and p-value will compare treatment to control.

[b] P value will compare treatment to control via Wilcoxon Rank Sum test (WMW).

7.5 Tertiary Efficacy Outcomes

1. proportion of participants with NIHSS ≤2 at 24 hours
2. change from baseline to 24-hour NIHSS
3. 90-day EQ-5D

7.6 Analysis of Secondary/Tertiary Outcomes

The difference between each active treatment group and control will be reported as absolute unadjusted mean difference (95% CI) and adjusted difference (95% CI) for continuous outcomes, and absolute difference in proportions (95% CI) and Risk/Odds Ratio (95% CI) for binary outcomes.

For Binary outcomes each active treatment arm will be compared to control in a two-sample test of proportions (or Fisher's exact test if events are rare).

The ordinal analysis of the 90-day mRS will use a Wilcoxon rank sum test to compare each treatment groups to control. For continuous outcomes, each active treatment arm will be compared to control in a two-sample t-test if normality assumption holds; otherwise, the rank sum test.

Missing secondary/tertiary outcomes will be imputed using the hot deck method matched by age and baseline NIHSS, time to symptom onset, treatment group, ET, and earlier collected time points of the missing outcome measure. Multiple imputation methods may be conducted if the total amount of missing data exceeds 5%.

7.7 Imaging Data

The baseline ASPECT score, presence and location of LVO at baseline for ET patients will be determined by the central imaging readers for all enrolled patients and reported in the Table of baseline characteristics (Table 1). The presence of ICH, PH-1, and PH-2 on the 24 hour CT will be determined by the central imaging reader for all enrolled patients and reported as Safety Outcomes (see below).

Other imaging data (e.g., M2 occlusions or collateral scores will be noted and included in exploratory analyses to be determined).

7.8 Subgroup Analysis: Gender, Race, Ethnicity

Recruitment and retention of women and minorities will be monitored by the DSMB and will be provided to the NINDS at fixed intervals and in the Final Progress Report.

Although we do not anticipate differential treatment effects based on sex, race, or ethnicity, our analyses will explore clinically important differences due to sex/race/ethnicity (See Subgroup analysis section 7.2). A clinically important interaction of the treatment effect by sex, race, or ethnicity, regardless of the statistical significance, will be reported to the NINDS in the Final Progress Report and to the scientific community in the primary paper.

7.9 Endovascular Therapy (Concomitant Therapy)

We expect ~30% of subjects will undergo Endovascular Therapy (ET). ET is a potentially confounding therapy (i.e., intervention arm patients may show improvement and be less likely to need ET).

We will closely monitor for confounding due to treatment effect as a possible reason for any differences in rates of ET between groups. A comparison of the proportion of participants who have groin puncture may not detect differences in treatment groups since the thrombectomy rates may still differ (i.e., a more effective intervention may obviate need for thrombectomy). As such, we will monitor the proportion of participants who have thrombectomy as a secondary outcome. If there is a statistically significant difference in the proportion of patients who receive thrombectomy by treatment group (in particular, a higher proportion in the control arm receive ET), then the intervention(s) may be reducing the need for thrombectomy. In this scenario, inference about the causal relationship between intervention arms and outcome will be explored using path diagrams and structural equation modeling which takes into account confounding of treatment with ET.

In the event that there is a difference in the thrombectomy rate, or difference in Modified TICI score, exploratory analyses in ET subjects will use either general linear regression with a logit link for binary outcomes or multiple linear regression for continuous outcomes to compare treatment groups after adjusting for age, baseline NIHSS, type of IV thrombolysis (rtPA/TNK), time to IV thrombolysis, presence of LVO (Yes/No), location of LVO (defined above), and time to groin puncture.

A subgroup analysis of the primary outcome will be conducted by ET versus non-ET groups as described above.

8 Safety Analyses

8.1 Primary Safety Outcome

- sICH within 36 hours from randomization (3% expected rate)

8.2 Secondary Safety Outcomes

- parenchymal hemorrhage types 1 (PH-1) and 2 (PH-2) (determined by the central imaging readers) within 36 hours from randomization
- any ICH on brain imaging within 36 hours from randomization (determined by the central imaging readers)
- Major hemorrhage (requiring >2 units of packed red cells) other than intracranial hemorrhage within seven days
- All-cause mortality within 90 days of randomization (20% expected rate)

8.3 Analysis of Safety Outcomes

Please refer to the MOST Safety Monitoring Plan for details of DSMB reporting, medical safety monitor, and endovascular safety monitor.

For each safety outcome, the proportion of patients will be reported by treatment group. At the end of the trial, each active treatment arm will be compared to control using Fisher's exact test. Kaplan Meier curves and log rank tests will be used to compare time to 90-day mortality. At the end of the trial, we will compare each safety outcome by treatment groups by ET subgroup (Yes/No) in order to assess the safety of the combination of Eptifibatide/Argatroban/ET using Fisher's exact test for the 3 x 2 table of treatment group and ET status (and with the ET subgroup comparing active to control if significant). Similarly, a subgroup analysis of sICH will be conducted by baseline ASPECTS score (< 6 or 6-10).

At the end of the trial, SAEs will be reported by MedDRA body system. The proportion of patients in each body system will be reported by treatment group. Fisher's exact test will be used to test for differences versus control for each body system.

9 Sample Size Determination

The minimum clinically significant difference was defined as 0.4 points on the utility scale primary outcome, which corresponds to ~45% of the effect for IV rt-PA over placebo in the NINDS study.[43] With a maximum N=1200, this design provides at least 80% power to detect a treatment effect when the overall true utility benefit is 0.4 for one active arm in non-ET subjects (Table 8). If both treatment arms are equally effective at 0.4 utility above the control arm, power is 89%. The properties of the study design have been investigated through clinical trial simulations that assumed 30% of patients receive ET, and the treatment effect for ET patients is 50% of the effect in non-ET patients, corresponding to 25% of the treatment effect of IV rt-PA

over placebo. The impact of ET on treatment effect is unknown but we assumed the worst scenario, i.e., marked attenuation of effect. Also, only 10-30% of IV rt-PA eligible patients are eligible for ET,[10,12] and clinical practice data from Get With the Guideline hospitals and StrokeNet suggest ~11-24% of IV rt-PA eligible patients undergo ET. We have assumed the upper end of possible ET subjects. Thus, our estimated power represents a conservative scenario for ET proportions and impact of ET on treatment effect. Table 8 shows power when only one treatment arm is effective. Power increases if <30% of patients receive ET. Although the maximum sample size is N=1200, the simulations conducted indicate the average sample size under the complete null scenario is 756 (SD=322) and under the scenario with all active arms equal to 0.4 is 1048 (SD=188). The type I error probability (incorrectly identifying treatment(s) to "win" that are truly no better than control) for the complete null scenario is <0.025.

Power for Varying Treatment Effects if only One Arm is Truly Effective			
	Utility Values Above IV rt-PA Alone		
	Scenario 1	Scenario 2	Scenario 3
Assumed True Effect Overall (70% non-ET and 30% ET subjects)	0.26	0.34	0.425
Assumed True Effect for non-ET subjects	0.30	0.40	0.50
Assumed True Effect for ET subjects (50% of non-ET effect)	0.15	0.20	0.25
Pr{Effective arm wins} (Power)	0.52	0.80	0.95
Pr{Other arm wins} (Power)	0.011	0.008	0.014
Numbers in gray are the utilities (difference in the expected treatment minus control). Numbers in white are the power under 3 different scenarios (determined by simulation).			

10 Appendix A: MOST Design and Analysis Plan

Berry Consultants
Statistical Innovation

Design and Analysis Plan for the MOST Trial
October 20, 2020

1.0 Background

This is a plan for a Phase III trial to explore the efficacy of Argatroban and Eptifibatide in combination with thrombolysis in treating stroke patients. One dose of each study drug will be compared to a control arm with respect to their ability to improve subjects' 90-day scores on the modified Rankin scale (mRS). Improvement will be quantified using patient-centered utility scores for mRS values.

The trial is a Bayesian adaptive design that includes multiple key features:

1. adaptive sample size ranging from 500 to 1200 patients;
2. initial consideration of two active arms, with the ability to stop an arm for futility or safety and continue as a 1:1 comparison between control and one of the active arms;
3. response-adaptive randomization is used to favor promising active arms;
4. frequent interim analyses can result in the trial stopping early for futility or for expected success;
5. a utility function on 90-day mRS scores to reflect patient and society valuation of outcome health states;
6. mRS scores are analyzed adjusting for baseline stroke severity as measured by the NIH Stroke Scale. The relationship between stroke severity and outcomes is modeled flexibly;
7. a longitudinal model relating 30-day mRS scores to 90-day mRS scores is utilized to bring more information from patients without 90-day data to stopping and adaptive allocation decisions.

The structure of this document is as follows: the next section describes the design in general terms, Section 3 elaborates on the subject population and the final analysis, and Section 4 fills in further details about decisions made during the trial. Section 5 presents operating characteristics for the design obtained using simulation and based on a variety of assumed truths about the effectiveness of the drugs. Appendices elaborate on the statistical models used in the final analysis and in the interim analyses and on default assumptions about the subject population.

2.0 Design Overview

For the first 150 subjects, the randomization probabilities for the three arms remain fixed at 1:1:1 for the two active arms and the control/placebo arm. An interim analysis occurs at the 150th enrolled subject, and randomization probabilities for the active arms are adjusted, with the arm having the higher predictive probability of a successful final analysis assigned higher allocation probabilities. Interim analyses continue every 4 weeks.

Beginning with the first interim analysis after the 150th subject is enrolled, the design implements response adaptive randomization between two active arms (based on predictive probability) where control gets the same randomization probability as the maximum active arm.

After 500 subjects have been enrolled, the stopping rule for futility is based on 20% predictive probability and there is a shift to equal randomization between control and one or two active arms. From this point until the end of the trial, all arms remaining in the trial have equal allocation probabilities.

Interim analyses occur after 700 and 900 patients have been enrolled to allow for

- stopping one active arm for futility (based on 5% predictive probability based on three arms continuing)
- stopping the trial altogether for futility (based on 5% predictive probability based on two arms continuing)
- stopping the trial for expected success of one or both active arms.

The trial enrolls at most 1200 subjects. When final data for all enrolled subjects are available, the final analysis is conducted, and it may result in showing a significant benefit for one or two active arms.

Stopping decisions are based on Bayesian predictive probabilities, and in particular the predictive probability of a successful final analysis. Details about these predictive probabilities will be given in Section 4.

3.0 Study Population, Primary Endpoint, and Statistical Test

3.1 Entry criteria

The trial enrolls subjects with initial NIHSS scores of 6 or larger.

3.2 Treatment arms

Three treatment arms are under consideration:

1. Control arm with iv-thrombolysis only;
2. Argatroban in addition to iv-thrombolysis;
3. Eptifibatide in addition to iv-thrombolysis;

Of the first 150 enrolled subjects, one-third will be assigned to the control arm and to each of the active arms.

3.3 Primary Endpoint

The primary endpoint for this trial is the 90-day mRS score. We choose to analyze this standard endpoint by converting the mRS scores into weights that directly reflect patient and society valuation of outcome health states. We then model a subject's weighted mRS score as normally distributed with expected value depending on initial NIHSS and treatment assigned.

The weights assigned to the possible mRS scores are shown in Table 1 below. These weights were obtained through a synthesis of studies (O. Rivero-Arias, et al, "Mapping the Modified Rankin Scale (mRS) Measurement into the Generic EuroQol (EQ-5D) Health Outcome," Medical Decision Making 2010 30:341, and K.-S. Hong and J.L. Saver, "Quantifying the Value of Stroke Disability Outcomes: WHO Global Burden of Disease Project Disability Weights for Each Level of the Modified Rankin Scale: Supplemental Mathematical Appendix," Stroke 2009 40:3828-3833). Both these studies assigned utility values and confidence intervals to mRS scores; these are also shown in Table 1. We renormalized these utilities to a scale where an mRS of 6 implies a utility of 0 and an mRS of 0 implies a utility of 10. The two scales are quite similar, and we take the mean of the renormalized utilities to obtain our own weights. The second study reported more precise estimates, so in some cases the consensus value is closer to its value.

mRS	0	1	2	3	4	5	6
Rivero-Arias et al	10	8.7	7.3	6.0	2.8	-0.1	0
Hong & Saver	10	9.5	7.9	6.7	3.5	0.1	0
MOST Trial	10	9.1	7.6	6.5	3.3	0	0

Table 1: weights used for 90-day mRS score

Relative to an approach that dichotomizes the 7 possible mRS scores into two possibilities, weighting the 7 Rankin levels by utilities improves the precision of the scale as a measure of disability. The weighted approach should also not be confused with an approach based on the raw mRS scores, which would erroneously treat each single-point increase in mRS as equally valuable to the subject.

Figure 1 below gives an example of what treatment effects look like for this endpoint. Each vertical bar depicts a probability distribution: the height of the darkest blue represents the probability of an mRS of 0, the darkest red shows the probability of an mRS of 6, etc. The leftmost bar shows the results of the control arm for the NINDS tPA study (Tissue plasminogen activator for acute ischemic stroke. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. The New England journal of medicine 1995;333:1581-7), which represents an expected utility of 5.01, and the second bar shows the tPA arm in the NINDS study, which represents an expected utility of 5.91. The remaining bars show distributions that represent improvements of 0.1 to 0.9 as compared to the tPA arm.

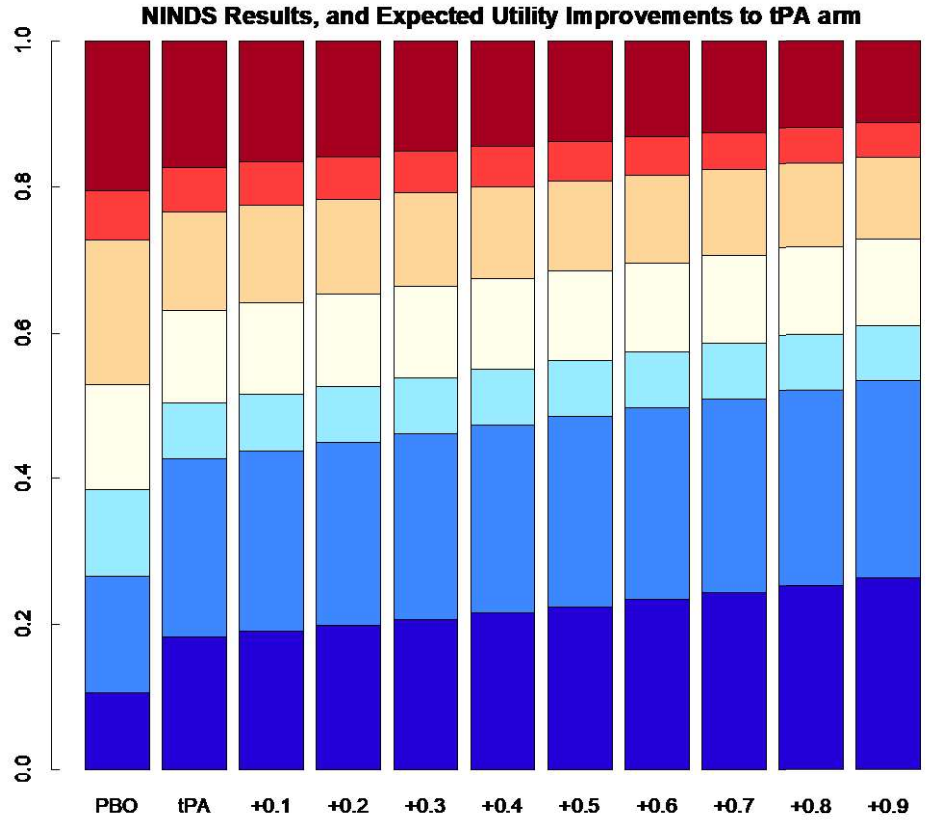


Figure 1: graphical depiction of mRS distributions. The first two bars depict the two arms in the NINDS study, and the remaining nine bars depict further improvements to expected utility. Dark blue represents the probability of an mRS of zero, dark red represents the probability of an mRS of 6, and so on.

	Expected Utility	mRS=0	mRS=1	mRS=2	mRS=3	mRS=4	mRS=5	mRS=6	0-1
NINDS Control	5.01	0.11	0.16	0.12	0.14	0.20	0.07	0.21	0.27
NINDS tPA arm	5.91	0.18	0.24	0.08	0.13	0.13	0.06	0.17	0.42
tPA + 0.1	6.01	0.19	0.25	0.08	0.13	0.13	0.06	0.17	0.44
tPA + 0.2	6.11	0.20	0.25	0.08	0.13	0.13	0.06	0.16	0.45
tPA + 0.3	6.21	0.21	0.25	0.08	0.13	0.13	0.06	0.15	0.46
tPA + 0.4	6.31	0.22	0.25	0.08	0.12	0.13	0.05	0.14	0.47
tPA + 0.5	6.41	0.22	0.26	0.08	0.12	0.12	0.05	0.14	0.48
tPA + 0.6	6.51	0.23	0.26	0.08	0.12	0.12	0.05	0.13	0.49
tPA + 0.7	6.61	0.24	0.27	0.08	0.12	0.12	0.05	0.13	0.51
tPA + 0.8	6.71	0.25	0.27	0.08	0.12	0.12	0.05	0.12	0.52
tPA + 0.9	6.81	0.26	0.27	0.08	0.12	0.11	0.05	0.11	0.53

Table 2: translating treatment effects on the utility scale into changes in probabilities of the mRS outcomes. The point estimate of the effect of tPA in the NINDS study is 0.9 units of utility. The table shows what further improvements to expected utility above tPA might do.

3.4 Primary Analysis

The final analysis is Bayesian and includes a flexible normal dynamic linear model (NDLM) to account for different expected outcomes as a function of initial NIHSS. This is a flexible spline-like model that will capture that the average weighted mRS score in the control group is a (possibly non-linear) function of initial NIHSS. Meanwhile the average effect of a given treatment d, θ_d, is the difference in the expected utility for the active treatment minus control. This average treatment effect, θ_d, is assumed to be equal over all values of initial NIHSS. Details of the statistical model are given in Appendix C1. For each active arm, the hypothesis test is

$$H_0: \theta_d \leq 0$$
$$H_A: \theta_d > 0$$

The treatment effect θ_d is given a vague prior, θ ~ N(0,2.5^2): in particular, the prior probability that a drug is beneficial is the same as the prior probability that it is harmful. Note: The largest treatment effect we considered in the simulations (an average of 0.50 unit benefit) is relatively small in comparison to the prior standard deviation. If there is a high posterior probability that the treatment effect θ_d is positive, the treatment is declared to be efficacious. The posterior probability is conditional on the final results for all subjects enrolled in the trial. If two active arms remained in the trial all the way to the end, this posterior probability is computed for each, and potentially both arms can be declared to be successful.

3.5 Thresholds for a Successful Trial

The trial is successful if in the final analysis, the posterior probability of a positive benefit is at least 0.985. That is, success is claimed if

$$\Pr(\theta_d > 0) \geq 0.985$$

This threshold is chosen to control the experiment-wise type I error probability at 0.025.

3.6 Longitudinal Modeling of 30-day mRS Scores

Subjects with 30-day data but no 90-day data are also included in the interim analyses, and their data contribute to the adaptive allocation probabilities and the stopping and arm selection decisions. A subject with only 30-day data is less influential than a subject with complete data, because the statistical model in effect takes into account the fact that different 90-day outcomes are still possible. The relationship between 30-day and 90-day data is initially assumed to be unknown, and is updated as more data from the trial come in. Consequently, the trial will not make irreversible decisions as a result of incorrect beliefs about the relationship.

Berry Consultants
Statistical Innovation

4.0 Prospectively Planned Interim Analyses

The first interim analysis takes place after 150 subjects have been enrolled. Subsequent interim analyses take place every four weeks (28 days) until 500 patients are enrolled, and less frequently thereafter.

4.1 Predictive Probabilities

Decisions made as a result of interim analyses are based on Bayesian predictive probabilities using the statistical model defined in Appendix C2. The predictive probability of a successful final analysis is calculated based on different assumptions about the remaining subjects to be enrolled.

First, for each active arm, we assume that the remainder of the trial consists of 1:1 randomization between that arm and the control arm, and calculate the predictive probability that the remaining patients generate data that lead to a significant result (high posterior probability of a positive treatment effect). These predictive probabilities are used in futility decisions and response adaptive randomization probabilities. For convenience we will refer to these predictive probabilities as *1:1 predictive probabilities.*

Second, we assume that the remaining subjects to be accrued are allocated 1:1:1 to the two active arms and to the control arm, and calculate the predictive probability that those subjects generate data resulting in a significant final analysis for the Argatroban arm, and respectively for the Eptifibatide arm. We will refer to these predictive probabilities as *1:1:1 predictive probabilities*.

Third, we also calculate the predictive probability that an arm would have a successful final analysis if enrollment were stopped immediately, based on predictions of results for subjects enrolled in the trial but without final data. This predictive probability is used to determine whether enrollment should be stopped early for expected success. We will refer to these predictive probabilities as *expected success predictive probabilities.*

4.2 Interim Monitoring for Early Futility and Expected Success.

Although the design assumes the trial cannot stop for futility or drop an arm before 500 subjects, recommendations by the DSMB (such as for safety concerns) supersede this assumption. Interim analyses before 500 subjects update response adaptive randomization probabilities.

After 500 subjects have been enrolled, one of the arms can be dropped if its 1:1:1 predictive probability is less than 20% or the trial can be stopped altogether if both active arms have 1:1 predictive probability is less than 20%.

After 700 and 900 subjects have been enrolled, one of the arms can be dropped if its 1:1:1 predictive probability is less than 5% or the trial can be stopped altogether if the 1:1 predictive probability is less than 5%.

When 700 patients and 900 patients have been enrolled the trial may stop for expected success. Beginning with the first analysis after this time, the expected success predictive probabilities are calculated. If an active arm has an expected success predictive probability of at least 99%, that arm stops for expected success, and its final analysis is conducted when all subjects enrolled are followed up for 90-day data. If it is the last remaining active arm, the trial stops altogether, and all enrolled subjects are followed up to determine the success of the trial.

4.3 Response-Adaptive Randomization

During the response-adaptive randomization regime, which begins at 150 subjects and ends when the 500th subject has been enrolled, the control arm's randomization probability is set to be equal to the higher of the randomization probabilities for the active arms. The allocation probabilities for the two active arms are set to be proportional to their 1:1 predictive probabilities. For example, suppose that the two active arms have 1:1 predictive probabilities of 0.2 and 0.8 respectively. Setting the control arm randomization probability equal to that of the arm with 0.8 predictive probability, the three arms have randomization probabilities of 0.8 / (0.8 + 0.2 + 0.8) = 0.444 for control and the better of the two active arms, and 0.2 / (0.8 + 0.2 + 0.8) = 0.111 for the other active arm.

5.0 Operating Characteristics

In this section we present operating characteristics. These results are obtained through simulation as illustrated in Section 5, and are based on 1000 simulated trials per scenario, except for the null scenario whose results are based on 10000 simulated trials.

In all simulations, we assumed that 30% of patients receive endovascular therapy (ET) in addition to the study drug or standard thrombolysis. The treatment effect for ET patients is 50% of the effect in non-ET patients. The impact of ET on treatment effect is unknown but we assumed the worst scenario, i.e., attenuation of the treatment effect. Patients with ET are assumed to have the same initial NIHSS distribution as those without.

5.1 Operating characteristics when both active arms are equivalent

First we present estimates of power and sample size characteristics in scenarios in which both active arms have the same effect: specifically, an increase of zero (a null scenario), 0.3, 0.4, and 0.5 units for the non-ET subjects and 50% of these same effects for the ET subjects. (The assumed overall treatment effects were 0.255, 0.34, and 0.425 for non-ET and ET subjects combined.) These effects benefit all injury severities equally with respect to expected utility. Examples of the scenarios are pictured in Figure 2 below, which shows the assumed distribution of mRS as a function of initial NIHSS (shown on the x-axis), for the control arm (left plot), for an arm with a 0.5 unit effect (center plot), and for an arm with a 1.0 unit (right plot; this effect is unrealistically large). For example, examining the heights of the dark blue bars on the far left of the plots show that the probability of an mRS of zero for subjects with an initial NIHSS of 6 treated with

the control arm is assumed to be almost 60%. An expected utility effect of 0.5 points raises this to about 70%, while an effect of 1.0 points raises it to about 85%.

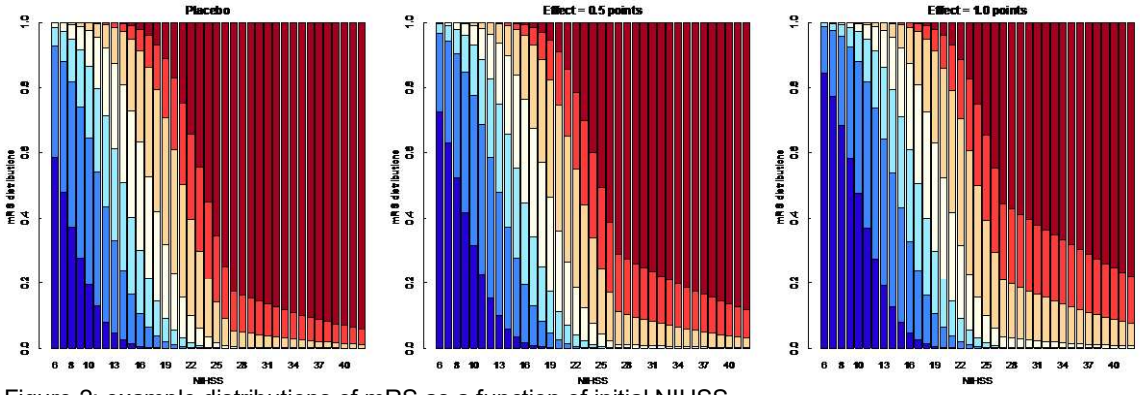


Figure 2: example distributions of mRS as a function of initial NIHSS.

Since the trial involves two different active drugs and potentially both of them can be successful, power has multiple aspects:

- the probability that at least one active arm is successful;
- the probability that the Argatroban arm is successful;
- the probability that the Eptifibatide arm is successful (this should be the same as the probability that an Argatroban arm is successful for the scenarios examined in this section);
- the probability that both active arms win.

Results based on 1000 simulated trials for each scenario, except with 10000 simulated trials for the null scenario, are shown in Table 3. Simulations of null scenarios are presented in Section 5.3. Here we note that the Type I error in this single null scenario is less than 0.025. If the overall true effect is 0.425 units for both active arms, the design has high power of 97% to be successful for some arm, and each drug has a 17% chance of being unsuccessful.

Assumed True Effect Overall (30% ET and 70% non-ET patients)	Null	0.255	0.34	0.425
Assumed True Effect For non-ET patients	0	0.30	0.40	0.50
Assumed True Effect for ET patients (50% of effect w/o ET)	0	0.15	0.20	0.25
Pr{Some Arm Wins}	0.0190	0.695	0.890	0.974
Pr{ A Arm Wins}	0.0089	0.453	0.644	0.832
Pr{E Arm Wins}	0.0107	0.459	0.678	0.833
Pr{A Arm and E Arm both win}	0.0006	0.217	0.432	0.691

Table 3: power-related operating characteristics for scenarios in which both active arms are equally effective. ET=endovascular

Further operating characteristics are shown in Table 4 below. The table displays the probability of a futility stop at exactly 500 subjects, and the probability that the design reaches full enrollment of 1200. The average and standard deviation of total sample sizes are also shown, as well as the averages for trials that were ultimately unsuccessful.

Assumed True Effect Overall (ET and non-ET patients)	Null	0.255	0.34	0.425
Pr{Futility at 500}	0.56	0.086	0.045	0.013
Pr{N = 1200}	0.34	0.725	0.555	0.363
Mean N	756	1080	1048	1004
SD(N)	322	218	188	155
Mean N (failed trials)	748	974	900	831

Table 4: sample size-related operating characteristics.

5.2 Operating characteristics when arms differ

In this section we present operating characteristics for scenarios in which the Argatroban arm has a benefit of up to 0.5 units, while the Eptifibatide arm has no benefit, to explore how successful the design is at identifying the most promising arm. We have simulated scenarios in which Eptifibatide is less effective than Argatroban and not the reverse, but the two drugs are treated exchangeably, so those simulation results also apply to the analogous cases where Eptifibatide is the superior drug.

We see from Table 5 that if one of the active arms has an overall effect of 0.425 units of utility, the design has approximately 95% power. If one drug is ineffective, the probability that the more effective drug will be successful is increased relative to Table 3 because the ineffective drug will likely be dropped. The probability that the ineffective drug is successful is small.

Assumed True Effect Overall (ET and non-ET patients) for Arm A only (Arm E assumed to be 0)	0.255	0.34	0.425
Assumed True Effect for Arm A For non-ET patients	0.30	0.40	0.50
Assumed True Effect for ET patients for Arm A (75% of effect w/o ET)	0.15	0.20	0.25
Pr{Some Arm Wins}	0.523	0.796	0.951
Pr{Arm A wins}	0.521	0.796	0.950
Pr{Arm E wins}	0.011	0.008	0.014
Pr{ A Arm and E arm both win}	0.009	0.008	0.013

Table 5: power characteristics for scenarios where the active arms differ.
ET=endovascular therapy

5.3 Type I Error

In this section we explore several null scenarios to demonstrate evidence of control of Type I error. The design is too complex to allow analytical control of Type I error, and the number of null scenarios is too large for exhaustive simulation study. Instead, we identify four characteristics that define null scenarios and vary those characteristics to create a population of scenarios. The scenarios are described in Table 6. Scenario #1 is the same null scenario as discussed in Section 5.1.

Scenario	NIHSS Distn	*f*(mRs)	Effect Size	Longitudinal Model	Accrual rate
1	Default	Standard	0	Default	Default
2	Default	Standard	0	Default	Fast
3	Default	Standard	0	Default	Slow
4	Default	Standard	0	Uniform	Slow
5	Default	Standard	0	Perfect	Slow
6	Default	Standard+	0	Default	Slow
7	Default	Standard–	0	Default	Slow
8	Default	Flat	0	Default	Slow
9	Default	Jagged	0	Default	Slow
10	Uniform	Standard	0	Default	Slow
11	Lower	Standard	0	Default	Slow

Table 6: Definition of null scenarios in Type I error simulations.

The other null scenarios differ by the following characteristics.

- Accrual rate. In the standard scenarios, the accrual rate is 5 patients per week. In Scenario 2, we assume an accrual rate of 10 patients per week, and in Scenario 3, we assume an accrual rate of 2.5 patients per week. In many designs that feature early stopping for expected success with endpoints only observed after a delay, slow accrual rates tend to inflate Type I error. This is because relatively few patients have incomplete data at the time of the decision to stop for expected success, so that expected success stopping decisions are made with information relatively close to the data in the eventual primary analysis. Since slow accrual rates are the most challenging for designs to control Type I error, we also use the 2.5 patients per week rate in scenarios 4 through 11. Note that 2.5 patients per week is an unrealistically slow accrual rate for the trial, which would under that assumption require more than nine years to enroll a full 1200 patients, so this represents a very conservative set of scenarios to evaluate Type I error.
- Longitudinal model. In most of our scenarios, 30-day mRS is a good but not perfect predictor of 90-day mRS; see Appendix C3.3. We also simulate null scenarios with extreme versions of the longitudinal model assumption. In Scenario 4, 30-day mRS is simulated so that all of its possible values (i.e. 0 through 5 if 90-day mRS is not 6, or 0 through 6 if 90-day mRS is 6) are equally likely; in this scenario 30-day data are not useful for predicting 90-day data. In Scenario 5, 30-day mRS is simulated to always be exactly equal to 90-day mRS; in this scenario the longitudinal model should learn that 30-day data are highly predictive of 90-day data and make predictions accordingly.
- Distribution of mRS given initial NIHSS (the column describing this variable is denoted “f(mRS)” in Table 6). In most scenarios, the distribution of 90-day mRS given initial NIHSS is as shown in Figure C2 in the appendix. We also consider four other sets of distributions. In Scenarios 6 and 7, the relationship between initial NIHSS and 90-day mRS is similar to Figure C2, but in Scenario 6, patients are systematically healthier and in Scenario 7, patients are systematically less healthy; see Figure 3. In Scenario 8, we consider a “flat” scenario, the case where all patients have the same distribution of 90-day mRS regardless of initial NIHSS. In Scenario 9, we consider a more “jagged” scenario, where NIHSS 6 through 11 have one distribution, NIHSS 12 through 26 have another less favorable distribution, and NIHSS 27 through 42 have another unfavorable distribution. The distributional assumptions in Scenarios 8 and 9 are shown in Figure 4.
- Initial NIHSS Distribution. In Scenario 1 and most of the other null scenarios, small initial NIHSS scores are assumed to be most common, see the distribution in Figure C1 in the appendix. In Scenario 10, all initial NIHSS scores are equally likely. In Scenario 11, the initial NIHSS distribution is tilted further toward small values; see Figure 5. Instead of being based on the exponential distribution with mean 10, it is based on the exponential distribution with mean 7.5. For example, under the standard NIHSS distribution, 20% of patients have an initial NIHSS of 21 or greater, while only 13% of patients do under Scenario 11.

  We do not vary the fraction of patients who receive endovascular therapy in the Type I error simulations, because we assume that the only effect of endovascular

therapy is to reduce the treatment effect by a factor, so that all assumptions about endovascular fraction lead to identical null scenarios.

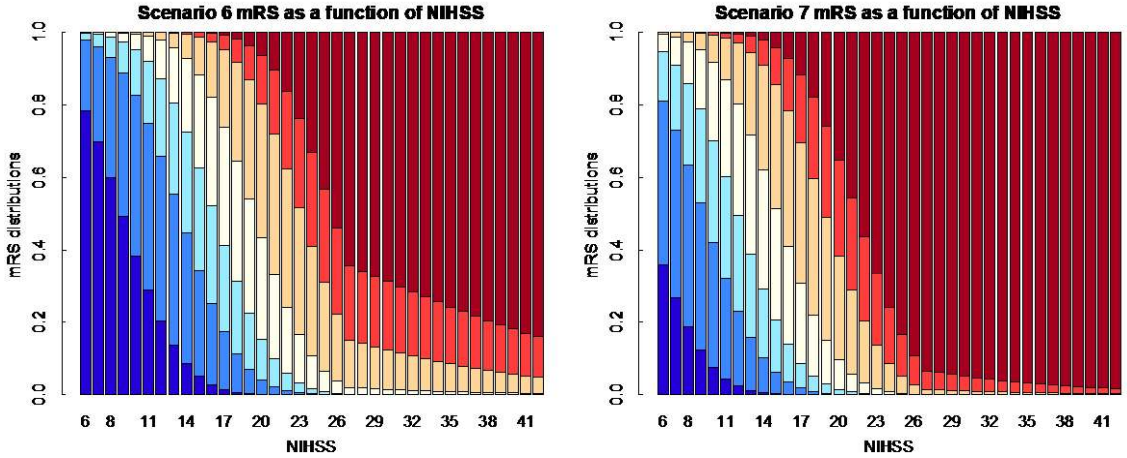


Figure 3: Distributions of mRS as a function of NIHSS in scenarios 6 and 7. In Scenario 6, patients are likelier to have good outcomes compared to scenario 1, and in Scenario 7, patients are likelier to have poor outcomes.

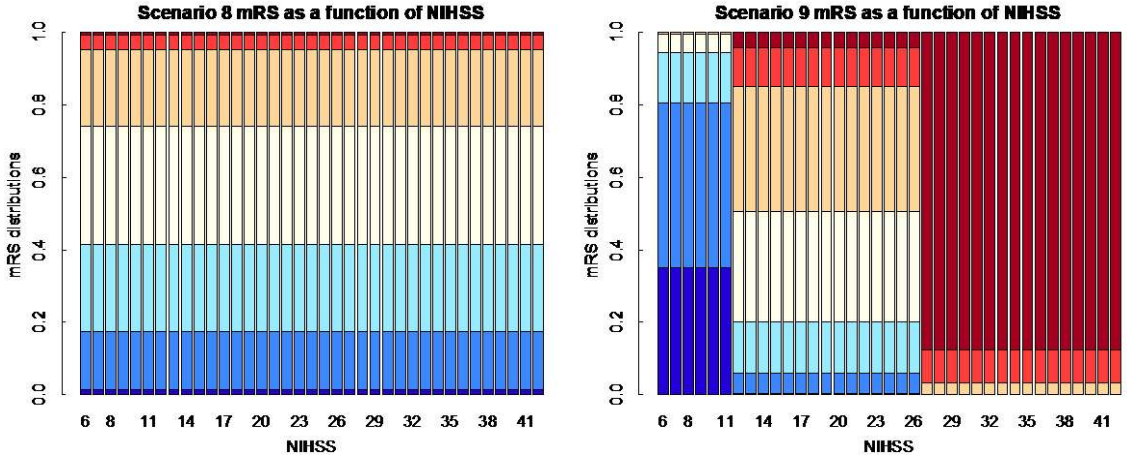


Figure 4: Distributions of mRS as a function of NIHSS in scenarios 8 and 9. In Scenario 8, all patients have the same distribution of mRS regardless of their initial NIHSS score. In Scenario 9, there are three different distributions of mRS depending on initial NIHSS, and there are abrupt changes in the distributions.

The Type I error operating characteristics are given in Table 7. For each scenario, the table shows the probability that at least one arm wins (i.e. reaches the threshold of 98.5% posterior probability of a benefit), the probability that arm A wins, the probability that arm E wins, and the probability that both active arms win. Scenario 1 results are reproduced from Table 3. For all 11 null scenarios, the probability of a Type I error is less than 0.025. The probability that a particular active arm attains statistical significance is less than 0.014 for all these scenarios.

Scenario	Pr{Some Arm Wins}	Pr{A Wins}	Pr{E Wins}	Pr{A and E Both Win}
1	0.0190	0.0089	0.0107	0.0006
2	0.0229	0.0138	0.0102	0.0011
3	0.0205	0.0105	0.0108	0.0008
4	0.0213	0.0106	0.0115	0.0008
5	0.0222	0.0104	0.0126	0.0008
6	0.0211	0.0119	0.0108	0.0016
7	0.0199	0.0116	0.0094	0.0011
8	0.0208	0.0104	0.0114	0.0010
9	0.0211	0.0126	0.0098	0.0013
10	0.0202	0.0091	0.0118	0.0202
11	0.0194	0.0104	0.0096	0.0006

Table 7: Type I error operating characteristics. For all 11 null scenarios considered, the probability of some arm winning is less than 0.025.

Berry Consultants
Statistical Innovation

Appendix C1: Statistical Model for Final Analyses

Denote the j'th subject's 90-day mRS by S_j, and her resulting weight score by Y_j; $Y_j = W_k$ if $S_j = k$. Further denote the jth subject's initial NIHSS by I_j and the treatment to which she was randomized by d_j (d_j is either 0 for control or 1, 2 for the active arms). Define, for $6 \leq i \leq 42$ and $d \in \{0,1,2\}$,

$$\Pr\{Y_j = \mathrm{W}k \mid I_j = i, d_j = d\} = p_k(i,d).$$

For the purposes of the final analysis, the $p_k(i,d)$ are used to define the expected values of the weight scores. The expected values of the weight scores, Y_j, are modeled as Gaussian, with expected values that depend on i, with a common treatment effect θ_d with $\theta_0 = 0$ by convention, and with variances σ_d^2 that depend on the treatment:

$$E(Y_j | I_j = i, d_j = d) = \sum_{k=0}^{6} p_k(i,d) W_k = \phi_i + \theta_d,$$

for all initial NIHSS scores i. We model the ϕ_i flexibly, and assume that they come from a second order normal dynamic linear model (NDLM). Specifically, the prior distribution for the ϕ_i assumes that for $8 \leq i \leq 42$, we have

$$\phi_i \sim Normal(2\phi_{i-1} - \phi_{i-2}, \tau^2).$$

This form of the normal dynamic linear model encourages the ϕ_i to be linear.

We use the following prior distributions:

$\theta_d \sim Normal(0, 2.5^2)$ $(d = 1,2)$

$\sigma_d^2 \sim Inverse\ Gamma\ (1, 10)$ $(d = 0,1,2)$, and

$\tau^2 \sim Inverse\ Gamma\ (10, 0.005)$.

The final analysis is performed with either one or two active arms. We evaluate the posterior distribution of the parameters of this model using the Gibbs sampler. Conditionally on the other parameters, (ϕ, θ) have a multivariate normal distribution, and the remaining parameters have inverse Gamma conditional distributions.

The primary output of the final analysis is the posterior probability that $\theta_d > 0$, for any d's that remain in the trial. If this probability is at least 0.985, the trial is considered to be a success. The threshold for defining significance is chosen so that the design has Type I error no larger than 0.025. Criteria for success (critical value for posterior probability of a positive benefit) was inflated from 0.975 to 0.985 to account for the two study drugs and the repeated interim looks (e.g. the trial can be stopped when data look favorable enough that a success is likely).

Appendix C2: Statistical Model for Interim Analyses

The statistical model used during the trial to compute predictive probabilities to determine allocation probabilities and make the decision to stop for futility or expected success, is more detailed than the final analysis model. We use a longitudinal model to impute values of final endpoints for subjects for whom we have 30-day mRS scores but not 90-day scores; we estimate the probability distribution of final endpoint values given early endpoint values. Another major change is that we also estimate the distribution of initial NIHSS scores for enrolled subjects; for predicting whether the trial will be successful it is critical to be able to forecast what kinds of subjects will appear in the future.

Whereas in the final analysis we use a noninformative prior with no information about the overall level of the ϕ_i or the overall slopes of the ϕ_i as a function of i, we now use the following prior distributions:

$\phi_6 \sim N(5, 2.5^2), and\ \phi_7 \sim N(\phi_6, 0.25^2)$.

Writing Y_j^{30} for the 30-day mRS value for the jth subject, we estimate the probabilities $\lambda_{mk} = \Pr\{Y_j = k \mid Y_j^{30} = m\}$ using a multinomial model with prior distributions

$$(\lambda_{m0}, \lambda_{m1}, \lambda_{m2}, \lambda_{m3}, \lambda_{m4}, \lambda_{m5}, \lambda_{m6}) \sim Dirichlet(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$$

for $m = 0,1,2,3,4,5$. We have $\lambda_{66} = 1$ and $\lambda_{6k} = 0$ for $k < 6$. As mentioned earlier, the longitudinal model plays no role in the final analysis. The parameters of the longitudinal model are updated at each interim analysis and are based on all subjects with complete 30-day and 90-day data to that point. We note that we are not using data from other studies to inform the parameters of the longitudinal model. We use the same longitudinal model for all arms (i.e. we pool the data for all subjects to estimate the probability distribution of 90-day outcome given 30-day outcome). The final piece of the statistical model for interim analyses is the model for the initial NIHSS distribution $\Pr\{I_j = i\} = \iota_i$, which is also a Dirichlet-multinomial model with prior distribution

$$(\iota_6, \iota_7, \dots, \iota_{42}) \sim Dirichlet\ (\frac{1}{3}, \frac{1}{3}, \dots, \frac{1}{3}).$$

The prior distributions for the σ_d^2 are as specified in the description of the final analysis.

During an interim analysis, we estimate the parameters $(\phi, \theta, \sigma^2, \tau^2, \lambda, \iota)$ of this model using Gibbs sampling. We then use these samples to estimate several predictive quantities. First, for each active arm in the trial, we calculate the probability that the trial would end with a significant result if we assigned all remaining subjects 1:1 to control and that active arm, and enrolled subjects up to the maximum sample size. This calculation consists of the following steps: for a given Markov chain Monte Carlo (MCMC) sample,

1. Using the λs, impute 90-day endpoint values for the subjects enrolled and with 30-day data.

2. Simulate random initial NIHSS scores and treatment assignments for the subjects yet to be enrolled, using the ι's and assuming that all remaining subjects are enrolled 1:1 to control and that active arm. Augment this list of subjects with the subjects included in the trial who have not yet provided 30 day data.
3. Calculate the probability, given that list of subjects, that final 90 day data will result in a significant trial.

Steps 1, 2, and 3 will be repeated for each of a large number of MCMC samples. Next, we compute the average of the resulting probabilities. These probabilities will be used to decide whether to make arm selection decisions and futility stopping decisions.

Second, we calculate the probability of a successful final analysis if the trial assigned all future subjects to control and the two active arms equally 1:1:1. This probability is used to decide whether to drop one active arm.

Finally, we calculate the probability of a successful final analysis if the trial were to stop enrollment at once and then wait for final data for all enrolled patients. This probability is based on predicting the final data for enrolled patients with no data yet, as well as those with 30-day data only.

Appendix C3: Default Population Assumptions

In this appendix we present the assumptions about the population that were used in the simulations. First we present the assumed distribution of NIHSS scores for enrolled subjects. Next we present the assumed distribution of mRS as a function of NIHSS for control subjects. Finally we present the assumed distribution of 90-day mRS given 30-day mRS scores.

Appendix C3.1: Initial NIHSS Distribution

In Figure C1 we display the assumed distribution of initial NIHSS scores. The distribution is based on an exponential distribution with mean 10 but with values less than 6 or more than 42 omitted. This assumption was chosen to be roughly consistent with Reeves et al (Distribution of National Institutes of Health Stroke Scale in the Cincinnati/Northern Kentucky Stroke Study**;** Mathew Reeves, et al; *Stroke.* 2013; 44:3211-3213).

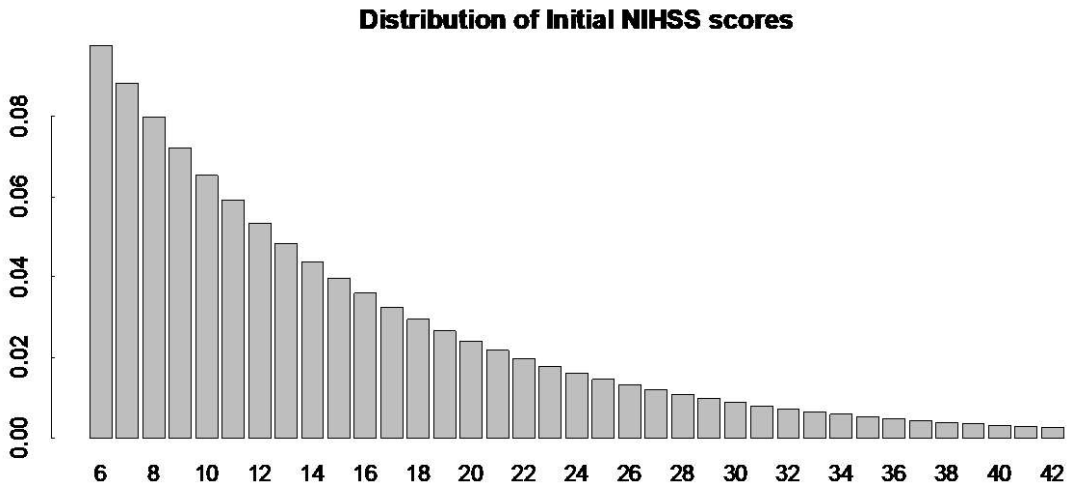


Figure C1: the assumed distribution of initial NIHSS scores for enrolled subjects. Lower scores are considerably more likely.

Appendix C3.2: Distribution of 90-Day mRS Given Initial NIHSS

Figure C2 below displays the distributions of 90-day mRS given initial NIHSS assumed for the control arm. Good outcomes are expected to be very likely for NIHSS of 6 and very infrequent for initial NIHSS larger than 26.

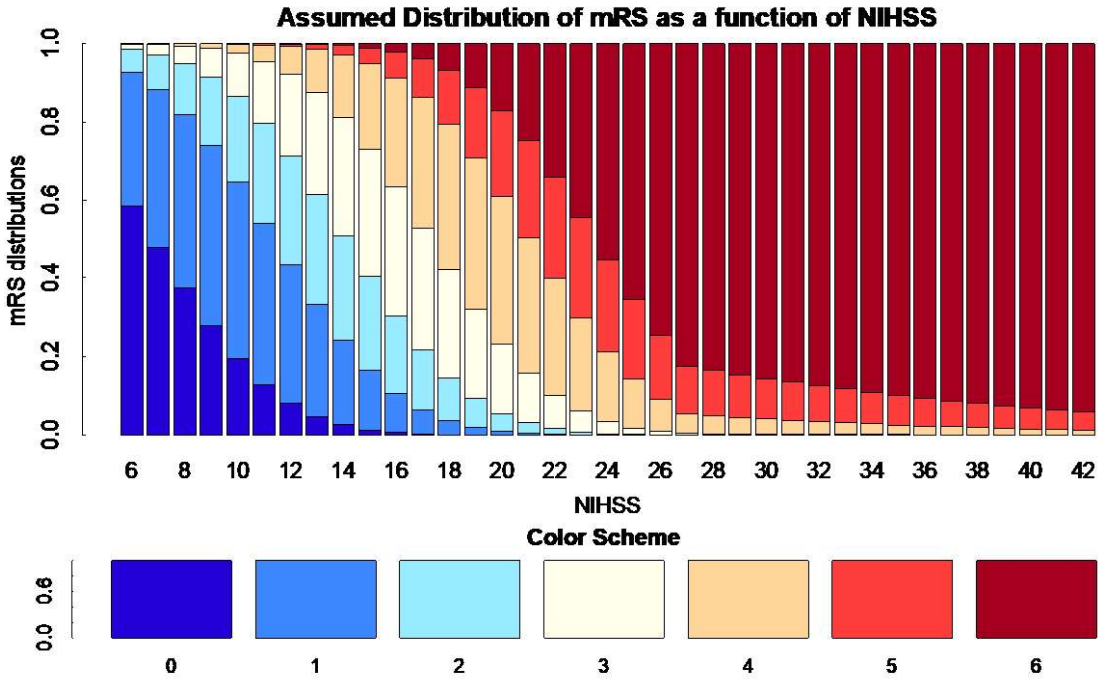


Figure C2: distributions of 90-day mRS values conditional on initial NIHSS.

Appendix C3.3: Distribution of 90-Day mRS Given 30-Day mRS

Table C1 below displays the assumed conditional distributions of 90-day mRS scores given 30-day mRS scores. These were taken from Ovbiagele, Lyden, and Saver (2010; Ovbiagele B, Lyden PD, Saver JL, Disability status at 1 month is a reliable proxy for final ischemic stroke outcome. Neurology 2010;75:688-92).

30 \ 90	0	1	2	3	4	5	6
0	0.78	0.18	0.02	0.003	0.003	0	0.009
1	0.24	0.65	0.075	0.02	0.01	0.001	0.01
2	0.06	0.38	0.45	0.09	0.02	0	0.009
3	0.02	0.12	0.34	0.44	0.06	0.01	0.01
4	0.003	0.02	0.06	0.29	0.53	0.05	0.04
5	0	0.001	0.001	0.007	0.06	0.87	0.06
6	0	0	0	0	0	0	1

Table C1: conditional distribution of 90-day mRS given 30-day mRS.