Berry Consultants
Statistical Innovation

Statistical Plan for the MOST Trial
September 23, 2014

1.0 Background

This is a plan for a Phase III trial to explore the efficacy of Argatroban and Eptifibatide in combination with rt-PA in treating stroke patients. Two doses of each combination will be compared to a control (rt-PA only) 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 1500 patients;
2. initial consideration of four active arms, with the ability to transition to either a 1:1 comparison between control and one of the active arms, or a 1:1:1 allocation between control, one combination dose with Argatroban, and one combination dose with Eptifibatide;
3. response-adaptive randomization is used to favor promising active arms;
4. interim analyses are conducted every four weeks early in the trial, and every ninety days later in the trial, and they 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 stroke severity as measured by baseline score on 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 illustrates an example trial and explains in detail how the trial would proceed given a particular random sequence of patients and outcomes, including the analyses performed during interim analyses. Section 6 presents tentative operating characteristics for the design obtained using simulation and based on a variety of assumed truths about the effectiveness of the drugs. Sections 7 and 8 elaborate on the statistical models used in the final analysis and in the interim analyses and on default assumptions about the subject population.

Berry Consultants
Statistical Innovation

2.0 Design Overview

For the first 150 subjects, the randomization probabilities for the five arms remain fixed at 1/3 for the control arm and 1/6 for each of the active arms. An interim analysis occurs at the 150th enrolled subject, and randomization probabilities for the active arms are adjusted, with arms performing well enough to demonstrate superiority over control assigned higher probabilities. Interim analyses continue every 4 weeks.

Beginning with the first interim analysis after the 300th subject is enrolled, the design has the option of dropping two or three of the active arms and continuing with either

- 50-50 randomization between control and the most effective dose of Argatroban
- 50-50 randomization between control and the most effective dose of Eptifibatide
- 1:1:1 randomization between control, the most effective dose of Argatroban, and the most effective dose of Eptifibatide.

After 500 subjects have been enrolled, the decision must be made to either stop for futility or to shift to equal randomization between control and one or two active arms as above. From this point until the end of the trial, all arms remaining in the trial have equal allocation probabilities.

The trial enrolls at most 1500 subjects. From the time of the reduction to one or two active arms until complete data for the 1500th subject are obtained, interim analyses occur only every 90 days. These interim analyses can result in stopping the trial for futility. If two active arms remain, the design can elect to drop one of them and proceed with 50-50 randomization between the control and the single remaining active arm. Furthermore, the design can elect to stop an arm for expected success beginning when 200 subjects have been enrolled after the start of the evenly allocated regime. (This first opportunity for expected success occurs no earlier than n=500 overall. It occurs at n=700 overall if the evenly allocation regime begins at n=500). If all active arms that are included in the evenly allocated regime are stopped for either futility or expected success, enrollment ends. 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.

Berry Consultants
Statistical Innovation

3.0 Study Population, Primary Endpoint, and Statistical Test

3.1 Entry criteria

The trial enrolls acute ischemic stroke patients with initial NIHSS scores of 6 or larger that receive intravenous tPA (IV rt-PA).

3.2 Treatment arms

Five treatment arms are under consideration:

1. Control arm with IV rt-PA only;
2. Low dose of Argatroban in addition to IV rt-PA;
3. High dose of Argatroban in addition to IV rt-PA;
4. Eptifibatide in addition to reduced dose IV rt-PA;
5. Eptifibatide in addition to standard dose IV rt-PA.

Of the first 150 enrolled subjects, one-third will be assigned to the control arm and one-sixth apiece 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 and 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
This Trial	10	9.1	7.6	6.5	3.3	0	0

Table 1: Utility weights used for 90 day mRS scores.

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 a mRS of 0, the darkest red shows the probability of a mRS of 6, etc. The leftmost bar shows the results of the control arm for the NINDS rt-PA 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 rt-PA arm in the NINDS study, which represents an expected utility of 5.91. The remaining bars show distributions that represent additional improvements of 0.1 to 0.9 as compared to the rt-PA arm (eg. rt-PA + 0.1, rt-PA + 0.2, etc.).

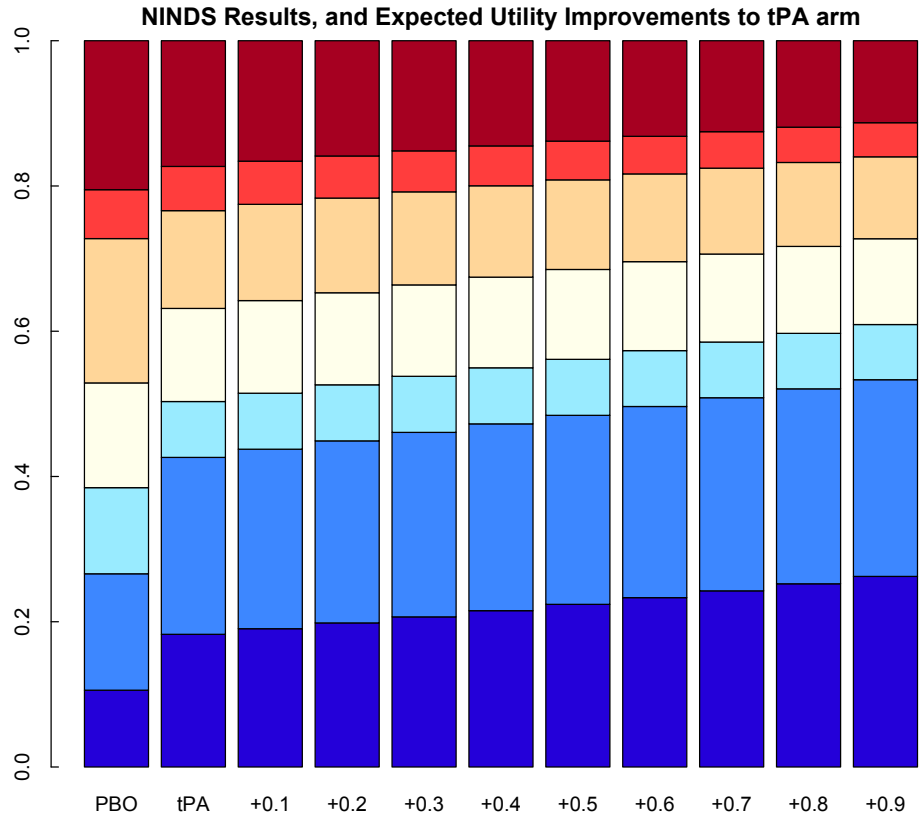

Figure 1: Graphical depiction of mRS distributions. The first two bars depict the two arms in the NINDS study (PBO = Placebo), and the remaining nine bars depict further improvements to expected utility. Dark blue represents the probability of an mRS of zero, dark red represents the probability of an 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 rt-PA arm	5.91	0.18	0.24	0.08	0.13	0.13	0.06	0.17	0.42
rt-PA + 0.1	6.01	0.19	0.25	0.08	0.13	0.13	0.06	0.17	0.44
rt-PA + 0.2	6.11	0.20	0.25	0.08	0.13	0.13	0.06	0.16	0.45
rt-PA + 0.3	6.21	0.21	0.25	0.08	0.13	0.13	0.06	0.15	0.46
rt-PA + 0.4	6.31	0.22	0.25	0.08	0.12	0.13	0.05	0.14	0.47
rt-PA + 0.5	6.41	0.22	0.26	0.08	0.12	0.12	0.05	0.14	0.48
rt-PA + 0.6	6.51	0.23	0.26	0.08	0.12	0.12	0.05	0.13	0.49
rt-PA + 0.7	6.61	0.24	0.27	0.08	0.12	0.12	0.05	0.13	0.51
rt-PA + 0.8	6.71	0.25	0.27	0.08	0.12	0.12	0.05	0.12	0.52
rt-PA + 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 rt-PA in the NINDS study is 0.9 units of utility. The table shows what further improvements to expected utility above rt-PA 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 baseline NIHSS. This is a flexible spline-like model in which the average weighted mRS score in the control group is a (possibly non-linear) function of the initial NIHSS. Meanwhile the average effect of a given treatment d, θ_d, is assumed to be equal over all values of initial NIHSS. Details of the statistical model are given in Section 7.1.

The treatment effect θ_d is given a vague prior probability distribution, $\theta \sim N(0,2.5^2)$: in particular, the prior probability that a drug is beneficial is the same as the prior

probability that it is harmful. 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 in the trial assigned to the control arm and to the selected active arm. 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 0.986. This threshold is chosen to control the total Type I error probability of the trial (i.e. the probability that any of the four active arms achieves a successful result) at the 0.025 level. If the design included just one active arm and did not have the potential to stop for success before the maximum sample size, then the posterior probability of a positive benefit would be closely related to the classical p-value of the hypothesis test of no benefit. In this case, the final analysis could declare success with a posterior probability of at least 0.975 of a positive benefit, and the Type I error of the design would be very close to the nominal level of 1 – 0.975 = 0.025. In the actual design, one or two active arms are selected from a total of four, and the trial can be successful with **as few as 700 subjects**, so the Type I error probability is inflated from the nominal level. However, the design requires at least 200 subjects after the arms are selected, and the final analysis takes place after final data are obtained from all enrolled patients as opposed to at the time a stopping decision is made, and both of these features help reduce the amount of inflation.

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. However, 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.

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). For all interim analyses between 150 and 500 subjects, no early stopping is allowed, but response-adaptive randomization alters the allocation probabilities for the four active arms.

Berry Consultants
Statistical Innovation

Also, the design may elect to drop two or three of the active arms and shift to an equal randomization regime at any interim analysis with between 300 and 500 subjects. If this shift occurs, future interim analyses are conducted every 90 days.

4.1 Predictive Probabilities

Decisions made as a result of interim analyses are based on Bayesian predictive probabilities using the statistical model defined in Section 7.2. 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, response adaptive randomization probabilities, and decisions to shift to an evenly allocated regime. For convenience we will refer to these predictive probabilities as ***1:1 predictive probabilities***.

Second, for each pair of active arms consisting of one Argatroban and one Eptifibatide arm, we assume that the remaining subjects to be accrued are allocated 1:1:1 to those 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. If and when the decision is made to shift to an evenly allocated regime, these predictive probabilities are used to decide whether one or two active arms participate. 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 Shifting to Equal Randomization

Beginning with the first interim analysis after enrollment of the 300th subject, the trial has the potential to shift to the equally allocated regime. If this shift has not yet happened, all the 1:1 predictive probabilities are evaluated, and if any one of them exceeds 90%, the decision is made to shift, and the arm with the largest predictive probability is included in the equally allocated regime. The other arm for that drug is dropped. The 1:1:1 predictive probability is calculated for the selected arm and the two arms for the other drug, and if either of those two arms has a two-arm

predictive probability of at least 50%, the arm with the larger predictive probability is also included in the equally allocated regime.

4.3 Interim Monitoring for Early Futility and Expected Success.

No futility stopping is possible before the 500th subject.

After the enrollment of the 500th subject, an interim analysis takes place, and a final decision is made whether to execute an equally allocated regime. Since this is the last opportunity, the threshold for the decision is reduced: if any active arm has a single-arm predictive probability of at least 50%, the equally allocated regime begins with the arm with the largest 1:1 predictive probability included. In this case, an arm for the other drug will also be included if it *also* has a ≥50% 1:1:1 predictive probability. If at the 500th subject, all four active arms have 1:1 predictive probabilities below 50%, the trial stops for futility.

During the equally allocated regime, if two active arms are involved, one of those arms can be dropped if its 1:1:1 predictive probability is less than 5%. If this occurs, the trial proceeds using 1:1 randomization for the control and the surviving active arm. If only one active arm remains, the trial stops for futility if its 1:1 predictive probability is less than 5%.

When 200 patients have been enrolled during the equally allocated regime, the trial gains the ability to stop for expected success. Beginning with the first analysis after this time, the expected success predictive probabilities are calculated. If a surviving 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 at the time of its stop are followed up for 90-day data, and based only on the subjects that were enrolled at that time. 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. If another arm remains after one arm stops for expected success, the trial continues with 1:1 randomization between control and the remaining arm.

4.4 Response-Adaptive Randomization

During the response-adaptive randomization regime, which begins at 150 subjects and ends when the shift to an equally allocated regime begins, the control arm retains a 1/3 allocation probability throughout. The allocation probabilities for the four active arms are set to be proportional to their 1:1 predictive probabilities. For example, suppose that the four active arms have 1:1 predictive probabilities of 0.2, 0.4, 0.6, and 0.8 respectively. The 2/3 total allocation probability assigned to active arms is divided up proportionally to the predictive probabilities, and in this case the four arms have allocation probabilities of (1/15, 2/15, 3/15, 4/15) respectively.

5.0 An Example Trial

In this section we walk through an example of how a trial with this design might proceed. We show the data that might be available for a series of interim analyses, and how these data contribute to the decisions made during the trial. These are simulated data similar to that generated for the purpose of estimating operating characteristics using simulation. For example, Figure 2 shows a potential data set for the first interim analysis, conducted after 150 enrolled subjects. We show two plots per interim analysis. The left plot shows the mRS data as a function of initial NIHSS (shown on the x-axis), maturity of data (larger plot character means final 90-day data, smaller character means 30-day data), and treatment assigned (shown using color: black represents control-Placebo (PBO), two shades of red represent the two Argatroban arms – A1 & A2, and two shades of blue represent the two Eptifibatide arms – E1 & E2). It also shows the model estimates of expected utility as a function of initial NIHSS using color-coded solid lines (point estimates) and dashed lines (95% credible intervals). Sample sizes for each arm are shown in the plot legend. The right plot shows the predictive probabilities used in the decisions, using the same color scheme to indicate the active arms. The leftmost set of bars shows the expected success predictive probabilities, the middle set of bars shows 1:1 predictive probabilities, and the rightmost set of bars shows the 1:1:1 predictive probabilities.

In Figure 2 itself, the data points show that good results have been typical for subjects with small NIHSS scores, but this is especially true for the two Eptifibatide arms (light blue and dark blue) and the larger dose of Argatroban (dark red), while subjects with poor results were predominantly on low dose Argatroban or control. The statistical model estimates that the two Eptifibatide arms and the high dose of Argatroban are approximately equally effective, as their curves are plotted on top of each other, while low dose Argatroban is estimated to be slightly worse than control. Predictive probabilities of eventual success at 1500 subjects are around 70% for the three promising arms, while the underperforming active arm has no better than a 20% predictive probability of success. The data are not so favorable as to make it likely to achieve a successful result should accrual stop at once, however: these predictive probabilities are in the range of 10-15%. No early stopping decisions are possible at this early stage anyway, but the optimistic predictive probabilities are used in response adaptive randomization, so that few subjects enrolled during the next 28 days will be assigned to low dose Argatroban.

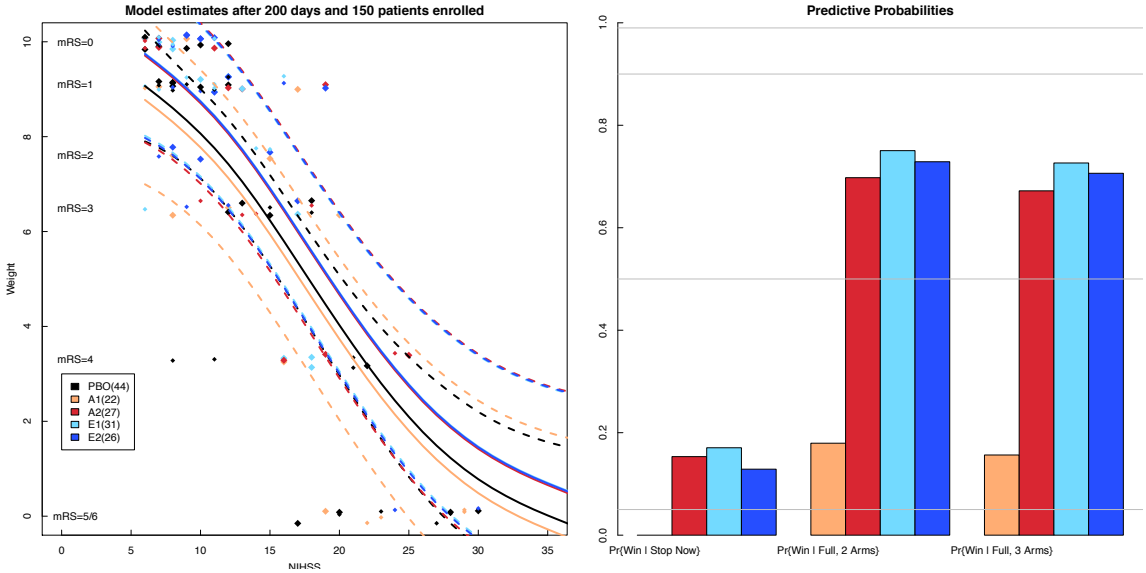

Figure 2: Results of **first** interim analysis, conducted after 150 patients enrolled. Left plot shows mRS (Y-axis) as function of NIHSS score (X-axis). PBO – placebo. A1 & A2 – argatroban arms; E1 & E2 – eptifibatide arms.

The second interim analysis takes place 28 days later, and results are similar to the first analysis. The third interim analysis is shown in Figure 3. Only two subjects have been randomized to low dose Argatroban since the first interim analysis, as compared to 13 to 22 for the other active arms, and that arm is still estimated to be ineffective. Some separation between the other arms has occurred, though: both Eptifibatide arms have high predictive probabilities of eventual success, and high dose Argatroban's predictive probabilities are down to under 60%. It is too early for a shift to the evenly allocated regime, but for reference, if that decision were made now, the evenly allocated regime would start and involve the two larger doses. This is because high dose Eptifibatide has a greater than 90% predictive probability given 2-arm allocation, and this is enough to trigger the shift, while high dose Argatroban has greater than a 50% predictive probability assuming 3-arm allocation, which is enough for that arm to be included as well.

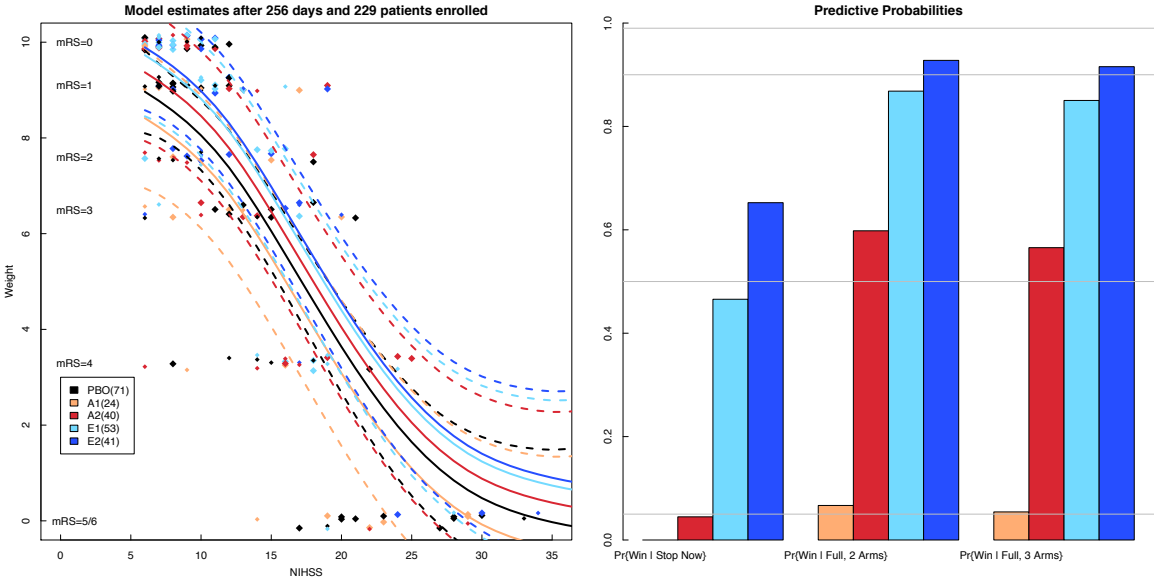

Figure 3: Results of **third** interim analysis; total enrolled = 229.

We move ahead to the fifth interim analysis, shown in Figure 4, which takes place with 311 enrolled subjects: since this is more than 300, it is legal for the first time to shift to an evenly allocated regime. The lower dose of Eptifibatide has overtaken the higher dose, and it has better than a 90% predictive probability of a successful final analysis based on 1500 subjects, so the evenly allocated regime begins. High dose Argatroban (dark red color) has better than a 50% 1:1:1 predictive probability, so this arm also proceeds to the new regime. Only 24 subjects were ever allocated to low dose Argatroban, while 60 subjects were allocated to the much better performing high dose Eptifibatide arm. Since the new regime begins with 311 subjects, neither active arm can stop for expected success until 511 subjects are enrolled, and interim analyses occur only every 90 days instead of 28.

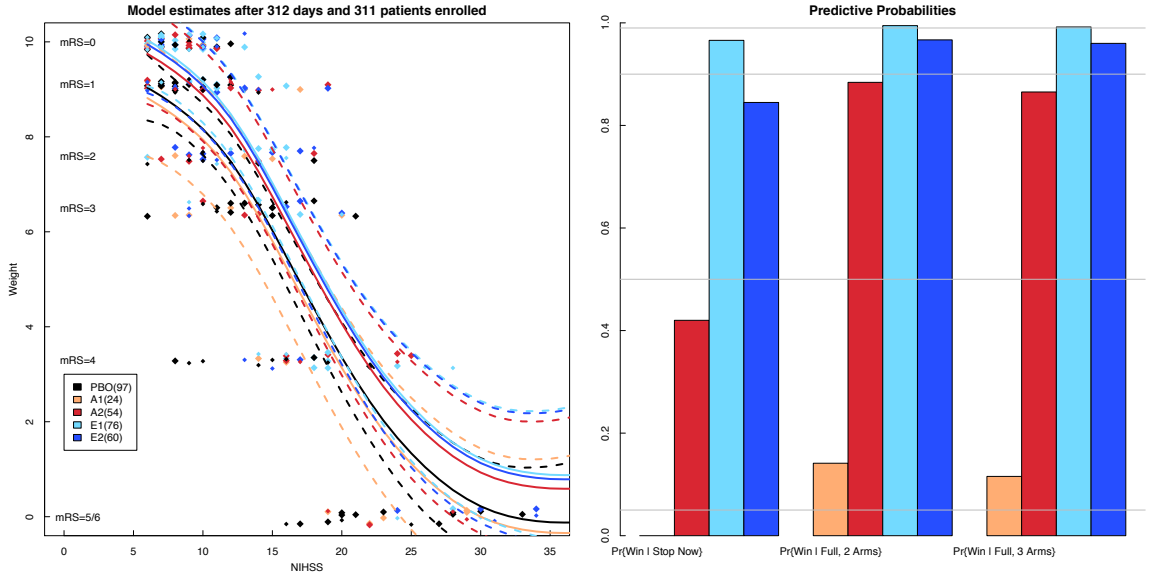

Figure 4: Results of **fifth** interim analysis, which triggers a shift to the evenly allocated regime with low dose Eptifibatide and high dose Argatroban. Total enrollment = 311.

The first opportunity for an arm to stop for expected success occurs with 638 subjects enrolled, which is the eighth interim analysis, and is shown in Figure 5. The rule is that if the predictive probability of a successful final analysis assuming enrollment stops immediately is at least 99%, that arm can be stopped, and as shown by the leftmost light blue bar, the low dose of Eptifibatide meets this threshold. After 161 subjects have been assigned to this arm, no more subjects will be assigned to it, and its success or failure is determined when all subjects currently enrolled have been followed up for their 90-day outcomes. (The higher dose meets the threshold as well, but it has already been dropped from the trial.) Since the scientific question about the benefit of high dose Argatroban is still not settled, the trial continues with 1:1 randomization between control and this remaining arm. Since the trial has enrolled more than 500 subjects, futility stopping is now a possibility. If the middle dark red bar drops below 5% indicating small probability of a successful trial given full enrollment, the high dose of Argatroban stops for futility, ending the trial, but it is well above this threshold.

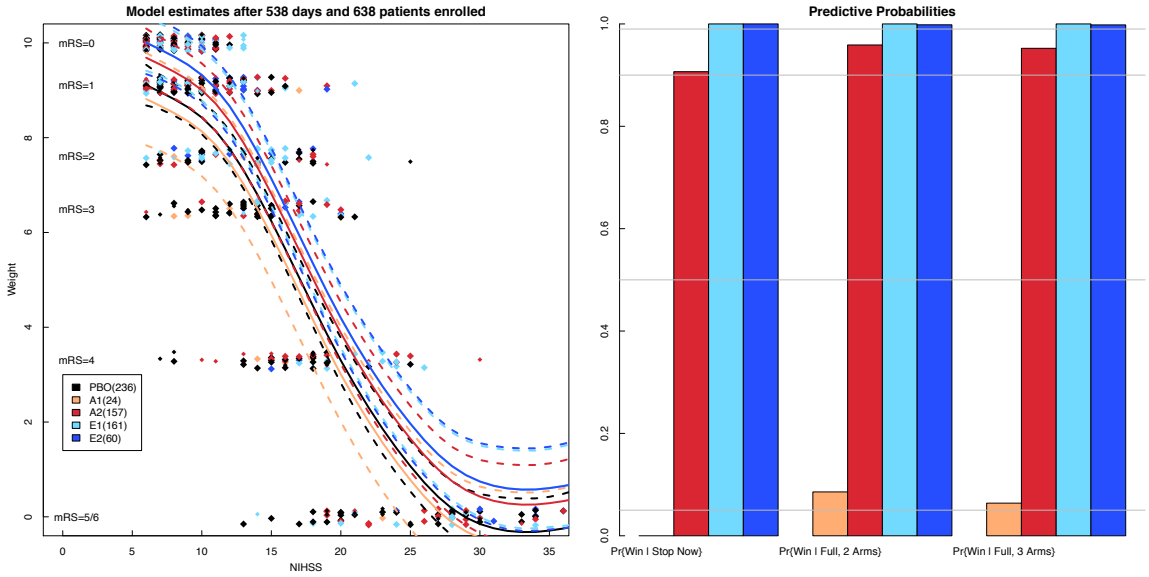

Figure 5: Results of **eighth** interim analysis, which results in the stopping of the low dose Eptifibatide arm (light blue) for expected success and the continuation of the trial with 1:1 randomization between control (PBO) and high dose Argatroban (dark red).

Finally, Figure 6 shows the final interim analysis. The leftmost red bar has exceeded the 99% threshold, indicating that the model expects a successful result if enrollment stops immediately, so the trial concludes at this point, with a final total sample size of 916 subjects, including 343 on control and 328 on high dose Argatroban. The solid curves on the left plot show that the estimated treatment effect is larger for low dose Eptifibatide than for high dose Argatroban, but after collecting final data on all enrolled subjects, the Argatroban arm is successful as well.

These data were simulated under a scenario where both Eptifibatide arms had a true benefit of 0.75 units of expected utility, the high dose of Argatroban had a true benefit of 0.5 units, and the low dose of Argatroban had zero benefit.

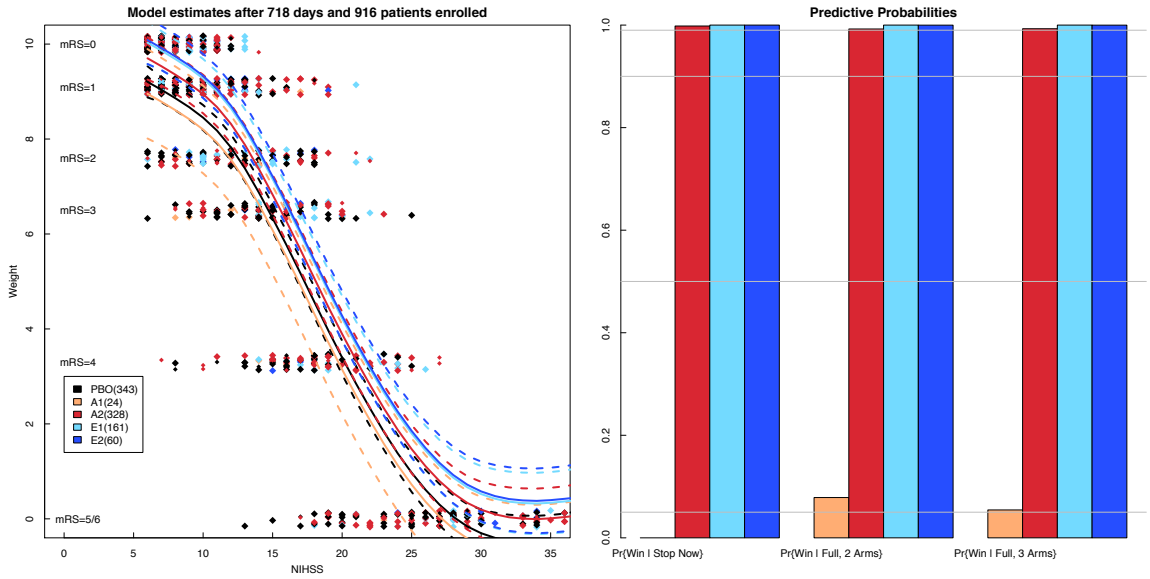

Figure 6: Results of the **tenth** interim analysis. The remaining active arm (high-dose Argatroban – red color) exceeds the predictive probability threshold for early stopping due to expected success, and enrollment stops.

6.0 Operating Characteristics

In this section we present tentative operating characteristics. These results are obtained through simulation as illustrated in Section 5, and are based on 500 simulated trials per scenario.

6.1 Operating characteristics when all active arms are equivalent

First we present preliminary estimates of power and sample size characteristics in scenarios in which all four active arms have the same effect: specifically, an increase of zero (a null scenario), 0.25, 0.30, 0.35, 0.40, 0.50, 0.75, and 1.0 units. These effects benefit all injury severities equally with respect to expected utility. Examples of the scenarios are pictured in Figure 7 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%.

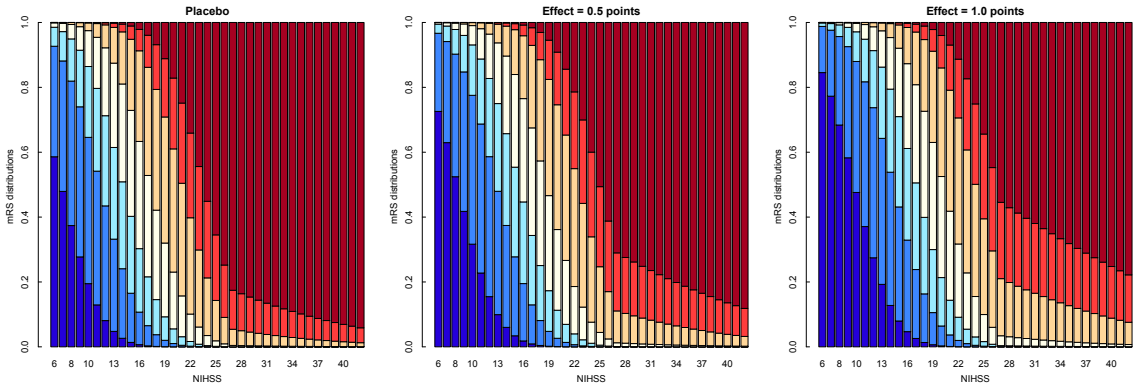

Figure 7: Example distributions of mRS as a function of initial NIHSS.

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 an Argatroban arm is successful;
- the probability that an 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 one arm of each drug both win.

Results based on 500 simulated trials for each scenario are shown in Table 3. Simulations and analysis have not yet been extensive enough to claim control of Type I error, but the numbers in the Null column are consistent with 0.025 Type I error. For the remaining scenarios, an effect of 0.75 units is large, and it is very likely that some arm will be successful, but there is a small but non-negligible probability that one drug or the other will not be successful. If the true effect is 0.5 units for all active arms, the design has high power of 98% to be successful for some arm, and each drug has a 10% chance of being unsuccessful.

Assumed True Effect Difference in Utility	**Null**	**0.25**	**0.30**	**0.35**	**0.40**	**0.50**	**0.75**	**1.00**
Pr{Some Arm Wins}	0.018	0.60	0.78	0.89	0.94	0.98	1.00	1.00
Pr{Some A Arm Wins}	0.009	0.41	0.60	0.72	0.79	0.90	0.99	0.999
Pr{Some E Arm Wins}	0.009	0.41	0.60	0.72	0.79	0.90	0.99	0.999
Pr{Some A Arm and some E Arm both win}	0.000	0.23	0.42	0.54	0.65	0.81	0.98	0.998

Table 3: Power-related operating characteristics for scenarios in which all four active arms are equally effective. Pr=Probability.

Berry Consultants
Statistical Innovation

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 elects to run an evenly allocated regime. The average and standard deviation of total sample sizes are also shown, as well as the averages for trials that were ultimately successful and those that were unsuccessful. Note that if no arm has any benefit (the null case), the design correctly elects to stop for futility at 500 subjects 63% of the time, and the overall average number of subjects in the null case is 738, with a standard deviation of 347 subjects.

Assumed True Effect Difference in Utility	**Null**	**0.25**	**0.30**	**0.35**	**0.40**	**0.50**	**0.75**	**1.00**
Pr{Futility at 500}	0.63	0.17	0.11	0.05	0.04	0.01	0.00	0.00
Pr{Even allocation regime}	0.37	0.83	0.89	0.95	0.96	0.99	1.00	1.00
Mean N	738	1196	1207	1197	1118	963	712	639
Standard Deviation(N)	347	375	340	299	284	222	112	39
Mean N (successful)	1395	1342	1287	1234	1142	968	712	639
Mean N (failed trials)	729	981	917	897	762	673	----	----

Table 4: Sample size-related operating characteristics. These include the probabilities of stopping for futility at 500 subjects (row 1) and of conducting an evenly allocated regime –row 2 (these probabilities will sum to one). The last four rows display the average number of subjects enrolled and the standard deviation, and the average number of subjects when only successful trials are considered, and when only failed trials are considered. Pr=Probability.

6.2 Operating characteristics when arms differ

In this section we present operating characteristics for scenarios in which some arms have benefits of up to 0.4 units, while other arms have no benefit or a smaller benefit, to explore how successful the design is at identifying the most promising arm. The six scenarios are denoted by a four-digit descriptor consisting of four 0's, 2's and 4's; the four digits correspond to the four active arms and indicate the size of the treatment effect: 0 means a benefit of 0 units of expected utility, 2 means 0.2 units, and 4 means 0.4 units. For example, code "0244" means that the small dose of Argatroban has no effect, the large dose of Argatroban has an effect of 0.2 units of utility, and both doses of Eptifibatide have effects of 0.4 units of utility. We have simulated scenarios in which Eptifibatide is more 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 Argatroban is the superior drug. Similarly, we have simulated cases in which the larger dose is the more effective, but the results also apply to cases where the smaller dose is preferable.

We see from the first five columns in the first row of Table 5 that if two of the four arms have effects of 0.4 units of utility, the design has approximately 81-85% power, regardless of whether the two good arms correspond to the same drug or to different drugs. When one arm of each drug has a 0.4 unit effect, there is approximately a 41% chance that both drugs will be successful. If one drug is more effective than the other, the probability that the more effective drug will be successful is not strongly dependent on the qualities of the weaker drug. If one arm of a particular drug is half as effective as the other arm, the weaker arm is successful approximately one-sixth as often as the more effective arm.

Scenario	**2424**	**2244**	**0244**	**0404**	**0044**	**0024**	**0004**
Pr{Some Arm Wins}	0.836	0.816	0.824	0.848	0.810	0.652	0.668
Pr{Some A Arm and some E arm both win}	0.416	0.222	0.145	0.398	0.014	0.006	0.014
Pr{Arm A1 wins}	0.077	0.124	0.003	0.001	0.007	0.004	0.007
Pr{Arm A2 wins}	0.549	0.124	0.159	0.622	0.007	0.004	0.007
Pr{Arm E1 wins}	0.077	0.395	0.403	0.001	0.405	0.094	0.004
Pr{Arm E2 wins}	0.549	0.395	0.403	0.622	0.405	0.556	0.664

Table 5: Power characteristics for scenarios where the active arms differ. Pr = probability. The 4-digits correspond to 4 active arms and indicate the size of treatment effect in utility units: 2=0.2 & 4=0.4.

Table 6 contains further operating characteristics for these scenarios related to sample size for these scenarios. In addition to the summaries in Table 4, this table shows the average numbers of subjects allocated to each arm. In general, subjects are allocated more often to arms with larger benefits. Any arm with zero benefit is expected to be allocated to roughly 100 subjects or fewer. The '0244' scenario assigns more subjects on the average to the arm with a benefit of 0.2 units than it does to either of the arms with 0.4 unit benefits, because the two arms with larger benefits split the opportunities with each other

Scenario	2424	2244	0244	0404	0044	0024	0004
Pr{Futility at 500}	0.09	0.09	0.10	0.11	0.12	0.19	0.20
Pr{Even allocation regime}	0.91	0.91	0.90	0.89	0.88	0.81	0.80
Mean N: Control	441	462	451	410	415	410	399
Mean N: Arm A1	123	165	84	70	99	96	108
Mean N: Arm A2	222	165	203	247	99	96	108
Mean N: Arm E1	123	183	195	70	205	153	82
Mean N: Arm E2	222	183	195	247	205	260	296
Mean N: Total	1130	1157	1129	1046	1022	1015	993
Standard Deviation(N)	323	427	338	318	311	350	350
Mean N (successful)	1188	1290	1195	1112	1098	1123	1117
Mean N (failed trials)	831	921	771	675	700	813	743

Table 6: Sample size-related operating characteristics for scenarios where the active arms differ. These include the probabilities of stopping for futility at 500 subjects (row 1) and of conducting an evenly allocated regime –row 2 (these probabilities will sum to one). Also shown are the arm-specific average sample sizes in the third through seventh rows. The last four rows display the average number of subjects enrolled and the standard deviation, and the average number of subjects when only successful trials are considered, and when only failed trials are considered. Pr=Probability.

7.0 Statistical Models

7.1 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 through 4 for the active arms). Define, for $6 \leq i \leq 42$ and $d \in \{0,1,2,3,4\}$,

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

For the purposes of the final analysis, we model the $p_k(i,d)$ 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:

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

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:

$\theta_d \sim Normal(0, 2.5^2)$ $(d = 1,2,3,4)$

$\sigma_d^2 \sim Inverse\ Gamma\ (1, 10)$ $(d = 0,1,2,3,4)$, and

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

The final analysis is performed with only one or two active arms, so only one or two of the θ_d's are involved in the final model. 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 0.986, 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. Type I error probability is inflated above the nominal value for two reasons: first, early data are used to select the active arm(s) that can appear in the final analysis, and then those data are used again in the computation of the posterior probability of a benefit. Second, the trial can be stopped when data look favorable enough that a success is likely. If these inflation factors were not present, the critical value for posterior probability of a positive benefit would be closer to 0.975.

7.2 Statistical Model for Interim Analyses

The statistical model used during the trial to compute predictive probabilities to determine allocation probabilities, choose active arms to move on to the equal allocation regime, and 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 patients 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 remaining 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 subject accrual is restricted to subjects in the enriched population. 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.

One may choose to repeat step 2 multiple times for a given MCMC sample. Compute the average of the resulting probabilities. These probabilities will be used to decide whether to enter the equally allocated regime, 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 two active arms, one for each drug. This probability is used to decide whether to include a second active arm in the equally allocated regime, and whether to drop one active arm from the equally allocated regime when it includes two active arms.

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.

8.0 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.

8.1 Initial NIHSS Distribution

In Figure 8 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).

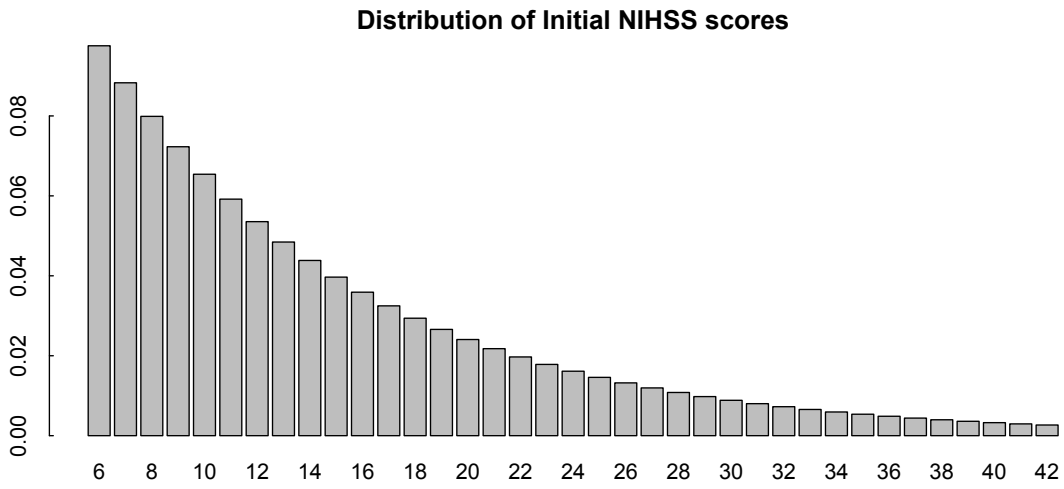

Figure 8: the assumed distribution of initial NIHSS scores for enrolled subjects. Lower scores are considerably more likely.

8.2 Distribution of 90-Day mRS Given Initial NIHSS

Figure 9 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.

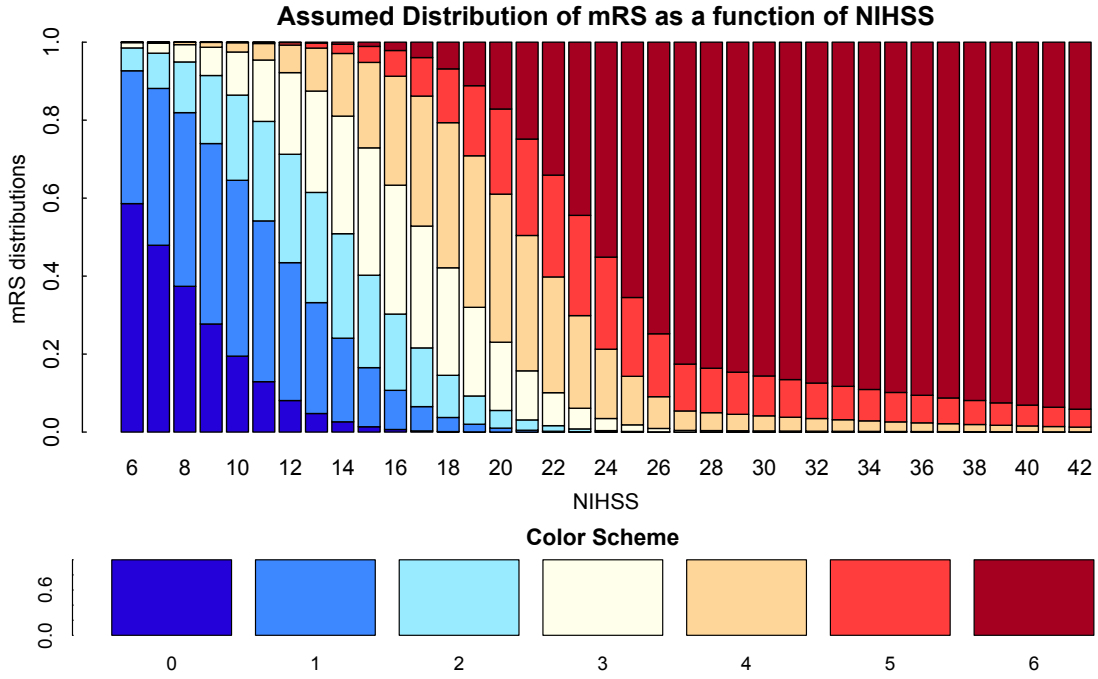

Figure 9: distributions of 90-day mRS values conditional on initial NIHSS.

8.3 Distribution of 90-Day mRS Given 30-Day mRS

Table 7 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 7: Conditional distribution of 90-day mRS results given 30-day mRS score.