Applying a Phase II Futility Study Design to Therapeutic Stroke Trials

Yuko Y. Palesch, PhD; Barbara C. Tilley, PhD; David L. Sackett, MD; Karen C. Johnston, MD, MSc; Robert Woolson, PhD

Background and Purpose—Most large, randomized phase III efficacy trials of therapeutic agents in ischemic stroke have failed to find treatment benefit. We determined whether some phase III studies could have been avoided if preceded by smaller single-arm phase II studies to evaluate the futility of proceeding to phase III.

Methods—To provide examples of the application of phase II methodology, we obtained primary outcome data for the active treatment group of 6 phase III ischemic stroke therapy trials. For each study, we estimated the sample size number required for a multistage single-arm study using parameters specified in the original study. We evaluated outcome data for the first number of subjects in the phase III study treatment arm ordered by enrollment dates. We compared the proportion of favorable outcomes to prespecified stopping criteria derived from a single-arm phase II futility design. If the observed proportion of favorable outcomes was less than the stopping criterion, we declared the treatment not sufficiently effective to warrant further evaluation in phase III.

Results—We identified 3 trials as futile in phase II; none of 3 showed treatment efficacy in phase III. In the 3 remaining phase II trials in which we did not show futility, one showed efficacy in phase III.

Conclusion—Single-arm phase II futility studies have been underused in stroke research, but provide a strategy for discarding treatments likely to be ineffective in phase III trials. (Stroke. 2005;36:2410-2414.)

Key Words: clinical trials, phase II ischemia stroke

Drug development in humans usually evolves over 3 phases. In phase I, a new drug is tested for toxicity in small numbers of healthy volunteers or patients with advanced disease. Promising drugs proceed to a phase II assessment of futility and are separated into those with provisional evidence of efficacy or those with no realistic prospect of efficacy. Phase II studies provide information about side effects and toxicities in the types of patients for whom the treatment is intended, determine the logistics of administration, and provide estimates of treatment costs. Drugs that remain promising after phase II generally proceed to full-scale phase III clinical trials. Both the regulatory and clinical communities regard well-designed, well-executed, randomized, concurrently controlled phase III clinical trials to be the ultimate proof of the efficacy or lack thereof for therapeutic interventions.

Neuroprotective agents that proved efficacious in animal models of ischemic stroke have been tested extensively in phase III randomized trials.1 Kidwell et al examined publications through December 1999 and found 88 trials testing neuroprotective drugs plus 26 testing combinations of neuroprotective and rheologic/antithrombotic agents.2 In all of these trials, many large and costly, the experimental treatments were considered futile.

Futility studies have been useful in the phase II evaluation of cancer treatments.3–6 The proportion of positive outcomes in a single, treated group is compared with the minimally worthwhile proportion of success expected by the drug’s proponents. We applied this strategy to therapeutic agents for ischemic stroke to determine whether phase II futility studies could have prevented some of the futile, large, and costly phase III trials dominating this field.

Materials and Methods

General Strategy

Consistent with outcomes used in many stroke trials, we focus on binary (eg, “favorable” or “unfavorable”) outcomes. We first estimate the proportion of favorable outcomes in untreated controls. Often, this information is derived from historical case-series or control groups in previous trials, just as one would in designing a phase III trial. Therefore, we refer to this value as p*, the reference proportion of favorable outcomes for the single-arm phase II futility study. We then take the minimum clinically meaningful effect size, Δ, used in designing a phase III study, and use Δ as the “minimally worthwhile improvement” in the proportion of favorable outcomes for the futility study. Using this conceptual approach, we design a single-arm phase II study in which all patients receive the investigational drug. The proportion of favorable outcomes from this study

Received February 9, 2005; final revision received May 2, 2005; accepted May 12, 2005.

From the Department of Biostatistics, Bioinformatics & Epidemiology (Y.Y.P., B.C.T., R.W.), Medical University of South Carolina, Charleston, SC; Trout Centre at Irish Lake (D.L.S.), Canada; and the Department of Neurology and Health Evaluation Sciences (K.C.J.), University of Virginia.
Correspondence to Yuko Y. Palesch, PhD, Department of Biostatistics, Bioinformatics & Epidemiology, Medical University of South Carolina, P.O. Box 250835, 135 Cannon St, Suite 303, Charleston, SC 29425. E-mail palesch@musc.edu
© 2005 American Heart Association, Inc.

Stroke is available at http://www.strokeaha.org

DOI: 10.1161/01.STR.0000185718.26377.07
is compared with $p^*$. If the proportion is too low, we would consider it futile to proceed to a phase III trial.

In formal statistical terms, the hypotheses in a single-arm phase II futility study are as follows:

$$H_0: p_a - p^* \geq \Delta \text{ versus } H_1: p_a - p^* < \Delta,$$

where $p_a$ is the hypothesized proportion of treated subjects with a favorable outcome and $p^*$ and $\Delta$ are defined previously. If we reject the null hypothesis that a “minimally worthwhile” improvement exists, we conclude the benefit of the new treatment is less than what we would want, and it is futile to proceed to further testing in a phase III trial. If we fail to reject the null hypothesis that a minimally worthwhile improvement exists, we conclude there is insufficient evidence of futility, and the treatment deserves further testing in a phase III trial to determine its efficacy.

As highlighted in Table 1, the difference in the direction of the futility hypotheses from those of the traditional phase III randomized trials affects the interpretation of type I ($\alpha$) and type II ($\beta$) error probabilities. In a futility analysis, we want to minimize our risk of drawing a false-negative conclusion and miss a potentially effective agent, i.e., we want to minimize $\alpha$. However, a futility trial is still a “pilot” study, and therefore, we select an appropriate level of $\alpha$ that would not require too large a sample size. We are less concerned about drawing false-positive ($\beta$ error) conclusions that ineffective treatments may be effective because treatments that are not determined to be futile in phase II would be tested further in phase III trials with smaller error probabilities at the expense of larger sample sizes.

As in phase III trials, we can conduct interim analyses in single-arm phase II futility studies. The same adjustments for “multiple looks” at interim data apply with appropriate adjustment of the type I error probability. Several authors have developed multi-stage designs for phase II studies allowing early stopping for futility.5–8 O’Brien and Fleming’s strategy for multiple testing, commonly used in phase III trials, is equally applicable in phase II studies.8 Their stopping boundary uses only a small portion of the overall type I error probability early, so we are less likely to declare futility prematurely. As a result, the final nominal $\alpha$ value used for testing at the end of phase II is close to the overall $\alpha$ level. Interim analyses are particularly relevant to phase II trials in stroke as sample sizes tend to be larger than for most phase II trials in cancer.

### Methods

To illustrate the use of the single-arm phase II futility study design using interim analyses, we obtained data provided by primary investigators from a convenience sample of completed phase III randomized stroke treatment trials. For illustrative purposes, we selected these studies to include some negative phase III trials. A positive phase III trial, and a negative phase III trial with marginal statistical significance. We had no prior knowledge of the conclusions that might be drawn from our simulated single-arm futility studies of these trials.

For each study example, Table 2 summarizes the projected and actual sample sizes, main inclusion criteria, favorable outcomes, types I and II error probabilities, hypothesized effect size $\Delta$, and the phase III trial results. Outcome data only in the active treatment arms of the phase III trials were evaluated in our hypothetical phase II study.

In our phase II futility studies, we chose a one-sided $\alpha$ of 0.10 because we wanted to keep required sample sizes small and were willing to tolerate a 10% chance of rejecting an effective treatment that could produce $\Delta$, the magnitude expected in the original phase III trial on which we based the phase II design. We were willing to accept a greater chance of carrying an ineffective treatment forward to phase III testing and set $\beta$ to 0.15 at $p_a = p^*$.

For each simulated single-arm phase II futility study, we used the EAST 3 (Cytel Software Corp.) software to estimate the required sample size and to generate the stopping criteria or the threshold required for taking the treatment to a phase III trial. We analyzed the data for futility after one third, two thirds, and all required patients had been enrolled using the O’Brien and Fleming6 boundaries.

In each futility study, we listed the outcomes for treated patients in the corresponding phase III trial in chronologic order until we achieved our required phase II sample size. We calculated the cumulative proportion of favorable outcomes at each of the 3 analysis stages. We compared the proportion of favorable outcomes at each stage to the proportion of favorable outcomes corresponding to the threshold for each stage. If the observed proportion of favorable outcomes was less than or equal to the threshold, we rejected the null hypothesis and concluded further testing of the treatment was futile.

### Results

In our simulated phase II single-arm studies, we demonstrated futility for 3 treatments, and all 3 were judged ineffective in the original phase III trials. The phase III trial of fosphenytoin for acute ischemic stroke began in 1995. Based on the prespecified $\Delta$ for favorable outcome at 3 months using the modified Rankin Scale (mRS) score, the required sample size was approximately 600 subjects (Fosphenytoin Study presented by W. Pulsinelli at the 1999 American Academy of Neurology Meeting). After 4 years, 462 patients had been entered; the study was prematurely terminated as a result of lack of efficacy (ordinal logistic regression analysis, $P=0.87$). For the single-arm phase II futility study, we defined favorable outcome as a dichotomized mRS score of 0 or 1 and used a $\Delta$ considered appropriate for the phase III trial (Poole RM, personal communication). Using this $\Delta$, our phase II single-arm study would have been terminated for futility at its first interim analysis after evaluating 19 patients (3% of the original projected sample size for its phase III trial and 4% of the number enrolled in before the phase III trial was abandoned; see Figure, a).

The Phase III ATLANTIS Part B trial of alteplase in acute ischemic stroke began in December 1993. Its primary favorable outcome was a National Institutes of Health Stroke Scale (NIHSS) score of 0 or 1 (indicating no neurologic functional deficit) 90 days after stroke symptom onset. ATLANTIS investigators sought a $\Delta$ of 9% in favorable outcomes, requiring a sample size of 968 patients.★★ The trial was terminated after enrolling 613 patients when the observed $\Delta$ was only 2% ($P=0.65$). Our phase II single-arm study would have been terminated for futility at its final analysis after evaluating 169 patients (18% of the required sample size for phase III and 28% of the number enrolled in phase III before the trial was abandoned; see Figure, b).

For the Phase III RANNTAS trial of tirilazad mesylate in 1993 to 1994, the favorable outcome was a combination of 2 functional measures (Barthel Index $\geq 60$ and Glasgow Out-
TABLE 2. Summary of Stroke Studies Evaluated

<table>
<thead>
<tr>
<th>Study Name (y)</th>
<th>Estimated No. Required for Original Trial (No. actually enrolled)</th>
<th>Main Inclusion Criteria</th>
<th>Parameter for a “Favorable” Outcome†</th>
<th>Parameters as Hypothesized in the Phase III Trial</th>
<th>Results as Observed in Phase III trial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fosphenytoin (1995 to 1999)</td>
<td>600 (462)</td>
<td>Drug within 4 hr of symptom onset</td>
<td>mRS=0/1</td>
<td>( \alpha = 0.05; \beta = 0.05 )</td>
<td>( p = 0.05 ) ( p = 0.65 )</td>
</tr>
<tr>
<td>ATLANTIS Part B9 (1993 to 1998)</td>
<td>968 (613)</td>
<td>Ages 18 to 79; drug within 3 to 5 hr of symptom onset; no ICH</td>
<td>NIHSS 1 or less</td>
<td>( \alpha = 0.05; \beta = 0.20 )</td>
<td>( p = 0.35 ) ( p = 0.44 )</td>
</tr>
<tr>
<td>RANTTAS10 (1993 to 1995)</td>
<td>1130 (660)</td>
<td>Age 18 or more; drug within 6 hr of symptom onset; no ICH</td>
<td>BI 60 or more and GOS=1/2 independently</td>
<td>( \alpha = 0.05; \beta = 0.20 )</td>
<td>( p = 0.66 ) ( p = 0.68 )</td>
</tr>
<tr>
<td>TOAST11,12 (1990 to 1998)</td>
<td>1300 (1281)</td>
<td>Age 18 to 85; drug within 1 to 24 hr of symptom onset; prestroke BI 12 or more; no ICH</td>
<td>GOS=1/2 and mBI 12 or more for each subject</td>
<td>( \alpha = 0.05; \beta = 0.10 )</td>
<td>( p = 0.50 ) ( p = 0.60 )</td>
</tr>
<tr>
<td>ECASS-II13 (1996 to 1998)</td>
<td>800 (800)</td>
<td>Age 18 to 80; drug within 6 hr of symptom onset; no ICH</td>
<td>mRS=0/1</td>
<td>( \alpha = 0.05; \beta = 0.20 )</td>
<td>( p = 0.30 ) ( p = 0.40 )</td>
</tr>
<tr>
<td>NINDS tPA Part 214 (1992 to 1994)</td>
<td>320 (333)</td>
<td>Drug within 3 hr of symptom onset; no ICH</td>
<td>Global statistics (GS) based on: NIHSS=0/1, mRS=0/1 BI 95 or more GOS=1</td>
<td>( \alpha = 0.05; \beta = 0.05 )</td>
<td>( p = 0.20 ) ( p = 0.40 )</td>
</tr>
</tbody>
</table>

†(m)BI indicates modified Barthel Index; GOS, Glasgow Outcome Scale.
‡P value for the ordinal logistic regression analysis using the raw mRS score.
§p* = Assumed control group proportions with a favorable outcome in the all randomized sample.
¶Assumed control group proportion with a favorable outcome for the dichotomized mRS. The trial was designed using assumptions about the mRS whereas the statistical analysis at the end of the trial was based on the global statistical test (GST).

Outcome Scale (mRS ≤ 2). To detect a Δ of 8% in their favorable outcome, the required sample size was 1130 patients. The trial was terminated for futility after randomizing 660 patients. Our phase II study would have been terminated for futility at its final analysis after evaluating 189 patients (17% of the sample size for phase III and 29% of the number enrolled in phase III before termination; see Figure, c).

Analyses of our single-arm phase II studies in the other 3 cases (one testing a heparinoid in TOAST11,12 and 2 testing alteplase in ECASS-II13 and NINDS tPA14 trials) indicated we could not declare these treatments futile to test these treatments in phase III (Figure, d through f). Two phase III trials (TOAST and ECASS-II) failed to demonstrate the hypothesized Δ in favorable outcomes, but the third trial demonstrated a worthwhile improvement across multiple rating scales.

The observed proportion of favorable outcomes from the futility studies were within 4% of the observed proportion of favorable outcome in the actively treated group in the respective phase III studies (p<0.05 in Table 2) with the exception of TOAST, in which the phase III study result was 6% higher than the futility study result.

Discussion
Our post hoc phase II analysis of the data from randomized stroke trials suggests a futility design may be useful for early and efficient identification of ineffective therapies with a minimal number of patients and at minimal cost. Phase II study results can help determine whether to abandon a treatment after phase II or continue to phase III. However, we stress that a single-arm phase II futility cannot by itself provide sufficient evidence to rule a promising treatment efficacious. Phase II studies are not designed or intended to test the efficacy hypothesis. As observed in the TOAST and ECASS II studies, failure to find futility in a single-arm phase II futility study would not guarantee a positive phase III trial. In general, evaluating the phase II futility design as one component of a multi-phase strategy allows for evaluation of alternative agents, treatment risks, the practicality of treatment administration on a large-scale trial, and funding availability.

One could take issue with the use of historic data as a reference for phase II futility studies. Temporal changes in other aspects of patient management, changing criteria for response assessment, variations in data quality, and variations...
in protocol adherence can distort estimates of the reference proportion ($p^*$). These same difficulties apply to hypothesizing the control group proportions in phase III trials as well, although in a phase III trial, one would still have a valid test of the hypothesis. In the phase II design, if $p^*$ is incorrect, one can erroneously conclude that an effective drug is futile or an ineffective drug is potentially effective. If the hypothesized $p^*$ is in doubt as a result of changes over time in the natural history or other concerns, it may be worthwhile to include a second small calibration group of placebo patients, not for making a direct comparison between groups, but to ascertain the validity of the hypothesized control group proportion. If the value observed in the calibration group is substantially different than the hypothesized $p^*$, the phase II study may need to be redone using the calibration group’s proportion as the new $p^*$. We do not recommend or advocate the use of historic controls in a definitive phase III or phase IIB study, but historical controls data could provide valuable information for determining $p^*$ for earlier stages of drug development such as the phase II futility study discussed here.

The observed $p_{\text{CTL}}$ proportions (shown in the last column of Table 2) in the 6 phase III trials we studied have a wide range of values (0.32 to 0.737). The heterogeneity of these observed values might be the result of the differences in the eligibility criteria (especially different upper age limits and different durations from symptom onset to treatment) and the timing
and specification of primary outcome measure of the studies (NIHSS score, mRS, Barthel Index, and Glasgow Outcome Scale).

The proposed futility evaluation approach does not directly address toxicity, and toxicity may determine the feasibility of a phase III trial. Usually, toxicity is more directly addressed in the design of phase I trials. Thall et al. have proposed an approach combining phase I and II trials, and their approach could be considered in future studies of new agents in which phase I data are not available.

Finally, in planning a single-arm futility study, the choice of p*, Δ, or pα, and α and β has a large impact on the determination of futility. In general, p* and Δ or pα may be estimated based on the values used to estimate sample size for a phase III trial (as we have done in our exercise). The errors (α and β) should be chosen based on the investigators’ level of comfort with the risk of having false-negative or false-positive conclusions from the futility study. For example, an investigator may be willing to risk a higher type II error for a treatment that may prevent disability from hemorrhagic stroke in which no known cure exists to date, whereas a lower type II error may be required for an expensive, invasive, risky procedure for mild ischemic stroke in which an alternative treatment exists. In our exercise, we chose to use Δ that were clinically meaningful, the same measures as used in the phase III trials from which we received the data. When designing a phase II futility study, investigators should choose a value as close as possible to the Δ they would use in the future phase III trial to provide a reasonable test of the futility hypothesis.

The single-arm phase II futility study design approach has been used recently in an NINDS-funded trial of intravenous and intraarterial tPA treatment for ischemic stroke. Results of that trial have been published. The trial used the data on the placebo arm of the NINDS tPA study to obtain p*. More recently, the phase II design has been used in studies of patients with Parkinson disease.

In summary, we have adapted the single-arm phase II futility study design commonly used in oncology to the evaluation of therapeutic agents in stroke. We found that single-arm phase II futility studies could help investigators avoid 3 large, expensive phase III randomized trials of treatments for ischemic stroke. Based on the reduction in sample size, this phase II strategy could permit the testing of a wider array of promising treatments at a fraction of the cost of taking all treatments directly to phase III trials.

Acknowledgments

Funding for the project was provided in part by KAI N01-NS-4-2320. The TOAST Study was funded by the National Institute of Neurological Disorders and Stroke (NINDS) grants R01-1-NS-27863 and R01-NS-27960. Additional support for TOAST Study, including a supply of the study drug, was given by Organon Inc. The NINDS t-PA Study data were collected under #N01-NS-2-2343. A supply of study drug was provided by Genentech. Dr. Johnston is supported by the NINDS grant K23NS02168. The authors thank Drs. R. Michael Poole and Manfred Wilhelm for providing access to the Fosphenytoin, ATLANTIS Part B, and ECASS II datasets, as well as the TOAST, RANTTAS, and the NINDS t-PA Study Investigators for the use of their respective datasets. Finally, the authors greatly appreciate the insightful comments and suggestions provided by the reviewers of the manuscript.

References
