Proof-of-Principle Phase II MRI Studies in Stroke
Sample Size Estimates From Dichotomous and Continuous Data

MR Stroke Collaborative Group*

Background and Purpose—Since the failure of a number of phase III trials of neuroprotection in ischemic stroke, the need for smaller phase II studies with MRI surrogates has emerged. There is, however, little information available about sample size requirements for such phase II trials and rarely enough patients in single studies to make robust estimates. We have formed an international collaborative group to assemble larger datasets and from these have generated sample size tables for MRI-based infarct expansion as the outcome measure.

Methods—Twelve centers from Australia, Europe, and North America contributed data from patients with hemispheric ischemic stroke. Infarct expansion was defined from initial diffusion-weighted images and later fluid-attenuated inversion recover or T_2 images. Sample size estimates were calculated from data on infarct expansion ratios treated as dichotomous or continuous variables. A nonparametric approach was used because the distribution of infarct expansion was resistant to all forms of transformation.

Results—As an example, a 20% absolute reduction in infarct expansion ratio (≤ 1), 80% power, and $\alpha = 0.05$ requires 99 patients in each arm. To achieve an equivalent effect size with a continuous approach requires 61 patients.

Conclusions—These tables will be useful in planning phase II trials of therapy with the use of MRI outcome measures. For positive studies, biologically plausible surrogates such as these may provide a rationale for proceeding to phase III trials. (*Stroke*. 2006;37:2521-2525.)

Key Words: magnetic resonance imaging ■ neuroprotection ■ sample size ■ stroke

Although stroke is the second largest contributor to global mortality, there are only limited numbers of interventions of proven benefit for its prevention or acute treatment.[1] Neuroprotectants are attractive therapeutic candidates because of their low toxicity, and they could be administered while the patient is en route to the hospital. However, despite convincing evidence of neuroprotection in animal models of cerebral ischemia, such agents have failed to live up to expectations in human phase III clinical trials.[2] Indeed, these failures have led investigators to reconsider the role of phase II trials from testing dosage and safety alone to providing a signal of efficacy with surrogate MRI outcome measures.[3–8] Such "proof-of-concept" studies are perceived to be appealing if they can be performed with small numbers of patients.

The surrogate measure for improved stroke outcome that seems biologically plausible and that has gained acceptance is the attenuation of infarct expansion from its initial volume on diffusion-weighted images (DWI) to final infarct volume, usually defined by T_2-weighted images. This expansion can be defined as either the ratio of final to initial ischemic volume or final minus initial volume. This surrogate approach has been successfully applied in animal models to assess the efficacy of therapy.[2] In humans, infarct expansion is correlated with poor clinical outcome, and expansion is attenuated with recombinant tissue plasminogen activator (rtPA) therapy, the only agent of proven benefit.[5,9–11] Furthermore, it is likely that the surrogate is more sensitive than are clinical outcomes, because infarct expansion has been shown to be significantly attenuated while there was only a trend toward improved clinical outcomes.[10] Clearly, provision of a signal of efficacy by this means would provide investigators with the reassurance that therapeutic translation from animal models to humans would be more likely. This would provide better justification for a phase III trial.[4,7]

In estimating sample size requirements with MRI measures, one of the main problems has been to accrue enough untreated patients. To address this issue, we have formed an international collaborative group and have pooled data from a large number of centers. We then generated sample size estimates for the MR surrogate measures.

Methods
Inclusion and Exclusion Criteria
Centers from Australia, Europe, and North America were invited to contribute data on patients with the following inclusion criteria:

---

onset of hemispheric ischemic stroke within 24 hours, adequate initial DWI images, outcome images (fluid-attenuated inversion recovery or T_2) within 1 week to 3 months, and availability of clinical outcome scores at the time of imaging. Patients were excluded on the basis of current recommendations, such as (1) initial ischemic tissue was <5 mL,[3] (2) initial ischemic tissue was outside the middle cerebral artery territory, (3) follow-up images or clinical data were not available, or (4) patients had received rtPA.[3] Patients were not excluded if they had received neuroprotectants from negative, randomized, controlled trials (on clinical end points). Clinical data collected included the initial and outcome National Institutes of Health Stroke Scale (NIHSS) scores and outcome modified Rankin Scale (mRS) scores. The initial NIHSS score was obtained at the time of the initial MRI and the mRS score, at 3 months. Data on outcome NIHSS scores at 3 months were available for a small number of patients.

Definitions

Initial ischemic volume was defined as tissue encompassed by the initial DWI volume (bright signal on DWI and low signal on apparent diffusion coefficient maps). Outcome infarct volume was defined as tissue encompassed by the T_2-weighted sequence at 1 week to 3 months (bright signal on T_2-weighted images in locations corresponding to the initial ischemic tissue). In all cases, the outcome infarct volumes included the volume of hemorrhagic transformation. The infarct expansion ratio (IER) was defined as the ratio of final infarct to initial ischemic tissue volumes. For example, if the final infarct volume was 60 mL and the initial ischemic volume was 30 mL, then the IER was 2.0. Expanding infarct was defined as an infarct with an IER >1 (final infarct volume greater than initial ischemic volume). ΔProportion was defined as the difference between the proportion (as a percentage) of expanding infarcts in the control and treatment groups.

Statistical Analysis

The data were tested for heterogeneity of the log IER between centers by ANOVA. We tested for an association between poor clinical outcome (defined as an mRS score >2) and infarct expansion, either as a continuous (logarithmically transformed) or a dichotomized variable into expanding and nonexpanding infarcts by logistic regression. Additionally, the receiver operating characteristic curve was used to display the global performance of infarct expansion for predicting poor clinical outcome.[12] The area under the receiver operating characteristic curve is a measure of how informative infarct expansion is for this purpose.

Sample sizes for were calculated for 2 different hypothetical analysis methods (dichotomous and continuous methods). The first was to divide patients into those with expanding infarcts and those without. The proportion of expanding infarcts would then be compared between study arms. Sample sizes for this method were calculated according to standard methods implemented in the Stata statistical package with our data to estimate the expected proportion of expanding infarcts in the control group for various values of Δproportion.

The second method used the Wilcoxon rank-sum test to compare all expansion ratios between study arms. For this purpose, it was necessary to make assumptions regarding both the distribution of expansion ratios and the detailed effect of treatment. We assumed a simple model of treatment: that the IER would be attenuated by a fixed factor between 0.5 and 0.95 relative to what would be observed without treatment. This was assumed to be independent of initial ischemic volume or other factors. To simulate the distribution of infarct values, we used a bootstrap approach, sampling with replacement from our data. At least 4000 bootstrap replicates were used, sufficient to obtain an accuracy of ±1% of power (95% prediction interval). In addition to sample size, we estimated the parameter $P_{noether}$, the probability that a randomly selected control patient would have a greater IER than a random treatment patient (see Walters and Campbell[13] for details). We also estimated the corresponding value of Δproportion to allow comparison of sample size requirements between methods and the odds ratio (OR) for expanding infarcts as

a familiar measure of effect size. For both methods of analysis, power was varied between 80% and 90%, and the time window from symptom onset to MRI varied from ≤3 to ≤24 hours. The type I error rate, α, was set at 0.05; tests were assumed to be 2 sided.

Results

Twelve centers from Australia, Europe, and North America submitted volume and clinical data for 259 patients with ischemic stroke accrued during a 3-year period. There were 121 men, with a mean and median age of 69 and 70 years, respectively (range, 20 to 93 years). Seventy patients who had an acute DWI volume of ≤5 mL were excluded.[3] This resulted in 189 remaining patients available for analysis with a mean±SD age of 69±11 years, 48.6% of whom were men. There were 39 patients who had their initial MRI performed within 3 hours (20.6%), 118 within 6 hours (62.4%), 171 within 12 hours (90.5%), and all 189 within 24 hours (100%). The mean±SD and median values for acute DWI volumes were 42±48 mL and 21 mL; for outcome, T_2 volumes were 83±75 mL and 60 mL; and for IER, the values were 3.25±4.44 and 1.59, respectively. The proportions of subjects with an mRS score ≤2 at 3 months were 49% for the ≤3-hour group, 42% for the ≤6-hour group, and 50% for the ≤12-hour group. In the interest of brevity, we discuss the results in terms of the 6-hour group only because this group represents the majority of the data. Sample size calculations for the 3-hour and 12-hour time windows are provided as supplementary data (supplemental Tables II and III, available online at http://stroke.ahajournals.org).

For the 6-hour window, there was no evidence of overall heterogeneity by site for initial volumes <5 mL either included (*P*=0.64) or excluded (*P*=0.62). Moreover, no individual site was significantly different from the others in these cases (minimum probability values of 0.40 and 0.10, respectively), which is a more stringent test of homogeneity owing to the 10 comparisons made.

Relation Between Surrogate Measures and Clinical Outcome

There was a statistically significant relation between IER and poor clinical outcome, as defined by an mRS score >2 (*P*<0.001; OR, 3.36, 95% CI, 1.86 to 6.07). This means that for every log unit increase in IER, there was a 3.36 times increase in the odds of a poor clinical outcome. The association remained when infarct expansion was dichotomized (*P*=0.001; OR, 4.75; 95% CI, 1.93 to 11.7). The area under the receiver operating characteristic curve for IER in predicting poor clinical outcome was 0.741 (95% CI, 0.652 to 0.830).

Sample Size for Dichotomized Data

We assume that our data, from patients who were either not treated with neuroprotective agents or were enrolled in clinical trials with negative clinical outcomes, are representative of untreated patients. Hence, we posit that the proportion of control IER ≤1 will be 25.4%. The sample size estimate for an absolute therapeutic effect size of 20% (so that the proportion with an IER ≤1 among those actively treated is 45.4%), 80% power, and α=0.05 (2 sided) was 99 patients in each arm (see Table 1 for sample size estimates at other therapeutic effect sizes). This effect size of 20% was chosen

TABLE 1.   Sample Size Estimates

ΔProportion	Sample Size	
Absolute ΔProportion With Expanding Infarct, %	80% Power, $\alpha=0.05$, 2 Tailed (Each Arm)	90% Power, $\alpha=0.05$ (Each Arm)
2.5	4989	6651
5	1302	1729
7.5	602	796
10	351	463
15	166	218
20	99	128
25	66	85
30	48	61
35	36	46
40	28	36

Values are the number of patients in each arm for the dichotomous method with use of the difference in proportions of patients with no infarct expansion and good clinical outcome (mRS score ≤2; 6-hour time window, n=118). The proportion of subjects (from natural history data) with an infarct nonexpansion ratio ≤1 is 25.4%.

as a conservative figure, given the ≈50% reduction shown in a recent thrombolytic trial.[14]

Sample Size for Continuous Data

The sample size estimates for analyzing IER as continuous data by the Wilcoxon rank-sum test were smaller than for the corresponding tests of differences of proportions, by ≈30%. If one assumes that treatment has a 35% effect on IER (equivalent to an ≈19% difference in proportions; OR, 1.40; $P_{noether}=0.65$), 61 patients per arm would be sufficient to achieve 80% power with $\alpha=0.05$ (2 sided; Table 2). The sample size estimate was similar when the infarct volume difference (final minus initial ischemic volume) was used as a surrogate instead of the IER (data not shown). We have analyzed the data including those patients with initial an DWI

volume of <5 mL and found no difference (Supplemental Table I, available online at http://stroke.ahajournals.org).

Discussion

This study represented pooled data from many centers to address the important question of the sample size required to perform a phase II proof-of-concept study with an MRI surrogate outcome measure. We approached this question by determining an appropriate type of MRI surrogate that would best measure the biological effects of the drug. Second, we used the statistically significant relation between infarct expansion and clinical outcome to support our proposal that this measurement is a candidate surrogate. Third, we considered the distribution of observed IER and infarct volume differences and examined various methods to calculate sample size. Fourth, we generated a practical set of sample size tables with variations in expected absolute therapeutic effect and power and noted that the calculation with the dichotomous data resulted in slightly smaller sample sizes.

We have used the term "surrogate marker" rather than "biomarker" because the former has gained currency for these MRI outcome measures and the latter has become commonly associated with serum markers altered by the stroke process. We recognize that the term "surrogate measure" requires adherence to a number of criteria that cannot, as yet, be completely fulfilled with infarct expansion as a measure. For example, some agents may impact positively on clinical outcome not necessarily due to an effect on the surrogate measure. Much of the early use of surrogates was in oncology, where clinical outcome was difficult to observe because of its infrequency or long duration.[15,16] With accelerated drug development in mind, a set of criteria for surrogacy was developed that included several key elements: (1) use of the surrogate is biologically plausible, (2) a statistical relation can be established between the surrogate and true outcome measures, and the therapeutic response is valid for both (3) true and (4) surrogate outcome measures.[17] For our MRI surrogate, criteria 1 and 2 may be fulfilled without difficulty

TABLE 2.   Sample Size Estimates

Infarct Growth Attenuation Factor*	$P_{noether}$†	Absolute Change in Proportion With Expanding Infarct, %	OR for Expansion With Treatment	80% Power, $\alpha=0.05$ (Each Arm)	90% Power, $\alpha=0.05$ (Each Arm)
0.95	0.523	3.4	0.76	2650	3450
0.90	0.542	8.5	0.76	720	960
0.85	0.563	11.0	0.76	325	430
0.80	0.583	11.8	0.75	195	250
0.75	0.603	15.2	0.74	120	160
0.70	0.623	17.0	0.72	84	115
0.65	0.645	19.4	0.71	61	79
0.60	0.673	23.7	0.70	44	58
0.55	0.698	30.0	0.70	32	42
0.50	0.725	33.9	0.70	24	31

Values are the number of patients in each arm with the use of continuous data for IERs and the nonparametric difference in ranks (Wilcoxon rank-sum test) based on data from 118 patients with a DWI volume ≥5 mL seen within a 6-hour time window.

*We assume that use of a hypothetical neuroprotectant drug results in an infarct ratio smaller by this factor compared with placebo.

†$P_{noether}$ is the probability that a randomly selected control patient will have a greater IER than a random treatment patient.

but not criteria 3 and 4. In the case of acute stroke phase II trials of neuroprotection, there is a good argument to consider the first 2 criteria only, given the current absence of an agent of proven benefit in phase III clinical trials. Should a proven agent become available, then this stance would need to be reviewed.[15,16]

We performed sample size calculations for infarct expansion by both dichotomous and continuous methods to determine which would be more useful. The advantage of using the continuous approach over that of the dichotomous method is that the whole range of data can be used. The sample size table is intended as a guide to planning phase II trials, because the sample size estimates by previous investigators for studies of neuroprotection were probably unrealistic, as they were based on large therapeutic effect sizes.[4,5] However, as mentioned earlier, these large effect sizes may be reasonably expected in trials of reperfusion with agents such as rtPA. The aggregation of large, imaging datasets enables reasonably precise estimates of sample sizes to be made as a framework for treatment effects. Because of the difficulties in accruing patients for such studies, collaborations such as the MR Stroke group are essential.

In this study, we were unable to address the issue as to whether the use of infarct expansion as a surrogate will allow investigators to use significantly smaller sample sizes than when using clinical outcome measures, because the therapeutic effects of putative agents on the surrogate are unknown. Whether investigators find the surrogate approach to be a more convenient proof-of-concept technique than clinical trials remains to be seen. It seems likely that the numbers required when infarct expansion is used as a surrogate will be significantly smaller for a number of reasons. First, we have established that there is a relation between infarct expansion and clinical outcomes, with an OR of 3.36, although we regard this as approximate, because we were unable to adjust for covariates such as baseline NIHSS, age, and diabetes. A more precise relation will be established in a separate publication. Second, there are a number of examples where similar MRI surrogates with reasonably small numbers have been used with positive results while clinical outcomes have been negative. Specifically, in our earlier study of infarct attenuation with thrombolysis (tPA), clinical outcome measures for the whole group (unlike the mismatch group, which showed positive clinical outcomes) were not significantly different, but surrogate outcomes were.[10] In the DIAS study with desmoteplase as the thrombolytic agent, in the overall group there was significant improvement in the reperfusion rate (49.3% versus 19.2%, $P=0.0054$) and a smaller nonsignificant effect on favorable outcome (22.2% versus 38.7%, $P=0.0640$).[14] In all of these examples, the influences of the agents tested on the surrogates were sufficiently biologically plausible to lead the investigators to progress to phase III studies.

Our study has a number of limitations. First, we used pooled infarct volume measurements, which had been quantified by different techniques from a number of institutions, and some patients were involved in failed trials of neuroprotection. For the latter, we believe that this would have little impact on the observed IERs. For the former, it would be ideal to prospectively collect data with the same imaging protocol and analyzed at 1 center. A prospective study of similar size to perform a similar feat would take several more years to perform and require funding. It is comforting to note that despite the use of data from a number of centers, there was no significant heterogeneity in IER between centers. Second, we used a broad time window and differing MRI sequences to define outcome infarct volume. Although there is no currently accepted definition of the optimal time for measuring outcome infarct on either DWI or T_2-weighted sequences, these times are currently recommended to minimize the influence of either edema or atrophy.[18] An important issue for the MR Stroke group will be to provide consensus on the appropriate MRI sequences and timing of such so that future data can be easily compared. Third, we did not include reperfusion as an influence on infarct expansion. It is well established that this is a confounding factor, and it seems logical that sample sizes might be further reduced by considering only those with an initial perfusion-weighted imaging/DWI mismatch.[19] Given that methods of calculating perfusion maps are not standardized at present, the results of perfusion-weighted imaging/DWI mismatch from different centers are not comparable. We are exploring the possibility of combining raw MRI data in an electronic medium to allow reanalysis by common methods. Furthermore, the sample size estimates may need to be increased by up to 20% if the effect of patients lost to follow-up is not taken into account.[10] Finally, these tables address the sample size required for inclusion in the trial, rather than the number needed to recruit before screening inclusion and exclusion criteria. Despite the increasing availability of MRI scanners, the number of centers that can perform MRI studies in acute stroke is small. For the purpose of a proof-of-concept study, this small number is not unreasonable.

In summary, we have provided sample size tables for infarct expansion on MRI as a surrogate for trials of therapy in acute ischemic stroke. The use of a biologically plausible surrogate such as this in a positive phase II study may provide investigators with an adequate rationale to proceed to phase III studies.

Appendix

*MR Stroke Group
Cochairs: Geoffrey A. Donnan, Stephen M. Davis.
Coordinator: Thanh G. Phan.
Statistical analysis: John Ludbrook, Graham Byrnes, Thanh G. Phan.
Writing group: Thanh G. Phan, Geoffrey A. Donnan, Stephen M. Davis, Graham Byrnes.

Contributing Members
Australia: Royal Melbourne Hospital (Mark Parsons, Alan P. Barber, Stephen M. Davis), Austin Health (Geoff Donnan, Thanh G. Phan, David C. Reutens), Royal Brisbane Hospital (Stephen E. Rose, Jonathan Chalk).
Canada: Foot Hills Hospital (Andrew M. Demchuk, Shelagh B. Coutts, Jessica E. Simon, Anna Tomanek).
Germany: University Hospital, Hamburg Eppendorf (Joachim Roether, Cornelius Weiller, Jens Fiehler, Gotz Thomalla, Thomas Kucinski), Heidelberg (Peter D. Schellinger, Werne Hacke), Mannheim

(Achim Gass, Kristina Szabo, Michael Hennerici), Düsseldorf (Mario Siebler), and Berlin Charite (Arno Villringer, G.J. Junge-Hülsing).

Spain: Hospital Universitari Doctor Josep Trueta (Salvador Pedraza, Antoni Dávalos), Hospital Clnico Universitario (Jose Castillo).

United States: Stanford University Medical Center (Gregory W. Albers, Maarten G. Lansberg, Vincent N. Thijs, Roland Bammer, Michael E. Moseley, Michael Marks).

Other Members
Steve Warach, Alison Baird, Chelsea Kidwell, Jeff Saver, Greg Sorensen, Marc Fisher (United States), Norbert Nighoghossian (France), Keith Muir (UK).

References

1. Adams HP Jr, Adams RJ, Brott T, del Zoppo GJ, Furlan A, Goldstein LB, Grubb RL, Higashida R, Kidwell C, Kwiatkowski TG, Marler JR, Hademenos GJ. Guidelines for the early management of patients with ischemic stroke: a scientific statement from the Stroke Council of the American Stroke Association. *Stroke*. 2003;34:1056–1083.
2. Recommendations for standards regarding preclinical neuroprotective and restorative drug development: Stroke Therapy Academic Industry Roundtable. *Stroke*. 1999;30:2752–2758.
3. Warach S. New imaging strategies for patient selection for thrombolytic and neuroprotective therapies. *Neurology*. 2001;57:S48–S52.
4. Warach S, Pettigrew LC, Dashe JF, Pullicino P, Lefkowitz DM, Sabounjian L, Harnett K, Schwiderski U, Gammans R; Citicoline 010 Investigators. Effect of citicoline on ischemic lesions as measured by diffusion-weighted magnetic resonance imaging. *Ann Neurol*. 2000;48:713–722.
5. Barber PA, Parsons MW, Desmond PM, Bennett DA, Donnan GA, Tress BM, Davis SM. The use of PWI and DWI measures in the design of 'proof-of-concept' stroke trials. *J Neuroimaging*. 2004;14:123–132.
6. Baird AE, Lovblad KO, Dashe JF, Connor A, Burzynski C, Schlaug G, Straroselskaya I, Edelman RR, Warach S. Clinical correlations of diffusion and perfusion lesion volumes in acute ischemic stroke. *Cerebrovasc Dis*. 2000;10:441–448.
7. Davis SM, Donnan GA. Neuroprotection: establishing proof of concept in human stroke. *Stroke*. 2002;33:309–310.
8. Fisher M. Recommendations for advancing development of acute stroke therapies: Stroke Therapy Academic Industry Roundtable 3. *Stroke*. 2003; 34:1539–1546.
9. Arenillas JF, Rovira A, Molina CA, Grive E, Montaner J, Alvarez-Sabin J. Prediction of early neurological deterioration using diffusion- and perfusion-weighted imaging in hyperacute middle cerebral artery ischemic stroke. *Stroke*. 2002;33:2197–2203.
10. Parsons MW, Barber PA, Chalk J, Darby DG, Rose S, Desmond PM, Gerraty RP, Tress BM, Wright PM, Donnan GA, Davis SM. Diffusion- and perfusion-weighted MRI response to thrombolysis in stroke. *Ann Neurol*. 2002;51:28–37.
11. Rother J, Schellinger PD, Gass A, Siebler M, Villringer A, Fiebach JB, Fiehler J, Jansen O, Kucinski T, Schoder V, Szabo K, Junge-Hulsing GJ, Hennerici M, Zeumer H, Sartor K, Weiller C, Hacke W. Effect of intravenous thrombolysis on MRI parameters and functional outcome in acute stroke <6 hours. *Stroke*. 2002;33:2438–2445.
12. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.
13. Walters SJ, Campbell MJ. The use of bootstrap methods for estimating sample size and analysing health-related quality of life outcomes. *Stat Med*. 2005;24:1075–1102.
14. Hacke W, Albers G, Al-Rawi Y, Bogousslavsky J, Davalos A, Eliasziw M, Fischer M, Furlan A, Kaste M, Lees KR, Soehngen M, Warach S. The Desmoteplase in Acute Ischemic Stroke (DIAS) Trial: a phase II MRI-based 9-hour window acute stroke thrombolysis trial with intravenous desmoteplase. *Stroke*. 2005;36:66–73.
15. Boissel JP, Collet JP, Moleur P, Haugh M. Surrogate endpoints: a basis for a rational approach. *Eur J Pharmacol*. 1992;43:235–244.
16. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*. 1998;54:1014–1029.
17. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*. 1989;8:431–440.
18. Donnan GA, Davis SM. Neuroimaging, the ischaemic penumbra, and selection of patients for acute stroke therapy. *Lancet Neurol*. 2002;1:417–425.
19. Thijs VN, Somford DM, Bammer R, Robberecht W, Moseley ME, Albers GW. Influence of arterial input function on hypoperfusion volumes measured with perfusion-weighted imaging. *Stroke*. 2004;35:94–98.