Outcome measures for clinical trials in neurotrauma

M. ROSS BULLOCK, M.D, PH.D., RANDALL E. MERCHANT, PH.D., SUNG C. CHOI, PH.D., CHARLOTTE B. GILMAN, B.S.N., JEFFREY S. KREUTZER, PH.D., ANTHONY MARMAROU, PH.D., AND GRAHAM M. TEASDALE, F.R.C.P., F.R.C.S.

Division of Neurosurgery Medical College of Virginia, Richmond, Virginia

Under the auspices of the American Brain Injury Consortium and the Joint Section of Neurotrauma and Critical Care of the American Association of Neurological Surgeons, the authors have reviewed and formulated opinions based on the evidence on protocol design and the outcome measures used for clinical trials in patients with a severe or moderate traumatic brain injury (TBI). First, in view of the heterogeneity of the population under study, the authors suggest that block randomization and stratification should always be used in the design of neurotrauma trials. Second, although the Glasgow Outcome Scale (GOS) remains the most widely used and accepted instrument for TBI trials, the authors believe the eight-point expanded scale that has recently been designed will ultimately provide greater discrimination, and narrower categories and will ultimately prove superior for detecting more subtle changes in outcome. Furthermore, the authors recommend, in view of the profound cognitive impairment in survivors of TBI, that neuropsychological tests be explored further as an adjunct to the GOS. Future research should focus on the development of more sensitive and specific surrogate outcome measures such as magnetic resonance imaging, neurochemical, neuropsychological, and quality of life measures in order to detect a neuroprotective effect in patients with TBI.

KEY WORDS • outcome measures • neurotrauma • Glasgow Outcome Scale

More than 40 years have passed since the first therapeutic trials were designed to test the ability of medications to improve outcome after severe TBI. To date, at least 14 large Phase III and extended Phase II trials have been performed, and few of these have yielded positive results (Table 1). Unfortunately, many of these large and expensive trials have been terminated prematurely; based on futility analyses and as a result of early termination, they have failed to provide any conclusive result, either positive or negative (Tables 1 and 2).

Over the last 10 years, there have been tremendous advances in laboratory modeling of the different aspects of TBI, and there are now more than 100 studies yielding positive results using different TBI models with various putative neuroprotective agents, many of which, nonetheless, have failed to be effective in clinical trials. During the same period, the authors of several clinical trials have reported positive results for other acute neurological indications, such as spinal cord injury and stroke.13

In 10 of the most recent Phase III TBI trials, the designs and outcome measures have been remarkably similar, relying heavily on the dichotomized GOS performed at either 3 or 6 months as the primary outcome measure. Most TBI trials have also focused on the same patient group: those with an enrollment GCS score of 4 to 8, preserved evidence of brainstem function, and in many cases, a CT-documented lesion. This homogeneity of trial design and the consistent failures cause us to question several issues concerning the trial design: are we, for instance, simply repeating the same mistakes (Table 3)?

A number of factors may be responsible for the failure of these neurotrauma Phase III trials to demonstrate clear evidence of efficacy: 1) the compound or the dosage of the compound may be ineffective; 2) the test compound may fail to reach the receptor or other site of action in the brain; 3) the compound may be administered at the wrong time or dosage to mediate the pathophysiological process it was designed to affect; 4) the trial population group may be inappropriate for demonstrating the maximum effect of the agent; and 5) the outcome measures used to determine drug effect may lack the sensitivity to detect a change.

Many attempts have been made by neurosurgical organizations, in the US and Europe, to address this absence of compound-related effectiveness in clinical trials. Con-
sortia of regulated and enthusiastic academic and nonacademic head injury centers have been organized both in the US (the ABIC) and in Europe (European Brain Injury Consortium). These consortia, together with recently published guidelines to standardize head injury management, have helped achieve a high degree of uniformity and excellence in the care of severely head-injured patients.6,28,39

The purpose of this report, which was formulated under the auspices of ABIC and the Joint Section of Neurotrauma and Critical Care of the American Association of Neurological Surgeons, is to review the evidence and current expert opinion that exist in the area of outcome measures for clinical trials to treat a severe or moderate head injury.

REVIEW OF OUTCOME MEASURES IN CLINICAL TRIALS

Characteristics of Phase III Prospective Randomized Placebo-Controlled Trials

Prospective randomized placebo-controlled trials are undoubtedly the most powerful tool for determining the effect of new therapeutic interventions in clinical care. Compared with other indications, neurotrauma PRPCTs have historically been small to moderate in size, but they have generally been more difficult and expensive to conduct because of the population being studied. This is because of the high level of intensive care that these patients require, the heterogeneous nature of their injuries, and the substantially large amount of intercenter variability in patient populations and management seen in neurotrauma PRPCTs. Although randomization strategies are designed to reduce the effect of these confounding factors, simple randomization has failed to allow for the matching of study groups adequately in at least one recent trial, the Tirilazad Severe Head Injury European Trial.

Block Randomization

Strategies such as block randomization have been remarkably effective recently. For example, the recent Cambridge Neuroscience Phase III PRPCT in which investigators used the noncompetitive glutamate antagonist, Cerestat, demonstrated enormous variability in mortality rates among centers because some had enrolled very few patients with a severe TBI. Nevertheless, the use of a block randomization strategy achieved almost a perfect balance in terms of demographic features, CT-documented injury, and some clinical variables in the two arms of the trial. Stratification based on certain clinical criteria at the time of enrollment, such as the motor score of the GCS, will help to balance the confounding variables between the drug-treated and control groups included in a trial with an enrollment of 1000 patients. Strategies such as block randomization and stratification should, therefore, always be used in the design of neurotrauma trials in view of the heterogeneity of the population being studied.

Stratification of the Population

Theoretical statistical considerations suggest that stratification at enrollment is highly desirable to balance the distribution of clinical characteristics in patients with TBI. In small trials, the impact of stratification may be greatest because the number of patients per treatment arm will be small and the consequences of imbalance will be greater. Stratification does not require any increase in sample size and, in terms of continuous outcomes, may reduce the

<table>
<thead>
<tr>
<th>Authors &amp; Year</th>
<th>Agent</th>
<th>No. of Patients, Country</th>
<th>Outcome/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ward, et al., 1950</td>
<td>atropine</td>
<td>940, US</td>
<td>uncontrolled, no benefit</td>
</tr>
<tr>
<td>Schwartz, et al., 1985</td>
<td>mannitol vs pentobarbital</td>
<td>59, Canada</td>
<td>randomized crossover permitted, mannitol group had better outcome</td>
</tr>
<tr>
<td>Eisenberg, et al., 1988</td>
<td>barbiturates (therapeutic)</td>
<td>73, US</td>
<td></td>
</tr>
<tr>
<td>Teasdale, et al., 1992</td>
<td>nimodipine (HIT I)</td>
<td>255, UK &amp; Finland</td>
<td>DBPRPCT, no benefit</td>
</tr>
<tr>
<td>Muizelaar, et al., 1993</td>
<td>PEG-SOD free radical scavenger (Phase II)</td>
<td>94, US</td>
<td>DBPRPCT, ICP lower, outcome better (p &lt;0.01)</td>
</tr>
<tr>
<td>Wolf, et al., 1993</td>
<td>troxethamine</td>
<td>149, US</td>
<td></td>
</tr>
<tr>
<td>Muizelaar, et al., 1994</td>
<td>PEG-SOD (Phase III), 3 dose levels</td>
<td>463, US</td>
<td>DBPRPCT, outcome 9% better in drug vs. placebo (p = 0.15)</td>
</tr>
<tr>
<td>Clifton, 1995</td>
<td>moderate hypothermia</td>
<td>~360, US</td>
<td>DBPRPCT, terminated after interim analysis; analysis awaited</td>
</tr>
<tr>
<td>Harders, et al., 1996</td>
<td>nimodipine (L-channel antagonist) Phase II</td>
<td>123, Germany (only traumatic SAH)</td>
<td>DBPRPCT, 55% relative reduction in bad outcome at 6 mos (p &lt;0.002)</td>
</tr>
<tr>
<td>Bolland, et al., 1998</td>
<td>Synthaleno eliprolid, (SL 82) Phase II</td>
<td>453, France</td>
<td>DBPRPCT, better outcome in “brain swelling” patients (p &lt;0.01)</td>
</tr>
<tr>
<td>Marshall, et al., 1998</td>
<td>tirilizad</td>
<td>1128, EU &amp; Australia</td>
<td>DBPRPCT, no benefit (reports awaited)</td>
</tr>
<tr>
<td>Marmarou, et al., 1999</td>
<td>Bradycor (bradykinin receptor antagonist)</td>
<td>133, US</td>
<td>DBPRPCT, no benefit but 10% trend toward better outcome</td>
</tr>
<tr>
<td>Morris, et al., 1999</td>
<td>Selfitol (CGS-19755) Phase III</td>
<td>266, US &amp; Israel, 426, EU &amp; Australia</td>
<td>DBPRPCT, both terminated due to excess mortality in concomitant stroke trials; no benefit (report awaited)</td>
</tr>
<tr>
<td>Stewart, et al., 1999</td>
<td>CGS 19755 (glutamate NMDA antagonist)</td>
<td>113, US &amp; UK</td>
<td>DBPRPCT, ICP lower</td>
</tr>
</tbody>
</table>

* DBPRPCT = double-blind, PRPCT; EU = Europe; NMDA = N-methyl-D-aspartate; UK = United Kingdom.
sample size needed for a specified treatment effect and power. It may be that the minimum number of patients in any single stratum should not be fewer than about 30 (15/treatment arm) for a dichotomous outcome, or approximately 10 for a continuous outcome measure such as the DRS, although such a minimum has not yet been clearly defined.

In many recent trials, the motor score component of the GCS has been used as a stratification variable, and this has worked extremely well in all the trials. Stratification of up to three variables may be used in a trial with more than 1000 patients. Currently, stratification based on CT classification has been proposed based on the six-point TCDB classification, first described in 1991.29 This would ensure balance between the placebo- and drug-treated groups for such variables as acute subdural hematoma, which is known to worsen outcome, or diffuse injury, in which mortality rates are as low as 6 to 15%.

Recently, it has been proposed that stratification should be considered based on the probability of neurological deterioration. In the current management of head injury, approximately 10 to 15% of cases will deteriorate by two or more scores on the GCS at some time between injury and the 14th day. Marshall and colleagues29 have proposed that such “neuroworsening” increased the mortality rate by fivefold. Six predictive factors define the population at risk for neuroworsening: ICP above 20 mm Hg, age greater than 40 years, CT scan TCDB classification of 5 or 6, evidence of subarachnoid or intraparenchymal bleeding, and effacement of basal cisterns observed on CT scanning. These six characteristics define the population most at risk for deterioration due to harmful secondary insults. Therefore, they may also define the population best targeted by neuroprotective agents.

**Statistical Power**

It appears that some of the previous trials may have been designed with insufficient power. For example, the series of PEG-SOD trials demonstrated an 8% absolute difference in the proportion of patients with favorable outcome between the drug compared with placebo categories, but only approximately 600 patients were enrolled in the trial (300/treatment group), the difference was not significant for the primary outcome variable—that is, the 3-month GOS score.34,44 It is imperative that TBI trials be powered adequately, based on the outcome measure used and the magnitude of difference to be detected (usually 900–1000 patients, or more).

**Primary Outcome Measures for Phase III Severe Head Injury Trials**

Almost all participants in the Phase III trials in severe TBI have used the GOS, which is usually dichotomized into the upper two and lower three categories for the purposes of statistical comparison (Table 4). Of the eight neu-

---

**TABLE 2**

Clinical trials in neurotrauma (1993–1997) with positive findings

<table>
<thead>
<tr>
<th>Authors &amp; Year</th>
<th>Drug/Agent</th>
<th>No. of Patients, Study Type &amp; Country</th>
<th>Outcome</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muizelaar, et al., 1993</td>
<td>PEG-SOD, free radical scavenger</td>
<td>94, Phase II, US</td>
<td>ICP lower, outcome better</td>
<td>Phase III trial negative, but 8% better trend</td>
</tr>
<tr>
<td>Clifton, et al., 1993</td>
<td>moderate hypothermia, ~33°C</td>
<td>46, Phase II, US</td>
<td>ICP lower, outcome better</td>
<td>Phase III Trial halted, analysis ongoing</td>
</tr>
<tr>
<td>Shiozaki, et al., 1993</td>
<td>mild hypothermia, ~34°C</td>
<td>33, Phase II, Japan</td>
<td>ICP lower, outcome better</td>
<td>cerebral “metabolic indicators” improved</td>
</tr>
<tr>
<td>Grumme, et al., 1995</td>
<td>triamcinolone early megadose 200 mg, &lt;4 hrs, 120 mg/day</td>
<td>396, Phase III (contusions + GCS &lt;8, n = 93), Germany</td>
<td>18% lower mortality in contusions + GCS &lt;8 (p &lt;0.01)</td>
<td>only post hoc subgroup analysis was positive</td>
</tr>
<tr>
<td>Harders, et al., 1996</td>
<td>nimodipine calcium channel blocker</td>
<td>123, Phase II, Germany</td>
<td>55% reduction in bad outcome (p &lt;0.002)</td>
<td>only traumatic SAH selected, any GCS</td>
</tr>
<tr>
<td>Marion, et al., 1997</td>
<td>moderate hypothermia, ~33°C for 24 hrs, w/in 10 hrs</td>
<td>82, Phase II, US</td>
<td>early ICP lower (p &lt;0.01), outcome better in GCS 5–7</td>
<td>Phase III trial negative</td>
</tr>
</tbody>
</table>

**TABLE 3**

Design of recent clinical trials in neurotrauma

<table>
<thead>
<tr>
<th>Drug or Agent</th>
<th>Mechanism</th>
<th>Company or Sponsor</th>
<th>Inclusion Criteria</th>
<th>Enrollment Window</th>
<th>Primary Outcome Measure(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-CPP-ENE (EAA-494)</td>
<td>competitive NMDA antagonist</td>
<td>Novartis (Europe)</td>
<td>GCS Score 3–8</td>
<td>12 hrs</td>
<td>GOS</td>
</tr>
<tr>
<td>CP-1 01,606</td>
<td>polyamine NR2b NMDA antagonist</td>
<td>Pfizer</td>
<td>GCS Score 4–8</td>
<td>8 hrs</td>
<td>GOS</td>
</tr>
<tr>
<td>CI 1009</td>
<td>t-calcium channel blocker</td>
<td>Parke-Davis</td>
<td>GCS Score 4–8 all types</td>
<td>20 hrs</td>
<td>GOS (DRS rejected by FDA)</td>
</tr>
<tr>
<td>BAY X3702</td>
<td>5HT1A agonist hyperpolarizes neurons</td>
<td>Bayer (Europe)</td>
<td>GCS Score 4–8</td>
<td>12 hrs</td>
<td>GOS</td>
</tr>
<tr>
<td>nimodipine</td>
<td>t-calcium channel blocker</td>
<td>Bayer</td>
<td>SAH due to trauma, GCS Score 4–15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypothermia</td>
<td></td>
<td>NIH</td>
<td>GCS Score 4–8, all types</td>
<td>8 hrs</td>
<td>GOS</td>
</tr>
</tbody>
</table>

* FDA = Food and Drug Administration.
rotrauma Phase IIb and Phase III trials in progress or in final planning during the last quarter of 1997, the GOS was used in all. In one case, the use of this outcome measure was mandated by the US Food and Drug Administration (Table 3). Dichotomization of GOS scores is usually performed for clinical reasons and for simplicity of interpreting the difference of outcomes between two trial arms. Unfortunately, the GOS may lack sufficient sensitivity to allow for discrimination of more subtle improvements that could be achieved using an investigational drug, and the use of a dichotomized GOS may in a way magnify this lack of sensitivity.26 Nevertheless, the use of the GOS has been cited in the neurotrauma literature on more than several hundred occasions, and it remains the most widely used and accepted instrument available to us today (Fig. 1).

**Bimodal Distribution of the GOS**

The GOS was devised to permit a coarse comparison in head injury outcome across different centers.21,22 At the time of its design, the effect of the bimodal distribution of head injury outcome on neuroprotection drug trials was not a significant issue. As shown in Fig. 2, the majority of patients with a severe TBI will fall into the two extremes of the GOS: those with good recovery and those who die. These two outcome categories will account for approximately 70% of patients, and the remainder will be distributed among the middle three categories (vegetative state, severe disability, and moderate recovery). It is, however, within these three categories that patients in whom drug-related effects most likely are demonstrated will be found. In cases classified as either good outcome or dead, the patients may be less likely to experience drug-related effects because their brains may be too mildly injured or so severely injured, respectively, to be amenable to the therapeutic benefit of any pharmaceutical agent.

Experience has shown that during the recovery period between the 3- and 6-month outcome points, considerable improvement in patient-related function may occur, yet during this time period it is unusual for patients to move fully from one category into another simply because these categories lack the sensitivity required.10 For this reason, a modification to the GOS was proposed by Jennett and colleagues22 in 1981, designed to widen the scale numerically to eight points (Table 4). This eight-point scale was designed to provide greater discrimination and narrower categories to enable detection of more subtle changes in outcome. Unfortunately, this advantage was found to be offset by greater interobserver variability in allocating a patient to a particular category. Consequently, a more rigorously structured questionnaire has been proposed by the Glasgow group to increase reliability of the eight-point scale and make outcome determination more precise. This approach awaits further validation in larger series of patients.40,44

**The DRS**

First devised and popularized by Hall, et al.,19 in 1985, the DRS was proposed to quantify the stages of recovery through rehabilitation of patients after a severe head injury. It is an ordinate, non-linear scale with 30 points and thus, intrinsically, has greater sensitivity than the GOS. Its major disadvantages are that it requires more specialized training for its implementation by the rater, and the inter-rater variability is high. It also underestimates the severity of outcome compared with the GOS (Table 5).13 Recently, pooled data derived from two different neuroprotective drug trials performed under the auspices of the ABIC were evaluated to test the concordance between the DRS and GOS, which was found to be rather poor. In particular, analysis indicated that there was complete overlap in the DRS score for the severely disabled and vegetative groups in the GOS. In addition, analysis of the probability values in the two studies showed that the DRS is less sensitive than the GOS. Overall, the DRS is, therefore, considered a less useful primary outcome measure compared with the GOS in trials involving patients with severe TBI.

**Neuropsychological Outcome Measures**

It is generally accepted that outcome after severe brain injury is multidimensional and difficult to measure, particularly because there is no baseline from which one can judge the extent of recovery.

Neuropsychological assessment is a logical method to determine outcome after either moderate or severe TBI. Such assessments, however, must be used in parallel with expected mortality in severe injury, and importantly, 10 to 15% of patients will be unable to undergo testing following severe TBI because they are too cognitively impaired. Although psychometric testing is extensively used in evaluating head-injured patients for return-to-work placement and to determine directions for rehabilitation, it has never been used successfully as an outcome measure in a neuroprotection clinical trial. This has partly been because psychologists could not agree on an optimum test battery. The other difficulty is that in a group with a higher mortality rate but fewer disabled survivors better neuropsychological scores will be demonstrated than in groups with a lower mortality rate, and therefore greater number of disabled survivors. In an attempt to further improve assessment of outcome, participants in the ABIC adopted a specific set of neuropsychological tests that would add
dimensions of cognitive function, motor skill, and QOL measures to the analysis of outcome. Those involved in the TCDB have cautioned against testing being too long or complicated. Of the 1030 patients accrued in the TCDB, for example, fewer than 100 completed the battery of tests, which was very extensive. As a result of these concerns, the ABIC selected four neuropsychological tests that would limit the testing period to approximately 30 minutes.

In addition, the ABIC adopted a QOL measure, termed the Neurobehavioral Functional Inventory, which included indices of depression, somatic complaints, memory/attention, communication, aggression, and motor function. The four neuropsychological tests, coupled with the Neurobehavioral Functional Inventory, comprised the standard ABIC battery. The tests and areas addressed are shown in Table 6. Although the skills of a neuropsychologist are desirable, they are not essential for administration of the test battery. With appropriate training, this battery can be administered by nurses or other trained medical personnel.

In a test of utility, these five ABIC-devised measures were used in conjunction with the GOS and DRS in two clinical trials designated A and B. The combined patient accrual of Trials A and B totaled 371 patients, and of these 230 were eligible for testing at 3 and 12 months were eligible at 6 months (Table 7). Based on the data summarized in Table 7, it is seen that approximately 60% of patients were capable of fully completing the test battery whereas 19.6 and 14.9% of patients were capable of completing only a portion of the tests at 3 and 6 months, respectively. Approximately 20% of patients refused testing, were unavailable, or could not be tested. Those in whom tests could not be performed accounted for 11.6% of patients at 6 months. Interestingly, of 191 patients in whom the GOS was either good or moderate, only 72% were capable of fully completing the test battery, signifying a major degree of cognitive impairment, even in the group with favorable outcome.

Cognitive Impairment of Placebo Groups

In Trial A of the ABIC, all patients with a GCS score of 4 to 8 and abnormal findings demonstrated on CT scans were included in the trial. In Trial B, all patients with a GCS score 3 to 8 and at least one reactive pupil were included. For the five-test neuropsychological battery, the percentile scores of the placebo-treated groups of Trials A and B ranged between the 40th and 60th percentile, where the 100th percentile was equivalent to the mean score for a normal population. It is thus clear from these data that survivors of severe TBI were profoundly impaired and did not approach average normal levels even at 6 months postinjury.

The CNPS

To simplify further the analysis of neuropsychological test scores, a CNPS was defined, which averaged all of the percentile scores of the individual tests of the ABIC battery. The use of the CNPS is illustrated in the comparison of placebo groups of Trials A and B (Fig. 3). The tendency toward a lower CNPS in Trial B is due to its inclusion of more severely injured patients (GCS Scores 3–8 [Trial B] compared with GCS Scores 4–8 [Trial A]). Of most importance, however, is the reduced CNPS observed in patients in the “good” or “moderate” GOS categories. These data illustrate that patients with GOS-classified good recovery are profoundly impaired and function at 40 to 60% of average normal level. Moreover, the CNPS in patients with moderate outcome signifies even greater cognitive deficits.

It is our recommendation, in view of the profound cognitive impairment in survivors of severe TBI, that neuropsychological tests be incorporated as an adjunct to more global measures of outcome in clinical trials.

Psychometric Testing for Moderate Brain Injury

No data regarding the use of psychometric testing in patients with moderate TBI have been published. Neuropsychological testing may be the only viable method for
measuring outcome in these patients because mortality rates are low (approximately 4%), and outcome in the majority of patients falls into the good or moderately disabled categories on the GOS at 6 months.

“Surrogate” Outcome Measures

A surrogate outcome measure has been statistically defined as a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the null hypothesis based on the true endpoint. In other words, the surrogate endpoint must accurately depict the true effect of treatment on the outcome of the patient, as demonstrated by the primary endpoint. Surrogate measures have revolutionized clinical trial design and permitted the rapid advances in combination therapy, as has been seen in acquired immune deficiency syndrome research, for example. Here, the CD4+ lymphocyte counts and viral load measures change over weeks and months in response to therapies and closely correlate with clinical status, mortality, and morbidity.

In neurotrauma, surrogate measures have not been useful in trial design, and research efforts need to be undertaken to determine optimum surrogate measures after TBI, where “final” outcome may only become apparent 6 to 12 months after injury. Several possible surrogates have been proposed and are discussed in the subsequent sections.

Intracranial Pressure Measurement and the TIL

The TIL was first proposed and validated as a means of cross comparing the severity of injury and different therapy protocols in different institutions in the TCDB study. Based on evidence derived from animal studies and theoretical considerations, an ICP lowering effect may be postulated for many of the drugs, particularly glutamate antagonists, that have been evaluated for severe TBI. The collection and cross comparison of ICP data, in which data were manually entered into a form and later entered into a computer for data manipulation, have been surprisingly successful across large numbers of centers in numerous Phase II and III trials. Unfortunately, however, in the Phase II trial of the competitive glutamate antagonist, Selfotel, a difference in ICP was seen in favor of the treatment group, but this failed to carry over into the larger Phase III trial. The same was seen in the Sterling PEG-SOD trial; the authors found lower ICP in Phase II, but this did not correlate with significant evidence of better outcome in Phase III trial.

For neuroprotective agents, which are thought to have an ICP lowering effect, the TIL is an appropriately sensitive tool at the Phase II level to help determine whether a biological effect of the agent may be present. For the Phase II trial with the free radical scavenger PEG-SOD, the apparent ICP lowering effect was interpreted as an indication of biological effect of the drug. For the first-generation N-methyl-D-aspartate antagonists, behavioral changes in healthy individuals and stroke patients, were interpreted as a “biological effect” indicator that the drug was interacting with receptors. In the studies with Selfotel and cerestat, behavioral changes in healthy individu-

| TABLE 5 |
| Comparison of rating on GOS and DRS |
| GOS Score | Good Recovery | Moderate Disability | Severe Disability |
| no disability | 24 | 7 | 0 |
| mild | 1 | 8 | 0 |
| partial | 0 | 6 | 0 |
| moderate | 0 | 3 | 21 |
| moderately severe | 0 | 0 | 4 |
| severe | 0 | 3 | 1 |

| TABLE 6 |
| The ABIC neuropsychological test* |
| Test | Area Addressed |
| Rey Complex Figure | visuoconstruction & memory |
| Controlled Oral Word Association | oral fluency |
| Symbol Digit Modalities (oral) | sustained attention |
| Grooved Pegboard | fine motor dexterity |
| Neurobehavioral Functioning Inventory | behavior/QOL |

* Listed in order of administration.

Fig. 3. Bar graph showing the CNPS for two clinical trials. The CNPS averages the percentile scores of the ABIC neuropsychological battery. Note that CNPS for survivors of head injury are far below the level of a normal population. This is also true for patients with good and moderate GOS scores. Study B included patients with GCS scores of 3, which accounts for the lowered CNPS compared with Study A.
Outcome measures for neurotrauma trials

als were used as a dose determinant for subsequent stroke studies.5,24,33

**Microdialysis of Extracellular Fluid**

More recently, microdialysis of brain extracellular fluid has been used. In patients with severe TBI, the technique has been used to measure penetration of study drug into cerebral tissue. For drugs that may be active against putative mediators of brain damage, the mediator could be measured by microdialysis as a surrogate endpoint. Other indicators of brain damage such as brain-specific creatine phosphokinase, have been measured in cerebrospinal fluid, but the correlation with outcome has been poor.35

**Magnetic Resonance Spectroscopy**

In the future, MR spectroscopy may provide a powerful surrogate measure for neuroprotectant drugs. For example, N-acetyl aspartate has been shown to correlate with the viability of neuronal populations and it can readily be measured by MR spectroscopy.9 Diffusion-weighted imaging, a measure of the amount of cytotoxic edema present on MR images, may also prove to be useful and has been widely advocated as a surrogate measure for neuroprotectant drug effects in stroke in both animal models and human studies.36,38,45

**Brain Oxygen Measurement**

More recently, participants at approximately 20 neurotrauma research centers have been working with continuous measurements of brain oxygen by using newly available, minimally invasive, Clark electrode systems (Licox GmbH, Germany; Paratrend/Neurotrend; Diagnostics, HighWycombe, United Kingdom). These systems offer a readily available surrogate marker for therapies designed to increase regional cerebral blood flow or oxygen transportation to tissues. Possible therapies include perfluorocarbons, diasprin-linked stroma-free hemoglobin, and allosteric hemoglobin modifiers.17,18,42

**Positron Emission Tomography and Receptor Binding**

Positron emission tomography or single photon emission tomography can be used to demonstrate receptor binding or receptor displacement by a drug. This could serve two purposes: demonstration of brain penetration of drug and definition of optimum dosing in a small number of patients.

**Role of Phase II Trials—Safety**

Many of the failures of Phase III neurotrauma trials have been attributed to inadequate attention to prior Phase II evaluation. Because there is intense competition among pharmaceutical companies to develop neuroprotectant agents, the temptation exists to proceed to a Phase III study before a complete analysis of Phase II data has been performed. In the case of some agents, no Phase II trials have been conducted prior to Phase III trials. Fortunately, the intensive reporting system regarding adverse events, which is in place in the pharmaceutical industry, has allowed the detection of any potentially harmful effect of neuroprotectant drugs; to date, no agent has been shown to be harmful either at Phase II or Phase III levels, and in this regard, those Phase II studies that have been performed have been successful, at least in demonstrating drug-related safety and tolerance.

Recently, the authors of four Phase II trials conducted in Japan and the US have shown a beneficial effect of moderate hypothermia when using a variety of surrogate markers.12,27,32,37 The stronger the range of Phase II evidence of this type, the greater the likelihood of success in a Phase III evaluation; however, in a Phase III trial of prophylactic hypothermia in the United States, no benefit was seen.

**Futility Analysis**

In view of the extremely high costs of neurotrauma trials and the compelling desire of the pharmaceutical industry to know as soon as possible whether the potential exists for a trial to show a favorable outcome, interim analysis is uniformly performed in TBI trials. In addition, clinical investigators do not like exposing patients to the possible risk of an experimental agent if there is little or no likelihood of benefit. To retain sufficient statistical power, enrollment of an increased number of patients is necessary if interim analyses are to be performed. Interim analyses of this type, therefore, weaken the overall conclusion in a trial and when they result in the termination of a trial, as has recently been the case for Selfotel, Cerestat, and hypothermia then no definitive statement can be made regarding the efficacy or lack of efficacy of the agent being investigated. Scientifically this is undesirable because no clear answer can be given regarding a potential drug-related effect, only that it is unlikely to achieve the requisite 10% difference, a level of efficacy that may be impossible to achieve.

**Sequential Analysis**

The sequential analysis technique proposed by Whitehead43 has been widely accepted in clinical trials in oncology and acquired immune deficiency syndrome. Using this technique, there is no penalty in terms of increased number for diminished statistical power, even though the outcome data are being analyzed continuously throughout the trial period. In this way, each case is analyzed as soon as it reaches its designated outcome point. By constructing preset significant limits for beneficial and negative outcome, the number of patients enrolled in a trial may be minimized (Fig. 4). Thus this strategy offers the benefit to the pharmaceutical industry of knowing the outcome of a trial (either efficacious or dangerous) at the earliest possible moment, without degrading its statistical validity. This method was used in a large Phase III trial with the Novartis compound EAA494 in Europe (for mortality only). This method would seem to have a strong rationale for neurotrauma clinical trial construction.

**Outcome Determination in Phase III Neurotrauma Trials**

Choi and colleagues10 have studied data obtained in a large cohort of patients with severe TBI from the TCDB to determine the time course of recovery based on GOS score (Fig. 5). They found that approximately 10% of cases changed by one GOS level between 3 and 6 months...
post injury. Thus, it is often appropriate to conclude outcome at 3 months in drug studies in which the GOS is used. This was used for the PEG-SOD trials and for the Parke-Davis/Neurex SNX III trial.

If sequential analysis techniques are used, there is a great advantage in rapidly determining the trial outcome, and a 3-month outcome point for neurotrauma trials is superior to a 6-month outcome for this purpose. Clearly, however, the final outcome for the purposes of determining the level of permanent disability should be delayed until the 6-month time point. If ultimate outcome is the best endpoint in a trial, then a 6-month GOS score is a better measurement because of the documented improvement in patients between 3 and 6 months after TBI. On the other hand, it is possible that an efficacious treatment may hasten recovery so that improved outcome might be present at 3 months and disappear at 6 or 12 months if control patients recover equally but more slowly.

For the present, we do not know the optimum time for outcome determination. We recommend that data be collected at least at 3 and 6 months after TBI and probably also at 12 months to try to resolve this question.

**Population Targeting Within Neurotrauma Trials**

Whereas our knowledge of pathophysiological mechanisms advances, it becomes clear that not all patients who have sustained neurotrauma will develop the same pathological processes postinjury. For example, early ischemia is observed primarily in patients with acute subdural hematoma and those with large contusions. Recent microdialysis studies have shown that patients who have sustained secondary ischemic injuries and focal contusions are primarily those in whom glutamate release into the extracellular space is demonstrated. Thus it is more appropriate to target glutamate antagonist drugs to the patient population in which the release of the neurotransmitter is more likely to occur.

Similarly, agents designed to increase focal cerebral blood flow or oxygen delivery may be more suited for application in patients with ischemia. The concept thus emerges that for a particular neuroprotectant agent attempts should be made to define the pathological mechanisms adequately in animal models and subsequently in Phase II trials. Phase III trials can then be designed to focus on the patient population most likely to benefit. With knowledge of alterable pathological mechanisms, surrogate endpoints also may be used to measure or establish efficacy. A large body of information obtained from recent Phase III trials in which investigators studied different neuroprotectants has shown that the mortality rate is low (that is, 5–7%) in patients with CT scanning–documented “diffuse injury II.” Patients with diffuse injury II, however, comprise approximately 38 to 40% of the total patient group in neurotrauma trials. These are the patients most likely to make an uneventful recovery and likewise in whom the pathomechanisms will fail to demonstrate sensitivity to drug therapy.

Thus there is a scientific rationale to focus the agent on the population at risk. In a recent trial of the bradykinin antagonist Bradycor the investigators focused specifically on patients with contusions larger than 20 mm in diameter, and the authors found a strong trend toward beneficial outcome—28% better outcome in the treated group in a recent Phase II study.

Similarly, post hoc analysis of the data derived from a large Phase IIb European trial with the NR2B subset selective glutamate antagonist, eliprodil, (unpublished data) has shown that in drug-treated patients an impressive 34% better outcome was achieved in the subgroup of patients with mass lesions and diffuse brain swelling. Although trials with specifically targeted subpopulations of neurotrauma patients are more difficult to perform because the characteristics of the targeted population need to be strictly defined for the investigators, they may yield a positive outcome in a smaller number of patients; additionally they also offer the advantage that patients in whom a specific pathomechanism is not likely to be demonstrated are consequently not exposed to the agent in question. Currently, the CT scan TCDB classification has proven to be the easiest criterion for selection of the targeted population in the study.

**Quality-of-Life Assessment as an Outcome Measure**

Quality-of-life assessment and costs of therapy (or phar-
macroeconomics) are important endpoints for neurotrauma trials, although these will always occupy the role of secondary outcome measures. Many authorities argue that these pharmaco-economic indicators should be reserved for Phase IV clinical trials in which cost comparisons with established therapy are the focus. The need, however, to reduce costs and time to marketing, has induced several pharmaceutical companies to include these endpoints in Phase III studies as “add-on” endpoints. The most popular instruments at present are the SF-36, the general health questionnaire 30, and the Functional Independent Measure.11,16,20,31,41 These three outcome measures correlate quite closely with the GOS and offer increased sensitivity because these scores have numerical ordinate linear scales with possible scores of 20 or 30, allowing smaller numbers of patients to be used to determine outcome.

These outcome measures have been validated on both sides of the Atlantic in large populations with several different states of disease.1,11,12,13 The SF-36 has been compared with the GOS, and a significant relationship was found across all categories (Spearman rank coefficients 0.51–0.68, p < 0.01).3,13 It was especially reassuring that after severe head injury, those patients with the good GOS-documented outcome had SF-36 scores that were very close to those observed in the healthy population (standard scores 0.07 ± 0.56) compared with patients with severe and moderate disability (0.30 ± 3.47).

Role of the NIH in Neurotrauma Trial Sponsorship

The NIH comprises 22 institutes, centers, and offices that support clinical trials, and many of these trials are related to neurological disorders. Blindness, deafness, mental health, neurodegenerative disorders, and injury are all represented, but the majority of NIH support for neurotrauma clinical trials comes from NINDS. This institute has sponsored fewer neurotrauma trials than the pharmaceutical industry; however, its role is to sponsor studies that may fall outside areas of commercial interest but that may contribute significantly to our understanding (as well as treatment) of head and spinal cord injury. An example of one such study is the recently completed trial in which investigators at the Baylor College of Medicine compared therapy directed at cerebral perfusion pressure and conventional ICP-directed therapy. The NINDS also sponsored a Phase III trial of hypothermia in severe head injury14 (see the subsequent section).

The impact of NINDS-sponsored trials in areas of clinical neuroscience can rival or exceed that of the pharmaceutical industry trials. Recent successful examples include: intravenous tissue plasminogen activator thrombolysis for occlusive stroke, the North American Symptomatic Carotid Endarterectomy Trial, and the National Acute Spinal Cord Injury Trials II and III. Although translation of research results into clinical practice is slower than one might hope, it is likely that the rate with which stroke and trauma patients are evaluated and treated will change as a result of the work reported in these studies.

The NINDS has a commitment to continued funding for high-quality clinical trials. Currently, this institute supports most trials through the investigator-initiated “R01” mechanism, although contracts (N01) and cooperative agreement (U01) are also used. Other institutes prefer to use contract mechanisms, because the involvement of institute staff is often considerable. Overall, NINDS spends approximately 7% of its extramural budget ($542 million in 1996) on clinical trials. A working group was recently established to facilitate the support of clinical trials in neurological disorders, and two new mechanisms have been put in place. The NINDS pilot clinical trial grant (R01) and the NINDS planning grant for clinical trials (R21) are available, and applications are accepted on the regular receipt dates (the 1st day of February, June, and October). The pilot grant may be used to determine factors such as side effects and dosages (Phase II and design evaluation) in patients; it is not meant to support infrastructure. The planning grant should be used to prepare a manual of operations, to conduct planning sessions for investigators, or to offer training experience for personnel involved in the trial. Both mechanisms should clearly address the timely need for the Phase III trial.

The NINDS also supports conferences and workshops by using an investigator-initiated R13 mechanism. The review process for such funding can be expedited but most often requires up to 9 months before funding is obtained.

The NIH Hypothermia Trial

Three hundred ninety patients were enrolled in the American Hypothermia Trial, coordinated by Dr. Guy Clifton.14 The trial reported a combined 27% overall mortality rate for the control and hypothermia-treated groups, and the pattern of complications and the causes of death have been similar to those reported in other severe head injury trials. No beneficial effect for prophylactic hypothermia was seen, but ICP was lower in the cooled patients.14

CONCLUSIONS AND RECOMMENDATIONS

1) Carefully designed, adequately powered, Phase II trials with a variety of mechanisms and safety endpoints and dose escalations when appropriate should be completed and analyzed before undertaking Phase III efficacy studies.

2) Evidence of effective brain penetration and biological effects (where appropriate for the agent in question) should be demonstrated before embarking on Phase III trials.

3) The heterogeneity of the neurotrauma patient population and the wide intercenter variability means that more complex trial design may be required in future Phase III trials. These include: stratification based on the motor component of the GOS, which has been successful in previous trials; stratification based on the CT classification; and block randomization, which has been highly effective in minimizing the effect of intercenter variability.

4) Larger numbers (usually > 1000) of patients should be enrolled in neurotrauma trials, based on power analysis, for the GOS.

5) The 6-month GOS score is currently the most appropriate primary outcome measure. Three-month GOS scores should also be determined. An expanded GOS with greater sensitivity within the middle outcome categories, which are most crucial to determining drug effect, should be evaluated.
6) In studies of moderate TBI, the ABIC five-test neuropsychological battery, CNPS, is the most appropriate outcome measure. In studies of severe neurotrauma this five-test battery should be included as a secondary outcome measure, which may be used to improve the sensitivity of the GOS in the future.

7) Neuropsychological measures of outcome parallel functional measures such as GOS and DRS scores.

8) Significant impairment is present even in severe TBI patients with a “good outcome” on the GOS. They score in the 40th to 60th percentile when using the ABIC CNPS battery, where 100% is considered the average normal level.

9) Sequential analysis should be performed to eliminate the need for interim and futility analyses, which weaken the statistical power of a Phase II study.

10) The strategy of targeted patient population selection has been adopted in many neurotrauma trials and is theoretically appealing. This concept, however, awaits validation where a pathophysiological mechanism and an appropriate agent can be identified. In trials in which the pathomechanism is less clearly defined, broader enrollment criteria are necessary. Post hoc analysis of Phase II trial data may allow for better targeting patient populations in Phase III trials.

11) Future research should be focused on development of sensitive and specific surrogate outcome measures that can be applied to Phase II trials and used to detect a neuroprotective effect. These may include MR imaging and neurochemical, neuropsychological, imaging, ICP, and QOL measures.

References


Neurosurg. Focus / Volume 13 / July, 2002


Manuscript received November 9, 2001. Accepted in final form January 15, 2002. Address reprint requests to: Ross Bullock, M.D., Ph.D., Division of Neurosurgery Medical College of Virginia, Box 980631, MCV Station, Richmond, Virginia 23298-0631. email: robullock@hsc.vcu.edu.