Surgical Decision Making in Brain Hemorrhage
New Analysis of the STICH, STICH II, and STITCH(Trauma) Randomized Trials

Barbara A. Gregson, PhD; Patrick Mitchell, FRCS; A. David Mendelow, FRCS(SN)

**Background and Purpose**—The STICH (Surgical Trial in Lobar Intracerebral Haemorrhage) I and II trials randomized patients with spontaneous intracerebral hemorrhage (ICH) to early surgery or initial conservative treatment. Both were nonsignificant; possibly because surgery has minimal effect on recovery, or because surgery benefits some and harms others. We introduce a new nonparametric method of analysis. The method is then applied to data from a third trial, STITCH(Trauma) (Surgical Trial in Traumatic Intracerebral Haemorrhage), which addressed a similar surgical question in head-injured patients.

**Methods**—Data from 1541 patients from the STICH trials were analyzed using (1) standard meta-analysis of prognosis-based dichotomized outcome and prespecified standard subgroups of Glasgow Coma Scale (GCS): 3–8, 9–12, and 13–15; (2) new nonparametric regression of ranked Extended Glasgow Outcome Scale against ranked GCS and ranked volume; and (3) analysis (1) repeated using categories identified by analysis (2).

**Results**—Standard meta-analysis showed more favorable outcomes, although nonsignificant, with surgery if presenting GCS was 9–12 (spontaneous ICH odds ratio, 0.70 [95% CI, 0.48–1.03; P=0.07]; traumatic odds ratio, 0.48 [95% CI, 0.18–1.26; P=0.14]). Ranked analysis showed a similar pattern of results for both spontaneous and traumatic ICH. Surgery was harmful for small lesions with increasing benefit for larger volumes. With GCS, surgery had little effect at either ends of the spectrum but suggested a beneficial effect in the range 10 to 13 (identified graphically). Repeating the meta-analysis with this categorization showed significant benefit for surgery (spontaneous odds ratio, 0.71 [95% CI, 0.51–1.00; P=0.05]; traumatic odds ratio, 0.16 [95% CI, 0.05–0.51; P=0.002]).

**Conclusions**—The nonsignificant results observed in the STICH trials are because of mixing patients who benefit from surgery with those who are harmed. Patients with a GCS 10–13 or a large ICH are likely to benefit from surgery. Our analysis showed a similar effect on traumatic ICH/contusion data and promises to be a valuable tool.

**Clinical Trial Registration**—URL: [http://www.isrctn.com/](http://www.isrctn.com/). Unique identifiers: ISRCTN19976990 (STITCH), ISRCTN22153967 (STICH II), and ISRCTN19321911 (STITCH[Trauma]). (Stroke. 2019;50:1108-1115. DOI: 10.1161/STROKEAHA.118.022694.)

**Key Words:** cerebral hemorrhage ■ conservative treatment ■ decision making ■ Glasgow Outcome Scale ■ surgery

Randomized controlled trials are the prevailing paradigms of clinical research, and surgical evacuation of intracerebral hematomas (ICHs) has been the subject of 16. 1–16 The more powerful trials and meta-analyses6,11,17,18 have repeatedly failed to displace the null hypothesis though have often shown nonsignificant findings. That notwithstanding, there is an ongoing interest in the question with further trials ongoing or planned (MISTIE III [Minimally Invasive Surgery and rtPA for Intracerebral Hemorrhage Evacuation], CLEAR IV [Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Hemorrhage], and STITCH III [International Surgical Trial in Intracerebral Haemorrhage]).

The null results of trials and meta-analyses led to the hypothesis that the effect of surgery, if any, is too small to detect with current trial data. The amount of randomized data now in hand has limited the size of such an overall benefit to the extent that, were it there, it would be of doubtful clinical significance. These analyses used an assumption common to much clinical research: there is a uniform response to a treatment for the population under study. This assumption is particularly weak with surgical treatments because in addition to the patient-specific variables that affect medical treatments, there is also variation between surgeons. Mechanistic lines of argument have led to a different theory, one commonly held by surgeons and one that predates all trial data. This theory is that response to surgery is not uniform from one case of ICH to another but varies with clot volume and clinical condition.

Continuing medical education (CME) credit is available for this article. Go to [https://cme.ahajournals.org](https://cme.ahajournals.org) to take the quiz.

Received June 29, 2018; final revision received January 31, 2019; accepted March 4, 2019.

From the Neurosurgical Trials Group, Newcastle University.

The online-only Data Supplement is available with this article at [https://ahajournals.org/doi/suppl/10.1161/STROKEAHA.118.022694](https://ahajournals.org/doi/suppl/10.1161/STROKEAHA.118.022694).

Correspondence to Barbara A. Gregson, PhD, Neurosurgical Trials Group, Institute of Neuroscience, Newcastle University, Wolfson Research Centre, Campus for Ageing and Vitality, Westgate Rd, Newcastle upon Tyne, NE4 5PL; Email barbara.gregson@ncl.ac.uk

© 2019 The Authors. *Stroke* is published on behalf of the American Heart Association, Inc., by Wolters Kluwer Health, Inc. This is an open access article under the terms of the Creative Commons Attribution Non-Commercial-NoDerivs License, which permits use, distribution, and reproduction in any medium, provided that the original work is properly cited, the use is noncommercial, and no modifications or adaptations are made.

**Stroke** is available at [https://www.ahajournals.org/journal/str](https://www.ahajournals.org/journal/str)

DOI: 10.1161/STROKEAHA.118.022694
In this article, we present a combined analysis of trial data specifically aimed at testing the hypothesis: Is the neutrality of trial data a consequence of little effect or a consequence of mixing cases who benefit from surgery with those that are harmed?

Two trials used similar methodology to investigate whether patients had a better outcome with a policy of early surgery to evacuate spontaneous ICH compared with initial conservative treatment. The STICH (Surgical Trial in Lobar Intracerebral Haemorrhage) included patients with supratentorial ICH at least 2 cm in diameter and with a Glasgow Coma Scale (GCS) of at least 5. In STICH II (Surgical Trial in Lobar Intracerebral Haemorrhage), these criteria were refined to include only patients with a lobar hemorrhage within 1 cm of the cortex surface of between 10 and 100 mL and where the Motor GCS was 5 or 6 and the Eye score of the GCS 2 or more.

Patients were excluded from both if there was an ictothemorrhagic lesion (aneurysm or arteriovenous malformation) underlying the hemorrhage, or if they had severe preexisting comorbidity that would mean that they could not achieve a Good Recovery on the Glasgow Outcome Scale (GOS) at 6 months even if they made a full recovery from the hemorrhage. Neither trial displaced the null hypothesis but both showed a nonsignificant benefit with surgical evacuation.

We used the combined STICH and STICH II data sets to investigate the relationship between outcome and the effect modification of variables that may influence the response to surgery following spontaneous ICH: neurological condition quantified by GCS and clot volume. We undertook standard meta-analysis with standard prespecified bandings: GCS 3–8, 9–12, and 13–15.

Methods
The data used in this analysis will be submitted to the VISTA database in the future but is available from the authors (online-only Data Supplement).

The first STICH trial randomized 1033 patients; STICH II randomized 601; outcome was not available for 69 patients from STICH, 18 from STICH II. A further 6 patients are lost to analyses involving the distinction that instead of dividing the patients into 2 groups according to how they were allocated in the trials, the patients are divided into 2 groups on the independent variable (×) scale. The limitation of nonparametric statistics with regard to CIs led us to analyze these differences using a new statistic and that is the area between the curves. This is a direct measure of the difference between the positions of the 2 curves summed over the entire range of the independent variable being considered. How clinically relevant this statistic is has yet to be determined as it has not previously been used but it is readily related to the outcome scales. Bootstrapping is a numerical method that makes no parametric assumptions. It is based on deriving a distribution for the statistic under study by doing the above-described analysis repeatedly on the data, but with the distinction that instead of dividing the patients into 2 groups according to how they were allocated in the trials, the patients are divided at random. Then after a million iterations enough area measurements have been generated to create a frequency distribution of areas which is centered on zero and is approximately normal. This distribution can then be used to compare with the area found on the initial data analysis and so determine what proportion of randomly generated areas are less than or more than the area we have found on analysis. This test is inherently 2-tailed because the area is not dependent on which of the 2 regression lines is the higher at any particular point.

We performed a sensitivity analysis to determine how sensitive the P values were to the robustness feature in the MATLAB Lowess function. We did this by selecting a random subsample of patients and then using a random noise or wiggling in the regression curves with correspondingly larger areas between the curves and any systematic trend will be more
easily masked by these larger areas so shorter spans will lead to lower \( P \) values. At the other end of the spectrum, 100% span will give \( P \) values identical to a Wilcoxon rank-sum test unless the regressions lines cross. The main area of interest was in spans between 35% and 65%.

**Meta-Analysis Using New Cut Points**

This ranked regression system has the advantage that it can show when the lines cross each other, that is, when one treatment is superior over one range of a variable and the other is superior over another range. So, for example, we found surgery to be superior for large clot volumes and conservative to be superior for small. The crossing point of the regression lines gives the cut point between these ranges.

**Meta-Analysis**

Figure 1A through 1D shows the results of the meta-analyses. All the studies individually show a nonsignificant trend.

**Results**

**Meta-Analysis**

Figure 1A through 1D shows the results of the meta-analyses. All the studies individually show a nonsignificant trend.
Figure 1A shows that the combined result from meta-analysis of the spontaneous ICH trials gives an overall odds ratio (OR) of 0.87 (95% CI, 0.70–1.09; \(P=0.23\)). Figure 1B shows that for traumatic ICH the OR is 0.57 (95% CI, 0.31–1.08; \(P=0.09\)). Figure 1C and 1D is the subgroup analyses looking at the 3 bands of the GCS. There is a wide variation in the ORs observed for the GCS band 3 to 8 with the results dominated by STICH showing no treatment effect. For GCS band 13 to 15 (the most conscious patients), the results are similar between the 2 studies giving a combined OR of 0.92 (95% CI, 0.68–1.23; \(P=0.56\)). For the intermediate GCS band 9 to 12, both studies give similar results with the combined OR for the spontaneous studies is 0.70 (95% CI, 0.48–1.03; \(P=0.07\)). In traumatic ICH, there is a tendency for improved outcome with surgery in both GCS band 3 to 8 (OR, 0.50 [95% CI, 0.11–2.38; \(P=0.38\)]) and GCS 9 to 12 band (OR, 0.48 [95% CI, 0.18–1.26; \(P=0.14\)]) but again no treatment effect in GCS band 13 to 15.

**Ranked Analysis**

Figure 2A through 2D shows the results of the ranked analyses. In these graphs, all patients are located on the y axis according to their GOSE ranking (where GOSE varies from 1 to 15, with 1 being the worst outcome and 15 being the best). The x axis is simply all patients ranked. Equal spacing on the axes equals equal numbers of patients. On the y axis nonparametric regression values are plotted using Matlab 8.5.0.197613 (R2015a) malowess with robustness and spans given below. Spans were determined by inspection. The area between the lines was used as the statistic for \(P\) value calculations. A, Data combined from the STICH and STICH II trials plotting ranked presenting GCS against ranked GOSE. Span was 0.4, \(P=0.04\). B, Data from the STITCH(Trauma) trial (Surgical Trial in Traumatic Intracerebral Haemorrhage) plotting ranked presenting GCS against ranked extended GOSE. Span is 0.4, \(P=0.012\). (Continued)
Meta-Analysis Using New Cut Points

The meta-analysis of the spontaneous and traumatic ICH data sets undertaken using the new banding for GCS identified by the rank analysis is shown in Figure 3A and 3B. The OR for GCS band 10 to 13 in the spontaneous data set is 0.71 (95% CI, 0.51–1.00; \( P = 0.05 \)) and in the traumatic data set is 0.16 (95% CI, 0.05–0.51; \( P = 0.002 \)).

Discussion

Analysis of the data from 1541 randomized and monitored spontaneous ICH patients suggests that patients with an intermediate presenting GCS in the range of 9 to 12 may benefit from surgery whereas those with higher or lower GCS may not. This finding is consistent with surgical opinion which generally holds that in patients with a high GCS, the risks between 1 for dead and 8 for upper good recovery) and on the x axis according to their parameter ranking (GCS or volume). The plots are nonparametric regression of patients in the 2 groups: surgical and conservative. It is of interest that despite the combined STICH data set being much larger and of a different pathology to the STITCH(Trauma) data set, both show the same basic relationships. For presenting GCS, surgery may be more beneficial in an intermediate range and not at the extremes. Though this pattern was common to both data sets it was not significant for spontaneous ICH (spontaneous \( P = 0.4 \), traumatic \( P = 0.012 \)). The maximum benefit from surgery seems to be for patients with a GCS of 10 to 13. For hematoma volume, surgery is detrimental for small volumes and beneficial for large volumes with benefit increasing the larger the volume.
of surgery are poorly justified by small potential benefit; and patients with a low GCS have a poor prognosis with or without surgery which consequently has little to offer.

The dichotomized method used for the primary analysis of the trials limits our ability to study how particularly GCS and volume interact with surgery. At the extremes of both GCS and volume, any effect of surgery would be expected to lie within the extreme ranges of the GOSE (death, upper good recovery) rather than between them and dichotomized outcomes that split the scale in the middle will not pick these up. To do so, we need to use the full GOSE rather than dichotomize it but this leads to a problem with the common method of relating variables—regression. Regression is a parametric method that relies on the assumption that variables are ordinal (each point on the scale is strictly less than the one below it and greater than the one above it) and uniform (and by the same amount). The ordinal assumption can be made for GOSE (because for each point on the scale, a patient at that point can do everything at the point below them plus something else) but not the uniform assumption.

The GCS is neither clearly ordinal nor uniform. It is made up of 3 subscales each of which can reasonably be regarded as ordinal but combining them means that it is not true that for every point on the GCS patients have worse coma than for the point above. However, in practice we consider the GCS to be ordinal because its correlation with outcome in large data sets is ordinal.

We, therefore, use a regression method based on rank that only requires the ordinal and not the uniform assumption. Each subject is allocated a rank on the GOSE and GCS scales and on lesion volume. Then nonparametric regression of ranked GOSE versus another ranked variable is plotted. Significance was tested with bootstrapping. Comparing areas under the 2 lines is related to the Wilcoxon ranked-sum test. This is fair in cases where the lines do not cross. Where they do cross it is not satisfactory because significance will be lost by combining a disadvantage in one region with a benefit in another irrespective of how large the effect is and whether or not it is real. Therefore, we use the statistic of area between the lines.

The standard meta-analysis found that patients with an intermediate presenting GCS in the range of 9 to 12 appeared to benefit from surgery whereas those with higher or lower GCS did not. The ranked analysis suggested that this benefit accrued for patients with a GCS in the range 10 to 13 and that this may be a better window to investigate in future trials as it was derived directly from the data rather than depending on an arbitrarily selected GCS range. The post hoc analysis confirmed that a GCS range of 10 to 13 identified those patients most likely to benefit from surgery.

**Figure 3.** Forest plots using new categorization of Glasgow Coma Scale (GCS; ≤9, 10–13, and 14–15). **A.** Spontaneous intracerebral hemorrhage (ICH) studies. **B.** Traumatic ICH study.

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Early Surgery</th>
<th>Initial Conservative</th>
<th>Odds Ratio</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Events</td>
<td>Total</td>
<td></td>
<td>M-H, Fixed, 95% CI</td>
</tr>
<tr>
<td>GCS 3 - 9</td>
<td></td>
<td></td>
<td></td>
<td>M-H, Fixed, 95% CI</td>
</tr>
<tr>
<td>STICH (2005)</td>
<td>118</td>
<td>137</td>
<td>117</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.75 [0.17, 3.36]</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>156</td>
<td>159</td>
<td></td>
<td>1.13 [0.62, 2.06]</td>
</tr>
<tr>
<td>Total events</td>
<td>132</td>
<td>132</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneity: Chi² = 0.34, df = 1 (P = 0.56); I² = 0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: Z = 0.39 (P = 0.70)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCS 10 - 13</td>
<td></td>
<td></td>
<td></td>
<td>M-H, Fixed, 95% CI</td>
</tr>
<tr>
<td>STICH (2005)</td>
<td>139</td>
<td>192</td>
<td>176</td>
<td>223</td>
</tr>
<tr>
<td></td>
<td></td>
<td>81</td>
<td>130</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.72 [0.43, 1.21]</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>322</td>
<td>351</td>
<td></td>
<td>0.71 [0.51, 1.00]</td>
</tr>
<tr>
<td>Total events</td>
<td>220</td>
<td>265</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneity: Chi² = 0.01, df = 1 (P = 0.92); I² = 0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: Z = 1.97 (P = 0.05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCS 14 - 15</td>
<td></td>
<td></td>
<td></td>
<td>M-H, Fixed, 95% CI</td>
</tr>
<tr>
<td>STICH (2005)</td>
<td>89</td>
<td>139</td>
<td>85</td>
<td>133</td>
</tr>
<tr>
<td></td>
<td></td>
<td>79</td>
<td>148</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.01 [0.63, 1.60]</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>287</td>
<td>272</td>
<td></td>
<td>1.01 [0.72, 1.41]</td>
</tr>
<tr>
<td>Total events</td>
<td>169</td>
<td>159</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneity: Chi² = 0.00, df = 1 (P = 1.00); I² = 0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: Z = 0.03 (P = 0.97)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>765</td>
<td>782</td>
<td></td>
<td>0.88 [0.79, 1.00]</td>
</tr>
<tr>
<td>Total events</td>
<td>520</td>
<td>556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneity: Chi² = 3.12, df = 5 (P = 0.68); I² = 0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: Z = 1.13 (P = 0.26)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for subarous differences: Chi² = 2.76, df = 2 (P = 0.25); I² = 27.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Table 3](http://ahajournals.org)
For hematoma volume, the finding was of a negative effect of surgical evacuation at low volume passing through neutrality to a positive effect as volumes increased. Again this is consistent with surgical opinion which generally holds that the effect of surgery in this situation is mediated by decompression and that the greater the volume of clot the greater the potential for decompression such that at small volumes, the risk of surgery is not justified but this reverses with larger volumes.

The STICH and STITCH(Trauma) trials were similar in design but they addressed different pathologies, in view of which it is striking how similar are the patterns seen in both with the ranked analysis. We consider that it is possible that this pattern reflects general principles that apply across neurological conditions and neurosurgical operations. Those being the larger the surgical target the greater the potential benefit of surgery and with mild disease the risks of surgery outweigh the benefits. An idea for which there is also support from animal models.27

Limitations

The nonparametric analysis is an exploratory post hoc analysis but has been undertaken on a large data set and offers better hypotheses for future trials in spontaneous and traumatic ICH.

The P values of the ranked analysis represent the probability that allocating cases from the whole population to groups at random would give an area between the lines of equal to or greater than that obtained by allocating cases according to the group they were randomized to in the trials. We consider all these P values to be marginal. A limitation of the method we used is that to some extent the P value is dependent on the selected span. At low span, more random wiggles in the lines increases the area between them. This reduces the proportion of that area that is because of systematic differences and so increases P values. At high span, key features of the line such as the boundary and crossing positions are more prone to bias. The conventional method of span selection, and the one we used is, by inspection. Regression plots are examined and the lowest span which leads to a reasonable suppression of random scatter is selected. It should be noted, however, that the lowest span which leads to a reasonable suppression of random scatter is selected. It should be noted, however, that this might be a more appropriate categorization of the GCS for future studies of ICH. Our ranked analysis method shows similar effects on traumatic ICH/contusion data and may prove to be a valuable tool in assessing the effects of treatments.

Conclusions

Meta-analysis favors the hypothesis that the neutral result of spontaneous ICH trials is because of the mixing of systematic positive and negative effects rather than there being a uniform small or absent effect. We identified that patients with an intermediate GCS (10–13) and a large spontaneous ICH may be more likely to benefit from surgery and that this might be a more appropriate categorization of the GCS for future studies of ICH. Our ranked analysis method shows similar effects on traumatic ICH/contusion data and may prove to be a valuable tool in assessing the effects of treatments.

Sources of Funding

The STICH I (Surgical Trial in Lobar Intracerebral Haemorrhage) was funded by the UK Medical Research Council, STICH II by the Efficacy and Mechanism Evaluation (EME) Programme of the National Institute for Health Research (NIHR) and STITCH(Trauma) (Surgical Trial in Traumatic Intracerebral Haemorrhage) by the Health Technology Assessment Programme of the NIHR.

Disclosures

All authors designed this analysis and wrote the article, the statistical analysis was undertaken by Dr Gregson and P. Mitchell. Dr Gregson and A. David Mendelow report research grant funding from National Institutes of Health (NIH) for work on the MISTIE III trial. A. David Mendelow and P. Mitchell report other research support from Newcastle Neurosurgery Foundation Limited for administrative staff support.

References


