Original Research

Agreement between core laboratory and study investigators for imaging scores in a thrombectomy trial

Robert Fahed,1 Malek Ben Maacha,2 Céline Ducroux,1 Naim Khoury,3 Raphaël Blanc,1 Michel Piotin,1 Bertrand Lapergue,4 for the ASTER Trial Investigators

ABSTRACT

Purpose We aimed to assess the agreement between study investigators and the core laboratory (core lab) of a thrombectomy trial for imaging scores.

Methods The Alberta Stroke Program Early CT Score (ASPECTS), the European Collaborative Acute Stroke Study (ECASS) hemorrhagic transformation (HT) classification, and the Thrombolysis In Cerebral Infarction (TICI) scores as recorded by study investigators were compared with the core lab scores in order to assess interrater agreement, using Cohen’s unweighted and weighted kappa statistics.

Results There were frequent discrepancies between study sites and core lab for all the scores. Agreement for ASPECTS and ECASS HT classification was less than substantial, with disagreement occurring in more than one-third of cases. Agreement was higher on MRI-based scores than on CT, and was improved after dichotomization on both CT and MRI. Agreement for TICI scores was moderate (with disagreement occurring in more than 25% of patients), and went above the substantial level (less than 10% disagreement) after dichotomization (TICI 0/1/2a vs 2b/3).

Conclusion Discrepancies between scores assessed by the imaging core lab and those reported by study sites occurred in a significant proportion of patients. Disagreement in the assessment of ASPECTS and day 1 HT scores was more frequent on CT than on MRI. The agreement for the dichotomized TICI score (the trial’s primary outcome) was substantial, with less than 10% of disagreement between study sites and core lab.

Trial registration number NCT02523261, Post-results.

INTRODUCTION

Clinical trials with an imaging-based primary endpoint frequently rely on an external core laboratory (core lab) for outcome adjudication, a practice that is endorsed by the Food and Drug Administration (FDA). Centralized imaging interpretation is key due to the role, the variability, the susceptibility to bias of imaging within the trial as well as modality-specific image quality considerations and overall trial design features.

Aspiration versus STEnt-Retriever (ASTER) was a randomized trial comparing thrombectomy and stent retriever for the endovascular treatment of acute ischemic stroke (AIS) in patients with anterior circulation large vessel occlusion. Using an imaging-based primary endpoint (ie, the rate of successful recanalization as primary outcome, defined by a Thrombolysis In Cerebral Infarction (TICI) score of 2b or 3 at the end of the procedure), the trial showed that thromboaspiration was not superior to stent retrievers. The core lab in the ASTER trial adjudicated the primary outcome (TICI score) as well as several other baseline and follow-up imaging and imaging scores which had also been assessed by the investigators at the study sites.

Although highly recommended by the FDA and PROBE guidelines for imaging trial design, the added benefit of a core lab adjudication has not yet been studied. Bias of study site investigators from being unblinded to the treatment arm is only presumed to affect results and disagreements with core lab assessment and has not been rigorously explored in any of the recent thrombectomy trials. We aimed to assess the interrater agreement between the unblinded study site investigators and the blinded core lab of the ASTER trial in adjudication of imaging scores.

METHODS

The study protocol was approved by the local ethics committee. For each patient, a core lab composed of four physicians with 10–20 years of experience in neuroradiology assessed the following scores:

- the baseline Alberta Stroke Program Early CT Score (ASPECTS)4 (when CT was performed at patient presentation) and diffusion-weighted imaging ASPECTS (DWI-ASPECTS) (when MRI was performed at presentation)5
- the presence of hemorrhagic transformation (HT) according to the European Cooperative Acute Stroke Study (ECASS)6 on day 1 imaging. The TICI score7 was assessed by a core lab composed of two interventional neuroradiologists with 10–15 years of experience as follows:
  - the intermediate TICI was defined as the TICI score after three passages of the first-line strategy determined by randomization (thromboaspiration or stent retriever)2
  - the final TICI was defined as the TICI score at the end of the procedure. The dichotomized TICI score (0/1/2a vs 2b/3) was used as the primary outcome of ASTER.
Each of these scores was also assessed by the investigators at the study sites (i.e., interventional neuroradiologists with various degrees of experience).

**Statistical analysis**

The interrater agreement for exact CT-ASPECT and DWI-ASPECT scores was assessed using Cohen’s unweighted kappa (κ) and weighted kappa (κw) statistics. The interrater agreement for dichotomized ASPECT scores (0–7 vs 8–10) was assessed with κ statistics. The agreement for the presence of HT and its ECASS classification, as well as the TICI score, was assessed with κ statistics. All analyses were performed with 95% confidence intervals obtained by 1000 bootstrap resampling. Slight (0–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and excellent (>0.80) agreement categories were defined according to Landis and Koch. All analyses were performed using R software Version 3.3.2 (R Foundation for Statistical Computing, Vienna, Austria) with a 5% significance level.

**RESULTS**

Initial characteristics of the study population have been previously published. In summary, the trial included 381 patients (mean age 69.9 years, 174 women (45.7%)), with a median ASPECT score of 7 (assessed by core lab) and a median time from symptom onset to groin puncture of 227 min.

The agreement for the exact ASPECT score was slight on CT (80 patients, κ=0.064 (−0.062–0.190)) and fair on MRI (259 patients, κ=0.291 (0.222–0.361)). A >1 point difference in ASPECT score occurred in 37/80 patients screened with CT (46.3%) and 56/259 patients screened with MRI (21.6%). When the ASPECT score was dichotomized for the analysis (0–7 vs 8–10), the interrater agreement reached a fair level on CT (80

![Figure 1](image_url)  
**Figure 1** Agreement between study site and core lab. (A) Agreement for CT-ASPECTs and DWI-ASPECTs (Cohen’s unweighted kappa and weighted kappa). (B) Agreement for detection of hemorrhagic transformation and its ECASS classification. (C) Agreement for intermediate and final Thrombolysis In Cerebral Infarction score. For each panel, a κ value >0.6 indicates substantial agreement.
The agreement for the diagnosis of HT was moderate on CT (198 patients, \( \kappa = 0.513 (0.399–0.627) \)) and excellent on MRI (134 patients, \( \kappa = 0.666 (0.571–0.761) \)). A disagreement on MRI was less substantial for all CT-based scores. There was less disagreement in the assessment of baseline ASPECTS and the presence of HT when assessed on MRI compared with CT. The level of agreement for TICI score was moderate (more than 25% disagreement rate). The agreement reached a substantial level when the score grading was dichotomized (0/1/2a vs 2b/3).

The disagreement rate was unexpectedly high for all scores in our study. However, several factors might explain such discrepancies. The limited reliability of ASPECTS/ECASS/TICI classifications has been previously demonstrated, with some studies suggesting that ASPECTS and ECASS might be more reliable on MRI than on CT. The assessment of ASPECTS on brain CT requires expertise, training (available on http://www.aspectsin-stroke.com), and optimal reading conditions with adequate adjustment of window and level of the image. The ASTER study investigators, who performed mostly emergency brain MRI (259/339 patients in the present study (76.4%)), may not have been comfortable with ASPECT scoring on brain CT, and they were not required to undergo the online training before the study. Moreover, they may have lacked time to adjust image calibration at patients’ arrival. Finally, the limited number of core lab members for ASPECT scoring ensured a uniform protocol for ASPECT scoring, whereas the numerous study site investigators across the country did not have the opportunity to agree on a consensual scoring process.

Although dichotomization of the ASPECT score led to increased agreement in line with previously published work, a substantial degree of disagreement persisted in our study, especially on brain CT. Our results provide evidence that the ASPECT score might not be a sufficient clinical tool to be used alone in clinical decision making.

Reading conditions also differ substantially: investigators at the study sites assess patients’ imaging individually and urgently in ‘real time’ whereas the core lab assessment occurs for all patients at one time after trial completion, outside of any emergency context, and in optimal reading conditions in terms of computer screen calibration and room lighting. The core lab adjudicators are aware of the future impact of their assessment on trial results, which could suggest the presence of a Hawthorne effect.

Finally, a major difference between study sites and core lab scorings relies on the blinding process for core lab adjudication. We can only speculate that the knowledge of study site investigators of each patient’s treatment arm may have subconsciously modified their behaviour in angiographic scoring. Nevertheless, this phenomenon could not happen in ASPECT scoring which was performed before randomization of treatment. Moreover, the TICI score, which might have been prone to this bias, had the lowest level of disagreement in our study.

Two previous thrombectomy studies (SWIFT and TREVO-2) used a core lab for the assessment of the primary outcome (angiographic recanalization score) and reported different proportions of recanalization rates between study investigators and core lab. In the SWIFT trial, the rates of successful recanalization according to the core lab were 69% (with Solitaire) and 30% (with Merci), while the rates according to study investigators were 83% and 48%, respectively. In the TREVO-2 study, the rates of successful recanalization according to the core lab were 86% (with Trevo) and 60% (with Merci), while the rates according to study investigators were 85% and 66%, respectively. Similar to ASTER, analysis of the primary outcome using scorings of the study sites did not modify the trial results.

However, the SWIFT and TREVO-2 studies did not calculate the proportions of disagreements or kappa values. In fact, two similar proportions of recanalization rates reported by two

---

**Table 1** Detailed results for agreement study. Agreement (95% CI) (Cohen’s unweighted kappa (\( \kappa \)) and proportion of similar answers for each CT-ASPECTS/DWI-ASPECTS, HT diagnosis, ECASS classification, and TICI score. Weighted kappa (\( \kappa_w \)) values and proportions of answers with \( \leq 1 \) point difference were also calculated for ASPECTS and DWI-ASPECTS.

<table>
<thead>
<tr>
<th>Score Type</th>
<th>Agreement Value</th>
<th>Similar Answers, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT-ASPECTS (n=80)</td>
<td>( \kappa = 0.064 (0.062–0.190) )</td>
<td>22 (27.5)</td>
</tr>
<tr>
<td>MRI (n=134)</td>
<td>0.168 (0.009–0.345)</td>
<td>43 (53.7)</td>
</tr>
<tr>
<td>Dichotomised CT-ASPECTS (0–7 vs 8–10)</td>
<td>( \kappa = 0.272 (0.021–0.523) )</td>
<td>57 (71.2)</td>
</tr>
<tr>
<td>DASH (n=259)</td>
<td>( \kappa = 0.291 (0.222–0.361) )</td>
<td>102 (39.4)</td>
</tr>
<tr>
<td>MRI (n=259)</td>
<td>0.607 (0.555–0.659)</td>
<td>203 (78.4)</td>
</tr>
<tr>
<td>Dichotomised DWI-ASPECTS (0–7 vs 8–10)</td>
<td>( \kappa = 0.666 (0.571–0.761) )</td>
<td>219 (84.5)</td>
</tr>
</tbody>
</table>

**HT diagnosis (yes/no)**
- CT (n=198): \( \kappa = 0.564 (0.441–0.686) \), 159 (80.3)
- MRI (n=134): \( \kappa = 0.835 (0.742–0.928) \), 122 (91.0)

**ECASS classification**
- CT (n=198): \( \kappa = 0.338 (0.213–0.463) \), 129 (65.1)
- MRI (n=134): \( \kappa = 0.513 (0.399–0.627) \), 86 (54.7)

**Dichotomised ECASS classification (0/HI1/HI2 vs PH1/PH2)**
- CT (n=198): \( \kappa = 0.326 (0.090–0.563) \), 171 (86.4)
- MRI (n=134): \( \kappa = 0.604 (0.434–0.774) \), 116 (86.6)

**TICI score**
- Intermediate (n=347): \( \kappa = 0.589 (0.522–0.655) \), 244 (70.3)
- Final (n=350): \( \kappa = 0.592 (0.519–0.665) \), 273 (78.0)

**Dichotomised TICI score (0/1/2a vs 2b/3)**
- Intermediate (n=347): \( \kappa = 0.751 (0.675–0.823) \), 310 (89.3)
- Final (n=350): \( \kappa = 0.659 (0.540–0.778) \), 321 (91.7)

ASPECTS, Alberta Stroke Program Early CT Score; DWI, diffusion-weighted imaging; ECASS, European Collaborative Acute Stroke Study; HT, hemorrhagic transformation; TICI, Thrombolysis In Cerebral Infarction.
readers is not synonymous with score agreement for each patient. For example, even if two readers report a 80% rate of successful recanalization in the same study cohort, the proportion of disagreements can be anywhere between 0% (readers agreed on every patient) and 40% (readers did not score an unsuccessful recanalization in the same patient each time). The kappa values and percentages of disagreements in the ASTER trial shed light on the presence of disagreements in a non-negligible proportion of patients. To our knowledge, this is the first analysis of agreement between core lab and study site investigators in imaging outcomes. Even if these discrepancies did not modify the ASTER trial or the results of other trials, our results suggest that the increased logistics and expenses related to the use of an external core lab seem justified in an imaging-based primary endpoint trial in order to harmonize the adjudication process and minimize scoring differences. In our study, discrepancies in imaging scores occurred less frequently in topics where study investigators (interventional neuroradiologists) had a higher expertise (TICI scoring). On the other hand, disagreement was more frequent in neuroradiology scores and classifications (ASPECT and ECASS) typically reported by diagnostic neuroradiologists. It could be argued, based on our results, that trial adjudication of interventional outcomes might be reliably reported by the site investigators who are also primary operators, whereas a core lab remains necessary for other imaging outcomes.

CONCLUSION
The scores assessed by the imaging core lab frequently differed from those reported by the study sites participating in the ASTER trial. Dichotomization of the ASPECT and TICI scores improved agreement between the core lab and the study investigators. Discrepancies were less frequent on MRI than on CT, suggesting that the use of MRI for imaging outcomes in future trials might lead to less frequent disagreement in adjudication.


Contributors RF: figures, study design, data collection, data analysis, statistical analyses, writing manuscript. MBM: data collection, data analysis, writing manuscript. CD: figures, data collection, data analysis, writing manuscript. NK: data analysis, writing manuscript. RB: data analysis, writing manuscript. MP: data analysis, writing manuscript. BL: study design, data collection, data analysis, writing manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent Not required.

Ethics approval Fondation Rothschild Ethics Committee.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data, analytic methods, and study materials will be made available to any researcher for the purposes of reproducing the results or replicating the procedure. Requests to receive these materials should be sent to the corresponding author, who will maintain their availability.

© Article author(s) (or their employer(s) unless otherwise stated in the text of the article) 2018. All rights reserved. No commercial use is permitted unless otherwise expressly granted.

REFERENCES