Interrater Agreement for Final Infarct MRI Lesion Delineation

Anders B. Neumann, MD; Kristjana Y. Jonsdottir, MSc, PhD; Kim Mouridsen, MSc, PhD; Niels Hjort, MD, PhD; Carsten Gyldensted, MD, PhD; Alberto Bazzi, MD; Jens Fiehler, MD, PhD; Roberto Gasparotti, MD; Jonathan H. Gillard, BSc, MD, FRCR; Marc Hermier, MD, PhD; Thomas Kucinski, MD, PhD; Elna-Marie Larsson, MD, DMSc; Leif Sørensen, MD; Leif Østergaard, MD, PhD

Background and Purpose—Lesion volume measured on follow-up magnetic resonance imaging (MRI) is commonly used as an outcome parameter in clinical stroke trials. However, few studies have evaluated the optimal sequence choice and the interrater reliability of this outcome measure. The objective of this study was to quantify the geometric interrater agreement for lesion delineation of chronic infarcts on T2-weighted and fluid-attenuated inverse recovery (FLAIR) MRI.

Methods—In a retrospective study of 14 patients, lesions on 90-day follow-up FLAIR and T2 fast spin echo MRI were outlined by 9 independent, blinded, experienced neuroradiologists. Voxel-wise interrater agreement was measured as (1) the volume of the intersection of individual rater’s lesion outlines relative to the mean lesion volume (overlap ratio) and (2) the Hausdorff distance between the lesion markings.

Results—Mean patient age was 64.4 years (range, 45 to 79). Lesion volumes on FLAIR were, on average, 2.5 mL greater than were T2 volumes (median; \( P<0.001 \)). We found considerable differences between raters’ lesion markings, but interrater agreement was consistently better on FLAIR than on T2 images, as measured by a greater overlap ratio (\( P<0.0001 \)) and a smaller Hausdorff distance (\( P<0.0001 \)) on FLAIR than on T2.

Conclusions—FLAIR should be used to quantify follow-up infarct size to minimize interrater variability. Our study suggests that imaging analysis performed by 1 or a few trained readers may be preferred. Future studies should address objective and preferably automated criteria for final lesion delineation. (Stroke. 2009;40:3768-3771.)

Key Words: biostatistics ▪ brain imaging ▪ cerebral infarct ▪ diagnostic methods ▪ magnetic resonance imaging ▪ outcomes ▪ clinical trial design

Lesion volume has been used as a surrogate outcome in several clinical stroke trials. It has been argued that the use of such imaging outcome markers allows the detection of efficacy signals in smaller patient cohorts than do clinical outcome scores such as the modified Rankin Score. With a potential to accelerate phase II trials, imaging markers may enable faster drug certification and faster development of treatment protocols.1 Several factors limit the reliable delineation of final lesion boundaries and hence, the accuracy and reliability of lesion volumes as a surrogate marker of stroke outcome. On T2 fast spin echo (T2FSE) magnetic resonance imaging (MRI) scans, hyperintensities in cortical lesions may be indistinguishable from surrounding cerebrospinal fluid (CSF). Likewise, subcortical lesions may be difficult to separate from adjacent regions of leukoaraiosis.

The perfusion-diffusion (PWI-DWI) mismatch model is widely recognized as a surrogate for salvageable tissue.2 With the use of so-called predictive algorithms, voxel-wise, multimodal, acute image information may be combined with outcome (infarcted, noninfarcted) to form quantitative infarct risk maps in subsequent patients.3,4 Although this approach may represent an even more sensitive means of detecting drug efficacy, it depends critically on not only lesion volume but also accurate lesion characterization at a voxel level. The importance of final lesion determination is contrasted by surprisingly few studies that have addressed the reliability of...
final lesion volume assessment, let alone any studies that have examined voxel-wise precision of final lesions outlining.

This study compared the final lesions markings performed by 9 expert raters on 90-day follow-up T2FSE and fluid-attenuated inverse recovery (FLAIR) images. We present lesion volume variability and quantify interobserver agreement in terms of lesion location according to a simple relative-overlap measure and a geometric measure of maximum distance between partially overlapping regions.

Methods

Patients

This study was part of a prospective MRI study of a consecutive series of stroke patients who received thrombolytic therapy at Aarhus University Hospital, Aarhus, Denmark, from April 2004 to March 2006. Fourteen patients (10 male; mean age, 64.4 years; range, 45 to 79) who presented with symptoms of acute stroke were selected for this analysis according to the following criteria: (1) no major, chronic ischemia present on baseline FLAIR, (2) MRI at day 90 was of acceptable technical quality (in particular, no head movement artifacts were present on baseline FLAIR, (2) MRI at day 90 was of acceptable quality, with the use of freely available image analysis software (MIRcro by Chris Rorden, 2005). All FLAIR images were first shown and delineated consecutively, followed by all 2 images in permuted order. The sequences were created by a computer algorithm that secured the presence of at least 4 images (median, 13) between each subject’s FLAIR and T2 images. Raters were allowed to take breaks for an unlimited time between any images, but no minimal time between T2 and FLAIR outlining was imposed. All raters evaluated the images in the same order, ie, performed identical tasks. Raters were allowed to freely adjust window/level settings, but no automatic preprocessing, such as thresholding, was applied.

Statistical Analysis of Agreement

We quantified the overall difference between the 2 follow-up modalities (FLAIR, T2) by calculating the average lesion volume for each patient and comparing average volumes between FLAIR and T2 images. Although a comparison of raters’ lesion volumes provides a standard measure of agreement, some disagreement might have been undetected. Two raters’ markings of the lesion (ie, region-of-interest [ROI]), might be of equal volume but be only partially overlapping. Therefore, we determined volume location agreement as the volume of the intersection of the 2 raters’ ROIs relative to the mean volume of the 2 ROIs. This overlap ratio will range from 0, meaning no intersection, to 1, implying identical ROIs (see Figure 1a).

Furthermore, to measure the maximal disagreement in lesion location, we also determined the Hausdorff distance between 2 raters’ ROIs. The Hausdorff measure is a standard geometric distance measure that quantifies the greatest distance between the 2 ROIs. Figure 1a illustrates how the Hausdorff distance is obtained (shown by an arrow). Figure 1b shows an example of the 2 measures of agreement. The figure shows the final infarct drawn by 2 raters for a specific subject. The boundary of the 2 infarcts is shown in blue (5.13 mL) and green (4.78 mL), respectively. The volume agreement is large (93.2%), but the overlap ratio is only 71.0%. Furthermore, a Hausdorff distance of 40.3 mm is observed by the thin “spike” involving a large segment of the parietal cortex (blue boundary). Analyses were performed with the MATLAB software package (Mathworks). For pairwise analyses, a Wilcoxon signed-rank test was applied.

Results

Great variation in volumes was recorded among the raters, and almost all patients had greater lesion volumes on FLAIR than on T2. The median lesion volumes for the 14 patients

**Figure 1.** Agreement measures. a, Two partially overlapping ROIs (blue and green). The overlap is marked by the double crossed area. The overlap ratio is the overlap area divided by the mean area of ROI 1 and ROI 2. The Hausdorff distance between ROI 1 and ROI 2, represented by the arrow, quantifies the greatest distance between the boundaries. b, This is a 91-day-poststroke FLAIR image of 59-year-old man. A volume agreement of 93.2% is seen between the blue ROI (5.1 mL) and the green ROI (4.8 mL). However, the overlap ratio is only 71.0%, expressing the visual impression of overlap. The “spike” in the blue ROI involves a large portion of the parietal cortex and is represented by a Hausdorff distance of 40.3 mm.
were as follows (in mL, FLAIR/T2): 5.9/4.7, 9.1/5.4, 10.4/4.5, 10.9/6.0, 8.7/8.9, 18.5/14.2, 21.7/32.2, 29.8/26.7, 52.9/54.2, 69.3/60.6, 72.2/62.5, 82.9/76.2, 81.8/80.5, and 132.7/126.5. In general, FLAIR volumes were greater than T2 volumes by 2.5 mL (median; interquartile range [IQR], 1.16 to 5.9 mL, \(P<0.001\)) and by 7.0% measured as a relative difference (FLAIR compared with T2, median; IQR, 3.3% to 31.4%, \(P<0.0001\)). This suggests that raters tend to delineate conservatively when faced with the less clear lesions on T2. To demonstrate the typical variation, we show an example of the intraslice variation (see Figure 2). Although raters obviously agreed on the center of the lesion, considerable disagreement is present on the border, resulting in notable interrater differences in lesion volume.

Median interrater agreement as determined by the overlap ratio was greater on FLAIR (0.78; IQR, 0.71 to 0.83) than on T2 (0.75; IQR, 0.59 to 0.83; \(P<0.0001\)). Figure 3a shows that in 9 of 14 patients, the median FLAIR overlap ratio was greater than on T2. Two patients with small volumes had low median agreement on both T2 and FLAIR, possibly caused by differences in border delineation, which have a greater effect on small lesions. However, 2 more patients with small volumes had similarly small T2 agreement but higher FLAIR agreement.

Median Hausdorff distance was 23.4 mm (IQR, 15.9 to 34.7 mm) on FLAIR but 25.8 mm (IQR, 16.9 to 40.8 mm) on T2, and FLAIR distances were significantly smaller than those on T2 (\(P<0.0001\)). Patient-wise, this measure also favored FLAIR, because the median FLAIR Hausdorff distance was smaller than on T2 for 10 of 14 patients (see Figure 3b).

**Discussion**

In the present study, we examined the 2 commonly used MRI sequences (FLAIR and T2) for imaging of chronic ischemic infarcts. It was found that lesion outlining on FLAIR images yielded significantly better interrater agreement than on T2 images, as assessed by the overlap ratio (\(P<0.0001\)) as well as by the Hausdorff distance (\(P<0.0001\)).

The 2 agreement measures considered in this study coincided by showing better interrater agreement on FLAIR images, which we speculate was caused by a more distinct border between tissue and CSF on FLAIR compared with T2. However, they provide complementary information on lesion overlap morphology in individual patients. For example, 1 patient had a median FLAIR overlap ratio of 0.85, which is smaller than the 0.88 on T2. On the other hand, the Hausdorff distance measure favored FLAIR (median of 22.5 mm) in this patient compared with T2 (median of 30.5 mm). In this way, it is possible to capture important information on the outlining strategy used by different experts, as shown in Figure 1 for another patient.

As pointed out by de Crespigny, surprisingly few stroke studies have evaluated interrater variability on follow-up MRIs,
Neumann et al  Interrater Agreement for Final Infarct MRI Lesions  3771

Despite its importance as a surrogate end point in clinical trials, Luby and colleagues found absolute interrater volume differences of 0.51 mL (median; IQR, 0.31 to 1.94) but calculated 62.52 as a percent deviation (median; IQR, 30 to 84) in 29 patients evaluated by FLAIR. T2-weighted images were not evaluated, although the authors questioned whether FLAIR was better than T2 images. As noted by Ay et al., an overabundance of small lesions was present in this study. Ritzl et al. reported interrater errors of 3 mL on FLAIR and 5 mL on T2 for 12 patients in a small study evaluated by 2 raters. A recent study by Ay and coworkers found a mean interrater variability of 8.3% (SD, 7.3%) among 58 patients with a mean final lesion volume of 39.2 mL (SD, 39.4 mL). Although the aforementioned studies compared lesion volumes only, lesion extent and location are crucial to accurately relate image findings, the so-called “real estate factor,” as well as the training of predictive algorithms that rely on voxel-wise classification of tissue outcome. Therefore, we currently perform a thorough comparison by applying methods that are pertinent to the spatial configuration of the outlined lesions.

The variations in our study might have been caused by the actual distribution of lesion volumes and the involvement of experts with different training backgrounds and from multiple institutions. Differences from other studies lie in the delineation procedure (either manually or semiautomatically) or the application of new and more-revealing interrater measures in our study. We consider that the major limitation of the current study is the relatively small number of patients. Second, the included subjects all had well-defined lesions, and cases that might have been more difficult to interpret were excluded, such as those with leukoaraiosis or agitated patients who might have induced movement artifacts on MRI. Also, raters only had access to the actual image to be outlined. Theoretically, access to baseline imaging data could reduce interobserver variability. Finally, randomization between modalities could be performed. Because outlining on FLAIR was better than on T2, a possible learning effect that would have favored T2 was very small.

In the study of Luby et al., consensus criteria for delineation had been determined by the readers beforehand, but this was not the case in our study or in the work by Ritzl et al. Interestingly, an MRI study of multiple sclerosis patients showed no change in interrater variability before and after a consensus meeting among readers, but to our knowledge, this has not been examined for cerebral infarctions detected by MRI. Indeed, our data suggest a notable difference in the perceptions of infarct extent, possibly reflecting a lack of a common definition of cerebral infarction. As pointed out by Saver, there is disagreement among the textbook definitions of infarct, and whether incomplete infarction should be included or not is under discussion. Although our study did not separate raters according to their classification of incomplete infarction, the existence of individual outlining strategies is supported by our data on 2 particular readers (No. 6 and 7), being represented in the top 3 of the smallest outlined lesion volumes in 12 of 14 patients (data not shown). Future work should establish a clear, common definition of cerebral infarction, with corollary MRI criteria to support this definition.

In future studies, allowing only 1 reader to delineate lesions may have the advantage of involving the same outlining strategy for all patients. However, a limitation of having only 1 reader is that this reader’s outlining strategy is not explicitly characterized. On the other hand, having several readers but only 1 reader per patient might not be recommended, because outlining strategies clearly differ greatly among readers. Reliable, automated algorithms for computerized lesion outlining, albeit technically challenging to devise, remain a crucial goal for improved sensitivity of clinical studies involving lesion size as an outcome measure.

In summary, the interrater variability reported here further underscores the need for automated algorithms or operational, objective criteria for the outlining of final infarct size. Furthermore, we propose that more revealing measures of interrater agreement, such as those presented in this work, should be applied as a component of quality assurance of lesion outlining procedures in future studies. Finally, we recommend using FLAIR rather than T2 for delineation of infarcts on MRI to minimize interrater variability.

Sources of Funding

The project was funded by the Danish National Research Foundation’s Center of Functionally Integrative Neuroscience, the EU Commissions Sixth Framework Program (A.B.N., N.H., K.Y.J., L.O.; contract 027294), and GlaxoSmithKline (K.M., educational grant) for financial support.

Disclosures

None.

References
