Interexaminer Difference in Infarct Volume Measurements on MRI
A Source of Variance in Stroke Research

Hakan Ay, MD; E. Murat Arsava, MD; Mark Vangel, PhD; Banu Oner, MD; Mingwang Zhu, MD; Ona Wu, PhD; Aneesh Singhal, MD; Walter J. Koroshetz, MD; A. Gregory Sorensen, MD

Background and Purpose—The measurement of ischemic lesion volume on diffusion- (DWI) and perfusion-weighted MRI (PWI) is examiner dependent. We sought to quantify the variance imposed by measurement error in DWI and PWI lesion volume measurements in ischemic stroke.

Methods—Fifty-eight consecutive patients with DWI and PWI within 12 hours of symptom onset and follow-up MRI on day-5 were studied. Two radiologists blinded to each other measured lesion volumes by manual outlining on each image. Interexaminer reliability was evaluated by intraclass correlation coefficients (ICC) and relative paired difference or RPD (ratio of difference between 2 measurements to their mean). The ratio of between-examiner variability to between-subject variability (variance ratio) was calculated for each imaging parameter.

Results—The correlation (ICC) between examiners ranged from 0.93 to 0.99. The median RPD was 10.0% for DWI, 14.1% for mean transit time, 18.9% for cerebral blood flow, 21.0% for cerebral blood volume, 16.8% for DWI/MTT mismatch, and 6.3% for chronic T2-weighted images. There was negative correlation between RPD and lesion volume in all but chronic T2-weighted images. The variance ratio ranged between 0.02 and 0.10.

Conclusion—Despite high correlation between volume measurements of abnormal regions on DWI and PWI by different examiners, substantial differences in individual measurements can still occur. The magnitude of variance from measurement error is primarily determined by the type of imaging and lesion volume. Minimizing this source of variance will better enable imaging to deliver on its promise of smaller sample size.

Key Words: acute stroke ■ diffusion-weighted imaging ■ MRI ■ neuroradiology

The mismatch between diffusion-weighted MRI (DWI) and perfusion-weighted MRI (PWI) indicates salvageable brain tissue and therefore is an attractive target for therapeutic intervention.1 Methods to quantify and visualize the DWI/PWI mismatch could clearly improve clinical research and patient care. For example, using imaging markers rather than the typical 90-day clinical outcome measurement provides shorter trials with smaller sample size and fewer patients lost to follow-up.2-4 A number of studies have already integrated lesion volume measurements on DWI and PWI as a surrogate outcome measure for assessing the effectiveness of new therapies.3,5-7 The reliable determination of the DWI, PWI, and mismatch volumes is of utmost importance for the validity of such studies. The measurement of lesion volume, however, is examiner dependent, as it partly requires visual identification and manual outlining. A number of studies have analyzed the intra- and interexaminer reliability of DWI and PWI lesion volume measurements using correlation analysis and demonstrated the presence of fair to excellent correlation.8-13 Correlation analyses measure the strength of a linear relation between 2 variables. However, in the design of prospective studies, a more relevant metric is the actual variability (agreement) among separate measurements. This variability becomes critical when a specific difference in volume is considered as a surrogate end point, selection tool, or prognostic marker in therapeutic trials or clinical stroke research. Because few studies have focused on this important metric, or on the variability in measurement of the DWI/PWI mismatch, we sought to calculate the magnitude of variance imposed by measurement error in DWI and PWI lesion volume measurements in a sample of patients with acute ischemic stroke.

Methods

Study Population
We retrospectively analyzed data collected as part of a prospective ongoing study evaluating the utility of DWI and PWI in predicting...
tissue risk of infarction (MRI Diffusion/Perfusion Mismatch in Human Acute Stroke). Consecutive patients with ischemic stroke who were admitted within the first 12 hours of symptom onset and who did not receive any thrombolytic treatment or investigational drugs were included. Each patient underwent 2 MRI studies, one obtained within 12 hours of symptom onset and the second on day-5 or later. In the first MRI, T2-weighted images, apparent diffusion coefficient maps (ADC), DWI, mean transit time (MTT) maps, cerebral blood flow (CBF), and volume (CBV) maps were obtained. The follow-up MRI included only T2-weighted sequences. The study was conducted at a single academic center, and the study protocol was approved by the local institutional review board.

Image Acquisition

MRI was performed on 1.5-T whole body scanners (GE Signa; GE Medical Systems; or Siemens Sonata; Siemens Medical Solutions). DWI was obtained using echo planar imaging (EPI) with a repetition time (TR) of 6000 ms to 10000 ms, an echo time (TE) of 78 ms to 101 ms, a field of view (FOV) of 22×22 cm, image matrix of 128×128, slice thickness 5 mm to 6 mm with 1 mm gap, and b values of 0 seconds/mm² and 1000 seconds/mm². Diffusion-weighted images were corrected for motion and eddy present distortions using the functional MRI of the brain (FMRIB) Linear Image Registration Tool (FLIRT 5.0; Oxford Centre for Functional Magnetic Resonance Imaging of the Brain).14 Average DWI maps as well as ADC maps were computed from these images. Perfusion-weighted images were acquired using dynamic susceptibility contrast EPI. Imaging parameters were TR 1500 to 1517 ms and TE 50 to 75 ms, with the same spatial resolution as for DWI. MTT and CBF maps were calculated using methods described previously.15,16 Fast spin-echo T2-weighted images were acquired with a TR of 4000 ms to 6500 ms, TE of 85 ms to 110 ms, FOV of 22×22 cm or 24×24 cm, acquisition matrix of 256×192 pixels or 320×256 pixels, and slice thickness of 5 mm to 6 mm with 1 mm gap.

Image Analysis

Two radiologists (B.O. and M.Z.), blinded to each other’s outlines, visually identified and sequentially outlined regions abnormal on DWI, MTT, CBF, CBV maps, and chronic T2-weighted images using a commercially available image display and analysis program (ALICE; Hayden Image Processing Solutions). Clinically relevant regions characterized by increased signal intensity on DWI and decreased intensity on the apparent diffusion coefficient maps were classified as acute infarct. Regions with hypoperfusion on PWI were characterized by increased signal intensity on DWI and decreased intensity on the apparent diffusion coefficient maps were classified as acute infarct. Regions with hypoperfusion on PWI were automatically produced by the software based on the slice thickness of mismatch tissue that eventually underwent infarction and was calculated as follows:

\[ \text{PML} = \frac{[(\text{Chronic T2 volume} - \text{DWI volume})/\text{MTT volume} - \text{DWI volume}) \times 100]}{\text{DWI volume}} \]

We calculated the “variance ratio” to quantify the contribution of variance imposed by measurement error (between-examiner variability) to the overall variance (variance of the average of the 2 measurements for each lesion or between-subject variability). This was estimated as follows:

\[ \text{Variance Ratio: } \frac{\text{SD}^2_{\text{measurement error}}}{\text{SD}^2_{\text{measured outcome}}} \]

The standard deviation for measurement error used in this calculation was obtained by plotting absolute difference between examiners for each subject across its mean as described by Bland and Altman.19 (Figure 1). To further demonstrate the importance of between-examiner variability, we estimated the proportion of sample size in a hypothetical research study that was attributable to the variance by measurement error, assuming an absolute minimum detectable difference in means of 20%, at the 0.05 2-tailed significance level, with a power of 0.8, for a 2-sample t test.

Statistical Analysis

Intraclass correlation coefficients (ICC) were computed to evaluate the correlation between examiners in lesion volumes measurements on DWI, MTT, CBF, CBV maps, and chronic T2 images. To quantify the magnitude of measurement error, the difference between examiners as a percentage of the mean volume per patient basis (relative paired difference or RPD) was calculated by the following formula:

\[ \text{RPD} = 100 \times \left( \frac{\text{volume}_{\text{examiner 1}} - \text{volume}_{\text{examiner 2}}}{\text{volume}_{\text{examiner 1}} + \text{volume}_{\text{examiner 2}}} \right)^{1/2} \]

Results

The study population included 36 male and 22 female patients. The mean age of the study population was 65 years (range, 18 to 96). The mean (±SD) time to initial MRI was 5.3 (±2.6) hours. Follow-up MRI was obtained after an
average of 37 (±69) days after the initial MRI. Cerebral infarcts were located within the middle cerebral artery territory in 4 patients, anterior cerebral artery territory in 3 patients, posterior cerebral artery territory in 4 patients, basilar artery territory in 3 patients, internal carotid artery territory in 1 patient, and in multiple territories in 2 patients. The stroke mechanism per the SSS-TOAST criteria was basilar artery territory in 3 patients, internal carotid artery patients, posterior cerebral artery territory in 4 patients, anterior cerebral artery territory in 3 patients, small artery occlusion in 1 patient, cardioaortic embolism in 25 patients, primary antiphospholipid antibody syndrome, and iatrogenic causes in 1 patient each, and undetermined causes in 13 patients (cryptogenic in 12 patients and other causes in 8 patients (acute arterial dissection in 5, and cerebral vasculitis, primary antiphospholipid antibody syndrome, and iatrogenic causes in 1 patient each), and undetermined causes in 13 patients (cryptogenic in 12 patients and multiple potential causes in 1 patient). The study population was composed of mainly large lesions rather than small subcortical infarcts. The image acquisition parameters did not differ across lesion volume quartiles in DWI, PWI, and final T2-weighted images.

Table 1 shows the between-subject mean (±SD) and median (interquartile range or IQR, 25% to 75%) volumes for each imaging sequence. There was very high correlation between examiners in volume measurements. The ICC values were equal to or greater than 0.95 for all of the images (Table 2). Despite high correlation, RPD analysis revealed that there were substantial differences in measurements per lesion basis depending on the type of imaging sequence and the size of ischemic lesion. Table 1 presents the mean (±SD) and median (IQR, 25% to 75%) RPD for each imaging sequence. There was a statistically significant negative correlation between RPD and lesion volume in all MRI modalities except for chronic T2-weighted images (Table 2). Because the distribution of RPD data were skewed toward smaller lesion volumes, RPD among examiners was calculated for each quartile of lesion volume (Figure 2). RPD was significantly different among quartiles of lesion volume in all MRI modalities (P=0.05 for DWI, P=0.02 for CBV and P<0.01 for MTT, CBF, and DWI/MTT mismatch), except for chronic T2-images (P=0.35).

The standard deviation of differences between 2 measurements for each lesion (between-examiner SD) was 5.1 mL for DWI, 27.0 mL for MTT, 29.4 mL for CBF, 11.3 mL for CBV, 4.8 mL for chronic T2-images, and 26.3 mL for DWI/MTT mismatch. Based on these standard deviations and standard deviations for averaged lesion volumes (between-subject SD) on Table 1, the variance ratio was calculated as follows: 0.02 for DWI, 0.05 for MTT, 0.10 for CBF, 0.06 for CBV, 0.01 for chronic T2, and 0.06 for DWI/MTT mismatch. Assuming an absolute minimum detectable difference in means of 20%, at the 0.05 2-tailed significance level, with a power of 0.8, for a 2-sample t test, the proportion of the sample size that was attributable to the variance of measurement error was 0.02 for studies using volume measurements on DWI, 0.05 for MTT, 0.10 for CBF, 0.06 for CBV, 0.01 for chronic T2, and 0.06 for DWI/MTT mismatch. If this calculation was repeated but for a study that was exclusively restricted to patients with infarcts on DWI within the first quartile in the present study (<4.0 mL), 28% of the estimated sample size was attributable to the measurement error (Table 3).

The PML analysis included 50 patients with DWI/MTT mismatch volume >20% of the DWI volume. The between-subject mean (±SD) and median (IQR, 25% to 75%) PML was 19.5% (±26.2) and 8.8% (0.9 to 27.1), respectively. The ICC between examiners for PML was 0.93 (P<0.01; Table 2). The mean (±SD) and median (IQR, 25% to 75%) RPD for PML were 40.2% (±57.9%) and 14.7% (0.0% to 40.1%), respectively. There was no change in RPD as a function of PML (P=0.21). Figure 1 shows that the agreement between examiners in PML measurements by plotting the difference in PML against the mean PML as displayed by Bland-Altman.
Plot. The between-examiner SD of measurement error for PML was 9.7 mL. The variance ratio for PML was 0.14.

Discussion

The present study provides quantitative evidence that is critical in designing and interpreting volume-based MRI studies in ischemic stroke. We found that volume measurements of abnormal regions on DWI and PWI by different examiners were highly correlated with each other, yet this high correlation did not translate into high agreement between examiners; correlation coefficients ranged between 0.93 and 0.99, but despite this, the measurement error defined as median RPD was 10.0% for DWI, 14.1% for MTT, 18.9% for CBF, 21.0% for CBV, 6.3% for chronic T2-images, 16.8% for DWI/MTT mismatch volume, and 14.7% for PML. The present study also systematically examined the interexaminer agreement rate as a function of infarct volume and found that measurement error significantly increased as the lesion volume decreased in most imaging sequences. For instance, RPD for DWI increased from 10% to 20% for lesions in the range of a typical lacunar infarct (<5 mL), suggesting that considerable amount of measurement error—on a percentage basis—can occur in studies dealing with small lesions.

The interexaminer agreement in DWI and PWI lesion volume measurements has been subject to several prior studies. Most of these studies used either kappa statistics or ICC to measure the reproducibility among examiners. Studies using kappa statistics reported good to excellent agreement rates whereas others that used ICC found correlation coefficients ranging from 0.84 to 0.99 for DWI\(^{8,10}\) and 0.87 to 0.99 for PWI.\(^{9,10}\) High correlation in volumetric stroke research is not unexpected because examiners measure the same lesions in a wide range of volumes. Both kappa statistics and ICC are measures of relation; they do not provide information for the magnitude of agreement (or variation) between examiners. As the present study also confirms, an excellent correlation does not necessarily denote excellent agreement.\(^{19}\)

There are various methods of defining the magnitude of agreement or measurement error. One approach is to estimate the percentage difference between group means (mean volume across all study subjects) for each examiner. This method is not always reliable. Assuming that there is no systematic error between examiners in such that one of the examiners always overestimates or underestimates lesion volumes with respect to the other, their group means will be very close to each other, when, in fact, there might be large differences between their volume measurements in individual subjects. For example, large differences were reported in a recent study dealing with infarct volumes in patients with stroke. Luby and coworkers reported <5% difference between group means of examiners for DWI and <1% for MTT maps, but also reported that the median of relative difference (RPD, see below) between volume measurements for each individual subject was 30% for DWI and 40% for MTT maps.\(^{13}\) A second approach to quantify measurement error is to calculate the difference between examiners as a percentage of estimated sample size attributable to measurement error.

![Figure 2. The figure shows mean relative paired differences (±95% confidence intervals) between examiners as a function of infarct volume. There was a statistically significant difference in mean RPD among quartiles for all images, except for chronic T2-weighted images and PML.](http://stroke.ahajournals.org/)

<p>| Table 3. The Impact of Measurement Error on Study Power for Various Quartiles of Infarct Volumes on DWI |
|-----------------|-----------------|-----------------|
| Mean (± SD) | Mean (± SD) | Percentage of Estimated Sample Size Attributable to Measurement Error |</p>
<table>
<thead>
<tr>
<th>Lesion Volume (mL)</th>
<th>Difference Between Examiners (mL)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1st quartile</td>
<td>1.7 (±1.3)</td>
<td>0.3 (±0.7)</td>
</tr>
<tr>
<td>2nd quartile</td>
<td>9.7 (±4.0)</td>
<td>1.0 (±2.0)</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>30.7 (±13.1)</td>
<td>0.6 (±3.6)</td>
</tr>
<tr>
<td>4th quartile</td>
<td>87.9 (±41.9)</td>
<td>1.7 (±9.4)</td>
</tr>
<tr>
<td>Overall group</td>
<td>32.1 (±39.7)</td>
<td>0.2 (±5.1)</td>
</tr>
</tbody>
</table>
of the mean volume (RPD).13,21 This method was used in the present study because relative terms are better suited to underline changes in measurement error as a function of another variable, such as infarct volume, than absolute differences. The 10% median RPD for DWI and 14% for MTT in the present study were much smaller than 30% and 40% rates reported by Luby et al.13 This was largely because the volume data in that study had been heavily skewed toward small infarcts, comparable to those in our first quartile; the RPD was also higher in the first quartile in the present study (Figure 2). Both studies, therefore, suggest that considerable amount of measurement error can occur in studies dealing with small lesions. A third approach, also known as Bland-Altman method, defines measurement error as the mean and SD of absolute difference between examiners for each subject (between-examiner SD).19 The major premise of this method is that it allows assessing the variance of measurement error. This, in turn, affects the study power and sample size because the variance used in power calculations is based on between-subject SD and has a component that comes from the measurement error. Because ischemic lesions occur in a wide range of volumes, volumetric MRI studies are inherently subject to high variance. Therefore, despite 6% to 21% relative difference (RPD) in measurements, the impact of additional variance imposed by measurement error on sample size was small; the proportion of sample size attributable to the variance from measurement error ranged from 0.01 to 0.14 depending on the type of image. In patients with small infarctes (<4 mL), this proportion was 0.28. Thus, volumetric measurements in stroke can be done with reasonable measurement error as long as such studies are not a priori limited to certain infarct groups such as very small infarcts. However, if a population of patients were expected to have very similar-sized lesion volumes, then measurement variance may become much more significant.

The present study is to first to assess the interexaminer agreement for a product of various MR images, the percentage mismatch, lost, or PML. Our results demonstrate that PML is a reliable marker for use as an imaging end point in stroke research. PML incorporates lesion volumes in acute and chronic images in a manner that allows the assessment of lesion growth as a function of territory at risk. The key benefit of using PML comes from the fact that, unlike individual lesion volumes in most imaging sequences, the interexaminer agreement does not change across the spectrum of PML values (Figure 2).

Although there is an intrinsic variance in lesion volume measurements on MRI, it is not inferior compared to other outcome scales used in therapeutic stroke trials. A direct comparison between MRI volumes and clinical outcome scales was not achievable in the present study because of unavailability of data on clinical outcome scores. Therefore, we estimated standard deviations of measurement error and measured outcome from studies published in the literature that listed per subject outcome scores for each rater. The between-examiner and between-subject standard deviations were 2.83 and 6.75 for NIHSS22 (based on retrospective assessment from medical records), 0.66 and 1.51,23 0.81 and 1.65,24 0.85 and 1.09,25 and 0.47 and 1.1125 for modified Rankin Scale, and 0.84 and 2.96 for Barthel Index,26 respectively. Assuming an absolute minimum detectable difference in means of 20%, at the 0.05 2-tailed significance level, with a power of 0.8, for a 2-sample t test, the proportion of sample size that was attributable to variance by measurement error was 0.18 for NIHSS, 0.17 to 0.60 for modified Rankin Scale, and 0.35 for Barthel Index. The proportion of sample size that was attributable to measurement error was only 0.14 for the PML. Although these calculations are based on diverse studies, and therefore cannot be regarded as definitive, they provide a hint of potential utility of volumetric measurements on MRI, such as PML.

In general, the between-examiner SD for absolute measurements was lower for initial DWI and final T2-weighted images than PWI. Likewise, RPD was higher for DWI and PWI maps than for T2-weighted images in the present study. The use of echo-planar imaging, which is characterized by lower resolution of acquisition matrix and sensitivity to degradation by susceptibility and other artifacts, has probably contributed to lower lesion conspicuity on PWI images, with respect to T2-weighted images. Higher spatial resolution of T2-weighted images may have also reduced the ambiguity between the lesion and the normal tissue. In addition, PWI is sensitive to inherent physiological differences of CBV and CBF in the gray and white matter, making differentiation between normal and abnormal tissue more difficult.16,27 It should also be noted that the lesion conspicuity differs among different perfusion maps; it is lower on CBV and CBF maps compared to MTT maps.16,27 Examiners in the present study were provided with brief clinical information to guide the examiners in outlining the clinically relevant lesion. Likewise, examiners evaluated each image in a sequential order (DWI-MTT-CBF-CBV-chronic T2). The order of lesion outlining was based on algorithms used in clinical patient evaluation in which DWI and MTT maps are initially viewed to identify the lower and upper bounds of the ischemic brain region, and CBF and CBV maps are evaluated later to assess the extent of perfusion failure within the region of interest defined by DWI and MTT. Providing clinical information and sequential image analysis can lead the examiner away from old lesions toward new and relevant lesions. This might produce volumes different than if such information were not available. The goal of this study was not to compare blinded evaluation methods but rather to estimate variance for the clinically relevant lesion, replicating the practical and routine evaluation of acute ischemic lesions. Our findings, therefore, are applicable to studies dealing with diagnosis, evolution, and prognosis of acute ischemic lesions.

Computer programs that automatically outline lesion borders are expected to eliminate examiner-dependent measurement error.28 Automated techniques, however, are far from perfect; currently available algorithms rely on thresholds that are calculated by taking the contralateral nonischemic hemisphere as the reference. The presence of acute or chronic ischemic lesions in the contralateral hemisphere may lead to miscalculations in thresholds. In addition, thresholds may vary because of differences between individuals and are subject to change depending on the interval between stroke onset and imaging. Moreover, physiological differences be-
tween gray and white matter may also be misinterpreted by automated techniques. There is currently much room for improvement for automated outlining techniques. As they become less examiner-dependent and thus less subject to human error, the variance ratio, and in turn, sample size needed to demonstrate a statistically significant difference will be smaller, or for a given sample size, the study power will be greater in stroke studies dealing with MRI lesion volumes.

The present study demonstrates that despite excellent correlation between examiners, substantial differences in volume measurements on MRI can occur. The magnitude of measurement error needs to be interpreted within the context of anticipated difference in clinical stroke research. For instance, because of measurement error, a true biological effect can be masked if the anticipated difference in volumetric studies is small. Likewise, stroke trials selecting patients a priori with respect to the presence of certain amount of DWI/MTT mismatch can be contaminated by cases without mismatch depending on the magnitude of measurement error. It is important to notice that interexaminer difference in measurement of study end points—both imaging and clinical—is a potential source of variance. Minimizing this source of variance may allow for more effective studies of stroke therapies with greater power or smaller sample size in the future.

Sources of Funding
This work was supported by US PHS NS38477. NIH grants RO1-NS38477-04 and P41-RK14075, and the MIND Institute.

Disclosures
None.

References
Interexaminer Difference in Infarct Volume Measurements on MRI: A Source of Variance in Stroke Research
Hakan Ay, E. Murat Arsava, Mark Vangel, Banu Oner, Mingwang Zhu, Ona Wu, Aneesh Singhal, Walter J. Koroshetz and A. Gregory Sorensen

Stroke. 2008;39:1171-1176; originally published online February 21, 2008; doi: 10.1161/STROKEAHA.107.502104
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2008 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/39/4/1171

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org/subscriptions/