Intra- and Interrater Reliability of Ischemic Lesion Volume Measurements on Diffusion-Weighted, Mean Transit Time and Fluid-Attenuated Inversion Recovery MRI

Marie Luby, MEng, MS; Julie L. Bykowski, MD; Peter D. Schellinger, MD, PhD; José G. Merino, MD; Steven Warach, MD, PhD

Background and Purpose—We investigated the intra- and interrater reliability of ischemic lesion volumes measurements assessed by different MRI sequences at various times from onset.

Methods—Ischemic lesion volumes were measured for intrarater reliability using diffusion-weighted (DWI), mean transit time (MTT) perfusion and fluid-attenuated inversion recovery (FLAIR) MRI at chronic (>3 days from stroke onset) time points. A single intrarater reader, blind to clinical information and time point, repeated the volume measurements on two occasions separated by at least 1 week. Interrater reliability was also obtained in the second set of patients using acute DWI, MTT and chronic FLAIR MRI. Four blinded readers performed these volume measurements. Average deviations across repeat measurements per lesion and differences between sample means between the two measurements were calculated globally, ie, across all sequences and time points, and per reader type for each sequence at each time point.

Results—There was good concordance of the mean sample volumes of the 2 intrarater readings (deviations were <4% and 2 mL globally, <2% and 2 mL for DWI, <6% and 7 mL for MTT, and <2% and 1 mL for FLAIR). There was also good concordance of the interrater readings (<5% and 2 mL globally).

Conclusions—Repeat measurements of stroke lesion volumes show excellent intra- and interrater concordance for DWI, MTT and FLAIR at acute through chronic time points. (Stroke. 2006;37:2951-2956.)

Key Words: acute stroke ■ brain imaging ■ magnetic resonance ■ neuroradiology ■ thrombolysis

Measurements of ischemic lesion volumes have been used in numerous imaging studies. Investigations of stroke outcomes and studies of the effects of therapeutic interventions including randomized clinical trials have included lesion volumes measured acutely with diffusion (DWI) and perfusion (PWI) weighted MRI and chronically with T2-weighted or fluid-attenuated inversion recovery (FLAIR) MRI. The use of volumetric measurements as an objective quantitative tool depends on intra- and interrater variability, but limited information is available on the reliability of these measurements across MRI sequence type and time from stroke onset. Reliabilities reported have been restricted to small sample sizes using manual or semiautomated techniques. Martel et al reported an intra- and interobserver repeatability coefficient defined as the 95% CI, of <6 mL for their semiautomatic method of measuring lesion volume using DWI (N=10). Ritzl et al reported an interobserver error of <3 mL for measurement of chronic FLAIR volume (N=8). Baird et al and Barber et al both reported an interrater variability of 5% with manual segmentation of lesion volumes on DWI. The current study provides intra- and interrater reliability estimates of lesion volume measurements, which may be relevant to stroke trial design and lesion data analysis.

Methods

Patients

This study is part of a prospective, natural history study of MRI in a consecutive series of tissue plasminogen activator (tPA)-treated patients at the National Institutes of Neurological Disorders and Stroke (NINDS) and Suburban Hospital, Bethesda, Md. The Institutional Review Board at NINDS and Suburban Hospital approved the study. From February 2000 through January 2005, 147 patients were treated with standard intravenous tPA; of those, 81 patients had an MRI before tPA treatment and were the subject of the intrarater analysis. Patients were eligible for this analysis if they received an MRI scan (“acute” scan) followed by standard intravascular tPA within 3 hours from stroke onset. The study design was to obtain serial MRI approximately 3 hours after the acute scan (“3-hour” scan), approximately 24 hours after the start of tPA (“24-hour” scan), and then at 5, 30, and 90 days after stroke onset (“chronic” scan). Chronic (>3 days from stroke onset) FLAIR volumes were taken from the latest available FLAIR for a given patient. Because of clinical care requirements, patient preferences, or...
death, all posttreatment MRI time points were not obtained in all patients. Only image time points and sequences with identifiable lesions were included.

A second confirmatory sample for intrarater analysis was obtained from August 1999 through July 2005 in 106 patients diagnosed with an acute cerebrovascular event, ischemic stroke, or transient ischemic attack. These patients had an acute MRI on average within 14 hours of onset and were not treated with tPA. The study design was identical to the tPA patients with the exception of the "3- and 24-hour" scans, which were not obtained. These same patients were also measured for the interrater analysis.

**Imaging Sequences**

MRI sequences were performed on a GE 1.5-T clinical scanner (Twinspeed; General Electric).

**Diffusion-Weighted Imaging**

In this study, the DWI spin-echo planar sequence included 20 contiguous axial oblique slices with b=0 and b=1000 s/mm², isotropically weighted, using repetition time (TR)/echo time (TE) artifact and produced apparent diffusion coefficient maps as necessary to identify positive DWI lesions. For MTT assessments, the reader paid particular attention to exclude hyperintensities attributable to the typical susceptibility artifacts adjacent to the paranasal sinuses. The reader assessed the MTT as not evaluable if the signal drop from the contrast did not produce at least a 10% drop or if there was significant patient motion causing inconsistency in the confirmation of the perfusion deficit. In evaluating FLAIR, the reader paid particular attention to preexisting chronic lesions present on the acute FLAIR images to avoid replication of these lesion areas in the current FLAIR volume calculations.

Measurements from each sequence were performed independently of the other time points except for the chronic FLAIR measurements in which reference to the acute FLAIR sequence aided identification of the index stroke. Within each time point, the reader had access to the other sequences to aid in accurate identification of the index lesion.

Patients’ intra- and interrater measurements from single slices of the acute DWI volumes are displayed in Figure 1. These same patients’ acute MTT and chronic FLAIR measurements from the corresponding single slices are displayed in Figures 2 and 3.

**Statistical Analysis**

Deviations between measurements were evaluated in two ways. The sample means, globally (across all sequences and time points) and within each reader type (intra- or interrater), time point, and sequence were compared for the two reads. The deviation per lesion was computed as the absolute value of the difference between the two readings; percent deviation per lesion was the per lesion absolute deviation divided by the average of the two readings. All volume data were skewed heavily toward mild. Cube root transformation of the raw volumes was performed to normalize the volumetric data. For example, for intrarater, acute DWI volume sample had a skewness of 2.1 and a kurtosis of 3.9; after transformation, the skewness was 0.7 and the kurtosis was −0.3. The absolute volume of the difference and the percentage deviation of the two readings, raw volume and cube root of volume, were calculated for each reader type and sequence and then averaged across all patients.

Bland-Altman plots were generated for the raw volume data to display the spread of data and the limits of agreement, specifically to
illustrate how many of the averaged data points lie within 2 SDs from the mean difference. The Bland-Altman plots were used to address the key question of whether one set of volumetric measurements is sufficiently representative or if two sets of measurements are required for providing the most accurate results. The 95% confidence limits are proposed as the repeatability coefficients of one type of measurement for another, i.e., one set of volumetric measurement is sufficient rather than requiring two sets in this particular study.

**Results**

In the intrarater analysis, DWI had measurable lesions in 68 patients at the acute time point, 62 at 3 hours and 41 at 24 hours. MTT images were measured in 62 patients at the acute time point, 38 at 3 hours and 24 at 24 hours. Chronic FLAIR images were measured in 46 patients. The volume statistics by sequence, DWI, MTT, and FLAIR are listed in Table 1. Only DWI, MTT, and FLAIR measurements in which at least one volume measurement is nonzero are reported in Table 1 and all subsequent tables and figures. The corresponding sample sizes are provided for each reader type, sequence, and time point. The correlation coefficients and cube root statistics between the two intrarater readings (N=68, 62, and 41) for DWI at acute, 3-hour, and 24-hour time points are reported in Table 1. Supplemental Figure I (panels A through C; available online at http://stroke.ahajournals.org) display the Bland-Altman plots for the DWI raw volume data for the intra-acute, 3-hour, and 24-hour time points. The upper and lower limits, shown as thick black solid lines, were calculated for each plot to represent ±2 SD from the mean. A total of 90%, 92%, and 90% data points were within these boundary limits for the DWI data, respectively. There were no significant changes with respect to time point in the reliability of the DWI data.

The confirmatory intrarater sample statistics are reported in Table 1. Supplemental Figure I (panel D) contains the Bland-Altman plot for the DWI raw volume data. At total of 96% data points were within the boundary limits.

The confirmatory intrarater sample statistics are reported in Table 1. Supplemental Figure I (panel D) contains the Bland-Altman plot for the DWI raw volume data. At total of 96% data points were within the boundary limits.

The mean volumes, absolute volume (mL) differences, and percent deviations along with the cube root statistics between the two interrater readings at acute DWI (N=103) time points are reported in Table 2. Supplemental Figure I (panel E) displays the interrater acute DWI Bland-Altman plot. A total of 94% data points were within the boundary limits. By excluding DWI lesions <10 mL in volume, the percent

**Sequence: Diffusion-Weighted Image**

The mean volumes, absolute volume (mL) differences, and percent deviations along with the cube root statistics between the two interrater readings at acute DWI (N=103) time points are reported in Table 2. Supplemental Figure I (panel E) displays the interrater acute DWI Bland-Altman plot. A total of 94% data points were within the boundary limits. By excluding DWI lesions <10 mL in volume, the percent

**Results**

In the intrarater analysis, DWI had measurable lesions in 68 patients at the acute time point, 62 at 3 hours and 41 at 24 hours. MTT images were measured in 62 patients at the acute time point, 38 at 3 hours and 24 at 24 hours. Chronic FLAIR images were measured in 46 patients. The volume statistics by sequence, DWI, MTT, and FLAIR are listed in Table 1. Only DWI, MTT, and FLAIR measurements in which at least one volume measurement is nonzero are reported in Table 1 and all subsequent tables and figures. The corresponding sample sizes are provided for each reader type, sequence, and time point. The correlation coefficients and cube root of the raw volume data are presented in supplemental Table I, available online at http://stroke.ahajournals.org.

The confirmatory intrarater sample had measurable lesions in 98 patients at the acute DWI, 72 patients at the acute MTT, and 29 patients at the chronic FLAIR time points. The volume statistics by sequence, DWI, MTT, and FLAIR are listed in Table 1.

In the interrater analysis, there were measurable lesions in 103 patients at the acute DWI, 77 patients at the acute MTT, and 29 patients at the chronic FLAIR time points. The volume statistics by sequence, DWI, MTT, and FLAIR are listed in Table 2 for the interrater analysis.

**Sequence: Diffusion-Weighted Image**

The mean volumes, absolute volume (mL) differences, and percent deviations along with the cube root statistics between the two interrater readings (N=68, 62, and 41) for DWI at acute, 3-hour, and 24-hour time points are reported in Table 1. Supplemental Figure I (panels A through C; available online at http://stroke.ahajournals.org) display the Bland-Altman plots for the DWI raw volume data for the intra-acute, 3-hour, and 24-hour time points. The upper and lower limits, shown as thick black solid lines, were calculated for each plot to represent ±2 SD from the mean. A total of 90%, 92%, and 90% data points were within these boundary limits for the DWI data, respectively. There were no significant changes with respect to time point in the reliability of the DWI data.

The confirmatory intrarater sample statistics are reported in Table 1. Supplemental Figure I (panel D) contains the Bland-Altman plot for the DWI raw volume data. At total of 96% data points were within the boundary limits.

The mean volumes, absolute volume (mL) differences, and percent deviations along with the cube root statistics between the two interrater readings at acute DWI (N=103) time points are reported in Table 2. Supplemental Figure I (panel E) displays the interrater acute DWI Bland-Altman plot. A total of 94% data points were within the boundary limits. By excluding DWI lesions <10 mL in volume, the percent
deviations (average, median, and SD of 25.01, 18.47, and 26.81, respectively) were comparable to those of the intrarater readings.

In addition to the measured samples, 11 and 12 acute DWI sequences were evaluated separately as normal, zero volumes, for the intra- and interrater readings. Included in the measured samples, there were 4 and 11 acute DWI sequences with one intrarater and interrater readings as zero volume. An additional five and four DWI sequences at 3 and 24 hours were evaluated separately as normal for both intrarater readings. There was one DWI sequence at 3 hours with one intrarater read as normal. There were three DWI sequences at 24 hours with one intrarater reading as normal.

**Sequence: Mean Transit Time**

The mean volumes, absolute volume (mL) differences, and percent deviations between the two intrarater readings (N=62, 38, and 24) for MTT at acute, 3-hour, and 24-hour time points are reported in Table 1. Supplemental Figure I (panels F through H) display the Bland-Altman plots for the MTT raw volume data for the intra- acute, 3-hour, and 24-hour time points. A total of 94%, 92%, and 92% data points were within the boundary lines for the MTT data, respectively. There were no significant changes with respect to time point in the reliability of the MTT data.

The confirmatory intrarater sample statistics are reported in Table 1. Supplemental Figure I (panel I) displays the Bland-Altman plot for the MTT raw volume data. At total of 99% data points were within the boundary limits.

The mean volumes, absolute volume (mL) differences, and percent deviations along with the cube root statistics between the two intrarater readings at acute MTT (N=77) time points are reported in Table 2. Supplemental Figure I (panel J) displays the interrater acute MTT Bland-Altman plot. A total of 95% data points were within the boundary limits. When excluding the MTT deficits <10 mL in volume, the percent deviation statistics improved but were still not comparable to the intrarater readings.

In addition to the measured samples, seven and 34 acute MTT sequences were evaluated separately as normal, zero volumes, for the intra- and interrater readings. Included in the measured samples, there were one and 21 acute MTT sequences with one intrarater and interrater readings as zero volume. An additional five and nine MTT sequences at 3 and 24 hours were evaluated separately as normal for both intrarater readings. There were three MTT sequences at 24 hours with one intrarater read as normal.

**Sequence: Fluid-Attenuated Inversion Recovery**

The mean volumes, absolute volume (mL) differences, and percent deviations between the two intrareadings (N=46) for chronic FLAIR time points are reported in Table 1. Supplemental Figure I (panel K) displays the Bland-Altman plot for the intrarater chronic FLAIR time points. A total of 91% data points were within the boundary lines. The confirmatory intrarater sample statistics are reported in Table 1. Supplemental Figure I (panel L) contains the Bland-Altman plot for the FLAIR raw volume data. A total of 89% data points were within the boundary limits.
The mean volumes, absolute volume (mL) differences, and percent deviations along with the cube root statistics between the two interrater readings at chronic FLAIR (N=29) time points are reported in Table 2. Supplemental Figure I (panel M) displays the interrater FLAIR Bland-Altman plot. A total of 93% data points were within the boundary limits. By excluding the FLAIR lesions <10 mL, the sample size was reduced to N=6; however, the percent deviations (average, median, and SD of 39.74, 39.73, and 27.69, respectively) were then comparable to those of the intrarater readings.

In addition to the measured sample, 12 chronic FLAIR sequences were evaluated separately as normal, zero volumes, for both intrarater readings. Included in the measured sample, there were five FLAIR sequences with only one intrarater reading as zero volume. There were no FLAIR sequences evaluated separately as normal, zero volumes, for both intrarater readings. Included in the measured sample, there were four FLAIR sequences with only one interrater reading as zero volume.

A subset of nine intrarater patients with all sequences and time points was calculated for comparison purposes to Table 1. The mean volumes (N=9) of the two intrarater readings for DWI were 35.3 mL and 35.9 mL, 46.47 mL and 42.14 mL, and 78.53 mL and 75.89 mL at acute, 3 hours, and 24 hour time points, respectively. The mean volumes (N=9) of the two intrarater readings for MTT were 164.56 mL and 162.04 mL, 134.73 mL and 138.07 mL, and 86.39 mL and 83.83 mL at acute, 3 hours, and 24 hour time points, respectively. The mean volumes (N=9) of the two intrarater readings for FLAIR were 58.05 mL and 58.59 mL. The volume statistics of this subset of nine patients was consistent with those of the larger group of patients contained in Table 1.

The global mean volumes (N=341) of the two intrarater readings were 63.0 mL and 60.9 mL. The absolute volume (mL) differences between the reads were 6.54±13.27/1.92 (0.4 to 6.86) (mean±SD/median [interquartile range {IQR} 25% to 75%]). The percent deviations between the reads were 26.94±46.51/11.14 (0.04 to 0.24).

The global mean volumes (N=198) of the two confirmatory intrarater readings were 30.5 mL and 28.4 mL. The absolute volume (mL) differences between the two readings were 5.84±26.52/1.06 (0.24 to 4.3) (mean±SD/median [IQR 25% to 75%]). The percent deviations between the reads were 48.37±66.26/18.28 (0.07 to 0.51).

The global mean volumes (N=209) of the two interrater readings were 28.9 mL and 27.6 mL. The absolute volume (mL) differences between the reads were 8.69±22.76/1.31 (0.34 to 5.98) (mean±SD/median [IQR 25% to 75%]). The percent deviations between the reads were 67.04±68.87/35.64 (0.18 to 0.87).

**Discussion**

The volume of injury, infarction, and ischemia has assumed an increasingly important role in stroke research in general and in stroke clinical trials in particular. Yet, relatively little attention has been paid to the reliability of volumetric measures and the contribution of measurement error to the overall variance of these outcomes in stroke studies. Because measurement error may potentially obscure true biologic effects and is potentially controllable by the investigator, it is important to understand the sources and magnitude of measurement error.

In the present study, technical variables that could affect lesion volume measurement such as the MRI scanner type, pulse sequence parameters, image processing, and analysis software were held constant. Because the intrarater reader’s experience performing stroke volume measurements with this software has been extensive over a period of years, the skill and consistency in measurement technique can be assumed to be constant over the period of the intrarater study. Thus, we have characterized the intrarater reproducibility independent of other sources of measurement error.

In principle, programs to measure lesion volumes in a completely automated fashion could eliminate this source of error. However, because regions of abnormality on these MRI sequences often lack sharp contrast boundaries with the normal brain and may overlap in signal intensity with nonpathologic structures, investigator judgment is needed to avoid inaccuracies. The intrarater study attempts to quantify reader subjectivity, the main concern with interpretations of the images by multiple readers. To reduce the measurement error, the intrarater reader trained the three additional intrarater readers in the same image processing techniques and software. The intrarater readers were also required to be familiar with the clinical reading of serial stroke MRI scans. Furthermore, the patients and imaging protocol used in the intrarater analysis were from the same time period, hospital, and stroke population as the intrarater analysis.

The intrarater repeatability seen for DWI was <1 mL at the acute and 3-hour time points and <2 mL at the 24-hour time point. The overall intrarater percent difference ranged from 2% to 5% for DWI. The intrarater percent difference was <5% for acute DWI. For DWI, the error of the lesion volume measurement was caused by two main sources. There were six outliers readily identified by the Bland-Altman plots in the intrarater DWI readings caused by reader differences in interpretation of subtle changes, differences in exclusion of sulcal areas, and in one case, misidentification of a chronic lesion as acute. Measurement error was also attributable to significantly smaller lesions seen for the intrarater readings.

The overall intrarater MTT percent difference ranged from 1% to 6%. The intrarater repeatability coefficient seen for MTT was <1 mL at the 3-hour time point but <7 mL and <6 mL at the acute and 24-hour time points, respectively. The intrarater repeatability coefficient seen for acute MTT was <1 mL. The intrarater MTT percent difference was <1%. The larger standard deviations seen with MTT are in part attributable to the larger lesions present in this sequence and the lower spatial resolution in the perfusion sequence acquisitions. This was most evidenced by the four outliers in the intrarater MTT readings, which are most readily identified by the Bland-Altman plots.

The intrarater repeatability coefficient seen for FLAIR was <0.6 mL and the overall percent difference of <1.5%. The most consistent lesion volume was seen with FLAIR attributable in part to the increased spatial resolution of this sequence compared with DWI and MTT as well as the resolution of edema at the chronic time point allowing for a more stable measurement. The intrarater repeatability coeffi-
cient seen for FLAIR was <2 mL but an overall percent difference of <50%. The decreased consistency in the interrater lesion volume measurement with FLAIR was mainly attributable to the significantly smaller lesions seen (approximately 6 mL on average). The two outliers displayed in the corresponding Bland-Altman plot in the interrater FLAIR readings were caused by one reader, including areas of cavitations that were excluded by the other reader. Overall, for the DWI, MTT, and FLAIR volumes by requiring a minimum lesion size, ie, 10 mL, the reproducibility of the measurements, especially percent deviation, is improved.

Quantitative volumes by a single expert reader can provide highly consistent and repeatable results of lesion volumes on DWI, MTT, and FLAIR. The variability seen with the interrater measurements increased compared with the intrarater measurements; however, they also demonstrated consistent and repeatable results. This study provides a resource of volumetric statistics from a large homogenous stroke population to researchers interested in potential selection of particular sequences and series of time points for further investigation.

**Sources of Funding**

This research was supported by the Division of Intramural Research of the National Institute of Neurological Disorders and Stroke, National Institutes of Health.

**Disclosures**

None.

**References**

Intra- and Interrater Reliability of Ischemic Lesion Volume Measurements on Diffusion-Weighted, Mean Transit Time and Fluid-Attenuated Inversion Recovery MRI
Marie Luby, Julie L. Bykowski, Peter D. Schellinger, José G. Merino and Steven Warach

Stroke. 2006;37:2951-2956; originally published online November 2, 2006;
doi: 10.1161/01.STR.0000249416.77132.1a
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2006 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/37/12/2951

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/