Reproducibility of Measurements of Cerebral Infarct Volume on CT Scans

H.B. van der Worp, MD; S.P. Claus, MD; P.R. Bär, PhD; L.M.P. Ramos, MD; A. Algra, MD; J. van Gijn, MD, FRCP; L.J. Kappelle, MD

Background and Purpose—Infarct volume is increasingly used as an outcome measure in clinical trials of therapies for acute ischemic stroke. We tested which of 5 different methods to measure infarct size or volume on CT scans has the highest reproducibility.

Methods—Infarct volume and total intracranial volume were measured with Leica Q500 MCP image analysis software, or with a caliper, on 38 CT scans of patients who participated in the Tirilazad Efficacy Stroke Study II (TESS II). The scans were performed 8 days (±2 days) after the onset of symptoms. The 5 methods tested were based on (1) semiautomated pixel thresholding, (2) manual tracing of the perimeter, (3) a stereological counting grid, (4) measurement of the 3 largest diameters, and (5) the single largest diameter. The measurements were performed independently by 2 observers; the first observer performed all measurements twice.

Results—The single largest diameter did not correlate well with infarct volume. Of the other methods, manual tracing of the perimeter of the infarct had the lowest intraobserver and interobserver variability: coefficients of variation were 8.6% and 14.1%, respectively. For total intracranial volume, manual tracing also provided the highest reproducibility: intraobserver and interobserver coefficients of variation were 3.3% and 4.9%, respectively.

Conclusions—Manual tracing of the perimeter is the most reproducible method for measuring the volumes of the infarct and the total intracranial space in multicenter trials of therapies for acute ischemic stroke. (Stroke. 2001;32:424-430.)

Key Words: cerebral infarction ■ computer-assisted image processing ■ stroke assessment ■ tomography, x-ray computed
selected from 3 subgroups on the basis of the following classification of the infarct: (1) cortical and subcortical; (2) subcortical only (diameter >1.5 cm); and (3) lacunar (diameter 1.5 cm).

Apart from the time frame and a ban on the use of contrast, the TESS II protocol provided no guidelines for the performance of CT scans. The scans in the present study came from 24 different centers in Europe, Australia, and New Zealand, and at least 13 types of machines were used (machine type was not given on every scan). Slice thickness of the different scans ranged from 3 to 10 mm, with the exception of one scan with a slice thickness of 16 mm. The number of slices per scan ranged from 7 to 28.

For volume measurements, the CT scan films were placed on a transillumination base under a camera. Each image was captured and digitized. Of each scan, the images were calibrated by means of the internal calibration marks. If necessary, contrast was enhanced. Three routines were developed with Leica Q500 MCP image analysis software: (1) The pixel thresholding technique7,8 (Figure 1A) was based on thresholds in signal intensity for both ICV and infarct. First, each intracranial pixel with an attenuation lower than bone was assigned to the ICV. Thereafter, each pixel with a signal intensity identical to or below that of the infarct was color coded. For each slice, this threshold in signal intensity was interactively determined by the investigator and depended on the density scale settings of the CT scan and the attenuation of the infarct. After color-coded compartments other than the lesion, such as cerebrospinal fluid, had been rejected by the investigator, the computer calculated the area of the infarct in the particular slice. Infarct was distinguished from cerebrospinal fluid on the basis of anatomy and signal intensity. (2) In the manual tracing technique6 (Figure 1B), the perimeter of the area of abnormal low attenuation was traced on each CT slice showing the infarct. The same procedure was performed for the total intracranial space. (3) The stereological counting grid,9 based on the Cavalieri principle12 (Figure 1C), is a single-lattice square grid with an interpoint distance of 1.5 cm superimposed over the brain image. In an earlier study of infarct volume measurement, this distance was chosen on the basis of pilot studies and published nomograms.9 For technical reasons, the interpoint difference was 1.4 cm in 4% of the cases and 1.6 cm in 2%. Each grid point within the intracranial space was assigned to either normal tissue or infarct. If a grid intersection fell on a boundary between 2 compartments, the point was assigned to the compartment in the lower left square. The points assigned to normal tissue or infarct were multiplied by the area of a single square. For each of the methods, all measured areas were multiplied by the slice distance to obtain the total volumes. (4) In the fourth method, the largest diameter (A) of the infarct and its largest perpendicular diameter (B) were measured with a caliper (Figure 1D). The third, vertical diameter (C) was determined by summing the thicknesses of the slices in which the lesion was visible. Infarct volume was calculated according to the formula 0.5\(A \times B \times C\).10 (5) Finally, the single largest visible diameter was measured.11

The observers, who had 7 and 4 years of experience in the field of vascular neurology, respectively, performed the measurements independently, but in case of doubt about the presence of an infarct, a decision could be made by mutual agreement. To avoid a learning effect during the study, the measurements were performed after a training session. To test the intraobserver variability, all scans were reassessed by one author (H.B. van der W.) after all first measurements had been completed. The time between the first and second assessments of a scan with use of the same method was at least 4 weeks.

Intraobserver variability was determined for each method by comparing data of one author (H.B. van der W.) at 2 different

![Figure 1. CT scan of a large subcortical and cortical infarct illustrating the tested methods to measure infarct volume: pixel thresholding (A); manual tracing of the perimeter of the infarct (B); use of a stereological counting grid (C); and measurement of the largest diameters (D).](image-url)
measuring sessions; interobserver variability was determined by comparing the values of the first measurements of 2 authors (H.B. van der W. and S.P.C.). Variability of measurements was analyzed according to the method of Bland and Altman,13 including scatterplots showing the difference between 2 measurements against their mean, and by calculating the coefficient of variation for each method. This measure is equal to the intraobserver or interobserver error (SD of the mean difference) times 100, divided by \( \sqrt{2} \) times the pooled mean values. Differences between the mean volumes of both the infarct and the total intracranial space as measured by the first investigator with the 4 different volume measurement techniques were compared by means of the paired Student’s \( t \) test.

Results

Seven of the 45 selected scans were excluded: 3 were incomplete, 1 did not show an infarct, 1 was contrast enhanced, 1 had no calibration marks, and 1 had several different calibration marks. The intraobserver and interobserver coefficients of variation of the infract volumes or sizes are presented in Table 1. Scatterplots showing the difference between 2 measurements against their mean for methods 1 through 4 are presented in Figure 2. Except for method 3 (stereological counting grid), the interobserver variability was larger than the intraobserver variability. For methods 1 through 4 there were no significant differences between the mean values of the infract volumes. Both the intraobserver and interobserver variability were smallest for the manual tracing method.

The reproducibility was highest for the 1-dimensional method 5, but Figure 3 shows that this method is not very accurate because there is no good correlation between the largest diameter and the volume of the infarct as measured with method 2. Figure 4 suggests a fair correlation between infract volumes as measured with semiautomated method 2 (perimeter tracing) and those measured according to the nonautomated method 4 (3 diameters); the coefficient of variation of these 2 methods was 23.1%.

Manual tracing was the only method that had a satisfactory reproducibility for the measurement of the volume of lacunar infarcts (Table 2). When the stereological counting grid was used, a lacunar infarct was missed in 36% of the cases because none of the grid intersections fell on the infarct.

Values for total intracranial volume are shown in Table 3. The average ICV measured with the counting grid (method 3) was 84 mL (95% CI, 28 to 139 mL) smaller than that measured with thresholding (method 1) and was 87 mL (95% CI, 34 to 140 mL) smaller than with the tracing method (method 2). For all 3 methods, the interobserver variability was larger than the intraobserver variability, but the reproducibility was much better than that for infract volume. The intraobserver variabilities for the pixel thresholding and manual tracing methods were comparable, but the latter method had the smallest interobserver variability.

The time needed for the complete measurement of a single CT scan was highly dependent on the total number of slices and on the number of slices showing the infarct, but it generally varied between approximately 10 and 30 minutes for each of the 3 semiautomated techniques. Measurements of infract volume were on average more rapidly performed with method 4 (3 diameters; approximately 3 minutes) than with each of the semiautomated techniques (2 to 10 minutes, depending on the number of slices showing the infarct).

Discussion

In this study we tried to find the most accurate and reproducible method for measuring cerebral infract volume on CT scans obtained in large multicenter clinical trials. In these trials, CT scans are likely to vary considerably in quality, density scales, calibration, and slice thickness. The scans in our study came from 24 different centers in Europe, Australia, and New Zealand that participated in the TESS II trial, and all these characteristics were indeed highly variable. Given these circumstances, the manual tracing method had the lowest intraobserver and interobserver variability for both infract volume and intracranial volume.

The variability of measurements was analyzed according to the method of Bland and Altman13 and by calculating the coefficient of variation for each method. The low variability of the manual tracing method was demonstrated by the small limits of agreement calculated according to Bland and Altman and by the low coefficient of variation. The mean difference in the scatterplot was nearly zero, showing that there were no systematic discrepancies between the first and second measurements. The limits of agreement in the intraobserver study for the manual tracing method indicate that differences up to 16 mL are possible for the larger infarcts (Figure 2b). Such differences may be acceptable in clinical trials that use cerebral infract volume as an outcome.
Figure 2. Scatterplots according to Bland and Altman\textsuperscript{13} presenting the intraobserver (a through d) and interobserver (e through h) variability of the 4 tested methods to measure infarct volume: pixel thresholding (a and e); manual tracing of the perimeter of the infarct (b and f); use of a stereological counting grid (c and g); and measurement of the largest diameters (d and h).
For all 3 semiautomated techniques, the reproducibility of ICV was much better than that of infarct volume. The most probable explanation is that the border of the intracranial space is better defined than that of most infarcts. The identification of the boundary of an infarct is therefore more strongly dependent on the interpretation of the investigator. Others have also found that interobserver agreement improves when compartments are better demarcated.9

Only manual tracing of the infarct proved to be reliable for measuring lacunar infarct volumes. Because of the small size of these infarcts, a small deviation in absolute measures readily corresponds to a substantial relative error. Irrespective of the question of whether measuring the volume of lacunar infarcts in acute stroke trials is meaningful, measurement on CT scans with one of the other tested methods is not worthwhile.

The variability in attenuation within the infarct may have contributed to the unfavorable results of the method based on pixel thresholding. This technique required the threshold to be set at such a level that the entire infarct was included. In this way, a considerable amount of normal tissue was primarily designated as infarct. Although erroneously classified areas had to be subsequently rejected by the investigator, normal tissue with a signal intensity similar to or lower than that of the infarct may have been included in the infarct to a varying extent.

For the stereological counting grid, the largest interpoint distance that produced the minimum error has been reported to be 1.5 cm.9 However, with this grid size we observed intraobserver and interobserver variabilities that were much higher than with the other methods. In addition, in 36% of the cases a lacunar infarct was missed because none of the grid intersections fell on the infarct. Both problems would probably be reduced if a smaller grid size was used, but this would make the method much more time-consuming.

Our stereological method differed slightly from the method that has been described before.9 In the previous study, the number of grid intersection points that fell on the infarct was divided by the total number of intersection points and then multiplied by the ICV that was obtained with manual tracing. In this study we directly multiplied each point with the area of each square and the slice thickness. Although this difference may lead to a minor difference in absolute infarct volumes, it will not affect the variability as expressed by a percentage for both variants of the method.

Obviously, it was not possible to compare the obtained volumes for intracranial space and infarct with the gold standard of measurement after autopsy. However, despite the differences in reproducibility, the mean infarct volumes as obtained with the 4 multidimensional techniques were virtually similar, suggesting accuracy. For each of the semiautomated methods, the mean values for intracranial volumes were within the normal range of 1100 to 1400 mL,14 but for unknown reasons the values obtained with the counting grid were substantially smaller than those obtained with the other methods.

Among the methods tested, measurement of the largest diameter of the infarct was the most rapid and reproducible.
The good intraobserver and interobserver variability may well be explained by the fact that only a single, 1-dimensional measurement had to be performed. However, the relation between the largest diameter of the infarcts and their volume was insufficient (Figure 3). This disadvantage is overcome by the inclusion of both largest perpendicular diameters and the slice thickness in a formula to obtain infract volume.10,15 This method has earlier been described to measure volumes of intracerebral hemorrhage.16 Despite its slightly lower reproducibility, the fair correlation between infract volumes as measured according to this technique and those obtained with manual tracing makes this simple method acceptable for investigators who do not have access to an image analysis system.

Although not formally tested, the time spent on the complete measurement of one CT scan was more or less the same for the 3 semiautomated techniques. This period of time was highly dependent on the total number of slices and on the number of slices showing the infract, but it generally varied between 10 and 30 minutes. In comparison with the large amount of time spent on inclusion, treatment, follow-up, and data collection of each patient included in an acute Stroke trial, the time involved in obtaining this additional outcome measure appears acceptable.

In recent years, the use of MRI for measurement of brain volume has expanded enormously.17–21 For this purpose, the advantages of MRI over CT include a better spatial resolution, earlier detection of ischemic tissue, and the possibility of performing different kinds of imaging in one sample.22 Some of these techniques put great demands on hardware, software, and personnel. In larger studies, these high requirements cannot always be met. Although MRI is the technique of choice for infract measurements in smaller studies performed in academic and other specialized centers, CT is more suitable in larger multicenter trials because of its higher availability.

In conclusion, manual tracing of the perimeter is the most reproducible method for measuring the volume of a brain infract and the total intracranial space in large multicenter trials of therapies for acute ischemic stroke. The method is rapid, reliable, and probably accurate, despite technical differences between the CT scans.

Acknowledgments
This study was supported by a grant from the Janivo Foundation to Dr van der Worp. Purchase of the Leica Q500 MCP image analysis system was made possible by a grant from Pharmacia & Upjohn. We thank Peter Sodaa for designing the routines for the volume measurements and for his assistance during the phase of data collection. We thank the TESS II investigators, Pharmacia & Upjohn, and Professor J.M. Orgogozo for giving us the opportunity to make use of CT scans of the TESS II trial.

References

### TABLE 3. Intraobserver and Interobserver Coefficients of Variation, Mean Volumes, and Limits of Agreement of the 3 Methods for Intracranial Volume Measurement in 38 CT Scans

<table>
<thead>
<tr>
<th>Method</th>
<th>CV, %</th>
<th>Volume</th>
<th>Limits of Agreement*</th>
<th>Interobserver</th>
<th>CV, %</th>
<th>Volume</th>
<th>Limits of Agreement*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Thresholding</td>
<td>3.0</td>
<td>1322±128</td>
<td>~119 to 97</td>
<td>7.1</td>
<td>1232±124</td>
<td>~267 to 251</td>
<td></td>
</tr>
<tr>
<td>2. Tracing</td>
<td>3.3</td>
<td>1325±119</td>
<td>~120 to 122</td>
<td>4.9</td>
<td>1232±131</td>
<td>~174 to 188</td>
<td></td>
</tr>
<tr>
<td>3. Counting grid</td>
<td>4.8</td>
<td>1238±114</td>
<td>~170 to 161</td>
<td>10.6</td>
<td>1237±132</td>
<td>~364 to 364</td>
<td></td>
</tr>
</tbody>
</table>

CV indicates coefficient of variation. Volumes are in milliliters; mean volumes are based on the 2 measurements for the respective analysis.

*The limits of agreement between the 2 measurements were calculated according to the method described by Bland and Altman13: mean difference between the 2 measurements = 1.96 × SD of the mean difference.
Reproducibility of Measurements of Cerebral Infarct Volume on CT Scans
H. B. van der Worp, S. P. Claus, P. R. Bär, L. M. P. Ramos, A. Algra, J. van Gijn and L. J. Kappelle

Stroke. 2001;32:424-430
doi: 10.1161/01.STR.32.2.424

Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2001 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/32/2/424

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/