Carotid Plaque Characterization by Duplex Scanning: Observer Error May Undermine Current Clinical Trials

J.A.C. Arnold, MSc; K.B. Modaresi, PhD; N. Thomas, PhD; P.R. Taylor, MChir; T.S. Padayachee, PhD

Background and Purpose—Clinical studies currently in progress are using subjective methods to characterize plaque morphology from ultrasound images. However, there are few studies on the intraobserver and interobserver variability of these classifications. This study was designed to assess these variables.

Methods—Grading of plaque morphology from ultrasound images, stored both digitally and to hard copy, was performed by 2 classification schemes. Interobserver agreement was determined by 4 observers. Within-observer agreement was performed at intervals for up to 6 months. Accuracy of the 2 methods was determined by comparison with histology.

Results—Within- and between-observer agreement was moderate to good for full-color digital image analyses, with pooled $\kappa$ values of $\kappa_p=0.49\pm0.10$ and $\kappa_p=0.62\pm0.07$ for the 2-category method and $\kappa_p=0.53\pm0.06$ and $\kappa_p=0.52\pm0.05$ for the 4-category method, respectively. Hard copy data analyses gave lower $\kappa$ values. The more experienced observers produced higher within-observer agreements and higher correlation with histology.

Conclusions—Reproducible grading of ultrasound images is not consistently achievable among experienced observers, and within-observer agreement may vary with time. The current subjective ultrasound characterization of carotid plaque morphology used in clinical trials may be associated with unacceptable levels of reproducibility in some centers. Variability between observers may be reduced by using the simpler 2-category grading of plaque morphology to interrogate full-color digitally stored images. Observer agreement should be audited regularly.

(Stroke. 1999;30:61-65.)

Key Words: atherosclerosis ■ carotid stenosis ■ ultrasonography, Doppler, duplex

The accepted criterion for surgical selection of patients with symptomatic carotid disease is a stenosis of $>70\%$ reduction in lumen diameter of the internal carotid artery (ICA). Current imaging of patients with ICA disease has focused on the accurate measurement of the severity of ICA stenosis. Increasingly, this has been achieved solely by noninvasive duplex ultrasound with color flow mapping. This modality also provides images of the artery wall and atherosclerotic plaque morphology, but the latter has not been used for selection of patients for surgery, although it is becoming increasingly recognized as important in the development of symptoms. For example, several authors have reported that plaques in the carotid and coronary arteries associated with large lipid pools or soft extracellular lipid are more prone to rupture and produce emboli and therefore cause symptoms.

Ultrasound has been used to evaluate plaque morphology, and, more recently, multicenter trials investigating patients with cerebrovascular disease have included plaque characterization in their protocols. The Asymptomatic Carotid Surgery Trial used a subjective method for ultrasound plaque characterization based on the approach described by Gray-Weale et al. The Asymptomatic Carotid Stenosis and Risk of Stroke study used the Gray-Weale method with additional categories for calcified and ulcerated plaques, while the European Carotid Plaque Study Group commented on both plaque echogenicity and surface characteristics. In all these studies, grading of plaque morphology was largely based on the ultrasound gray scale appearance, which was assessed subjectively by visual inspection of ultrasound images.

Thus, a range of criteria and classification schemes for assessing plaque morphology are currently in use, but there is no consensus on protocol or methodology between trials. Subjective grading has been compared with histology; however, a wide range of agreements has been reported. Furthermore, there are few studies on reproducibility, which is very important when amalgamating data in multicenter trials. In this study we determined the accuracy and intraobserver and interobserver agreement of plaque characterization performed by subjective, visual inspection of ultrasound images as currently used in clinical trials. We also investigated the influence of image storage media on image interpretation and reproducibility issues.
Carotid Plaque Characterization by Duplex Scanning

Subjects and Methods

Sixty patients undergoing carotid endarterectomy were entered into this study (mean age, 65±8 years; 21 female, 39 male); 51 were operated on for symptomatic and 9 for asymptomatic severe stenosis of the ICA. The mean percent diameter stenoses for these 2 groups were 79% (range, 60% to 99%) for symptomatic and 81% (range, 70% to 90%) for asymptomatic. All patients underwent duplex scanning of the ICA on the day before carotid surgery. Duplex examination was performed with a 4- to 7-MHz linear array transducer (HDI3000 Advanced Technology Laboratories). A standardized imaging format was adopted that included the following: linear gray scale map, dynamic range of 60 dB, medium persistence level, high line density, and medium frame rate. Optimum image focusing was achieved by positioning image focal zones at the anterior and posterior walls of the artery. B-mode ultrasound images of the internal carotid artery illustrating different grades of atherosclerotic plaques according to the 2 plaque morphology classification schemes. Images I through IV illustrate the 4 Gray-Weale classifications, and images 1 and 2 represent homogeneous and heterogeneous plaques, respectively. Arrows indicate the plaque border.

Results

Intraobserver Variability

Hard Copy

The within-observer agreement for grading carotid plaques using only the B-mode ultrasound information ranged from 61±17% (κ = 0.08 to 0.66) for the 2-category classification and gave an overall κ classification of moderate according to the κ value (Table 1). The addition of color flow information improved the overall agreement to good (κ = 0.65 ± 0.08). The 4-category classification showed poor (κ = 0.15 ± 0.06) or fair (κ = 0.34 ± 0.06) agreement when B-mode or combined B-mode and color images were analyzed, respectively. The range of agreement values for the 4 observers was wide and tended to improve with the more experienced observers.

Digital Images

The intraobserver agreement for grading plaques from digital images showed increased κ values, indicating moderate agreement for B-mode alone and with color for both the 2-category (κ = 0.45 ± 0.08 and 0.49 ± 0.10) and 4-category (κ = 0.48 ± 0.06 and 0.53 ± 0.06) classifications (Table 1).

Plaque morphology grading was repeated by observers 1 and 2 over a range of time periods (1 month, 5 months, and 6 months) for full-color images and demonstrated the following values for κ, standard error, and agreement for the 2-category classification: observer 1: 0.89±0.11 (96%), 0.65±0.18 (88%), 0.50±0.21 (85%); observer 2: 0.49±0.18 (77%), 0.42±0.16 (69%), 0.29±0.15 (62%).

Although the agreements were higher for observer 1, both observers showed a range of agreements over the time period studied.
Interobserver Agreement

Interobserver agreement between the 4 observers for the hard copy data is summarized in Table 2. For B-mode images alone, $\kappa_p$ values indicated moderate agreement ($\kappa_p=0.52\pm0.07$) for the 2-category approach, and, surprisingly, agreements were reduced to fair after the addition of color information ($\kappa_p=0.28\pm0.08$). Values for the 4-category system showed fair agreements with B-mode alone or with color ($\kappa_p=0.28\pm0.05$ and $0.24\pm0.05$).

Interobserver values for the digital images showed good to moderate agreement for the combined B-mode and color images according to both the 2-category ($\kappa_p=0.62\pm0.07$) and 4-category ($\kappa_p=0.52\pm0.05$) classification schemes (Table 2). This was converse to the agreements demonstrated for the hard copy data, in which color increased variability between observers. The highest overall agreement was observed with the 2-category method for grading color flow mapping images displayed in digital format.

Accuracy

The correlation between histology and grading of plaque morphology from combined B-mode color flow images was poor to fair according to $\kappa$ for the 4 observers (Table 3). The overall accuracy of the Gray-Weale approach compared with histology was $30\pm10\%$, whereas the 2-category classification showed improved accuracy of $72\pm12\%$.

Discussion

Measurement of ICA stenosis by duplex ultrasound has come under close scrutiny, and although the latter involves numerical measurement of velocity, sources of variability and protocol standardization remain contentious issues. Grading of plaque morphology by ultrasound is based on subjective evaluation and therefore is susceptible to variability in observer interpretation. Despite this, reports in the literature suggest that subjective grading of ultrasound plaques can be achieved with good reproducibility and high correlation with histology.

The present study was conducted with standardized protocols for image acquisition and standardized hard copy and digital image displays; however, a reproducibility study did not fully support previous findings. Intraobserver agreement for the 2-category classification achieved only moderate agreement for digital B-mode images. The addition of color flow information improved delineation of the plaque border and increased agreement for hard copy data but not for digital images. The 4-category classification showed moderate agreement for digital images but poor to fair agreement for hard copy data. The reduced agreement for the latter may be due to the reduced contrast present in hard copy images. This would have the greatest effect on the 4-category classification, which required more subtle interpretation of the relative contribution of echogenic and echolucent material.

### Table 1. Intraobserver Pooled $\kappa$ Values and Agreements for 4 Observers Grading Plaque Morphology From Hard Copy and Digital Images, Based on the 2-Category (Homogeneous and Heterogeneous) and 4-Category (Gray-Weale) Classifications

<table>
<thead>
<tr>
<th>Method</th>
<th>2-Category Method</th>
<th></th>
<th>4-Category Method</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa_p$ (±SE)</td>
<td>$\kappa$ Range (Mean ± SD)</td>
<td>$\kappa_p$ (±SE)</td>
<td>$\kappa$ Range (Mean ± SD)</td>
</tr>
<tr>
<td>Hard copy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-mode</td>
<td>0.35±0.09</td>
<td>0.08–0.66</td>
<td>66±17</td>
<td>0.15±0.06</td>
</tr>
<tr>
<td>B-mode and color</td>
<td>0.65±0.08</td>
<td>0.28–0.88</td>
<td>82±9</td>
<td>0.34±0.06</td>
</tr>
<tr>
<td>Digital</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-mode</td>
<td>0.45±0.08</td>
<td>0.05–0.68</td>
<td>75±9</td>
<td>0.48±0.06</td>
</tr>
<tr>
<td>B-mode and color</td>
<td>0.49±0.10</td>
<td>0.24–0.64</td>
<td>87±6</td>
<td>0.53±0.06</td>
</tr>
</tbody>
</table>

$\kappa_p$ indicates pooled $\kappa$.

### Table 2. Interobserver Pooled $\kappa$ Values and Agreements Between 4 Observers for Grading Plaque Morphology From Hard Copy and Digital Ultrasound Images, Based on the 2-Category (Homogeneous and Heterogeneous) and 4-Category (Gray-Weale) Classifications

<table>
<thead>
<tr>
<th>Method</th>
<th>2-Category Method</th>
<th></th>
<th>4-Category Method</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa_p$ (±SE)</td>
<td>$\kappa$ Range (Mean ± SD)</td>
<td>$\kappa_p$ (±SE)</td>
<td>$\kappa$ Range (Mean ± SD)</td>
</tr>
<tr>
<td>Hard copy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-mode</td>
<td>0.52±0.07</td>
<td>0.43–0.64</td>
<td>82±3</td>
<td>0.28±0.05</td>
</tr>
<tr>
<td>B-mode and color</td>
<td>0.28±0.08</td>
<td>0.02–0.51</td>
<td>73±7</td>
<td>0.24±0.05</td>
</tr>
<tr>
<td>Digital</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-mode</td>
<td>0.39±0.06</td>
<td>0.10–0.75</td>
<td>70±11</td>
<td>0.24±0.05</td>
</tr>
<tr>
<td>B-mode and color</td>
<td>0.62±0.07</td>
<td>0.47–0.84</td>
<td>88±5</td>
<td>0.52±0.05</td>
</tr>
</tbody>
</table>

$\kappa_p$ indicates pooled $\kappa$. 

$\kappa$ indicates pooled $\kappa$. 

Interobserver agreement between the 4 observers for the hard copy data is summarized in Table 2. For B-mode images alone, $\kappa_p$ values indicated moderate agreement ($\kappa_p=0.52\pm0.07$) for the 2-category approach, and, surprisingly, agreements were reduced to fair after the addition of color information ($\kappa_p=0.28\pm0.08$). Values for the 4-category system showed fair agreements with B-mode alone or with color ($\kappa_p=0.28\pm0.05$ and $0.24\pm0.05$). Interobserver values for the digital images showed good to moderate agreement for the combined B-mode and color images according to both the 2-category ($\kappa_p=0.62\pm0.07$) and 4-category ($\kappa_p=0.52\pm0.05$) classification schemes (Table 2). This was converse to the agreements demonstrated for the hard copy data, in which color increased variability between observers. The highest overall agreement was observed with the 2-category method for grading color flow mapping images displayed in digital format.

Accuracy

The correlation between histology and grading of plaque morphology from combined B-mode color flow images was poor to fair according to $\kappa$ for the 4 observers (Table 3). The overall accuracy of the Gray-Weale approach compared with histology was $30\pm10\%$, whereas the 2-category classification showed improved accuracy of $72\pm12\%$. 

Discussion

Measurement of ICA stenosis by duplex ultrasound has come under close scrutiny, and although the latter involves numerical measurement of velocity, sources of variability and protocol standardization remain contentious issues. Grading of plaque morphology by ultrasound is based on subjective evaluation and therefore is susceptible to variability in observer interpretation. Despite this, reports in the literature suggest that subjective grading of ultrasound plaques can be achieved with good reproducibility and high correlation with histology.

The present study was conducted with standardized protocols for image acquisition and standardized hard copy and digital image displays; however, a reproducibility study did not fully support previous findings. Intraobserver agreement for the 2-category classification achieved only moderate agreement for digital B-mode images. The addition of color flow information improved delineation of the plaque border and increased agreement for hard copy data but not for digital images. The 4-category classification showed moderate agreement for digital images but poor to fair agreement for hard copy data. The reduced agreement for the latter may be due to the reduced contrast present in hard copy images. This would have the greatest effect on the 4-category classification, which required more subtle interpretation of the relative contribution of echogenic and echolucent material.
Intraobserver agreement was not consistent when studied over a 6-month period, indicating that a single evaluation of agreement is not necessarily representative of the individual’s reproducibility in grading images. Although all observers in this study were experienced in ultrasound imaging, only 1 observer achieved consistently good agreement levels after a training period of 3 months. A recent study demonstrated similar variability in within-observer agreement for grading plaque morphology but higher values for between-observer agreement.

Interobserver variability showed a wide range of agreement values, with good agreement being achieved for color flow digital images graded according to the 2-category classification. The somewhat puzzling finding of lower intraobserver agreements compared with interobserver agreements may be explained by the variation in intraobserver agreement with time. Correlation with histology was disappointing, with almost no correlation in some instances; interestingly, higher values of accuracy were noted for the more experienced observers.

These findings suggest that the previously reported high values of agreement within and between observers are not easily reproducible. This may be due to one of several factors: the subjective nature of the classification process; the number of images that are routinely studied; the experience of the observers; the plethora of classification schemes; the variability in image storage and display media used; variations in image quality; the statistics used; and the distribution of plaque types within each study. This is illustrated by the improvements in agreement achieved by using the more simplistic 2-category classification and also by interrogating full-color digital data compared with hard copy images.

Comparison of results between studies is limited by the different statistical approaches used. For example, simple percentage values may suggest a good test, while more detailed $\kappa$ analysis will reveal inadequacies. In a study on intravascular ultrasound images of coronary arteries, a high overall agreement of 95% was reported; however, $\kappa$ values indicated good ability for identifying hard ($\kappa=0.67$) and soft ($\kappa=0.61$) lesions but an inability to detect lipid.

The implications of this study are that results of plaque characterization in clinical studies and multicenter trials could be seriously flawed. Improvements to clinical trials could be achieved by standardizing image parameters, using test objects with a range of echogenicity and echolucency to eliminate individual scanner variations, interrogating full-color digitally stored images, and using the simpler 2-category classification scheme to grade plaque morphology. Observers should not participate in multicenter trials using subjective analysis of ultrasound images until they have achieved high levels of intraobserver agreement that are verified throughout the trial period.

**Acknowledgments**

We are grateful to the Stroke Association and Impra for funding this research.

**References**


**TABLE 3. Pooled $\kappa$ and Accuracy for Comparing Plaque Morphology Graded From Digital Ultrasound Images With Histology, Based on the 2-Category (Homogeneous and Heterogeneous) and 4-Category (Gray-Weale) Classifications**

<table>
<thead>
<tr>
<th>Plaque Composition</th>
<th>2-Category Method</th>
<th>4-Category Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa_4$</td>
<td>Agreement (%)</td>
</tr>
<tr>
<td></td>
<td>(±SE)</td>
<td>(Mean ± SD)</td>
</tr>
<tr>
<td>B-mode</td>
<td>0.13±0.09</td>
<td>0.00–0.20</td>
</tr>
<tr>
<td>B-mode and color</td>
<td>0.28±0.09</td>
<td>0.12–0.51</td>
</tr>
</tbody>
</table>

$\kappa_4$ indicates pooled $\kappa$. 

References


Carotid Plaque Characterization by Duplex Scanning: Observer Error May Undermine Current Clinical Trials

J. A. C. Arnold, K. B. Modaresi, N. Thomas, P. R. Taylor and T. S. Padayachee

Stroke. 1999;30:61-65
doi: 10.1161/01.STR.30.1.61

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/30/1/61

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/