Visual Rating of Age-Related White Matter Changes on Magnetic Resonance Imaging

Scale Comparison, Interrater Agreement, and Correlations With Quantitative Measurements

P. Kapeller, MD; R. Barber, MD; R. J. Vermeulen, MD; H. Adèr, PhD; P. Scheltens, MD; W. Freidl, PhD; O. Almkvist, MD; M. Moretti, MD; T. del Ser, MD; P. Vaghfeldt, MD; C. Enzinger, MD; F. Barkhof, MD; D. Inzitari, MD; T. Erkinjunti, MD; R. Schmidt, MD; F. Fazekas, MD; for the European Task Force of Age Related White Matter Changes

Background and Purpose—To provide further insight into the MRI assessment of age-related white matter changes (ARWMCs) with visual rating scales, 3 raters with different levels of experience tested the interrater agreement and comparability of 3 widely used rating scales in a cross-sectional and follow-up setting. Furthermore, the correlation between visual ratings and quantitative volumetric measurement was assessed.

Methods—Three raters from different sites using 3 established rating scales (Manolio, Fazekas and Schmidt, Scheltens) evaluated 74 baseline and follow-up scans from 5 European centers. One investigator also rated baseline scans in a set of 255 participants of the Austrian Stroke Prevention Study (ASPS) and measured the volume of ARWMCs.

Results—The interrater agreement for the baseline investigation was fair to good for all scales (k values, 0.59 to 0.78). On the follow-up scans, all 3 raters depicted significant ARWMC progression; however, the direct interrater agreement for this task was poor (k, 0.19 to 0.39). Comparison of the interrater reliability between the 3 scales revealed a statistical significant difference between the scale of Manolio and that of Fazekas and Schmidt for the baseline investigation (z value, −2.9676; P=0.003), demonstrating better interrater agreement for the Fazekas and Schmidt scale. The rating results obtained with all 3 scales were highly correlated with each other (Spearman rank correlation, 0.712 to 0.806; P<0.001), and there was significant agreement between all 3 visual rating scales and the quantitative volumetric measurement of ARWMC (Kendall W, 0.37, 0.48, and 0.57; P<0.001).

Conclusions—Our data demonstrate that the 3 rating scales studied reflect the actual volume of ARWMCs well. The 2 scales that provide more detailed information on ARWMC progression remains problematic and may require modifications or extensions of existing rating scales. (Stroke. 2003;34:441-445.)

Key Words: magnetic resonance imaging ▪ white matter

A ge-related white matter changes (ARWMCs) are frequent incidental findings on T2-weighted MRI. Besides their correlation with advanced age and hypertension, they have been reported to be associated with various diseases such as migraines, transient ischemic attacks, ischemic stroke, primary intracerebral hemorrhage, and Alzheimer disease. Their significance in otherwise normal individuals is even less clear. To further explore these signal abnormalities, it may be preferable to use visual rating scales instead of more time-consuming quantitative measurements. However, existing rating scales are heterogeneous in regard to their range, exact morphological description, including the diameter of lesions, and consideration of the anatomic distribution of signal hyperintensities [periventricular changes (PVCs) or deep white matter changes (DWMCs)]. This may lead to considerably different results in the assessment of ARWMCs when a cohort of patients is evaluated by different scales. It can be expected that the assessment of ARWMC is still more variable among raters with different levels of experience or when scans have been collected from different sites. In
addition, no scale has been investigated so far in terms of its correlation with quantitative measurements of MRI hyperintensities. To further explore the clinical impact of ARWMCs, multicenter investigations will necessarily have to involve large sample sizes. Hence, homogenization and evaluation of visual ARWMC rating scales with respect to their intercorrelation and correlation with quantitative measurements are needed. The European Task Force on Age Related White Matter Changes is currently working on such improvements. In this effort, Wahlund and coworkers recently introduced a new rating scale extending earlier protocols.

In the present study, task force members with different levels of experience working in different centers met to test 3 different established rating scales in terms of their interrater agreement in evaluating scans from different sites. In addition, they wanted to determine the sensitivity of these scales for longitudinal data analysis and the correlation between visual rating and quantitative measurements of ARWMCs. Thus, one of the investigators performed visual rating and quantitative measurements in an additional set of MRIs from participants of the Austrian Stroke Prevention Study.11

Subjects and Methods

Baseline and follow-up MRI scans from 74 randomly chosen individuals from 5 different European sites (Amsterdam, Graz, Huddinge, Lille, Newcastle) were evaluated by 3 independent raters (R.B., R.J.V., P.K.) with different levels of experience. The field strength of the scanners used ranged from 1.0 to 1.5 T. T1- and T2-weighted scans in the axial plain were available for all patients. These were complemented by either scans obtained with fluid attenuation inversion recovery or proton density sequences to allow better separation of white matter hyperintensities and cerebrospinal fluid. The follow-up period ranged from 1.5 to 4 years. All individuals had participated in local dementia trials or in local studies on normal aging. Three widely used visual rating scales were evaluated: the scales of Manolio et al5 (the Manolio scale), Fazekas et al6 and Schmidt et al7 (the Fazekas and Schmidt scale), and Scheltens and coworkers8 (the Scheltens scale). More detailed information about the 3 scales is given in the Appendix. Although the Manolio scale provided global ARWMC assessment, the scales of Fazekas and Schmidt and of Scheltens account for DWMCs and PVCs separately.

For all 3 scales, templates were provided to homogenize the rating. Raters were not blinded to the study date, and progression was evaluated by direct comparison of baseline and follow-up scans. The rating was performed from filmed material.

One investigator (P.K.) also performed visual rating using these scales and quantitative volumetric measurements of ARWMCs in 255 scans of participants in the Austrian Stroke Prevention Study.11 Volumetric measurements were performed on a Philips Gyro View computer with cursor-controlled manual outlining of ARWMCs (DWMCs plus PVCs) on the computer screen to assess total absolute area. ARWMCs were measured in both cerebral hemispheres. Cerebellum and brain stem were not considered. This technique has been described elsewhere.15,16

Statistical Analyses

The agreement for ARWMC rating among observers was expressed by means of Cohen’s $\kappa$ statistic.17 Values $<0.40$ were considered to reflect poor agreement; 0.40 to 0.75, fair to good agreement; and $>0.75$, excellent agreement.18 In addition, the $\kappa$ values of the 3 scales were compared by calculating $z$ scores. The relationship between visual rating and quantitative volumetric measurements was assessed by the nonparametric Kendall $W$ test. For this correlation, PVC and DWMC scores of the Fazekas and Schmidt and the Scheltens scales were summed to provide a global ARWMC score.

The global ARWMC scores were correlated with the quantitatively assessed total ARWMC volume (DWMCs plus PVCs). Kendall $W$ values of 0 was considered no relationship between the variables; values equal to 1 were considered to reflect total agreement. The correlation of the rating results between the scales was assessed with the Spearman rank correlation. The Kruskal-Wallis test was used to compare the mean rating values for each scale and to exclude systematic errors resulting from significant differences between the 3 raters. The mean rating scores and the median for each rater and each scale were assessed with the Wilcoxon test. A comparison of the ratings of the baseline versus follow-up was determined with the sign test.

Results

The mean age of the 74 randomly chosen individuals was 73 years (range, 44 to 85 years). The mean rating values and SD for DWMCs at baseline ranged for the scale of Fazekas and Schmidt (rank, 0 to 3) from 1.59 (SD, 0.95) to 1.78 (SD, 1.01) and for the scale of Scheltens (rank, 0 to 24) from 6.32 (SD, 4.85) to 6.96 (SD, 5.41). The mean values for the Manolio global ARWMC rating (rank, 0 to 9) ranged from 2.54 (SD, 1.71) to 2.72 (SD, 2.27). For follow-up, the mean rating values in general increased by 2% to 6.3%, depending on the rater and scale used. In detail, rater 1 reported an increase of 5% with the Manolio scale, 4.3% with the Fazekas and Schmidt scale, and 3.75% with the Scheltens scale. Rater 2 rated 5.4% higher mean values with the Manolio scale, 6.3% higher mean values with the Fazekas and Schmidt scale, and 4.3% higher mean values with the Scheltens scale. Rater 3 depicted no increase with the Manolio scale, 2.6% increase with the Fazekas and Schmidt scale, and 2.2% increase with the Scheltens scale. For raters 1 and 2, these results were significant for all 3 scales (sign test, $P<0.01$), but rater 3 depicted significant progression only with the scales of Fazekas and Schmidt (sign test, $P=0.05$) and Scheltens (sign test, $P=0.01$). The Kruskal-Wallis test showed no significant differences in rating scores between the 3 raters. Table 1 gives the median values, ranges, means, and SD for all 3 raters and all 3 scales for baseline and follow-up as complementary information.

Although the interrater agreement between the 3 raters for baseline evaluation was fair to excellent for all 3 scales ($\kappa$, 0.59 to 0.78), the evaluation of agreement for the follow-up examinations was generally poor ($\kappa$, 0.09 to 0.39). Table 2 gives the results for all 3 scales and for each score obtained. The comparison between scales regarding interrater reliability showed a statistically significant difference between the scales of Manolio and Fazekas and Schmidt ($z$ value, $-2.9676; P=0.003$) for baseline evaluation, demonstrating better interrater agreement for the Fazekas and Schmidt scale. Comparison between the scales of Manolio and Scheltens and between the scales of Fazekas and Schmidt and Scheltens revealed no differences. For the follow-up, interrater reliability showed no difference between the 3 scales (Table 3). In addition, for every rater, there was a high correlation between the ARWMC ratings as obtained with investigated scales (Spearman rank correlation, 0.712 to 0.806; $P<0.01$).

The extended set of 255 subjects for whom baseline scans were evaluated with visual rating and quantitative volumetric measurements showed a mean white matter hyperintensity volume of 0.487 cm$^3$ (range, 0 to 24.5 cm$^3$; SD, 3.078). The
agreement of visual rating with quantitative volumetric measurements was significant for all scales (P < 0.001) and was better for the scales of Fazekas and Schmidt and Scheltens than for the scale of Manolio. Table 4 shows the Kendall W coefficients of concordance for all 3 scales.

### Discussion

This comparison provides further insight into the use of 3 commonly used visual rating scales for both cross-sectional and follow-up evaluation of ARWMCs. By using scans obtained from different sites and readers with different levels of experience trained at different sites, we attempted to simulate true multicentric study conditions. We also addressed the correlation between the rating results of the 3 scales and quantitative volumetric measurement.

All 3 raters showed high comparability in terms of the assessment and evaluation of ARWMCs with all 3 scales. However, the comparison of κ values between the scales demonstrated better interrater agreement between the scales of Fazekas and Schmidt and Scheltens than between those 2 scales and the Manolio scale. For the scales of Fazekas and Schmidt and Manolio, the difference was even statistically significant. The reason may be that the scales of Fazekas and Schmidt and Scheltens use a similar, well-defined approach in evaluating different aspects of ARWMsC, ie, DWMCs and PVCs, compared with the Manolio scale, which provides a global score.

Although the interrater agreement for the baseline evaluation yielded fair to good κ values, the 3 raters reached only poor agreement for follow-up examinations. Although all raters reported ARWMC progression as expressed by higher mean scores in the follow-up investigation, the interrater agreement remained unsatisfactory. This result appears to be of great practical importance concerning future ratings of ARWMC progression. We deliberately decided not to blind the raters to the scan date in this investigation. In clinical practice, it may often be very difficult or even impossible to blind raters to the scan order because of characteristic changes that occur over time, eg, different film layouts or scanner changes indicated on the film. Furthermore, not knowing about the scan order may result in an increase in noise within the obtained data, which may lower the sensitivity for assessing progression in patients with minor changes. We had thought that avoiding these problems might

### Table 2. Interrater Agreement for All 3 Scales for Baseline and Follow-Up Investigations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Baseline (κ)</th>
<th>Follow-Up (κ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manolio (ARWMC)</td>
<td>0.5988 (0.0457)</td>
<td>0.2643 (0.0758)</td>
</tr>
<tr>
<td>Fazekas (DWMC)</td>
<td>0.7808 (0.0409)</td>
<td>0.3934 (0.0103)</td>
</tr>
<tr>
<td>Fazekas (PVC)</td>
<td>0.6258 (0.0592)</td>
<td>0.1998 (0.0839)</td>
</tr>
<tr>
<td>Scheltens (DWMC)</td>
<td>0.6865 (0.0373)</td>
<td>0.3569 (0.0537)</td>
</tr>
<tr>
<td>Scheltens (PVC)</td>
<td>0.7142 (0.0385)</td>
<td>0.2005 (0.0718)</td>
</tr>
</tbody>
</table>

ARWMC indicates age-related white matter changes; DWMC, deep white matter changes; and PVC, periventricular changes.

### Table 3. Scale Comparison in Terms of κ Statistics

<table>
<thead>
<tr>
<th></th>
<th>Fazekas, z</th>
<th>Scheltens, z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manolio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>−2.9676 (P = 0.003)</td>
<td>−1.4867 (P = 0.1371)</td>
</tr>
<tr>
<td>Follow-up</td>
<td>1.0101 (P = 0.3124)</td>
<td>0.9968 (P = 0.3188)</td>
</tr>
<tr>
<td>Fazekas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>1.7036 (P = 0.0885)</td>
<td></td>
</tr>
<tr>
<td>Follow-up</td>
<td>0.3145 (P = 0.7532)</td>
<td></td>
</tr>
</tbody>
</table>

Based on the κ, z values have been determined to account for the different ranking of the 3 scales used. Shifting of the values in either direction from 0 expresses the difference between the raters. In case of the scales of Manolio and Fazekas and Schmidt, the difference was even statistically significant, demonstrating better interrater agreement for the Fazekas and Schmidt scale.
offset the error that can be introduced by a greater readiness for rating actual progression when the scan order is known. However, this was obviously not the case. All 3 raters unanimously depicted progression, albeit small, of ARWMCs between the first and second examinations, but there was considerable difference in the assessment of its extent. Depending on the scale, rater 2 gave 4.3% to 6.3% higher mean scores in the follow-up investigation, whereas rater 3 gave only 0% to 2.6% higher means. The low \( \kappa \) values for follow-up ratings appear even more worrisome because they suggest that all readers had the impression that ARWMCs progressed to some extent but that this was frequently stated for different patients. Of course, we must also consider that a comparison of scans is much more demanding than a cross-sectional rating. Even slight modifications in angulation between the baseline and follow-up investigation may influence the shape of a lesion, and the level of experience, which was different among the raters of this study, is likely to have a strong impact on the interpretation of such “changes.” We should also consider that the scales used to assess and evaluate ARWMCs here were designed for cross-sectional and not for longitudinal data analysis. Therefore, our results may also reflect the weakness of these scales in assessing minor progression. Data concerning the rating of progression are scarce and sometimes are published only in abstract form.\(^{19-22}\)

In a 3-year follow-up study addressing ARWMCs, Schmidt and coworkers\(^{20}\) reported high interrater agreement in cases of marked ARWMC progression compared with only poor agreement in cases of minor progression. They also indicated image quality and angulation as possible explanations. Clearly, further studies are needed to improve the rating of ARWMC progression, and the danger of overreading ARWMC progression by unblinded rating has to be kept in mind.

Many scales have been tested for their interrater reliability.\(^ {5,10,14,23} \) Some are even validated histopathologically.\(^ {24,25} \) However, the assumption that the ARWMC score correlates with the actual ARWMC volume has not been tested so far. This study documents for the first time this correlation. All 3 scales showed significant agreement with quantitative volumetric measurements. However, according to our analyses, the agreement was even closer for the scales of Fazekas and Schmidt and Scheltens than for the Manolio scale. This result provides evidence that scales that provide more detailed information on ARWMC better reflect their actual volume. If rating scales are considered to be used for ARWMC assessment instead of time-consuming and technically challenging quantitative measurement, this finding should be taken into consideration.

None of the assessed parameters showed a significant difference between raters even though the 3 raters were trained at different sites and had different levels of experience. The instructions given with the scales and the provided template images certainly helped to homogenize the rating.

In conclusion, our data demonstrate that the 3 rating scales studied reflect the actual volume of ARWMCs well. The 2 scales that provide more detailed information on ARWMC seemed preferential to that which yields more global information. The visual assessment of progression of ARWMCs remains problematic and may require modifications or extensions of existing rating scales.

Appendix

In the scale of Manolio, the total volume of PVCs and DWMCs is evaluated according a 0-to-9 point scale by using a template image and a text description. The scale of Fazekas and Schmidt separately accounts for PVCs and DWMCs; the DWMC rating accounts separately for number and extent of lesions. The number is given as 0 for no lesions, 1 for 1 to 4 lesions, 2 for 5 to 9 lesions, and 3 for >9 lesions. DWMC extent is given as 0 for no lesion, 1 for punctuate foci, 2 for beginning confluent foci, and 3 for confluent DWMC. PVC is scored as 1 for caps or pencil-thin lining, 2 for smooth halo, and 3 for irregular PVCs extending into deep white matter. The scale of Scheltens accounts separately for PVCs and DWMCs, evaluating the presence and extent of PVCs and DWMCs in different anatomic regions with a 0-to-6-point scale. PVCs are ranked as follows: 0= absent, 1= 1 to 5 mm, and 2= >5 mm and <10 mm in 3 regions (frontal, occipital, and along the ventricles). Resulting in a maximum of 6 points. DWMC rating addresses changes in deep white matter, in the basal ganglia, and in infratentorial locations according a 0-to-6-point scale: 0= absent, 1= <3 mm in ≤5, 2= >3 mm in ≥6, 3= 4 to 10 mm in ≤5, 4= 4 to 10 mm in ≥6, 5= <11 mm in >1, and 6= confluent. Deep white matter is separated into frontal, parietal, occipital, and temporal region scoring with a maximum of 24 points. Basal ganglia are divided into caudate nucleus, putamen, globus pallidus thalamus, and internal capsule, scoring a maximum of 30 points. Infratentorial foci are rated in the cerebellum, mesencephalon, pons, and medulla, scoring a maximum of 24 points. Examples of PVCs and DWMCs for all 3 scales were provided as template images to reach the best possible agreement.

References


Visual Rating of Age-Related White Matter Changes on Magnetic Resonance Imaging. Scale Comparison, Interrater Agreement, and Correlations With Quantitative Measurements


Stroke. published online January 30, 2003;

Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2003 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/early/2003/01/30/01.STR.0000049766.26453.E9.citation