White-Matter Changes on MRI as Surrogate Marker

PHILIP SCHELTENS, FREDERIK BARKHOF, AND FRANZ FAZEKAS

ABSTRACT. This article reviews the use of white-matter changes as identified on MRI as a surrogate marker in clinical trials on vascular dementia. As of yet, insufficient evidence is present to implement rating or volumetric assessment of the burden of white-matter changes as surrogate marker in such trials, in view of the limited sensitivity to change of current methods and the high measurement error.

KEYWORDS: Dementia; white matter; MRI; clinical trials; vascular; pH

The advent of morphologic imaging technologies for the brain such as magnetic resonance imaging (MRI) has provided new views on the parenchymal alterations associated with various disease processes and ageing. Especially with MRI, much interest has focused on changes of the white matter because of the exquisite sensitivity of this technique for damage to myelin and the recognition of a high rate of unexpected white-matter lesions (WML) even in the normal population.

Especially these WML, however, play a very important role in the development of vascular dementia (VaD). In the National Institute of Neurological Disorders and Stroke and the Association Internationale pour la Recherche et l’Enseignement en Neurosciences (NINDS-AIREN) international work group criteria for vascular dementia (Roman et al., 1993), white-matter changes are regarded to be sufficient to cause dementia. In a recent study (Scheltens & Kittner, 2000), it was found that the majority of the NINDS-AIREN-diagnosed VaD cases had small-vessel disease only. This has led a group of investigators to formulate specific criteria for “subcortical VaD” (Erkinjuntti et al., 2000).

From the Department of Neurology and the Alzheimer Centre (P. Scheltens, MD), and the Department of Radiology and the Image Analysis Centre (F. Barkhof, MD), Vrije Universiteit Medical Centre, Amsterdam, The Netherlands; and the Department of Neurology, Karl-Franzens University, Graz, Austria (F. Fazekas, MD).

Offprints. Requests for offprints should be directed to Philip Scheltens, MD, Department of Neurology, Vrije Universiteit Medical Centre, PO Box 7057, 1007 MB Amsterdam, The Netherlands. email: p.scheltens@vumc.nl
RATING SCALES

Various scores and rating systems have been used to grade the amount of WML (Scheltens et al., 1998). In a study by Mäntylä and colleagues (1999), the MRI signal hyperintensities of 395 poststroke patients were rated with various scales, and it became evident that the severity of the same white-matter lesions was appreciated quite differently by the various scales, and the relation with clinical findings such as cerebrovascular risk factors was also dependent on which scale had been used.

Evidently, the use of rating scales as an instrument for assessing WML is still far from optimal because commonly agreed concepts regarding their instrumentation and application are lacking. In parallel, there is consensus that more homogeneity in the evaluation of WML is needed to allow the pooling of data and comparison of results.

Alternatively, considering the rapid progress of image analysis technology, it would be imaginable to introduce quantitative measures such as a determination of the entire volume of hyperintense brain white matter. Automated techniques are being developed for such measurements and will soon be available. Absolute quantitation may therefore become part of study protocols that focus on these abnormalities be it in the context of natural history investigations or therapeutic intervention. However, such quantitation does not only require sophisticated tools for image analysis, it also needs MRI examinations performed with appropriate and uniform protocols at a high quality level. For multiple sclerosis, guidelines for the use of quantitative MRI measures for study purposes have already been formulated (Filipi et al., 1998), and the same may be needed for investigating WML. However, regarding the differences in hardware and software equipment available at different sites and their impact on image quality, large scale computerized image analysis for routine clinical purposes or even retrospective studies involving multiple sites are still not foreseeable. Also, simply measuring lesion volume may not account for a possibly different impact of different lesion types as assumed from MRI-histopathologic correlations (Fazekas et al., 1998).

It is in this context that rating scales can provide fast and standardized semiquantitative information. Visual scoring of WML can be applied to images of different quality and from different scanners as the trained interpreter can correct for variations in image contrast, resolution and to some extent even for differences in angulation. Problems of electronic data handling and transfer due to different imaging formats are also not encountered with visual analysis. Most important, however, visual assessment by means of a rating scale need not be limited to a semiquantitative analysis of the amount of abnormal brain tissue but can also include specific features of lesion location or texture that may be more important than the extent of abnormality itself.

The greatest disadvantage of the use of rating scales is their nonlinearity. Also, the criteria for scoring WML according to specific scales frequently remain poorly defined, which invites modifications and subjective interpretation. This leads not only to unnecessarily high differences in results between raters but is likely to increase even intrarater variability. Provision of
illustrative examples and common training sessions may be means to reduce this problem. The best solution for this, however, seems to be the installment of a central reading facility, using one or two experienced raters with established high intrarater reliability. It has been proven that such an undertaking can be performed on a large scale in clinical trials involving hundreds of scans on VaD and AD (Scheltens & Kittner, 2000), and some large scale VaD trials are now using this concept and are now ongoing.

Efforts of the European Task Force for Age-Related White-Matter Changes to improve upon WML rating scales have recently gone in two different directions. First, it was attempted to create and describe a rating scale that may be used with both CT and MRI. Secondly, a more formal assessment and comparison of three commonly used MRI rating scales for WML has been performed including comparison with quantitative data. Extending earlier suggestions, Wahlund and colleagues created a four-point scale (Table 1) for rating WML lesions in five different regions of the right and left hemisphere, the so-called Age-Related White Matter Changes (ARWMC) Scale (Wahlund et al., 2001).

The regions to be rated are the frontal area, which equals the frontal lobe—that is, frontal to the central sulcus, the parietooccipital area consisting of the parietal and occipital lobes together, the temporal area that equals the temporal lobe, the infratentorial area including the brainstem and cerebellum, and the basal ganglia including the striatum, globus pallidus, thalamus, internal and external capsules, and insula. Applying this scale to 77 pairs of CT and MRI images, the new ARWMC scale provided good interrater reliability for MRI ($\kappa = 0.66$) and moderate for CT ($\kappa = 0.48$). As expected, MRI was superior to detect small white-matter lesions. Important, however, was the fact that large lesions were detected equally well with both CT and MRI. In a further study, Kapeller and colleagues (2000) compared the performance of three different but frequently used MRI rating scales for WML. The scales tested were that of Manolio, a modification of the Fazekas Scale based on histopathologic data and on the Scheltens Scale. Comparisons showed a similar interrater reliability and, interestingly, the correlation between scales was also relatively high and rater independent.

### TABLE 1. ARWMC Rating Scale

<table>
<thead>
<tr>
<th>Rating score WM lesions</th>
<th>0</th>
<th>No lesions (including symmetrical, well-defined caps or bands)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>Focal lesions</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Beginning confluence of lesions</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Diffuse involvement of the entire region/with or without involvement of U-fibers</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rating score basal ganglia lesions</th>
<th>0</th>
<th>No lesions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>One focal lesion (5 mm or more)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>More than one focal lesion</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Confluent lesions</td>
</tr>
</tbody>
</table>
LONGITUDINAL STUDIES

So far, rating of WML has been used primarily for cross-sectional studies. To further prove the clinical significance of these abnormalities, longitudinal studies are now ongoing. These investigations will attempt to establish a temporal relationship between the increase of white-matter damage and clinical deterioration. Only limited experience exists on how WML progression could be evaluated with the use of rating scales (Veldink et al., 1998; Wahlund et al., 1996; Schmidt et al., 1999).

Wahlund and colleagues (1996) studied a cohort of 24 normal elderly individuals over a 5-year period. Only the deep WML and basal ganglia scores changed significantly over time, as measured by the Scheltens Scale. Veldink and colleagues (1998) studied 59 elderly individuals with mixed phenomenology and found that only deep WML changed significantly over a 2-year period, using the same scale. Schmidt and colleagues (1999) studied 273 community-dwelling elderly without neuropsychiatric disease. MR images were read by three independent raters, and the change of white-matter hyperintensities from baseline was assessed by direct image comparison. The change was graded as absent, minor, or marked.

Wahlund and colleagues (1996) studied a cohort of 24 normal elderly individuals over a 5-year period. Only the deep WML and basal ganglia scores changed significantly over time, as measured by the Scheltens Scale. Veldink and colleagues (1998) studied 59 elderly individuals with mixed phenomenology and found that only deep WML changed significantly over a 2-year period, using the same scale. Schmidt and colleagues (1999) studied 273 community-dwelling elderly without neuropsychiatric disease. MR images were read by three independent raters, and the change of white-matter hyperintensities from baseline was assessed by direct image comparison. The change was graded as absent, minor, or marked.

Minor change was defined as a difference of no more than one to four punctate lesions between both scans. A change was considered to be marked if there was a difference of more than four abnormalities or a transition to early confluent and confluent lesions. Combined ratings indicated lesion progression in 49 individuals (17.9%). Lesion progression was minor in 27 participants (9.9%) and was marked in 22 (8.1%). These studies show that deep white-matter changes, rated by the Scheltens scale in 2 and by the Fazekas scale in one study, significantly progress over time, but that the periventricular changes do not measurably progress. Explanations may include differences in angulation or image contrast, but also, the probably relatively slow progression of WML may play a role.

SURROGATE MARKER

In general, a surrogate marker needs to fulfill the criteria detailed in Table 2. When examining visual rating of WML for meeting these criteria, the issues of sensitivity to change and measurement error are the most critical. As of yet, insufficient evidence seems to be present to implement visual rating of WML as surrogate markers in clinical trials. Testing the sensitivity and reliability of instruments for determining WML evolution over time will be an important and timely task.

CONCLUSIONS

Despite a long history of rating scales for WML, development, refinement, and understanding of the implications of
such scales are still ongoing. Use of these scales including both cross-sectional and longitudinal designs and applications in clinical trials is now common. Given interrater and intrarater reliability, the optimal use seems to be in a centralized reading setting. Sensitivity for measuring changes over time seems to be too inadequate to be used in large-scale clinical trials aimed at slowing progression of these changes.

REFERENCES


