Reproducibility and variability of quantitative magnetic resonance imaging markers in cerebral small vessel disease.
François De Guio, Eric Jouvent, Geert Jan Biessels, Sandra Black, Carol Brayne, Christopher Chen, Charlotte Cordonnier, Frank-Eric De Leeuw, Martin Dichgans, Fergus Doubal, et al.

To cite this version:
Reproducibility and variability of quantitative MRI markers in cerebral small vessel disease

François De Guio1,2, Eric Jouvent1,2,3, Geert Jan Biessels4, Sandra E. Black5, Carol Brayne6, Christopher Chen7, Charlotte Cordonnier8, Frank-Eric De Leeuw9, Martin Dichgans10,11, Fergus Doubal12, Marco Duering10, Carole Dufouil13, Emrah Duzel14, Franz Fazekas15, Vladimir Hachinski16, M. Arfan Ikram17,18, Jennifer Linn19, Paul M. Matthews20, Bernard Mazoyer21, Vincent Mok22, Bo Norrving23, John T. O’Brien24, Leonardo Pantoni25, Stefan Ropele26, Reinhold Schmidt15, Sudha Seshadri27, Eric E. Smith28, Luciano A. Sposato26, Blossom Stephan29, Richard H. Swartz5, Christophe Tzourio13, Mark van Buchem30, Aad van der Lught31, Robert van Oostenbrugge31, Meike W. Vernooij17, Anand Viswanathan32, David Werring33, Frank Wollenweber10, Joanna M. Wardlaw12,34, Hugues Chabriat1,2,3

1University Paris Diderot, Sorbonne Paris Cité, UMRS 1161 INSERM, Paris, France; 2DHU NeuroVasc Sorbonne Paris Cité, Paris, France; 3AP-HP, Lariboisière Hosp, Department of Neurology, Paris, France; 4Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands; 5Department of Medicine (Neurology), Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Canada; 6Department of Public Health and Primary Care, Cambridge University, Cambridge, UK; 7Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; 8University of Lille, CHU Lille, Inserm U1171, Lille, France; 9Donders Institute for Brain, Cognition and Behavior, Radboud University Medical Center, Department of Neurology, Nijmegen, The Netherlands; 10Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilian-University (LMU), Munich, Germany; 11Munich Cluster for Systems Neurology (SyNergy), Munich, Germany; 12Department of Neuroimaging Sciences, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK; 13Inserm UMR897, University of Bordeaux, Bordeaux, France; 14Department of Cognitive Neurology and Dementia Research, University of Magdeburg, Magdeburg, Germany; 15Department of Neurology, Medical University of Graz, Graz, Austria; 16Department of Clinical Neurological Sciences, University of Western Ontario, London, Canada; 17Department of Radiology and department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands; 18Department of Neurology, Erasmus MC, University
Medical Center Rotterdam, Rotterdam, The Netherlands; 19Department of Neuroradiology, University Hospital Munich, Munich, Germany; 20Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK; 21CNRS-CEA UMR5296, University of Bordeaux, Bordeaux, France; 22Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China; 23Department of Clinical Sciences, Neurology, Lund University, Lund, Sweden; 24Department of Psychiatry, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK; 25NEUROFARBA Department, University of Florence, Florence, Italy; 26Centre for Healthy Brain Ageing, School of Psychiatry, University of New South Wales, Sydney, Australia; 27Department of Neurology, Boston University School of Medicine, Boston, USA; 28Department of Clinical Neurosciences and Hotchkiss Brain Institute, University of Calgary, Calgary, Canada; 29Institute of Health and Society, Newcastle University Institute of Ageing, Newcastle University, Newcastle, UK; 30Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands; 31Department of Neurology, Maastricht University Medical Center, Maastricht, The Netherlands; 32Department of Neurology, J. Philip Kistler Stroke Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, USA; 33Stroke Research Group, Department of Brain Repair and Rehabilitation, UCL, London, UK; 34Centre for Cognitive Ageing and Cognitive Epidemiology (CCACE), University of Edinburgh, Edinburgh, UK

Corresponding author: Pr Hugues Chabriat, Dept of Neurology, Hopital Lariboisière, 2 rue Ambroise Paré, 75010 Paris, France. Email: hugues.chabriat@aphp.fr

This study was supported by grants from Fondation Avenir and Fondation Leducq (Transatlantic Network of Excellence on the Pathogenesis of Small Vessel Disease of the Brain, http://www.fondationleducq.org).

This work follows from a project funded by the Centres of Excellence in Neurodegeneration (COEN), ref MRC MR/J006971/1, “Standards for determining the vascular contribution to neurodegeneration” in which many of the authors participated.

Word count
Abstract: 170
Full text: 7893
Figures: 4
Table: 1
**Abstract** *(200 words)*

Brain imaging is essential for the diagnosis and characterization of cerebral small vessel disease (SVD). Several magnetic resonance imaging (MRI) markers have therefore emerged, providing new information on the diagnosis, progression and mechanisms of SVD. Yet, the reproducibility of these SVD markers has received little attention despite being widely used in cross-sectional and longitudinal studies. This review focuses on the main SVD-related markers on MRI including: white matter hyperintensities, lacunes, dilated perivascular spaces, microbleeds and brain volume. The aim is to summarise, for each marker, what is currently known about: 1) its reproducibility in studies with a scan-rescan procedure either in single or multi centre settings; 2) the acquisition-related sources of variability; and, 3) the techniques used to minimise this variability. Based on the results, we discuss technical and other challenges that need to be overcome in order for these markers to be reliably used as outcome measures in future clinical trials. We also highlight the key points that need to be considered when designing multicentre MRI studies of SVD.

**Keywords:** MRI, cerebral small vessel disease, marker, brain volume, atrophy, white matter hyperintensities, microbleeds, lacunes, perivascular spaces, reproducibility, repeatability, variability
1. Introduction

Cerebral small vessel disease (SVD) is a major source of cognitive impairment and is the second most common cause of dementia in older aged people. SVD is typically sporadic and related to aging and vascular risk factors, but can also be caused by genetic conditions. The most frequent sporadic forms of SVD are arteriolosclerosis and cerebral amyloid angiopathy (CAA). A significant proportion of patients with SVD have associated neurodegenerative diseases such as Alzheimer’s Disease (AD), which can lead to a mixed dementia. In therapeutic trials for dementias, surrogate imaging markers of both pathophysiological processes should ideally be used to account for the potential mixed origin.

Pathological features of SVD are heterogeneous and out of reach of current in vivo imaging techniques. Nonetheless, parenchymal lesions presumed to be caused by small vessel changes have been adopted as markers of SVD, giving neuroimaging a central role for diagnosis and characterization. The spectrum of brain magnetic resonance imaging (MRI) manifestations of SVD is wide, ranging from minor white matter hyperintensities (WMH) often seen in population-based studies to numerous subcortical lesions. Indeed, MRI using various sequences allows an evaluation of several imaging markers related to SVD, including recent small subcortical infarcts, WMH, lacunes, enlarged perivascular spaces (PVS), cerebral microbleeds (CMB), and brain atrophy, as summarized by an international working group defining neuroimaging standards (STRIVE) for research into SVD.

The uncertain or lower reproducibility of these MRI markers, especially across centres, is a major concern in cross-sectional and longitudinal observational studies as well as in clinical trials. Indeed, a key challenge is to minimize the variability caused by technological and subject-related factors. Such variability may confound the detection of disease- or treatment-related MRI changes, thereby limiting the power to detect and follow the progression of imaging markers of SVD. From the patient being imaged to the final imaging marker of interest, which is quantified by an expert rater or an automatic image-processing algorithm, there are several sources of variability. These can include subject-related physiological changes or instrument-related factors such as field strength, head coil, gradients and sequence parameters (see Fig. 1). Each can affect the reproducibility of the marker. Additional variation may be introduced at the image post-processing stage, depending on which software is used for inhomogeneities correction or tissue segmentation, for example. The definition of lesions and the interpretation of images by several experts may be other sources of variability. The measurement of each MRI marker therefore depends on several factors, the effects of which need to be minimized, corrected or taken into account in a posteriori analyses for studies to be comparable.
The aims of this review are to summarise current knowledge regarding the reproducibility and variability of quantitative imaging markers of SVD including WMH, lacunes, PVS, CMB and brain volume in both single centre and multicentre contexts. To highlight technical solutions for common problems and to identify the remaining challenges that need to be overcome for such markers to be considered valid and reliable. Other SVD-related MRI markers, due to their acute manifestation (recent small subcortical infarct) or their so far non-quantitative assessment (cortical superficial siderosis, cortical microinfarcts) have not been examined in reproducibility studies to our knowledge and hence were not included in the literature search. Post-processing tools or expert ratings that also contribute to the global variability and accuracy of the final marker measurement are beyond the scope of this review. For each MRI marker of SVD we review: 1) its reproducibility in studies with a scan-rescan design in single or across multiple centres; 2) the acquisition-related sources of variability; and, 3) the techniques that can be used to minimise methodological limitations. We also discuss those factors that need to be considered to deal with image variability inherent in MRI markers in multicentre design studies such as clinical trials or longitudinal studies.

2. Methods

Search strategy and selection criteria

This review was designed to focus on the reproducibility and variability of well-defined MRI markers of SVD: WMH, lacunes, PVS, CMB and brain volume.

For each biomarker, taking into account the vocabulary heterogeneity of each marker (see supplementary material), a PubMed search has been conducted up to August 2015 14th with the associated terms “MRI” AND (“variability” OR “reproducibility” OR “reliability” OR “consistency”). Papers published before 2002 were not considered to focus on recent MR technology. Reviews, conference abstracts and articles not written in English were excluded from analysis. Articles from our personal databases were also incorporated to the search. To be included in the reproducibility summary (Table 1), studies had to include subjects who underwent a scan-rescan procedure either at the same centre or/and at different centres. Filtering was firstly based on review of the title and abstract for relevance. Studies were not mandatorily related to SVD (e.g., multiple sclerosis or AD) to include the largest number of studies about variability and reliability for the MRI markers of interest.

Term and metrics definitions

As the development and implementation of quantitative imaging biomarkers has been hampered in part by inconsistent use of terminology, an interdisciplinary group of radiologists,
statisticians, physicists, and other researchers worked to develop a comprehensive terminology to serve as a foundation for quantitative imaging biomarker claims\(^6\). Definitions agreed are reproduced in supplementary material.

In order to simplify the manuscript reading, we will only use the terms “within-centre reproducibility” (strictly speaking, the repeatability) to report results from scan-rescan procedures in a single centre; and “between-centre reproducibility” to report results from multicentre studies.

Several metrics are used to measure variability or reproducibility across different studies. A brief description of each is included in supplementary material.

### 3. Results

#### 3.1 Variation in the number of studies and in the evaluation of the reproducibility

**Number of studies**

There was large variation in the number of studies conducted for assessing the reproducibility and variability of each SVD MRI marker. Brain volume and brain atrophy measures have received the most attention (19 studies with within or/and between centre repeated scans of the same subjects)\(^7\)\(^-\)\(^22\). This is probably related to the significant imaging research efforts in AD community during the last decade in which cerebral atrophy was of primary concern. Particularly, the Alzheimer’s Disease Neuroimaging Initiative (ADNI) permitted the most important studies on this aspect, including repeated examinations of the same subject on the same or on different scanners \(^14\), \(^17\). Only five studies were found that evaluated the within-centre or between-centre variability of WMH measures\(^13\), \(^23\)-\(^26\) (Table 1). Few studies examined the potential sources of variability of WMH volumes, despite a large number of image processing techniques attempting to quantify them accurately. Finally, there were only four studies on CMB\(^25\), \(^27\)-\(^29\) and no study on the variability of PVS and lacunes. This probably reflects the absence of consensus until recently about radiological terminology and definitions as well as difficulties to automatically quantify these markers\(^30\). Indeed, computational segmentation of lacunes and perivascular spaces has not yet been validated in a specific study. Overall, the reproducibility of measure of longitudinal change, i.e. the evolution of a given marker as a function of time, has only been evaluated for brain volume change (brain atrophy) in 4 studies\(^9\), \(^17\), \(^19\), \(^22\) and not for other markers.
Evaluation of reproducibility

Several studies with a scan-rescan procedure selected in this review included different measures of variability in volumetric measurements that were not consistent across studies: intra-class coefficients, coefficients of variation, standard deviations, mean absolute values of differences. Hence, most of these studies were not directly comparable (Table 1). This prevented a quantified comparison between the multiple factors that may influence the measurements. It would be valuable for the community to harmonize measures of reproducibility and make sure that all important aspects of reproducibility are included.

3.2 Reproducibility and variability of MRI markers

3.2.1 White-matter hyperintensities (WMH)

Definition and interest in SVD

From a radiological perspective, WMH of presumed vascular origin have a higher signal intensity compared to the normal-appearing white matter on proton-density (PD)-weighted, T2-weighted and FLAIR (Fluid-attenuated inversion recovery) sequences and may appear iso- or hypointense on T1-weighted images. It is widely accepted that WMH are important for clinical outcome, in terms of cognitive and functional impairment. Many qualitative and quantitative techniques have been used to measure WMH. Qualitative approaches are based on visual rating scales completed by well-trained readers. Quantitative techniques involve image-processing algorithms to obtain volumetric measures or spatial distribution of WMH lesions. In this review, we only focused on quantitative techniques, as they offer a continuous measure that is a more reliable and sensitive alternative to visual rating scales.

Within-centre reproducibility

Several studies found good to excellent reproducibility of WMH segmentation and volumetry for different computational image analysis methods using a scan-rescan procedure in a single site (Table 1). For example, in the study recruiting the largest number of subjects, De Boer et al. assessed the within-centre reproducibility of WMH segmentation by comparing computational segmentation in 30 subjects scanned twice within a short interval at 1.5T. They found with their best segmentation pipeline a mean (sd) WMH volume difference expressed in % of intracranial volume of 0.01% (± 0.05%) and a coefficient of variation less than 6%.

Between-centre reproducibility (effect of various vendors, coils)
No report was found about between-centre reproducibility of WMH measures in large cohorts evaluated in several centres and using the same magnetic field.

**Magnetic field effect**

It is not clear how the magnetic field strength impacts the measurement of WMH volume beyond the expected improved resolution at higher field. Indeed, there are relatively few direct comparisons of different scanner field strengths in any situation. In 10 patients with SVD, Theysohn et al. showed that FLAIR images acquired at 7T highlighted WMH known from 1.5T with comparable extent. However, WMH were not adequately quantified in this study as neither number nor volume was assessed. On the contrary, in patients with multiple sclerosis, it has been reported that 3T scans showed a mean 10% higher WMH volume compared to 1.5T scans in total lesion volume based on PD-weighted images.

**Sequence effect**

WMH detection and automated quantification usually relies on the FLAIR sequence, which produces T2-weighted images with cerebrospinal fluid suppression. It has been recently shown that 3D FLAIR sequence parameters could be optimized to increase lesion detection and identification in multiple sclerosis compared with default 3D FLAIR or 2D FLAIR (Fig. 2). This approach might also be used in SVD. Indirectly, this study showed that MR parameters (echo time, inversion time) influence the tissue contrast between the normal white matter and white matter lesions on which the algorithms to segment WMH heavily depend.

Some pipelines have used an interleaved PD/T2 acquisition for WMH segmentation, especially as this dual spin echo sequence is more sensitive to WMH in the deep basal ganglia and thalamus than FLAIR and also allows better detection of perivascular spaces. Multisequence image analysis approaches to quantify WMH, e.g., which combine FLAIR with T2* or T1 or T2 or several sequences together, are increasingly used and emphasise the need to obtain several key standard sequences when studying SVD.

**Partial volume effects**

Partial volume effect is the combination of signals from different tissue types in a single voxel whose image intensity is thus dependent on proportions of each tissue. Partial volume effects with clinically feasible resolutions complicate delineation of the lesion borders. Particularly, smaller voxel sizes allow more accurate estimation and yield larger total lesion volumes, especially for patients with small lesions. For example, in 10 patients with multiple sclerosis, lesion volumes were reported to be on average 9% greater with 3 mm compared to 5 mm slice thickness.

**Other factors with some effects**
The approach to post-processing for bias field (or B1 inhomogeneities) correction can have an important effect on WMH volume quantification according to the computational method used.\textsuperscript{46}

**Reproducibility in longitudinal studies**

No study was found to assess the reproducibility of any longitudinal measurement of WMH.

### 3.2.2 Lacunes and perivascular spaces

**Definition and interest in SVD**

From a radiological perspective, lacunes of presumed vascular origin are round or ovoid, subcortical, fluid-filled cavities of between 3 mm and 15 mm in diameter, consistent with a previous acute small subcortical infarct or haemorrhage in the territory of one perforating arteriole.\textsuperscript{5} Number, size, shape and location are measures of interest for lacunes. To date, very few automatic image processing methods exist to segment lacunes, and most studies rely on semi-automatic computational approaches or manual segmentations by trained observers.\textsuperscript{37, 47-50}

PVS are fluid-filled spaces that follow the typical course of a small vessel as it goes through grey or white matter and have a signal intensity similar to CSF on all sequences.\textsuperscript{5, 51} Number, size, shape and location are measures of interest for PVS. Rating methods have been developed to describe these markers\textsuperscript{52, 53}. To date, very few computational image processing methods exist to segment PVS; though some are in development\textsuperscript{37, 54-56}. The first image processing methods for both lacunes and PVS detection were not fully robust and efficient, requiring time-consuming user manual intervention but new methods are more automated and much faster.

**Within- or between-centre reproducibility**

The literature is very sparse concerning the reproducibility of measurements describing lacunes or PVS.\textsuperscript{51} In 20 subjects scanned twice on a 1.5T scanner, one study assessed a good reproducibility for lacunar volume with the used dedicated software but the extent or numbers of lacunes per subject were not specified.\textsuperscript{26}

High-resolution images should be more sensitive for detecting lacunes and PVS because they are small. Therefore, their volume assessment may heavily depend on spatial resolution and geometrical inaccuracies induced by gradient nonlinearities.

**Reproducibility in longitudinal studies**

No study was found to assess the reproducibility of any longitudinal measurement for both lacunes and PVS.

### 3.2.3 Cerebral microbleeds

**Definition and interest in SVD**
From a radiological perspective, cerebral microbleeds (CMB) associated with SVD are small hypointense lesions on paramagnetic sensitive MR sequences such as T2*-weighted gradient-echo (GE) sequence, and are most frequently located in the cortico-subcortical junction, deep grey or white matter in the cerebral hemispheres, brainstem and cerebellum. CMB are associated with SVD and lobar CMB are often seen in patients in memory clinics who have CAA. Most studies suggest that they are associated with impaired cognitive function, although their direct clinical impact remains uncertain. Several rating scales have been designed to minimize observer variation along with several post-processing algorithms for semi-automatic or automatic quantification. We consider in this review all methods to study CMB as a few computational methods have been published and tested in reproducibility studies. Many MRI acquisition factors are expected to influence CMB detection as their size is a consequence of susceptibility effects that are highly dependent on MRI sequence parameters. In fact, hypointensities are typically larger than the physical size of the corresponding histopathological lesion (blooming effect).

Within- or between-centre reproducibility

No study was found to assess CMB detection reproducibility through a scan-rescan procedure at the same centre or between centres with the same imaging protocol. However, several studies examined various factors involved in CMB detection such as magnetic field strength or sequence parameters.

Magnetic field effect

Intravoxel dephasing due to susceptibility effects and associated signal loss in gradient-echo images increase with higher magnetic fields. Several studies hence demonstrated an increased number of CMB at higher field strength with an increase in contrast to noise ratio and CMB size as illustrated in Fig. 3. For example, mean increases from 1.5T to 3T were 48% for contrast to noise ratio and 12% for CMB size on 119 CMB detected in 4 subjects. In addition, more CMB are found at higher field strength; Theysohn et al. found 101 CMB at 7T (mean size: 3.2 mm) against 33 at 1.5T (mean size: 1.4 mm) in 10 subjects.

Sequence effect

3D T2*-weighted gradient-echo images reveal significantly more CMB in more subjects compared to 2D T2*-weighted gradient-echo images, primarily because of higher spatial resolution. In 200 older aged participants, Vernooij et al. reported 71 subjects with at least one microbleed (median number 2.5) using the 3D sequence compared to 42 subjects (median number 1.0) with the 2D sequence. Likewise, the susceptibility weighted imaging (SWI) technique, which uses phase images to enhance susceptibility-related contrasts in the resulting image, is particularly sensitive to paramagnetic and ferromagnetic substances. CMB are better delineated on SWI images compared to classic T2*
Another study also confirmed that SWI detected more CMB than conventional T2* images (prevalence of 40% vs. 23%, total number was 284 vs. 219 in 141 patients)\textsuperscript{76}. To note, the differentiation between CMB and vessels might be more difficult with SWI compared to T2* images although this is a matter of debate \textsuperscript{75,77}. A last point is that susceptibility-weighted sequences are different according to vendors (e.g., ‘SWI’ for Siemens, ‘SWAN’ for General Electric, ‘VenoBOLD’ for Philips) and only the Siemens ‘SWI’ technique uses phase data. Therefore, the effect of vendor’s susceptibility-weighted technique on CMB detection needs further assessment.

**Sequence parameters effect**

The echo time (TE) reflects the timescale on which dephasing occurs, with long echo times enabling more time for dephasing thereby enlarging the susceptibility effect. Thus, choice of TE directly impacts on the number and volume of microbleeds detected as well as the number of structures that may be mistaken for microbleeds (calcium or iron deposits, flow voids, etc.). Increasing TE improves the sensitivity to detect CMB but more artefacts are generated near air-tissues interfaces causing uncertainty about rating some lesions\textsuperscript{78,79}. Otherwise, thin MR imaging sections compared with thicker sections were associated with a small decrease in CMB diameter and a large increase of contrast (CMB signal relative to tissue signal) and thus an improved CMB detection (Fig. 3)\textsuperscript{28}, due to a reduction of partial volume effects.

**Reproducibility in longitudinal studies**

No study was found to assess the reproducibility of any longitudinal measurement for CMB.

### 3.2.4 Brain volume

**Definition and interest in SVD**

Brain volumetry refers here to measures obtained through computational segmentation of high-resolution anatomical images. Global brain volume, intracranial volume (classically used for normalization purpose) and brain parenchymal fraction (BPF, brain volume divided by intracranial volume) were considered in this review as these are important markers in SVD \textsuperscript{80,81,82,83-85}. For example, BPF has been associated with dementia rating scales and the modified Rankin score in CADASIL\textsuperscript{81,86}, a monogenic SVD\textsuperscript{2}. In sporadic forms, brain volume was notably associated with the presence and progression of lacunes\textsuperscript{87}. Gray matter, white matter or hippocampus volumes are sometimes specifically assessed and can be markers of interest in SVD studies, but were not included in this review as these structures depend more on post-processing stages. However, no study amongst any of the following has assessed the influence of SVD markers, particularly WMH, on the performance of any brain volume measurement methods.
Within-centre reproducibility

Overall, several studies reported a satisfactory within-centre reproducibility of brain volumetric measurements at 1.5T, 10, 13-15, 17, 18 or 3T10, 14, 15, 18 with the main brain image processing suites (FreeSurfer88, 89, SPM90, FSL91,BrainVisa92). Table 1 presents per study the numbers of centres, subjects, rescans per subject and main results for reproducibility. For example, in a large cohort of 671 subjects who underwent two successive 3D T1 weighted scans, the median of the relative absolute difference in brain volume was found to be less than 1%17. However, the 90th percentile of the absolute value of the difference was reported at 5.11%, which is more than twice the value that would be expected for a Gaussian distribution. This highlights the importance of studying several statistical outcomes to infer the reproducibility beyond the median or mean value of the difference. Thus, little variability is added when measuring brain volume in a single centre without changing the conditions of acquisition, at least regarding commonly reported disease-related changes. It is worth noting that most studies with a scan-rescan design included small samples (all n< 50 except Cover et al.17) and recruited mainly healthy subjects (two studies with mild cognitive impairment (MCI) and AD patients14, 17 and one study in multiple sclerosis16).

Between-centre reproducibility (effect of various vendors, magnetic field strengths, coils)

Effects of combining several MRI vendors, static magnetic field strengths or coil designs on brain volumetry can be assessed in studies where subjects are scanned several times in different centres (Table 1). As expected, compared to the within-centre variability, the between-centre volume variability was much higher at 1.5T7, 8, 10, 11, 14, 15, 18, 65 and 3T10, 12, 14, 15, 18. In one of the largest multicentre studies with a scan-rescan procedure (a subset of ADNI database in which 172 subjects were rescanned once in 59 different MRI sites on 17 different scanner types), the effects of scanner hardware and imaging protocol on volume quantification were analyzed14. The most striking result was that the variance in volumetric measures was 10 times higher when subjects were imaged on different scanners compared to acquisitions on the same scanners. For example, for brain volume, the median (standard deviation) of the absolute within-subject variability was 0.3% (0.5%) vs. 4.8% (4.6%) for repeated scans on the same scanner vs. on different scanners. This was essentially explained by the scanner-dependent geometrical inaccuracies and by the differences in the gray to white matter tissue contrast from variation in the imaging sequences used and RF coils (while this contrast is obviously impacting gray matter and white matter volumes, it also influences the total brain volume computation14). Such geometric distortions originate from gradient field nonlinearity (in theory, a gradient system is supposed to produce an incremental magnetic field varying linearly with distance from isocenter) and the static field inhomogeneity, whilst the contrast between brain tissues is determined by a variety of sequence parameters. Both factors are dependent on the scanner hardware.
(a combination of vendor, magnetic field and coil systems), which explained 30 to 50% of the variance of any volumetric measure in that study. In another study, 15 subjects underwent 4 scan sessions at 3 sites with two different vendors\textsuperscript{10}. Interestingly, combining data of different vendors and/or field strength did not significantly change the standard deviation of the volume differences across conditions relative to the test–retest reproducibility within a fixed MRI system. However, the mean volume difference did change within one standard deviation from zero for intracranial volume.

**Upgrades and repair**

Only a few studies have examined the effects of MR systems upgrades and repairs on volumetric measurements. It appears that mixing volumes derived from data acquired across a major scanner upgrade (such as replacement of the magnet, gradient system and software) do not show significant changes, although volumes of some specific structures may show a slight scanner specific volume bias\textsuperscript{10}. Similarly, consistent measures for brain structure volumes have been found before and after gradient coil replacement in a 3T system\textsuperscript{93}. However, inadvertent differences in sequence parameters for example may be introduced during the upgrade (human error)\textsuperscript{94} which can cause major variance between pre- and post-upgrade scans if not detected quickly. Scanner upgrade or replacement are almost impossible to avoid, and therefore centres should consider having longitudinal phantom and volunteer assessments for quality assurance and to calculate differences in brain volumetry across time that would allow conversion factors to be generated.

**Sequence effect**

Use of 3D inversion prepared T1-weighted fast gradient-echo sequences for brain tissue segmentation and volumetry is universal nowadays. It is widely used for its contrast properties and is a standard for most MRI vendors. It is therefore easy to implement in multicentre studies. Other sequences such as multi-echo fast low angle shot sequences have similar contrast properties and showed comparable volume reproducibility from Bland-Altman plot analysis than the MPRAGE sequence for example\textsuperscript{10}. However, comparison between volumes derived from both sequence types demonstrated significant biases in the mean volume difference, probably due to a differing $T_2^*$ sensitivity\textsuperscript{10}.

Isotropic voxels are superior to non-isotropic voxels for accurate estimation of brain volume. In sequences with non-isotropic resolution, the choice of image orientation (sagittal, coronal or axial) naturally yields volumetric differences when comparing datasets acquired with different image orientations\textsuperscript{95}. These differences may be due to partial volume averaging and susceptibility-induced geometric distortions. It is anticipated that these findings can be generalized to all MRI-derived metrics of brain volume.

**Contrast effect**
Variability of the tissue contrast, defined as the ratio between gray and white matter tissue image intensities on T1-weighted images, relates to the choice of MR sequence parameters such as echo time (TE), repetition time (TR), inversion time (TI) and flip angle (FA) along with the selection of field strength and coil. The contrast has been shown to have non-negligible influence on brain tissue segmentation and thus volumetry. For example, a 6% change in the average contrast led to a 2% change in the computed brain volume, which is of a similar size as the annual change in brain volume seen in the early stages of AD so could easily be interpreted as a disease effect. Notably, datasets with a higher contrast have a lower variability, but here the effect of different burdens of WMH have not been tested.

**Subject-related factors**

Physiological variability within the subject being imaged is a source of variability beyond the disease-related variability. It has even been suggested in a study of BPF variability in multiple sclerosis that the variability due to physiological and pathological causes is important and likely larger in magnitude than scan–rescan repositioning effects. The authors found that the variance assigned to physiological fluctuations was nearly two times greater than the variance due to patient repositioning during scan-rescan. This implies that scan–rescan experiments might only provide a lower bound on the true error in repeated volumetric measurements from MRI exams. It is well known that the hydration status of an individual can affect the measurement of brain volume. Specifically, lack of fluid intake for 16 hours has been found to result in a 0.55% decrease in cerebral volume and re-hydration an increase of 0.72% in brain volume. In the ADNI dataset and a large cohort of multiple sclerosis patients there was a significant effect of time-of-day on the BPF, with a greater brain volume in the morning. Similarly, brain volume varies with plasma sodium. These results suggest potential acquisition time and hydration biases should be randomized or statistically controlled to account for the day-to-day brain volume fluctuations.

**Factors with no or little effect**

Correction for so-called B1 inhomogeneities (variations of image intensity due to the radiofrequency field) was not found to substantially change the reproducibility of whole-brain volumetric measurements, unlike its effect on WMH and focal infarcts. Similarly, some studies found that greater acceleration factor in parallel imaging (this technique combines the signals of several coil elements in a phased array to reconstruct the image and finally improve the signal-to-noise ratio or reduce the scan time) had very little influence on test–retest accuracy. However, these studies tested the influence of acceleration factor in healthy controls. Results might be different if older aged subjects or patients move during the acquisition of the reference lines part of the parallel imaging
strategy. Finally, reproducibility of brain volumetric measures was also relatively unaffected by small differences in voxel geometry compared to the effects observed on measures of cortical thickness\textsuperscript{12}.

3.2.5 Brain atrophy

**Definition and interest in SVD**

Brain atrophy is the difference of total brain volume between two time points computed by a dedicated image processing software from high-resolution anatomical images. Quantitative measures derived from high-resolution anatomical images are considered over atrophy scales (not covered in the present review) as the quantitative measures have better sensitivity to change over time\textsuperscript{101}. Surprisingly, very few studies reported brain atrophy measurement in SVD cohorts\textsuperscript{102}, \textsuperscript{103}. Several methods are available, such as SIENA, which estimates the percentage brain volume change of the same subject between two time points, by registering the two brain images together and resampling into the space halfway between the two\textsuperscript{104}, \textsuperscript{105} and the Boundary Shift Integral (BSI), that is an automated method that determines the total volume through which the boundaries of a given cerebral structure have moved\textsuperscript{106}. One alternative method to assess brain atrophy is by performing tissue segmentation at both time points, but this approach suffers from a higher variability given the added variability of each measurement\textsuperscript{22}.

**Within-centre reproducibility**

Within-centre reproducibility of quantitative image-processing methods for atrophy assessment have been shown to be relatively high in several studies (Table 1)\textsuperscript{17,9,19}. For example, a study using the ADNI dataset reported a 0.35\% mean difference in 385 subjects acquired in more than 50 sites (105 healthy controls, 195 subjects with MCI and 85 with AD) undergoing two SIENA measurements 12 months apart\textsuperscript{17}. Further, the 90th percentile of the difference was much larger than expected (1.33\%) for a Gaussian distribution. These results should be compared to the annualized mean percentage brain volume change obtained in that study: -0.65\% ± 0.82\% for healthy controls, -1.15\% ± 1.21\% for MCI, -1.84\% ± 1.33\% for AD.

**Between-centre reproducibility (effect of various vendors, magnetic field strengths, coils)**

Some studies have examined the between-centre reproducibility of atrophy measurements using SIENA\textsuperscript{17,22,107} or other image-processing techniques\textsuperscript{22}. It has been shown that even with scanner of the exact same model, scanner drift (a shift in the magnet's field) and inter-scanner variability may cancel out longitudinal brain volume change\textsuperscript{44}. It is worth noting that compared to voxel-based morphometry approaches, SIENA, which corrects for differences in imaging geometry using the outer skull surface for both time points, reduced the effects of scanner drift and inter-scanner variability on longitudinal morphometric results\textsuperscript{107}. Similarly, other registration-based algorithms such as BSI
show better reproducibility compared to segmentation-based techniques such as Freesurfer or SIENAX\textsuperscript{22}. However, segmentation-based methods using atlases can be improved by using fine registration methods and age-relevant templates\textsuperscript{108}.

Evaluation of coil effects on volumetric measurements showed that a change from a 12- to a 32-channel coil was responsible for a shift of 0.5\% in the whole brain BSI\textsuperscript{100}, which appears unacceptable as it corresponds to the annual atrophy rate observed in healthy subjects\textsuperscript{109}.

Finally, analysing the back-to-back scans of the ADNI dataset in 200 pairs of images (118 subjects with 0-12 months pairs and 82 subjects with 6-24 months pairs), it has been reported that 3T was no more reproducible than 1.5T for the whole brain atrophy measure using SIENA\textsuperscript{110}.

**Contrast effect**

Atrophy measurement is strongly dependent on image contrast. Indeed, the error using BSI may exceed 100\% if image contrast properties dramatically differ between the two scans in a measurement pair\textsuperscript{111}. For example, in 10 volunteers, a mean brain volume change of -0.43\% has been measured versus a +1.85\% change when the second scan was acquired with a flip angle equal to 12° instead of 25° (the flip angle directly impacts the image contrast).

**Noise effect**

Noise is an unwanted random signal from various origins including: the subject's body, coils, electronics, nearby machinery, etc. The signal-to-noise ratio is often computed as the ratio of the average signal intensity over the standard deviation of the noise. Signal-to-noise ratio directly impacts BSI values as, for example, the position of the brain surface is seen to retract as noise in the second time point scan increases\textsuperscript{111}. Thus, the average percent volume change progressively diverges from zero with increasing noise mimicking increased brain atrophy.

**Subject-related factors**

Motion will have an impact on atrophy measurement. For example, it has been reported that, as the amplitude of the motion in the second scan increases, the mean absolute brain and ventricular BSI brain volume changes as well as the associated standard deviation increase progressively\textsuperscript{111}. However, the different levels of motion have been simulated in this latter study and the severity may have been overestimated.

Subject’s head positioning in the magnet also influences brain atrophy measures. Indeed, nonlinear gradient distortions associated with shifts in the long magnet’s axis can significantly decrease the accuracy and precision of MRI-derived measures of brain atrophy assessed by SIENA\textsuperscript{112}. These negative effects increase in magnitude with: (i) increases in the Z-distance (long magnet’s axis) between the brains to be compared at two time points; and, (ii) increases in the Z-distance from magnet isocenter to the center of the pair of brains to be compared\textsuperscript{112}. Typical gradient distortion field for a T1-
weighted acquisition is presented in Fig. 4. It shows the dramatic shifting experienced by voxels especially at the border of the brain and the need for accurate correction and repositioning of patients in longitudinal studies.

3.3 Techniques to manage variability

We have highlighted a large number of reasons why quantitative MRI markers of SVD may be variable beyond subject-related or disease-related factors. These sources of variability introduce significant barriers, and potential biases, to multicentre imaging studies of SVD. Fortunately, while adding variability to the measurements of markers is inherent to imaging studies, some techniques have been designed to reduce it.

First, an optimised MRI protocol with high contrast-to-noise ratio, minimal artefacts and good reproducibility must be determined. Hereto, a methodological approach has been proposed to target a protocol with the highest possible reproducibility based on the assessment of quality measures\textsuperscript{21}. The principle is to acquire several datasets including: within-subject within-centre datasets, between-subjects between-centre datasets and between-subject within-centre datasets. By analysing the summary data quality and quantitative measures extracted with dedicated and validated software the desired optimal MRI protocol can be determined.

Second, to account for the problem of upgrades in longitudinal studies, Jovicich et al. advise planning a 'system upgrade calibration study' as part of the design, with subjects scanned shortly prior to and immediately after the upgrade so as to derive a correct estimation for potential systematic biases\textsuperscript{10}.

Third, several corrective techniques based on the MR system characteristics\textsuperscript{113, 114} or using a dedicated phantom\textsuperscript{112, 115, 116} have been developed to reduce the effects of gradient distortion on imaging reproducibility. The advantage of using a phantom travelling from one site to another is to apply a unified correction independently of the vendor’s own corrections. Correction for gradient nonlinearity and intensity non-uniformity reduces the variance of longitudinal changes in brain volumes and improves the accuracy for detecting subtle volumetric changes\textsuperscript{117}. For example, in 208 subjects, the raw mean percentage of brain volume change was \(-0.59\%\) against \(-0.46\%\) when images were corrected for gradient nonlinearity and intensity non-uniformity\textsuperscript{117}. However, if the expected group difference far exceeds brain volume change related to geometric distortions, the extra time and expense associated to the phantom-based correction needs to be carefully evaluated.

Finally, the use of dedicated statistical methods is important in analysing multicentre MRI data. A simple\textit{ a posteriori} method to account for scanner hardware variability (combination of different
devices, coil types and field strengths) is to include the scanner hardware as a covariate in regression models. Moreover, it has been shown that T1-weighted regional anatomical brain volume data can be reliably combined across 1.5T and 3T centres with the application of an appropriate correction procedure (regression-based correction function when data varied in a linear and systematic fashion). More sophisticated statistical methods such as linear mixed-effect models including protocol as well as individual MRI acquisition parameters have also emerged to more accurately model longitudinal changes (BPF and volume of WMH). Another statistical approach is based on the statement that classical inference testing is inappropriate because it is designed to detect differences and not to prove similarity. Equivalence testing has then been applied to determine if MRI-based measurements obtained under variable conditions can be pooled. This could be generalized to all MRI-derived measurements. Otherwise, Moorhead et al. used scanner specific priors (probability mappings of tissue occupancy) in the segmentation pipeline to reduce tissue classification differences between scanners to improve statistical power. Interestingly, they found that the between scanner differences were not reduced to the level of within scanner variability, the ideal for scanner harmonisation.

Ultimately, central application of automatic quality assessment and image processing on raw data, using consistent techniques, can reduce the variability introduced by a posteriori image processing across centres.

3.4 Suggestions for future research

A between-centre variability study in a large group of subjects, representing a range of types and burdens of SVD lesions, to assess the effects of vendor and field strength (protocols as close as possible) on reproducibility of SVD lesions and brain volumes is missing. So are experiments examining the effects of changing coil, correcting for gradient distortion or field heterogeneities, upgrading the system, or varying classical parameters such as TE, TI or TR, on the final WMH volume. Finally, WMH appears to be a reliable marker in single centre studies where all acquisition and processing parameters are kept constant, but additional studies are missing to infer the true reproducibility of WMH measures in a multicentre context.

As for WMH, a multicentre study assessing the effect of scanner hardware (same magnetic field but different vendors and models) on CMB detection with a given optimized protocol would greatly add to the knowledge on the reliability of CMB detection. New data on methods of automated detection of CMB throughout different brain regions are needed. Also, the analysis of the potential effects of coil, and voxel size would be particularly useful. Finally, CMB can be used as a reliable marker in single
centre studies but additional efforts are needed to determine the sources of variability in multicentre data.

For lacunes and PVS, more automated and validated image processing algorithms are required, followed by assessment of reproducibility through both within and between centres acquisitions. In view of the likely interactions between acquisition protocols and performance of analysis methods, the assessment of all SVD features should be included in any study that assesses reliability of WMH and/or CMB and/or atrophy quantification.

4. Ten key points for assessing quantitative SVD markers in multicentre studies

Based on our review of the current literature regarding reproducibility and acquisition-related sources of variability of MRI markers in SVD and on personal experiences, we propose the following key points should be considered when designing or analysing a multicentre MRI study in SVD. Obviously, prior to acquisitions, the scientific question should guide which marker is the focus, although opportunities to assess all SVD markers should be obtained where possible. Optimising acquisition for contrast, signal-to-noise ratio and spatial resolution is particularly important when imaging elderly or cognitively impaired patient populations who will not tolerate prolonged acquisition times. Importantly, although it is essential to have accurate and reproducible measurements, the nature of the questions will be guiding the level of precision that is needed. Requirements might be different between interventional clinical trials with relatively short term duration and long-term population-based observational studies. For example, in some situations, the added noise due to multicentre variability will not alter a strong statistical association between one SVD-related MRI marker and one clinical outcome. Conversely, studying the quantitative evolution of one SVD-related marker during a short follow-up will require a very fine quantification that necessitates a variability caused by technological factors as small as possible. Hence we organized our recommendations as being a) advisable whatever the question guiding the study as they tackle effects being important and/or involve easy-to-follow techniques to manage variability, b) advisable in some conditions, i.e. when the expected effect of interest is small relatively to the unwanted noise caused by the multicentre setting.

a) strongly advisable and easy-to-follow recommendations

1. Automated quality checks for MRI sequence parameters are advisable (for each scan, comparison of the acquisition parameters with the intended protocol for each scanner). Also,
images should be monitored for motion and other artefacts such as magnetic susceptibility artefacts, aliasing artefacts or Gibbs artefacts. A minimum standard for quality control will have to be defined to exclude problematic data (for example, based on the severity of artefacts or from automatic measures of signal to noise ratio, background noise, etc.). Indeed, even if the computational method seems to run properly, artefacts can dramatically alter the data.

2. MRI gradient nonlinearities have to be corrected at least with corrections proposed in options by the manufacturer as they can affect dramatically measures of brain atrophy.

3. Subject’s positioning in the scanner is another key aspect and strict instructions for MR technicians may be particularly useful to reproduce confidently the position of the head, from one subject to another, and from the first time point to the following time points, in longitudinal studies.

4. Considering the hardware, the same coil has to be used during an MR protocol both in cross-sectional and longitudinal studies. When upgrades or repairs are needed during the study period, a strategy should be in place to scan some participants twice, before and after the upgrade, to estimate the potential bias for a posteriori analyses. Otherwise, the imaging protocol and hardware should be kept as constant as possible over the whole time of the study.

**b) recommendations to evaluate given the targeted trade-off between sensitivity and variability**

5. Within centre reproducibility is much higher than between-centre reproducibility. The need for a multicentre setup should then be carefully evaluated while taking into account specific aspects such as number of participants, subject transportation, data centralization, quality control and cost. For multicentre studies, the use of only one vendor or only some models can be an option.

6. In multicentre studies, protocol harmonization is a mandatory step to obtain similar sequences with acquisition parameters as close as possible. There is no way to get a perfect harmonization of MRI protocols between different vendors and models. Therefore, it is a trade-off between a high reproducibility of markers (few centres, same vendors) and a sufficient statistical power through the recruitment of a large number of subjects, generally in several centres with different scanner vendors and models. Specifying acceptable ranges for different acquisition parameters may help obtain satisfactory harmonization for large-scale clinical studies.
7. Spatial resolution is a key parameter in assessing correctly the variation of any MRI marker such as brain volumetry, longitudinal atrophy, volume of WMH or quantification of lacunes and PVS. Increasing spatial resolution decreases signal-to-noise ratio and increases acquisition duration. Sensitivity for quantifying small objects such as lacunes and PVS will be compromised if using voxels as big as the object itself or where there is a slice gap. Selection of optimal spatial resolution must therefore be targeted to the primary question.

8. When analysing multicentre MRI data, the use of statistical models to take into account the induced variability is recommended (adjusting on centre or using random effects models for example).

9. While some quantitative MRI markers of SVD are regularly assessed (brain atrophy, WMH volume), others are difficult to quantify given their small size in relation to conventional anatomical MRI (lacunes, PVS) or their strong dependence on MRI parameters (CMB). High-field MRI and detailed studies of these particular markers may be performed in studies nested within larger studies assessing more common markers in a clinical setting. For CMB characterization, SWI or quantitative susceptibility mapping are promising techniques. Since the total magnetic susceptibility of a CMB is an intrinsic physical property independent of imaging characteristics, mapping CMB by using quantitative susceptibility mapping may be an alternative approach more consistent over a wide spectrum of imaging parameters.

10. It is important for studies to acknowledge that variation is inevitable and they should detail the steps that have been made to minimise this variation.

Conclusion

There is a need to reduce image variability induced by technological choices (coil, MR sequence, MR parameters and options) as much as possible in large multicentre research studies and clinical trials using MRI in SVD. Indeed, the added variance due to the multicentre setup can be deleterious in various applications and should be minimised using multiple approaches. Protocol harmonization and a posteriori corrections can help in minimizing biases; minimisation on centre or scanner type may also be advisable. Even if several sources of variability have been previously evaluated, most studies have focused on brain volume measurements while few have focused on WMH
Further, less is known concerning the sources of variability when quantifying lacunes, CMB and PVS. Additional studies are warranted to investigate the variability of these makers, which is a key aspect for future longitudinal or multicentre studies. Other potential SVD-related MRI markers derived from new and multiple techniques (such as diffusion tensor imaging, arterial spin labelling, magnetization transfer mapping, blood-brain barrier permeability imaging, BOLD functional MRI, cortical morphology, cortical microinfarcts, cortical superficial siderosis) will also have to address these important questions regarding their reproducibility to facilitate large-scale, multicentre research.

**Author contribution**

FDG, EJ and HC designed the review. FDG collected and analysed the literature. FDG, EJ and HC wrote the initial draft of the review. JMW critically revised the manuscript. The manuscript was then sent to the imaging and small vessel disease experts involved in the COENgroup (network of Centres of Excellence in Neurodegeneration, [http://www.coen.org](http://www.coen.org)). All authors of the present manuscript agreed to contribute and revised carefully the manuscript. Finally, all authors accepted the current version.

**Disclosure/conflict of interest**

The authors declare no conflict of interest.

Supplementary material for this paper can be found at

http://jcbfm.sagepub.com/content/by/supplemental-data

**References**


Titles and legends to figures

Figure 1 title: Schematic view of the different sources of variability in a multicentre MRI study (adapted from www.imaios.com)

Figure 2 title: Enhanced lesion detection in 3D FLAIR sequence (courtesy of Polak et al.42)
Figure 2 legend: Sagittal slices from (a, d) 3D FLAIRED (a FLAIR sequence optimized in order to enhance detection of lesions); (b, e) GE 3D FLAIR; and (c, f) 2D FLAIR. Note increased lesion detection in the frontal lobe (1) deep white matter (2) and juxtacortical (3) regions with 3D FLAIRED. Improved deep white matter lesion (4) detection and resolution in 3D FLAIRED compared to other sequences. Cerebellar lesion (5) detected in 3D FLAIRED and GE 3D FLAIR, but uncertainly discerned in 2D FLAIR.

Figure 3 title: Effect of slice thickness, sequence and magnetic field on CMB detection (courtesy of Nandigam et al.28)
Figure 3 legend: The pairs of images illustrate comparisons of thick-section GRE (A) versus thin-section SWI (B); thin-section GRE (C) versus thin-section SWI (D); thick-section GRE (E) versus thin-section GRE (F) (all preceding images at 1.5T); and SWI at 1.5T (G) versus SWI at 3T (H). The black arrows in (A) and (B) illustrate a CMB prospectively counted on both sequences, whereas lesions denoted by white arrows were initially identified only on the SWI image. The black arrows in the remaining images highlight lesions on the paired images for comparison.

Figure 4 title: Typical gradient distortion field in a T1-weighted acquisition (courtesy of Caramanos et al.112)
Figure 4 legend: Sagittal, coronal, and axial views of a phantom-based gradient distortion field that was determined based on a T1-weighted acquisition. The color scale represents the distance (in millimeters) that a voxel moves because of gradient distortion between its “real” location and its “apparent” location on such an MRI scan that has not been corrected for the gradient distortion.
## Table 1. Summary of reproducibility studies (within-centre or/and between centre scan-rescan procedure)

<table>
<thead>
<tr>
<th>Study</th>
<th>Yearpublished</th>
<th>Marker of interest</th>
<th>Between centre</th>
<th>Within centre</th>
<th>Magneticfield</th>
<th>Analysistool</th>
<th>Between-centre reproducibility</th>
<th>Within-centre reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>First author</td>
<td></td>
<td></td>
<td>Nb subjectsres canned</td>
<td>Nb of MRI sites</td>
<td>Nb rescans per subject</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schnack</td>
<td>2004</td>
<td>ICV, BV, GMV, WMV</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1.5T, 1T own segmentation</td>
<td>BV: ICC = 0.98 BV: ICC = 1</td>
</tr>
<tr>
<td>Ewers</td>
<td>2006</td>
<td>GMV, WMV, CSFV</td>
<td>1</td>
<td>10</td>
<td>NA</td>
<td>NA</td>
<td>1.5T</td>
<td>CoV=5% for all volumes NA</td>
</tr>
<tr>
<td>Jovicich</td>
<td>2009</td>
<td>ICV</td>
<td>15</td>
<td>3</td>
<td>15</td>
<td>2</td>
<td>1.5T, 3T Freesurfer</td>
<td>NA Reproducibility error 2.56%</td>
</tr>
<tr>
<td>Reig</td>
<td>2009</td>
<td>ICV, GMV, WMV, CSFV</td>
<td>5</td>
<td>5</td>
<td>NA</td>
<td>NA</td>
<td>1.5T SPM-based</td>
<td>ICV: CoV=1.9%, ICC=0.99 NA</td>
</tr>
<tr>
<td>Wonderlick</td>
<td>2009</td>
<td>BV, GMV, WMV</td>
<td>11</td>
<td>2</td>
<td>NA</td>
<td>NA</td>
<td>3T Freesurfer</td>
<td>BV: ICC=0.99 NA</td>
</tr>
<tr>
<td>De Boer</td>
<td>2010</td>
<td>BV, GMV, WMV, CSFV</td>
<td>NA</td>
<td>NA</td>
<td>30</td>
<td>2</td>
<td>1.5T FSL, SPMS, kNN</td>
<td>NA</td>
</tr>
<tr>
<td>Huppertz</td>
<td>2010</td>
<td>BV, GMV, WMV, CSFV, ICV</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>1.5T, 3T SPM5</td>
<td>BV: CoV=3.78%; BV: CoV=2.97%</td>
</tr>
<tr>
<td>Kraugge</td>
<td>2010</td>
<td>ICV, BV, GMV, WMV</td>
<td>172</td>
<td>59</td>
<td>41</td>
<td>2</td>
<td>1.5T, 3T Fantasm</td>
<td>BV: median RAVD=4.76%, 90th percentile RAVD=12.34%; ICV: median RAVD=1.66%, 90th percentile RAVD=3.92% BV: median RAVD=0.30%, 90th percentile RAVD=1.09%; ICV: median RAVD=0.29%, 90th percentile RAVD=1.27%</td>
</tr>
<tr>
<td>Sampat</td>
<td>2010</td>
<td>BPF</td>
<td>NA</td>
<td>NA</td>
<td>18</td>
<td>2</td>
<td>1.5T own segmentation</td>
<td>NA</td>
</tr>
<tr>
<td>Cover</td>
<td>2011</td>
<td>Normalized brain volume</td>
<td>NA</td>
<td>NA</td>
<td>671</td>
<td>2</td>
<td>1.5T</td>
<td>SIENAX</td>
</tr>
<tr>
<td>Shokouhi (1.5T)</td>
<td>2011</td>
<td>ICV, GMV, WMV, CSFV</td>
<td>13</td>
<td>3</td>
<td>13</td>
<td>2</td>
<td>1.5T BrainVISA</td>
<td>BV: reliability=0.97 BV: reliability=0.99</td>
</tr>
<tr>
<td>Shokouhi (3T)</td>
<td>2011</td>
<td>ICV, GMV, WMV, CSFV</td>
<td>11</td>
<td>5</td>
<td>11</td>
<td>2</td>
<td>3T BrainVISA</td>
<td>BV: reliability=0.88 BV: reliability=0.96</td>
</tr>
<tr>
<td>Landman</td>
<td>2011</td>
<td>GMW, WMV</td>
<td>NA</td>
<td>NA</td>
<td>21</td>
<td>2</td>
<td>3T Java Imaging Science Toolkit</td>
<td>NA</td>
</tr>
<tr>
<td>Chalavi</td>
<td>2012</td>
<td>GMW, WMV</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3T Freesurfer, SPM</td>
<td>NA</td>
</tr>
<tr>
<td>Huang</td>
<td>2012</td>
<td>WMV, CSFV</td>
<td>6</td>
<td>2</td>
<td>NA</td>
<td>NA</td>
<td>3T FSL</td>
<td>WMV: ICC=0.62; CSFV: ICC=0.87</td>
</tr>
<tr>
<td>Pfefferbaum</td>
<td>2012</td>
<td>Supratentorial volume</td>
<td>114</td>
<td>2</td>
<td>NA</td>
<td>NA</td>
<td>1.5T, 3T Freesurfer</td>
<td>BV: ICC=0.81 NA</td>
</tr>
</tbody>
</table>
### Brainatrophy

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Methodology</th>
<th>Segmentation Algorithms</th>
<th>Imaging Parameters</th>
<th>Segmentation Software</th>
<th>CoV %</th>
<th>Coefficient of repeatabilityb</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith</td>
<td>2007</td>
<td>PBVC</td>
<td>NA</td>
<td>1.5T</td>
<td>SIENA, BSI</td>
<td>NA</td>
<td>SIENA: median RAVD=0.16%; BSI: median RAVD=0.17%</td>
<td></td>
</tr>
<tr>
<td>De Bresser</td>
<td>2011</td>
<td>PBVC</td>
<td>NA</td>
<td>1.5T</td>
<td>SIENA, SPM8, kNN</td>
<td>NA</td>
<td>coefficient of repeatabilityb=0.92% for SIENA, 2.64% for SPM8, 0.31% for kNN</td>
<td></td>
</tr>
<tr>
<td>Cover</td>
<td>2011</td>
<td>PBVC</td>
<td>NA</td>
<td>1.5T</td>
<td>SIENA</td>
<td>NA</td>
<td>PBVC: median RAVD=0.35%, 90th percentile RAVD=1.37%</td>
<td></td>
</tr>
<tr>
<td>Durand-dubief</td>
<td>2012</td>
<td>Atrophy</td>
<td>7 segmentation algorithms</td>
<td>1.5T</td>
<td>SEM: 1.77 for Freesurfer, 0.98 for BSI, 0.85 for SIENA</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### White-matterhypenintensities

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Methodology</th>
<th>Segmentation Algorithm</th>
<th>Imaging Parameters</th>
<th>CoV range</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wei</td>
<td>2002</td>
<td>WMH volume</td>
<td>NA</td>
<td>1.5T</td>
<td>2.6 - 7.5% according to pipeline</td>
<td></td>
</tr>
<tr>
<td>Scotte</td>
<td>2003</td>
<td>WMH volume</td>
<td>NA</td>
<td>1.5T, 3T</td>
<td>median RAVD: 12.5%</td>
<td>NA</td>
</tr>
<tr>
<td>De Boer</td>
<td>2010</td>
<td>WMH volume</td>
<td>NA</td>
<td>1.5T</td>
<td>CoV: 5.87%</td>
<td></td>
</tr>
<tr>
<td>Theysohn</td>
<td>2011</td>
<td>WMH grading</td>
<td>NA</td>
<td>1.5T, 7T</td>
<td>75% of subjects with equal grading at 7T compared to 1.5T, 13% superior and 13% inferior</td>
<td>NA</td>
</tr>
<tr>
<td>Ramirez</td>
<td>2013</td>
<td>WMH volume</td>
<td>NA</td>
<td>1.5T</td>
<td>ICC=1; mean AVD in 10 older subjects=213mm³</td>
<td></td>
</tr>
</tbody>
</table>

### Cerebral microbleeds

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Methodology</th>
<th>Segmentation Index and Diameter</th>
<th>Imaging Parameters</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stehling</td>
<td>2008</td>
<td>CMB number</td>
<td>25</td>
<td>1.5T, 3T</td>
<td>more CMB at 3T vs 1.5T</td>
</tr>
<tr>
<td>Nandigam</td>
<td>2009</td>
<td>CMB contrast index and diameter</td>
<td>4</td>
<td>1.5T, 3T</td>
<td>smaller contrast index and diameter at 1.5T vs 3T</td>
</tr>
<tr>
<td>Conijn</td>
<td>2011</td>
<td>CMB number</td>
<td>34</td>
<td>1.5T, 7T</td>
<td>more CMB at 7T vs 1.5T</td>
</tr>
<tr>
<td>Theysohn</td>
<td>2011</td>
<td>CMB number</td>
<td>20</td>
<td>1.5T, 7T</td>
<td>more CMB at 7T vs 1.5T</td>
</tr>
</tbody>
</table>

ICV: intracranial volume, BV: brain volume (GMV + WMV), GMV: gray matter volume, WMV: white matter volume, CSFV: cerebrospinal volume, BPF: Brain parenchymal fraction, WMH: white matter hyperintensities, CMB: cerebral microbleeds, ICC: Intraclass correlation coefficient, CoV: coefficient of variation, NA: not available, Reproducibility error: The group reproducibility error for each structure is derived averaging the reproducibility errors across subjects, where for each subject the error is estimated as the absolute test–retest volume percent change relative to the mean test–retest volume, kNN: k-Nearest-Neighbors segmentation method, RAVD: relative absolute value of the difference, PBVC: percent brain volume change, a reliability in this study has been computed as the ratio of the variance excluding the contribution from the factor (within-centre or between-centre) to the total variance, b coefficient of repeatability: 1.96 times the standard deviation of the differences between two measurements (Bland-Altman), SEM: standard error of measurement, AVD: absolute value of the difference.