1. Introduction

Multiple sclerosis (MS) is a disease of the central nervous system (CNS) that typically affects both the brain and the spinal cord (Compton and Coles, 2008). In addition to the well-known focal inflammatory, demyelinating lesions that are typically seen in the white matter (WM), MS also causes degeneration and consequent volume loss of grey matter (GM), which is often referred to as GM atrophy. The focus of the present paper will be on measurement of GM atrophy in the brain, which is currently widely accessible as opposed to spinal cord GM atrophy measurement. Nevertheless, GM atrophy in the spinal cord is an important topic, accurate measurement of which can contribute to better understanding MS and quantitatively characterizing its effects in individual patients in trials or clinical care. However, technical challenges in quantification and interpretation of brain grey matter atrophy in individual MS patients using MRI are currently hampering the measurement and interpretation of GM atrophy in MS. The focus is on measurements that are obtained in individual patients rather than on group analysis methods, because of their importance in clinical trials and ultimately in clinical care. We discuss the sources and possible solutions of the current challenges, and provide recommendations to achieve reliable measurement and interpretation of brain GM atrophy in MS.
improvements are needed to obtain high resolution images at reduced scan times and to increase coverage of the spinal cord, before broader application is feasible (Kearney et al., 2015).

Brain GM atrophy is typically measured in vivo from standard 3D T1-weighted images acquired by magnetic resonance imaging (MRI), using automated analysis methods. Studies have demonstrated volume decrease of subcortical GM structures and volume or thickness decrease of cortical regions (Bergland et al., 2012; Pareto et al., 2016). GM atrophy in the brain has been shown to be associated with cognitive impairment in MS (Nocentini et al., 2014; Riccietti et al., 2011). Furthermore, atrophy of the deep GM (DGM) structures, such as thalamus, has been shown to be strongly related to clinical and cognitive decline (Bermel et al., 2003; Houchens et al., 2007; Pagani et al., 2005).

This paper, after reviewing the measurement techniques, will discuss the most important challenges for reliable measurement and interpretation of brain GM atrophy in patients with MS. We divide these challenges into two categories. Challenges in the category “Pathology, physiology and treatment effects” are: 1) Unclear pathological substrate, 2) Evolution of GM atrophy, 3) Influence of physiological variability, and 4) Evaluation of treatment response. “Measurement challenges”, the second category, includes: 1) Influence of WM lesions, 2) Influence of atrophy itself, 3) Influence of other pathology, and 4) Technical variability. For each section, we provide specific recommendations (summarized in Box 1) to improve measurement and interpretation of GM atrophy in individual MS patients.

2. Review of measurement techniques

Various methods have been developed to measure anatomical changes of the brain. Some of these techniques produce single-subject measurements while others, such as voxel-based morphometry and vertexwise analyses, provide group-based statistical tests. In the current paper we are predominantly interested in the application of brain GM atrophy measures in individual MS patients, because single-subject measures are the most relevant outcome measures in clinical care and clinical treatment trials. Therefore, this review discusses cortical and deep GM atrophy measurement techniques with an emphasis on methods that produce single-subject results, while briefly reviewing relevant aspects of methods for group-level analysis used in the literature.

2.1. Cortical atrophy measurement techniques

There is convincing evidence that MS involves structural cortical changes resulting in atrophy (Charil et al., 2007; Lansley et al., 2013; Narayana et al., 2012; Prinster et al., 2006; Ramasamy et al., 2009; Sailler et al., 2003). Current MRI technology provides a valuable tool for detecting and monitoring cortical atrophy early in the disease (Chard et al., 2002; Chard et al., 2004; Dalton et al., 2004; De et al., 2003; Sastre-Garriga et al., 2004; Sepulcre et al., 2006). For reproducible and accurate estimation of cortical thickness (CTs), a number of automated methods have been developed, such as FreeSurfer (Dale et al., 1999; Fischl and Dale, 2000) (http://surfer.nmr.mgh.harvard.edu/fswiki/), CIVET (Zijdenbos et al., 2002), and CLADA (Cortical Longitudinal Atrophy Detection Algorithm) (Nakamura et al., 2011). For measurement of cortical volume, other methods can be used such as SIENAX (Smith et al., 2002) (cross-sectional pipeline of SIENA; Structural Image Evaluation using Normalization of Atrophy), SIENAX multi-time-point (SIENAX-MTP) (Dwyer et al., 2014), and SPM (Statistical Parametric Mapping software) (Ashburner and Friston, 2005) or IBASPM (Individual Brain Atlas using SPM) (Alemán-Gómez et al., 2006).

Automated methods for measuring CTs in MRI may be categorized as surface-based, voxel-based, or a combination of the two categories (Hutton et al., 2008). Currently, the majority of published studies that employ automated methods for measuring CTs have been using surface-based techniques for which the reliability has been investigated (Han et al., 2006; Lerch and Evans, 2005). With these methods, the thickness is calculated at each point on the extracted cortical surface and surface-based smoothing is applied to the results. The benefit of this type of smoothing is that it prevents the problem of averaging across different banks of sulci and gyri (Hutton et al., 2009). However, this is only the case if the surface has been accurately extracted. In general, these methods are limited by the difficulty in obtaining homogeneous GM boundaries due to the variation of contrast within the image that results from MRI (B1) field inhomogeneity. This problem has been partly solved by using 3D T1-weighted images and by implementing tools for intensity non-uniformity correction.

FreeSurfer (Dale et al., 1999; Fischl and Dale, 2000; Fischl et al., 2002) calculates the CTs after inflation of the folded cortical surface using tools for the analysis and visualization of structural MRI data (Dale et al., 1999; Fischl and Dale, 2000). It contains a fully automated structural imaging stream, although visual inspection and manual editing (if necessary) are recommended at several stages in the pipeline. For analysis of changes over time, FreeSurfer has a longitudinal pipeline (Reuter et al., 2012).

CIVET (Zijdenbos et al., 2002) is a series of algorithms for cortico-metric analysis of MR images including the extraction and analysis of cortical surfaces from MR images, and other volumetric and cortico-metric functions (Zijdenbos et al., 2002). In contrast to other software such as FreeSurfer, CIVET is a fully automated method, which can be an advantage for the analysis of large MR data sets.

CLADA (Zijdenbos et al., 2002) measures changes in CTs using explicit deformable models. Like FreeSurfer, CLADA allows the concurrent analysis of multiple time points, thereby improving quantification of within-patient change. The algorithm creates a deformable model of the cerebral cortex consisting of two explicit surfaces based on a combined image representing all time points, and then deforms the model for each individual time point (Nakamura et al., 2011). An advantage of CLADA compared to other methods is that it can be applied to 2D images with large slice thickness (3–5 mm), which can be advantageous for retrospective analysis of clinical or trial data.

Voxel-based cortical thickness (VBCT) measurements do not require the construction of a three-dimensional surface model (Han et al., 2006; Hutton et al., 2009). This approach is a complementary technique to the voxel-based morphometry (VBM) (Ashburner and Friston, 2000; Good et al., 2001), which uses a mass-univariate approach with voxel-wise comparisons of the local volume of GM across brain regions between two groups of subjects. VBCT maps may be particularly advantageous for the analysis of conditions that are associated with cortical thinning, because the local topography of the GM is used to assign an absolute metric to GM voxels (Hutton et al., 2009). In contrast, in VBM the local GM volume is confounded by the convolution of the brain in a given region. However, using both VBM and VBCT together might be a useful tool for understanding the topography and time-course of cortical atrophy in MS (Hutton et al., 2009).

Other methods such as SIENAX (Smith et al., 2002) (part of FSL, http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) and (IBA)SPM (Alemán-Gómez et al., 2006; Ashburner and Friston, 2005) primarily use intensity histogram based voxelwise GM partial volume estimation. These have been combined with anatomical atlases to separately quantify the total volume of GM in cortical regions (Dupuy et al., 2016; Horakova et al., 2008). SIENAX is the cross-sectional pipeline of the SIENA method (Smith et al., 2002). Based on voxel intensities it estimates partial volume fractions of GM, WM and cerebrospinal fluid (CSF) for each voxel. The longitudinal SIENA only quantifies overall brain volume change (based on the shift of the parenchyma-CSF border over time), and therefore does not measure GM or WM volume change separately. Two different extensions of SIENAX to perform direct longitudinal analysis of GM or WM atrophy have recently been developed, both of which combine cross-sectional and longitudinal modelling: SIENAX-MTP (Dwyer et al., 2014) and SIENA-XL (Battaglini et al., 2016).

SPM (http://www.fil.ion.ucl.ac.uk/spm/software/; (Penny et al., 2001; Holmes et al., 2001).
is a MATLAB toolbox that performs segmentation providing probability maps for GM, WM, and CSF. Individual Brain Atlas using SPM (IBASPM, http://www.thomaskoenig.ch/Lester/ibaspm.htm) is an SPM extension that uses the SPM segmentation and then warps a standardized atlas to the individual scan to classify the image into different anatomical structures, in this case the cortical GM (Aleman-Gomez et al., 2006). The GM volume of each structure is then quantified. Both non-linear registration and GM segmentation processes are performed using SPM subroutines; SPM12 includes symmetric registration procedures (Ashburner and Ridgway, 2012).

Longitudinal GM or WM volume changes can also be quantified using Jacobian integration (Ashburner et al., 1998; Nakamura et al., 2014). This is done by quantifying the total net amount of contraction or expansion of a selected region in an image during an accurate non-linear registration between the images of the first and last time points. Jacobian integration applied to longitudinal GM atrophy analysis has shown to reduce variability related to measurement error compared to other commonly used methods, thus potentially requiring smaller sample sizes (Nakamura et al., 2014).

2.2. Deep GM atrophy measurement techniques

DGM atrophy has been shown to be associated with the development of definite MS and disability progression in early relapsing remitting MS (RRMS) (Mesias et al., 2011; Rocca et al., 2010; Sepulcre et al., 2006; Zivadinov et al., 2013a; Zivadinov et al., 2013b). However, measurement of DGM atrophy remains challenging. To date, several available tools have been developed for segmenting DGM structures, some of which have been widely used in MS research. These methods are based on combinations of various approaches and techniques such as shape analysis (FIRST, FMRIB’s Integrated Registration and Segmentation Tool (Patenaude et al., 2011) and TOADS, Topology-preserving Anatomy-Driven Segmentation, https://www.nitrc.org/projects/toads-cruise/), voxelwise classification (FreeSurfer, TOADS), voxelwise partial volume estimation (SPM/IBASPM), topological constraints (TOADS) and brain atlases (IBASPM, FreeSurfer, TOADS). Here we describe the principles of each method.

FIRST (part of the FSL toolbox) is a fully automated model-based segmentation software (Patenaude et al., 2011). The model was trained to fit 15 different structures using 336 manually-labelled T1-weighted MR images. FIRST models the outer surface of each DGM structure as a mesh using models derived from the reference images and the local intensity profiles around the mesh. Finally, while taking into account the presence of neighboring structures, it labels voxels to assign them to different DGM structures.

FreeSurfer (Dale et al., 1999; Fischl et al., 1999) includes volumetric segmentation of DGM structures. The processing briefly includes registration to the Talairach space and intensity normalization (Sled et al., 1998), skull-stripping (Segonne et al., 2004), and labeling pre-defined brain regions including DGM and deep WM using a probabilistic atlas (Fischl et al., 2002; Fischl et al., 2004). DGM volume change over time can be analyzed through the longitudinal stream (Reuter et al., 2012), which uses an unbiased within-subject template that is created using inverse consistent registration (Reuter et al., 2010), improving reliability of volume change measurements (Reuter et al., 2012).

TOADS is another automated brain segmentation tool that includes segmentation of DGM structures (Bazin and Pham, 2007; Shiee et al., 2010). TOADS uses a topological atlas to constrain not only the topology of each brain structure, but also the relations between the structures. In addition, it uses a shape atlas and intensity based tissue classification.

As described above for cortical regions, the methods based on estimating voxelwise partial volume or probability such as SPM or SIENAX, can be used together with anatomical atlases as in IBASPM to extract GM volumes of subcortical regions.

2.3. Group-level analysis methods

Two types of group-level analysis methods are important in the context of this work. Voxel-based morphometry (VBM) is an extension of voxelwise segmentation-based techniques such as SIENAX or SPM, in which GM segmentation maps are transformed into a common space. Similarly, individual vertexwise cortical thickness maps can be transformed into a common space, for example in FreeSurfer. One can then investigate differences between groups and correlations with other variables for each voxel or vertex in the common space separately (followed by appropriate corrections for multiple testing). A major strength of such methods is that they allow the study of anatomical patterns of atrophy without any a priori selection of regions of interest. Salient regions and anatomical patterns can, therefore, be detected from the data itself. A limitation of such techniques is that in order to detect an effect, e.g. a common GM atrophy pattern in MS patients, there has to be substantial overlap between the true GM atrophy patterns of individual patients in the study. For example, if a patient group has reduced mean global GM volume compared to controls but the locations of GM atrophy in the brain vary, the voxel-based and vertexwise group-level analysis methods may not detect any significant atrophy patterns. This limitation should be considered when interpreting such group-level VBM or vertexwise studies.

3. Challenges of pathology, physiology and treatment effects

3.1. Unclear pathological substrate

Despite the importance of GM atrophy in MS, the underlying pathological substrate is not well known. The extent and severity of neuronal loss is less severe in MS than in a ‘classical’ neurodegenerative disease such as Alzheimer’s disease (Dutta et al., 2011; Haider et al., 2014; Peterson et al., 2001). Nevertheless, a recent post mortem MRI and histopathology study found that regional cortical atrophy is predominantly explained by a combination of (mild) neuro-axonal loss and neuronal shrinkage (Popescu et al., 2015). The cortical atrophy was not related to myelin loss (Popescu et al., 2015), despite (subpial) cortical demyelination being frequent and extensive in (progressive) MS (Geurts et al., 2005; Kutzelnigg and Lassmann, 2005). Another study found that neuronal and axonal density did not differ between subjects with normally myelinated GM and those with subpial cortical lesions (Klawer et al., 2015), suggesting that neurodegeneration in MS cerebral cortex is largely independent of cortical demyelination. It has also been shown that the observed cortical volume reduction of around 10% in MS patients is independent of the presence of cortical lesions (Pareto et al., 2016; Wegner et al., 2006). Although the work by Popescu et al., 2015 is valuable as a direct local association study between MRI atrophy measures and pathological characteristics, the number of patients is small and the results have to be confirmed independently. It is therefore imperative that more data is obtained on the pathological substrate of cortical and subcortical grey matter atrophy in MS.

3.2. Evolution of GM atrophy

It is not clear which brain regions are most likely to develop GM atrophy in the early phase of MS, whether the atrophic process is mainly primary or secondary, or to what extent its evolution differs between disease types (Bishop et al., 2017; Calabrese et al., 2007; Fisher et al., 2008). In part this arises from methodological issues such as unknown sensitivities of different measurement methods to atrophy in different GM regions. The concurrent evolution of focal lesions and other pathological changes, together with the varying and partly unknown effects of different treatments on GM atrophy, further complicate understanding the natural evolution of GM in MS.

In relapse-onset MS, GM atrophy has been observed already in the earliest phases of the disease (Bergland et al., 2012; Calabrese et al.,...
Moreover, GM atrophy may differ between disease types (Fisher et al., 2008; Sicotte et al., 2008), as well as between patients with and without evidence of disease activity (Nygaard et al., 2015). Some evidence also exists of early and eloquent GM atrophy in specific regions: for example, in primary progressive MS, the involvement of the cingulate cortex was found to occur throughout the disease course (Eshaghi et al., 2014). Also, significant GM volume loss occurred in the right precuneus in relapsing-remitting MS patients with progressive disability (Hofstetter et al., 2014). Moreover, atrophy of specific DGM structures, most notably the thalamus, also occurs in MS patients (Bishop et al., 2017; Houtchens et al., 2007; Schoonheim et al., 2015; Solomon et al., 2017). Atrophy of the thalamus appears to occur especially early and prominently, to be worse in men, and to be associated with cognitive decline (Schoonheim et al., 2015). However, it remains unclear to what extent these observations could be biased by the size and partially limited contrast with the neighboring WM of the thalamus. In general, more work is needed to validate the dynamic changes and anatomical patterns observed in previous studies. Specifically, the sensitivity of current techniques to measure atrophy of different anatomical regions should be quantitatively evaluated.

Although studies have shown that the relation between GM pathology and WM lesions is an important factor in MS, this relation remains to be fully elucidated (Geurts et al., 2012). Global WM lesion volumes are associated with both global GM atrophy (Charil et al., 2007; Roosendaal et al., 2011; Steenwijk et al., 2014) and local GM atrophy (Charil et al., 2007; Sailer et al., 2003). Furthermore, there are similarities between the average anatomical distribution of GM atrophy and the distributions of both existing WM lesions (Bodini et al., 2009; Muhla et al., 2013; Riccitelli et al., 2012) and new WM lesions (Battaglini et al., 2009; Bendfeldt et al., 2010). The relation of local GM atrophy to lesional and non-lesional damage in connected WM tracts has only been partially explored (Bergsland et al., 2015; Louapre et al., 2016; Steenwijk et al., 2015) and further studies are needed to clarify this.

How GM atrophy is related to other pathological changes inside the GM has been less investigated. Longitudinal studies, in larger groups and other clinical disease sub-types, are needed to fully answer the questions about temporal and causal relations between GM atrophy and other MS pathology, both in natural evolution and on treatment.

### 3.3. Influence of physiological variability

Effects of physiological variability could affect measurement of GM atrophy; brain volumes have been shown to be affected by, among other, food intake (Roberto et al., 2011), steroids, (excess) body fat (Janowicz et al., 2015), alcohol abuse (Thayer et al., 2016), and tobacco smoking (Sutherland et al., 2016). One study (Sampat et al., 2010) showed that, when using whole-brain volume change measures in an MS trial, physiological fluctuations may increase required sample sizes by a factor of 5 compared to sample sizes based on scan-rescan reproducibility error only.

Different physiological sources of variability also have been investigated separately. Brain volumes decrease systematically during the day, and by correcting for these diurnal fluctuations, required sample sizes in a typical trial could be reduced by 2.6% (Nakamura et al., 2015). Brain volumes also differed significantly between hydrated and dehydrated states induced by 16 h of no fluid intake (Duning et al., 2005). Although such extreme situations seem unlikely in MS clinical practice, the effect of hydration status should be taken into account. In all, clear guidelines towards diminishing the impact of physiological variations are needed before GM atrophy measurements can be considered in clinical practice.

### 3.4. Evaluation of treatment response

The effect of disease modifying therapies may be different for WM than GM atrophy measures (Filippi et al., 2014; Fisher et al., 2016; Perez-Miralles et al., 2015). Moreover, currently available therapies predominantly target processes involved in formation of new WM lesions rather than atrophy. Therefore, tissue-specific brain volume monitoring could provide differential insights on treatment effects and response (De Stefano et al., 2014) and help in the development of new treatments. WM volume changes after starting treatment are heavily dependent on baseline inflammation (Vidal-Jordana et al., 2016; Vidal-Jordana et al., 2013). Therefore, immediate decreases in WM volume after therapy onset need to be interpreted with great caution. This initial pseudo-atrophy effect does not seem to occur for GM (Prinster et al., 2006). In addition, GM volume changes have been shown to be relatively insensitive to baseline inflammation (Fisher et al., 2016; Vidal-Jordana et al., 2016; Vidal-Jordana et al., 2013). Whether a whole-tissue or a region-specific approach should be preferred, remains to be elucidated and would depend on the acquisition of an a-priori knowledge of the pathological substrate as well as the potential effects of any given drug on the specific regions.

Another important aspect is the influence of pre-existing damage on GM atrophy development. As indicated above (“Evolution of GM atrophy”), there is currently in general insufficient data on the relation between lesions and other damage in the WM on the one hand, and (subsequent) atrophy of the GM on the other. Also, specifically whether or not GM atrophy resulting from pre-existing WM damage may be halted by treatment remains to be investigated. Studies should be performed to provide such data, because otherwise interpretation of data obtained in patients on treatment may be incorrect. If the secondary atrophy due to pre-existing damage cannot be halted, a situation may arise in which in a treated patient new GM atrophy is adequately suppressed by treatment, but GM atrophy still progresses merely as a late result of pre-existing WM lesions. In such a case, in clinical treatment trials, these effects of pre-existing WM damage on current GM atrophy may lead to an underestimation of the effectiveness of the treatment under evaluation. Similarly, in a clinical setting, the treating physician may incorrectly conclude that the treatment is not effective on GM atrophy in that patient. It is therefore of great importance to obtain data to understand the relation between WM damage and GM atrophy and allowing, ultimately, the disentanglement of new GM atrophy from GM atrophy that results from pre-existing WM damage.

### 4. Measurement challenges

#### 4.1. Influence of WM lesions

Several papers (Battaglini et al., 2012; Chard et al., 2010; Gonzalez-Villa et al., 2017; Nakamura and Fisher, 2009) have shown that WM lesions that are hypointense on T1-weighted images can affect the assessment of GM volumes, as illustrated in Fig. 1. One paper noted an artificial GM volume underestimation in the presence of lesions with an intensity between GM and WM, also when lesions were correctly reclassified as WM (Nakamura and Fisher, 2009). Another study, using synthetic data, found that the extent and the direction of the GM estimation error depended strongly on the lesion load, on the degree of hypointensity, and on the partial volume model used in the segmentation algorithm (Battaglini et al., 2012).

The presence of T1 hypointense lesions, especially along tissue interfaces, severely affects non-linear registration of images (Di Perri et al., 2008; Sdika and Pelletier, 2009). In VBM-type experiments, these local distortions in registrations led to an overestimation of GM loss in the cortex and in the DGM structures (Ceccarelli et al., 2012).

The most commonly used solution to these WM lesion-induced problems is the so-called ‘lesion-filling’ or lesion ‘inpainting’, which replaces WM lesion voxels with intensities similar to normal WM, using parametric or numerical approaches to reproduce global or local WM intensity histograms (Battaglini et al., 2012; Chard et al., 2010;
Gelineau-Morel et al., 2012; Prados et al., 2014; Sdika and Pelletier, 2009). Lesion-filling significantly reduces the error due to the presence of lesions in evaluating cortical and DGM volumes (Prados et al., 2014). Moreover, lesion-filling significantly improves the accuracy of local CTh estimation, thereby also improving the whole-brain average CTh measurement (Magon et al., 2014). A limitation of the lesion-filling approach is that it requires accurate high-resolution segmentation of the lesions. Nonetheless, even with low-resolution lesion segmentations it is possible to measure the whole-brain GM volume from lesion-filled images with good accuracy, although analysis of smaller GM regions may be hampered by the remaining lesion voxels (Popescu et al., 2014). Because of the importance of this issue, integrated segmentation algorithms by which not only WM, GM and CSF but also lesions can be segmented could be beneficial.

4.2. Influence of atrophy itself

Paradoxically, the presence of brain atrophy distorts measurement of that same brain atrophy, by affecting various steps in the analysis pipelines used.

Many analysis pipelines involve the removal of non-brain tissue when measuring atrophy; the commonly used BET software (part of FSL) was found to require fairly extreme parameter settings to perform accurately in images of MS patients compared to manual brain extraction (Popescu et al., 2012). Besides BET, other methods have also been proposed to achieve accurate non-brain tissue in atrophic patient scans including multi-atlas skull stripping (Doshi et al., 2013), OptiBET (Lutkenhoff et al., 2014), BrainMAPS (Leung et al., 2011), BEAST (Eskildsen et al., 2012), and STEPS (Cardoso et al., 2013).

The accuracy of image co-registration, another step used in almost all methods, also can be affected by the presence of brain atrophy (Pereira et al., 2010), as illustrated in Fig. 2. Proposed approaches to overcome this problem include multi-channel registration using fractional anisotropy (FA) maps from diffusion tensor imaging (DTI) alongside T1-weighted images (Ceritoglu et al., 2009; Roura et al., 2015), and inclusion of a-priori knowledge of ventricular shapes in the presence of atrophy (Djamanakova et al., 2013). These are expected to improve GM atrophy quantification by better matching of templates, priors, atlases and regions of interest; these approaches should be evaluated correctly for better use and understanding.

As a very specific example, it has been shown that atrophy of the WM affects cortical surface curvature in MS (Deppe et al., 2014), which could lead to incorrect interpretations of GM atrophy metrics based on cortical surface reconstruction, depending on the method used.

4.3. Influence of other pathology

Beyond WM lesions and brain atrophy, focal GM lesions are prominent in MS (Geurts et al., 2005; Kutzelnigg et al., 2005). The presence of such GM lesions also could have a significant effect on GM volume estimations, as some of them may still be visible in the sequences frequently used for atrophy measurement such as MPRAGE (Bagnato et al., 2006; Nelson et al., 2014). There is also abundant evidence of diffuse GM and WM changes from post mortem and in vivo studies (Raz et al., 2010; Seewann et al., 2009; van Munster et al., 2015; Vrenken et al., 2010). The effect of these diffuse WM and GM abnormalities on quantification of atrophy measures, to the best of our knowledge, has not been investigated and could affect these measurements. First, it is interesting that even a small volume of WM lesions can have a substantial effect on GM partial volume estimates throughout the brain for segmentation techniques in which intensity distributions of different tissue classes are modelled (with or without priors), such as FSL-FAST (Popescu et al., 2014). Second, subtle age or disease related changes of the WM-GM contrast at the white matter – cortical border have been shown to influence cortical thickness measurement systematically in a way that can be corrected statistically, at least at group level (see Fig. 3) (Westlye et al., 2009). There is also evidence that diffuse GM pathological abnormalities are more widespread and possibly independent of atrophy (Khaleeli et al., 2007; Mallik et al., 2015). How these subtle GM and WM abnormalities affect segmentation and registration algorithms of the atrophy measurement pipelines is unknown and this warrants further research.

4.4. Technical variability

MRI-derived GM atrophy measurements can exhibit large variability due to causes related only to the MR image acquisition and analysis; these problems are not specific to MS but occur equally in other diseases that cause GM atrophy.

Concerning image acquisition, usually (near)isotropic 3D T1-weighted images with similar acquisition parameters are used to quantify brain atrophy. Nevertheless, remaining minor differences can systematically affect GM atrophy measurements. As an example, Fig. 4 shows how signal intensity and image contrast can change noticeably even when only a single sequence parameter (the inversion time) is slightly altered (from 450 to 400 ms). Similarly, different scanners can give different volumes or volume changes and this has been reported – this can occur even when using similar or even identical acquisition parameters (Durand-Dubief et al., 2012). Large systematic differences have been observed between scanners for CTh and GM volume measurements (Biberacher et al., 2016). Also, it has been shown that even with identical scanners and imaging protocols, systematic differences of GM volumes can occur (Takao et al., 2011).

Focusing on image analysis reproducibility, using the ADNI1 data (http://adni.loni.usc.edu/), which contains a within-session rescan, i.e.
without repositioning, the reproducibility error was non-negligible compared to typical atrophy rates for a commonly used whole-brain software, SIENA (Cover et al., 2011). Moreover, for both FSL-FIRST and FreeSurfer the within-session reproducibility error for measuring 1-year hippocampal volume change was of similar size as the typical atrophy rates (Mulder et al., 2014). A recent study (Cover et al., 2016) found that some image analysis methods exhibited better reproducibility of 1-year hippocampal volume change than others, leading to improved statistical power and smaller required sample sizes in treatment trials.

Reproducibility of caudate nucleus, putamen, amygdala, globus pallidus, and thalamus atrophy rates were found to be about 1.5–3.5 times higher than the mean measured volume change across 500 Alzheimer’s disease, mild cognitive impairment and healthy elderly subjects for both FreeSurfer and FSL-FIRST (Meijerman et al., 2018). In a smaller group in MS, 3 T imaging yielded slight improvements of subcortical volume reproducibility compared to 1.5 T by using FSL-FIRST (Chu et al., 2017).

With respect to differences between different image analysis methods, in a longitudinal study, robustness across sites was better for registration-based techniques than for segmentation-based techniques (Durand-Dubief et al., 2012), although the data suggests that the improved robustness across sites may come at the cost of a somewhat reduced sensitivity to the volume change between time points. The two most widely used methods for GM-VBM (FSL-VBM and SPM-VBM) gave substantially different results when comparing RRMS patients to healthy controls, which was especially noticeable when considering separate brain regions based on an anatomical atlas (Battaglini et al., 2009). For volume measures in individual MS patients, volumetric agreement between different analysis methods (FreeSurfer, SPM, and FSL) was only moderate to good for most DGM structures and cortical regions, and correlations with clinical and cognitive measures varied with the analysis method used (Popescu et al., 2016). Lastly, FreeSurfer GM volume and CTh measurements depend on operating system, FreeSurfer version and workstation type (Gronenschild et al., 2012). A benchmark dataset is needed to establish and test the robustness of analysis methods in the context of repeated scans within or between scanners. Data acquired in MS patients is of special interest, because that would allow users to concurrently investigate the effects of lesions and other pathology occurring in MS on the GM atrophy measurement.

5. Conclusions

We reviewed the state of the art regarding the most urgent challenges in the measurement of GM atrophy in MS, and provided recommendations to overcome them (summarized in Box 1). Difficulties associated with physiological variability or with technical variability between scans and image analysis methods extend beyond MS. Therefore, the broader brain research community would benefit from a benchmark dataset is needed to establish and test the robustness of analysis methods in the context of repeated scans within or between scanners. Data acquired in MS patients is of special interest, because that would allow users to concurrently investigate the effects of lesions and other pathology occurring in MS on the GM atrophy measurement.

Fig. 2. An example of image misregistration due to atrophy and a possible method for its improvement based on multi-channel mapping. Warped contours of lateral ventricles and surrounding regions are superimposed on T1-weighted image as color overlays. (a) The red outline indicates the contour of the lateral ventricle warped to the target image space by default parameters FNIRT. This registration compared to the actual location of the lateral ventricle (white dotted line) is poor. (b) Taking into account the large deformations necessary to co-register images in the presence of severe atrophy (method proposed by Djamanakova et al. (2013)) slightly improved the registration. (c) Using a dual-channel approach incorporating a coarse ventricle segmentation in the target image space improved the lateral ventricle registration substantially. Reproduced in modified form with permission from Djamanakova et al. (2013).

Fig. 3. The effect of grey/white contrast on cortical thickness measurement using FreeSurfer. Statistical p maps thresholded at p < 10⁻² superimposed on a template brain’s semi-inflated surface showing the results from GLMs testing the difference between Alzheimer’s disease (AD) and controls (Con). Warm colors denote areas with significantly thinner cortex in AD compared to controls. Adjusting for grey/white tissue contrast (bottom row) increases sensitivity to the AD-control differences in cortical thickness over large portions of the brain compared to results obtained when not adjusting for this contrast (top row). Adopted with permission from Westlye et al. (2009).
addressing these challenges adequately. Some disease-specific technical challenges also exist: while solutions are available to minimize the influence of WM lesions on brain atrophy, the influence of diffuse damage and of GM lesions remains to be elucidated. Such improvements are needed to allow evaluation of GM atrophy independent of other MS pathology. This would be achieved when the evolution of GM in untreated and treated patients, as well as its relation to the other MS pathology can be properly understood. This understanding and the technical improvements are also prerequisites to the use of GM atrophy measures in the clinical routine.

Acknowledgements

This paper is the outcome of the MAGNIMS workshop “Measuring cortical and deep gray matter atrophy in MS”, which was kindly supported by the Multiple Sclerosis International Federation (www.msif.org), the Dutch MS Research Foundation (www.msresearch.nl), the journal “Nederlands Tijdschrift voor Geneeskunde” (www.ntvg.nl), the Neuroscience Campus Amsterdam (www.neurosciencecampus-amsterdam.nl), and the VU University medical center. Frederik Barkhof is supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

References


