Journal of Neurology
https://doi.org/10.1007/s00415-018-8825-8

ORIGINAL COMMUNICATION

Within-patient fluctuation of brain volume estimates from short-term repeated MRI measurements using SIENA/FSL

Roland Opfer[1,2] · Ann-Christin Ostwaldt[2] · Christine Walker-Egger[1] · Praveena Manogaran[1,3] · Maria Pia Sormani[4] · Nicola De Stefano[5] · Sven Schippling[1]

Received: 12 January 2018 / Revised: 19 February 2018 / Accepted: 4 March 2018

Abstract

Background Measurements of brain volume loss (BVL) in individual patients are currently discussed controversially. One concern is the impact of short-term biological noise, like hydration status.

Methods Three publicly available reliability MRI datasets with scan intervals of days to weeks were used. An additional cohort of 60 early relapsing multiple sclerosis (MS) patients with MRI follow-ups was analyzed to test whether after 1 year pathological BVL is detectable in a relevant fraction of MS patients. BVL was determined using SIENA/FSL. Results deviating from zero in the reliability datasets were considered as within-patient fluctuation (WPF) consisting of the intrinsic measurement error as well as the short-term biological fluctuations of brain volumes. We provide an approach to interpret BVL measurements in individual patients taking the WPF into account.

Results The estimated standard deviation of BVL measurements from the pooled reliability datasets was 0.28%. For a BVL measurement of x% per year in an individual patient, the true BVL lies with an error probability of 5% in the interval $x\% \pm (1.96 \times 0.28)/(\text{scan interval in years})\%$. To allow a BVL per year of at least 0.4% to be identified after 1 year, the measured BVL needs to exceed 0.94%. The median BVL per year in the MS patient cohort was 0.44%. In 11 out of 60 MS patients (18%) we found a BVL per year equal or greater than 0.94%.

Conclusion The estimated WPF may be helpful when interpreting BVL results on an individual patient level in diseases such as MS.

Keywords Brain atrophy · MRI · SIENA · Multiple sclerosis · Reliability

Introduction

Magnetic resonance imaging (MRI)-derived whole brain atrophy is increasingly recognized as an important imaging marker of neurodegeneration in multiple sclerosis (MS) and has recently been recommended to be included in the no evidence of disease activity (NEDA) criteria [10]. Brain volume loss (BVL) in young, healthy individuals ranges between 0.1 and 0.3% per year [23]. For older individuals, higher values are deemed physiological; at an age range of 75–80 years the mean BVL extrapolated from cross-sectional data has been reported to be as high as 0.52% per year [19]. BVL beyond what is assumed physiological for a respective age group already occurs in the earliest stages of MS [3]. In a typical relapsing–remitting MS cohort BVL ranges between 0.5 and 1.35% per year. A cut-off of 0.52% per year (with a specificity of 95%) or 0.4% per year (with a specificity of 80%), has been determined to distinguish physiological from

✉ Roland Opfer
roland.opfer@jung-diagnostics.de

1 Neuroimmunology and Multiple Sclerosis Research, Department of Neurology, University Hospital Zurich and University of Zurich, Zurich, Switzerland

2 Jung Diagnostics GmbH, Hamburg, Germany

3 Department of Information Technology and Electrical Engineering, Swiss Federal Institute of Technology, Zurich, Switzerland

4 Biostatistics Unit, Department of Health Sciences, University of Genoa, Genoa, Italy

5 Department of Medicine, Surgery and Neuroscience, University of Siena, Siena, Italy

Published online: 16 March 2018

pathological BVL in MS [7, 15] using the FMRIB Software Library (FSL)-based structural image evaluation using normalization of atrophy (SIENA) toolbox.

Currently, the significance of measuring BVL in individual patients with MS is being discussed controversially [1, 24]. Biberacher and colleagues pointed out that with the currently available image processing tools, a sufficiently reliable estimation of BVL in individual patients only seems possible over periods of at least 5 years [2]. In addition to methodological constraints, other factors such as behavioral (e.g., alcohol consumption, smoking, diet and de-/hydration), genetics (e.g., apolipoproteinE expression), comorbidities (e.g., diabetes, cardiovascular risks) and pharmacological treatments can impact brain volume measurements and may contaminate any BVL caused by the disease [5, 9, 11]. In extreme cases such effects can be substantial. For instance, a significant brain volume change has been reported following rehydration [9, 14].

The SIENA toolbox is a widely used FSL application to quantify BVL, including recent clinical phase III trials [22]. To correctly assess and interpret measurements of BVL with SIENA on the individual patient level, it is paramount to understand the magnitude of the intrinsic measurement error of the technique as well as the impact of potential biological factors leading to short-term fluctuations of brain volumes. These effects need to be distinguished from BVL related to effects mediated though the disease, such as influx or efflux of inflammatory edema and neuro-axonal degeneration/atrophy in case of MS.

The intrinsic measurement error can be estimated by means of test–retest analyses of very short-term repeated scans, where the same subject is scanned twice within the same imaging session using identical scanner settings. Within these very short periods (i.e., minutes) and in the absence of dehydrating exercise, no circadian, chronic biological or disease-related factors should impact the results. Therefore, all deviations from zero under such settings are due to instabilities of the computational method and/or image acquisition. In recent publications, median measurement errors of about 0.15–0.20% have been reported for SIENA [4, 21, 22].

However, such test–retest measurements do not necessarily reflect potential daily or short-term (days/weeks) fluctuations of brain volumes due to biological factors as the ones mentioned above. Including such intervals could provide a setting much closer to clinical reality. An experimental design, covering daily fluctuations of brain volumes, would be to re-scan individuals repeatedly within a few days to a few weeks. Variability estimated from such analyses will not only reflect the intrinsic measurement error but also the short-term biological fluctuations of brain volumes, while at the same time excluding any effects related to disease and ageing.

Materials and methods

Within-patient fluctuation

We define the within-patient fluctuation (WPF) of BVL measurements to comprise the intrinsic measurement error of the method and the potential short-term (days/weeks) biological fluctuations of the brain volume.

Reliability datasets

We analyzed three publicly available MRI datasets to estimate the WPF.

Maclaren dataset This freely available MRI dataset contained data from three healthy subjects (2 males, 1 female; 26, 31 and 30 years old) [12]. For each subject, 20 MRI examinations were performed within a 31-day period on a 3 Tesla (T) General Electric (GE) MRI scanner. In each scanning session, every subject was scanned twice with repositioning of the subject between scans.

OASIS reliability dataset The OASIS reliability dataset was part of the cross-sectional Open Access Series of Imaging Studies (OASIS [13], http://oasis-brains.org/) and contains data from 20 healthy controls who received two MRI examinations on a 1.5 T Siemens scanner. Median age was 22 years (interquartile range (IQR) 20–25 years) and median interval between the two scans was 11 days (IQR 3–31 days).

Biberacher dataset The freely available MRI dataset from Biberacher et al. [2] contained data from two relapsing–remitting MS patients (both female; 29 and 24 years old). Within 3 weeks, patients received five or six MRI examinations, each time on three different 3 T scanners (Philips, Siemens and GE) with an interval of several days between scans.

For the present study, we used 3D pre-contrast T1-weighted images from all datasets. Information on the three datasets, as well as acquisition parameters are detailed in Table 1.

Longitudinal cohort of MS patients

A single scanner cohort of MS patients (Zurich dataset) was used to address the question whether after 1 year pathological levels of BVL can be identified in a relevant fraction of MS patients.

The dataset consisted of 60 MS patients (22 CIS, 36 RRMS and 2 PPMS patients) sampled from clinical routine who received at least two consecutive MRI examinations on a 3 T scanner (Philips Ingenia, Best, The Netherlands) as part of an observational study at the Neuroimmunology and Multiple Sclerosis Research Section at the University

Table 1 Patient and scan characteristics of the four cohorts investigated

Cohorts	Maclaren	OASIS	Biberacher	Zurich
Sample size	3	20	2	60
Disease type	Healthy individuals	Healthy individuals	RRMS patients	22 CIS patients 36 RRMS patients 2 PPMS patients
Female	1 (33%)	12 (60%)	2 (100%)	41 (68%)
Age (years)	26, 31 and 30	22 (20–25)	29 and 24	33.1 (29.3–39.6)
Disease duration (years)	–	–	5 and 5.5	0.3 (0.05–2.4)
EDSS	–	–	1 and 2	2 (1–2.5)
No. of scans per patient	40 scans (2 scans per session) within 1 month	2 scans within 1–31 days	6 scans/5 scans on three different MRI scanners each within 2–3 weeks	2 scans
No. of scans in the dataset	120	40	34	120
Total no. of BVL measurements	114	20	28	60
Scan interval	1 (max 3) days	11 (3–31) days	3 (3–4) days	1.3 (1.1–1.6) years
MRI scanner	3T GE Discovery	1.5T Siemens Vision	3T GE Signa 3T Philips Achiva 3T Siemens Verio	3T Philips Ingenia
Sequence protocol	Accelerated sagittal 3D IR-SPGR	Sagittal MPRAGE	BRAVO sagittal 3D IR-FSPGR (GE), sagittal MPRAGE (Philips and Siemens)	3D TFE 32 SENSE
Slice thickness (mm)	1.2	1.25	1	0.9
TR (ms)	7.3	9.7	8.2, 9, < 9	8–9.1
TE (ms)	3	4	3.2, 4, 2.45	3.7–4.2
TI (ms)	400	20	450, 1000, 900	

Further details on the first three cohorts are reported elsewhere [2, 12, 13]. Values are expressed as median (interquartile range) or frequency (percentage)

EDSS Expanded Disability Status Scale, *TR* repetition time, *TE* echo time, *TI* inversion time, *IR* inversion-recovery, *SPGR* spoiled gradient recalled, *MPRAGE* magnetization-prepared rapid gradient-echo, *FSPGR* prepared fast-spoiled gradient echo, *TFE* turbo field echo

Hospital Zurich (Switzerland). Acquisition parameters and details of the cohort are given in Table 1. Median age was 33.1 years (IQR 29.3–39.6 years) and median scan interval was 1.3 years (IQR 1.1–1.6 years). All patients gave written informed consent and the study was approved by the Ethics Board of the Canton of Zurich (KEK-ZH-Nr. 2013-0001).

Brain volume loss (BVL) with SIENA

High-resolution (3D) pre-contrast T1-weighted images for each subject were used to assess BVL (in %) of the whole brain using SIENA method [22] which is part of the FMRIB Software Library (FSL; http://www.fmrib.ox.ac.uk/fsl). The performance of SIENA can differ greatly depending on parameter settings and pre-processing steps [17]. For this work, SIENA (version 5.06) was deployed with FSL brain extraction tool (BET) using the configuration "-B -f 0.2 -m", which differ from the default settings (-f 0.5). Before applying SIENA, we used the FSL script fslreorient2std to reorient all images to match the orientation of the standard template images (MNI). This can be useful because some image formats use a convention for image orientation different from the FSL library. In addition to the reorientation, we performed a neck removal as recommended elsewhere [17] before starting SIENA. These described configurations were recommended as the parameter with the lowest variability in a previously published paper [17]. We estimated BVL for all subjects from all datasets and the entire pairs of timely consecutive MRI scans. In case of the Zurich cohort, BVL values were annualized (BVL per year) by dividing the measured BVL by the length of the interval between scans (in years).

Statistical analysis

For each of the 3 subjects in the Maclaren dataset, 20 MRI examination sessions were performed. In each session, each subject was scanned twice with repositioning of the subject between scans. WPF for each pair of timely consecutive MRI scans for each dataset was computed using SIENA with the parameters explained above. More precisely, the first scan of the first session was compared against the first

scan of the second session, the first scan of the second session was compared against the first scan of third session and so forth. The same pairs were used for the second scans of each imaging session. For each subject, we thus obtained $2 \times 19 = 38$ WPF measurements and altogether $3 \times 38 = 114$ WPF measurements. For the OASIS dataset 20 WPF measurements were obtained (20 subjects with 2 scans within 1–31 days). The Biberacher dataset resulted in 13 (5 times GE, 4 times Philips, and 4 times Siemens) WPF measurements for patient 1 and 15 (5 times GE, 5 times Philips, and 5 times Siemens) WPF measurements for patient 2. All data were pooled into one single dataset. We tested for a correlation (Pearson) between the scan interval (in days) and the magnitude of the WPF. As subjects contributed a varying number of scans, we performed an analysis using a linear mixed effect model. The 25 subjects as well as the 5 different scanners involved were used as random effects [model was 'BVL ~ 1 + (1|Subject) + (1|Scanner)']. Full covariance matrix was used in the mixed effect model. Despite the large number of WPF measurements analyzed, our cohort contained only 25 subjects, with an age range between 20 and 31 years. Therefore, an adjustment for age, sex or baseline brain volume was not possible with the chosen cohorts and beyond the scope of this manuscript.

The standard deviation (SD) of the model residuals is a suitable measure for the SD of the pooled WPF measurements since it takes repeated measurements into account. We, therefore, used $1.96 \times \mathrm{SD}$ as an estimate for WPF range. Assuming a normal distribution of the WPF, the interval $\pm 1.96 \times \mathrm{SD}$ contains 95% of the variability caused by the WPF. In addition, the 25th percentile, the median, the 75th and the 95th percentiles of the absolute WPF for each dataset individually and for the pooled data were computed to allow a direct comparison with published data.

Assessment of single subject BVL measurement

To use BVL for decision-making in individual patients in clinical routine, the WPF needs to be taken into account.

To derive a realistic estimate for the WPF from reliability datasets (short-term repeated measurements with an expected mean around zero) two main assumptions are imposed: first, the WPF is independent of the time interval between the scans and, second, the WPF does not depend on the absolute value of true biological BVL caused by a disease or by ageing.

To explain these assumptions, we provide a hypothetical example (Fig. 1). A patient has a follow-up scan after 1 year. The true (but unknown) biological BVL caused by ageing or by a disease in this example is 1.2%. For reasons detailed above, one cannot expect to measure exactly 1.2%

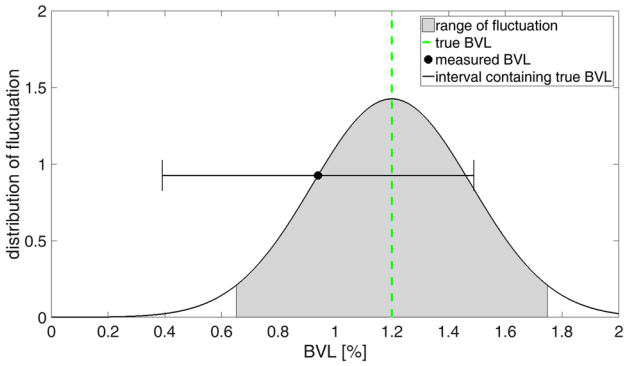

Fig. 1 Illustration of within-patient fluctuation (WPF) around an individual brain volume loss (BVL) measurement. Each measured BVL value lies within a certain distribution around the true BVL (in this hypothetical case 1.2%). By adding the fluctuation range of ±0.54% to the measured BVL value this interval will contain the true BVL with a probability of 95%

after 1 year, since WPF will impact the measurement. If we were to repeat the follow-up measurement multiple times within a short time period (days) and determine the BVL between the baseline scan and each of these repeated follow-up scans, we obtain a certain distribution of measurement results around the true BVL (Fig. 1). Based on the assumptions stated above, we can further expect that the distribution of the WPF around 1.2% is the same as the WPF around zero that we would derive from reliability datasets. Hence, 95% of all repeated BVL measurements at follow-up fall within the interval true BVL $\pm 1.96 \times \mathrm{SD}$.

The considerations so far are independent of the time interval between baseline and follow-up scan. Since BVL is a progressive process, the measured BVL will usually increase with the time elapsed between baseline and follow-up scan. To facilitate a comparison between different BVL measurements, BVL is usually annualized by dividing the BVL by the time interval (in years) between baseline and follow-up scan (BVL/year). For a BVL measurement x% after a certain time interval the true and usually unknown BVL lies within the interval $[x\% - 1.96 \times \mathrm{SD}\ \%, x\% + 1.96 \times \mathrm{SD}\ \%]$ with an error probability of 5%, and therefore, the true BVL is at least $x\% - 1.96 \times \mathrm{SD}\ \%$. The value $1.96 \times \mathrm{SD}$ is subtracted from the actual measurement as a safety margin. For the annualized BVL this means that the true BVL per year is at least $(x\ \% - 1.96 \times \mathrm{SD}\ \%)$/interval. The division by the scan interval in the latter expression originates from the mathematical properties of variances. If a random variable (in our case the WPF) is divided by a factor (in our case the interval), then the variance needs to be divided by the square of that factor and hence the SD needs to be divided by that factor.

Results

Reliability datasets

Results of the SIENA analyses from each dataset (Maclaren, OASIS reliability and Biberacher) are listed in Table 2 and presented as box plots with the WPF values for each of the datasets in Fig. 2. In addition to the analyses of the individual datasets, we also pooled all three reliability datasets. For the combined dataset, SIENA results were as follows: absolute median = 0.15%, absolute 75th percentile = 0.27%, 95th percentile = 0.48%. In Fig. 2, all measurements are shown as box plots. The grey area of the box plot represents the interquartile range (25th to 75th percentile) of the WPF while two outliers are marked with crosses. The two outliers are both part of the Biberacher dataset and feature values of − 1.8 and 1.07%, respectively. Both fixed effects (intercept and interval) were not statistically different from zero in the linear mixed effect model ($p = 0.94$ and $p = 0.46$). Estimate of the SD of the residuals was 0.28%. The WPF range can, therefore, be estimated by $1.96 \times 0.28\% = 0.54\%$. There was no correlation between the scan interval (in days) and the magnitude of the WPF ($r = 0.08$, $p = 0.26$). The scan interval was, therefore, not included into the model.

As mentioned before, BVL values of 0.40 and 0.52% per year have been suggested as possible cut-off values to distinguish physiological from pathological brain volume loss in MS [7]. Using 0.54% as the WPF range we can calculate the BVL per year that minimally needs to be measured in an individual subject to ensure that a true BVL per year exceeds the BVL cut-offs of 0.4 or 0.52% per year, respectively (according to the methods section). In Table 3, we provide these threshold values for intervals of 1.0, 1.5 and 2.0 years between the baseline and the follow-up scan. If the scan interval between baseline and follow-up is 1 year, a BVL per year of at least 0.94% needs to be measured in an individual patient to ensure that the true BVL is greater than 0.4% per year. If the time interval between baseline and follow-up is 2 years, a BVL of 0.67% per year needs to be measured to make the same statement.

Table 2 Median, 25th, 75th and 95th percentiles of the absolute brain volume loss (BVL) values for the three reliability datasets individually and combined

	25th	Median	75th	95th
Maclaren	0.05	0.14	0.24	0.43
OASIS	0.14	0.24	0.30	0.63
Biberacher	0.07	0.13	0.34	1.15
Combined	0.06	0.15	0.27	0.48

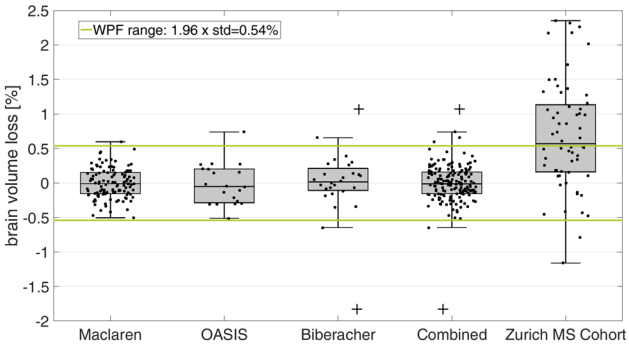

Fig. 2 SIENA results for all three reliability datasets individually (Maclaren, OASIS and Biberacher) and combined are shown as box plots. For easy comparison between within-patient fluctuation (WPF) and long-term brain volume loss (BVL) values, the SIENA results of the Zurich MS cohort (without scaling to the duration of the scan interval) are also displayed. The centerline in each box plot is the median of all measurements. The grey area of the box plot represents the interquartile range of the signed SIENA results. The horizontal green lines indicate the WPF range (±0.54%)

Longitudinal cohort of MS patients

We determined BVL in a group of 60 patients with follow-up MRI examinations from a cohort of the Neuroimmunology and MS Research Center in Zurich. For easy comparison between WPF and long-term BVL, Fig. 2 additionally shows the BVL measurements of the Zurich MS cohort (without scaling to the duration of the scan interval). Individual annual BVL values are presented as histograms in Fig. 3. In this cohort, the median BVL per year was 0.44% (interquartile range 0.12–0.80%). In 11/60 patients (18%) we found a BVL per year equal to or greater than 0.94%.

Table 3 Required brain volume loss (BVL) per year to ensure a true BVL of 0.4 or 0.52%

Scan interval in years	Minimal BVL per year that needs to be measured to ensure a true BVL per year of	
	0.4%	0.52%
1	0.94%	1.06%
1.5	0.76%	0.88%
2	0.67%	0.79%

The values given in this table represent the BVL per year values that need to be measured, to ensure with a 95% probability that the true BVL is higher than 0.4% per year or 0.52% per year, respectively (as proposed in [8]). Measured BVL per year values are given for scan intervals of 1, 1.5 and 2 years

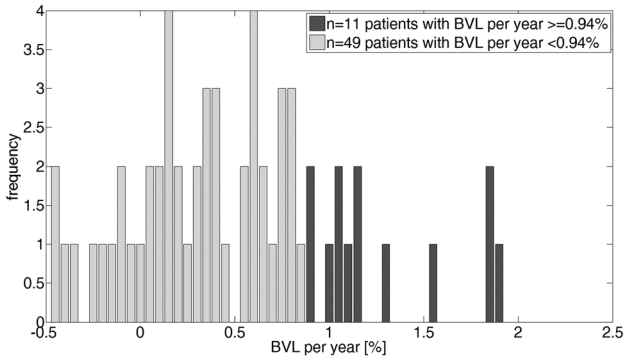

Fig. 3 Brain volume loss (BVL) per year of the 60 MS patients are shown as a histogram. The median BVL per year was 0.44% (IQR 0.12–0.8%). Eleven patients (18%) showed a BVL per year greater than 0.94% (marked as dark gray bars)

Discussion

We re-analyzed three publicly available independent MRI datasets from healthy subjects (Mclaren and OASIS dataset) as well as MS patients (Biberacher dataset), each acquired on the same scanners within a time interval of days to weeks. Given the short time intervals, we would not expect significant aging- or disease-related atrophy to contribute to the volumes assessed. Therefore, we anticipated that all deviations from zero in the SIENA results represent either measurement errors or short-term biological fluctuation, e.g., due to de- or rehydration or hormonal status and others.

The WPF range that we estimated in the present study derived from the three datasets was $\pm 0.54\%$. By definition (interval is defined as $\pm 1.96 \times \mathrm{SD}$, and therefore, contains 95% of the variability) there is still a 5% chance that BVL measurements from short-term repeated scans are even higher or lower, as can be seen in Fig. 1. We proposed to consider the WPF when assessing BVL measurements in individual patients. As explained in the methods section, the range of fluctuation needs to be accounted for when judging on the measured BVL as a safety margin. For example, for a measured BVL of 0.94% in an individual patient the true BVL lies within the interval [0.4%, 1.48%]. (with an error probability of 5%). Thus, the BVL value is at least 0.4% and can be considered pathological atrophy with a high certainty. The question arises as to what extent any such pathological atrophy can be expected in a contemporary cohort of MS patients. To answer this question, we assessed a group of 60 relapsing–remitting MS patients form an ongoing longitudinal, observational study with a median scan interval of 1.3 years. As shown in Fig. 3, 11/60 patients (18%) presented with an annual BVL that exceeded 0.94%. For this group of patients, brain atrophy could be identified with satisfying certainty after an average follow-up of 1 year and 3 months only. Given the fact that atrophy rates of MS patients and healthy participants overlap significantly [8] assessing pathological atrophy in 18% of early MS patients with relatively short follow-up can be considered substantial.

With increasing scan intervals, the impact of biological confounders as related to true pathological atrophy decreases (Table 3). This was deduced by the formula provided in the methods section and the assumption that the WPF is independent of scanning intervals. If, for example, we assume a yearly BVL of 0.4% per year for a hypothetical patient, the total BVL after 5 years will be approximately 2%. If we measure a total BVL of 2% after 5 years, we can expect the true BVL to be found within an interval of $2 \pm 0.54\%$. For the corresponding BVL per year, we obtain the interval $0.4 \pm 0.11\%$ according to the formula provided in the methods section. With longer intervals, the signal (BVL) increases, while WPF remains constant.

Our approach has several limitations. As discussed, we assume the WPF to be independent of the scan interval. This might not be completely true. Longer time intervals between baseline and follow-up scan might result in a higher chance that small variations in the scanner configuration occur (e.g., software updates) which can influence the variability of the measurement. However, changes in scanner configuration are always a potential source of measurement error for longitudinal data and should be known and controlled for by carefully checking the scanning parameters previous to the analysis. Reliable BVL measurements can only be expected if scanner, coil and scanning protocol remain unchanged. If this is controlled for, the signal-to-noise ratio of annualized BVL measurements improve with increasing interval between scans. This has been observed by others; correlations between BVL per year and clinical parameters (such as EDSS) become stronger for longer scanning intervals [18]. Another potential limitation of our study is the assumed independence of range of fluctuation from the subject. Age and disease type might well influence the WPF. In a recent study [16], it has been shown that test–retest variability of an atlas-based automatic segmentation method is higher in patients with Alzheimer's disease in comparison to healthy subjects. It was shown that the higher variability was caused by higher occurrence of motion artifacts among the patients' MRIs. Assessing image quality is of paramount importance. If artifacts are present, results need to be interpreted with caution. Our three reliability cohorts consist of relatively young adults (under 30 years) and the images did not contain motion artifacts. As a consequence, our results might not be transferable to older populations or other disease entities.

The aim of this study was to distinguish between WPF of brain volumes and real long-term effects on the brain volume related to disease or ageing. As explained in the introduction, there are many factors such as behavioral, genetics, comorbidities, inflammation, and pharmacological treatments, which can potentially contaminate any

BVL measurement. In case a BVL is clearly beyond the WPF, it is in the hands of the clinicians (the radiologist and treating neurologist) to carefully weigh the findings against such known confounders, a task, for which knowledge about the clinical status and treatment of the patient is paramount.

Since the datasets used in our study are publicly available, we can compare our results with those published by others. In a recent publication, Smeets et al. [20] used the Maclaren dataset to compare the performance of SIENA with a Jacobian integration-based approach ("MSmetrix"). The authors used a BET parameter of $f=0.1$ (instead of the default parameter $f=0.5$). For SIENA, the results were as follows: 25th percentile = 0.15%, median = 0.31%, 75th percentile = 0.75%, and maximum = 2.4%. Processing the same data (using an optimized pre-processing pipeline as detailed in the methods section) we obtained the following results for SIENA: 25th percentile = 0.05%, median 0.14%, 75th percentile = 0.24%, and maximum = 0.56% (see Table 2). The variability of SIENA results in our analysis was lower than the one reported in the paper by Smeets et al. [20] which underpins the importance of an optimized pre-processing pipeline when using SIENA.

In another study by Biberacher et al. [2], the Biberacher dataset was processed with two different SIENA pipelines. Our results turned out to be consistent with those obtained in the original publication. Biberacher et al. [2] concluded from their results that the measurement of individual patients' BVL with SIENA with satisfying reliability is only possible at scan intervals that exceed 5 years. The authors' conclusion seemed very much driven by the two extreme outliers within their measurements. In the present work, we showed that taking the WPF range in account, a reliable detection of pathological BVL is already possible after 1–2 years in a substantial proportion (approximately 20%) of MS patients.

In the pooled dataset, we found a median absolute BVL of 0.15%. These results are very consistent with the previously published test–retest results, which have been reported in the range of 0.15–0.20% for SIENA [4, 21, 22]. In these studies, patients were examined twice within a few minutes. Since within such short time periods biological factors should not influence the results significantly, one can assume that all deviations from zero within minutes up to a few hours are largely if not entirely due to endogenous noise of the computational method and/or image acquisition. With greater time intervals between scans, biological factors such as hydration status [9] and endocrine factors [6], as well as disease related factors (e.g., influx/efflux) of inflammatory edema in case of MS [6] may add to the measurement error. As indicated above, the median of the test/re-test variability estimated in our study is very similar to results reported for SIENA based on immediate repeated measures [4, 21, 22]. This indicates that the dominating factor for WPF is the noise of the computational method. The short-term biological factors seem to have only limited impact on the overall WPF.

Conclusion

Using optimized pre-processing, we examined three independent MRI reliability datasets to assess the within-patient fluctuation (WPF) range consisting of the magnitude of the intrinsic measurement error of SIENA and the impact of short-term biological fluctuations on brain volumes. The estimated WPF range may be helpful when interpreting BVL results on an individual patient level in diseases such as multiple sclerosis.

Acknowledgements The OASIS database is made available by the Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN), and supported by NIH grants P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584.

Funding S. Schippling and C. Walker-Egger are supported by the Clinical Research Priority Program of the University of Zurich.

Compliance with ethical standards

Conflicts of interest S Schippling reports receiving research Grants from Novartis and Sanofi Genzyme; and consulting and speaking fees from Biogen, Merck Serono, Novartis, Sanofi Genzyme, and Teva. P Manogaran and C. Walker-Egger received travel support from Sanofi Genzyme and Merck Serono. N de Stefano has received honoraria and consultation fees from Merck Serono S.A., Teva Pharmaceutical Industries, Novartis Pharma AG, BayerSchering AG, Sanofi-Aventis and Serono Symposia International Foundation. MP. Sormani received consulting fees from Biogen Idec, Merck Serono, Teva, Genzyme, Roche, Novartis, GeNeuro and Medday.

Ethical standards statement All patients gave written informed consent. The study was approved by the Ethics Board of the Canton of Zurich (KEK-ZH-Nr. 2013-0001).

References

1. Barkhof F (2016) Brain atrophy measurements should be used to guide therapy monitoring in MS—NO. Mult Scler 22:1524–1526
2. Biberacher V, Schmidt P, Keshavan A, Boucard CC, Righart R, Samann P, Preibisch C, Frobel D, Aly L, Hemmer B, Zimmer C, Henry RG, Muhlau M (2016) Intra- and interscanner variability of magnetic resonance imaging based volumetry in multiple sclerosis. NeuroImage 142:188–197
3. Chard D, Miller D (2009) Grey matter pathology in clinically early multiple sclerosis: evidence from magnetic resonance imaging. J Neurol Sci 282:5–11
4. Cover KS, van Schijndel RA, Popescu V, van Dijk BW, Redolfi A, Knol DL, Frisoni GB, Barkhof F, Vrenken H, neuGrid, Alzheimers Disease Neuroimaging I (2014) The SIENA/FSL whole

brain atrophy algorithm is no more reproducible at 3 T than 1.5 T for Alzheimer's disease. Psychiatry Res 224:14–21
5. De Stefano N, Silva DG, Barnett MH (2017) Effect of Fingolimod on brain volume loss in patients with multiple sclerosis. CNS Drugs 31(4): 289–305
6. De Stefano N, Silva DG, Barnett MH (2017) Effect of Fingolimod on brain volume loss in patients with multiple sclerosis. CNS Drugs 31:289–305
7. De Stefano N, Stromillo ML, Giorgio A, Bartolozzi ML, Battaglini M, Baldini M, Portaccio E, Amato MP, Sormani MP (2016) Establishing pathological cut-offs of brain atrophy rates in multiple sclerosis. J Neurol Neurosurg Psychiatry 87:93–99
8. De Stefano N, Stromillo ML, Giorgio A, Bartolozzi ML, Battaglini M, Baldini M, Portaccio E, Amato MP, Sormani MP (2015) Establishing pathological cut-offs of brain atrophy rates in multiple sclerosis. J Neurol Neurosurg Psychiatry
9. Duning T, Kloska S, Steinstrater O, Kugel H, Heindel W, Knecht S (2005) Dehydration confounds the assessment of brain atrophy. Neurology 64:548–550
10. Giovannoni G, Turner B, Gnanapavan S, Offiah C, Schmierer K, Marta M (2015) Is it time to target no evident disease activity (NEDA) in multiple sclerosis? Mult Scler Relat Disord 4:329–333
11. Hagemann G, Ugur T, Schleussner E, Mentzel HJ, Fitzek C, Witte OW, Gaser C (2011) Changes in brain size during the menstrual cycle. PLoS One 6:e14655
12. Maclaren J, Han Z, Vos SB, Fischbein N, Bammer R (2014) Reliability of brain volume measurements: a test–retest dataset. Sci Data 1:140037
13. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL (2007) Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci 19:1498–1507
14. Nakamura K, Brown RA, Araujo D, Narayanan S, Arnold DL (2014) Correlation between brain volume change and T2 relaxation time induced by dehydration and rehydration: implications for monitoring atrophy in clinical studies. Neuroimage Clin 6:166–170
15. Opfer R, Ostwaldt AC, Sormani MP, Gocke C, Walker-Egger C, Manogaran P, De Stefano N, Schippling S (2017) Estimates of age-dependent cutoffs for pathological brain volume loss using SIENA/FSL-a longitudinal brain volumetry study in healthy adults. Neurobiol Aging 65:1–6
16. Opfer R, Suppa P, Kepp T, Spies L, Schippling S, Huppertz HJ (2016) Atlas based brain volumetry: how to distinguish regional volume changes due to biological or physiological effects from inherent noise of the methodology. Magn Reson Imaging 34:455–461
17. Popescu V, Battaglini M, Hoogstrate WS, Verfaillie SC, Sluimer IC, van Schijndel RA, van Dijk BW, Cover KS, Knol DL, Jenkinson M, Barkhof F, de Stefano N, Vrenken H (2012) Optimizing parameter choice for FSL-brain extraction tool (BET) on 3D T1 images in multiple sclerosis. NeuroImage 61:1484–1494
18. Radue EW, Barkhof F, Kappos L, Sprenger T, Haring DA, de Vera A, von Rosenstiel P, Bright JR, Francis G, Cohen JA (2015) Correlation between brain volume loss and clinical and MRI outcomes in multiple sclerosis. Neurology 84:784–793
19. Schippling S, Ostwaldt A-C, Suppa P, Spies L, Manogaran P, Gocke C, Huppertz H-J, Opfer R (2017) Global and regional annual brain volume loss rates in physiological aging. J Neurol 264(3): 520–528
20. Smeets D, Ribbens A, Sima DM, Cambron M, Horakova D, Jain S, Maertens A, Van Vlierberghe E, Terzopoulos V, Van Binst AM, Vaneckova M, Krasensky J, Uher T, Seidl Z, De Keyser J, Nagels G, De Mey J, Havrdova E, Van Hecke W (2016) Reliable measurements of brain atrophy in individual patients with multiple sclerosis. Brain Behav 6:e00518
21. Smith SM, Rao A, De Stefano N, Jenkinson M, Schott JM, Matthews PM, Fox NC (2007) Longitudinal and cross-sectional analysis of atrophy in Alzheimer's disease: cross-validation of BSI, SIENA and SIENAX. NeuroImage 36:1200–1206
22. Smith SM, Zhang Y, Jenkinson M, Chen J, Matthews PM, Federico A, De Stefano N (2002) Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17:479–489
23. Takao H, Hayashi N, Ohtomo K (2012) A longitudinal study of brain volume changes in normal aging. Eur J Radiol 81:2801–2804
24. Zivadinov R, Dwyer MG, Bergsland N (2016) Brain atrophy measurements should be used to guide therapy monitoring in MS—YES. Mult Scler 22:1522–1524