NeuroImage 172 (2018) 450–460

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data

Christoph Vogelbacher [a,i,1], Thomas W.D. Möbius [b,1], Jens Sommer [c,i], Verena Schuster [a,i], Udo Dannlowski [d], Tilo Kircher [a,i], Astrid Dempfle [b], Andreas Jansen [a,c,*,i,1], Miriam H.A. Bopp [a,e,i,1]

[a] Department of Psychiatry and Psychotherapy, University Marburg, Marburg, Germany
[b] Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany
[c] Core-Unit Brainimaging, Faculty of Medicine, University Marburg, Marburg, Germany
[d] Department of Psychiatry and Psychotherapy, University Münster, Münster, Germany
[e] Department of Neurosurgery, University Marburg, Marburg, Germany
[i] Marburg Center for Mind, Brain and Behavior (MCMBB), Marburg, Germany

ARTICLE INFO

Keywords:
MRI quality assurance
Multicenter study
Major depression
Bipolar disorder
fMRI
DTI

ABSTRACT

Large, longitudinal, multi-center MR neuroimaging studies require comprehensive quality assurance (QA) protocols for assessing the general quality of the compiled data, indicating potential malfunctions in the scanning equipment, and evaluating inter-site differences that need to be accounted for in subsequent analyses.

We describe the implementation of a QA protocol for functional magnet resonance imaging (fMRI) data based on the regular measurement of an MRI phantom and an extensive variety of currently published QA statistics. The protocol is implemented in the MACS (Marburg-Münster Affective Disorders Cohort Study, http://for2107.de/), a two-center research consortium studying the neurobiological foundations of affective disorders. Between February 2015 and October 2016, 1214 phantom measurements have been acquired using a standard fMRI protocol. Using 444 healthy control subjects which have been measured between 2014 and 2016 in the cohort, we investigate the extent of between-site differences in contrast to the dependence on subject-specific covariates (age and sex) for structural MRI, fMRI, and diffusion tensor imaging (DTI) data.

We show that most of the presented QA statistics differ severely not only between the two scanners used for the cohort but also between experimental settings (e.g. hardware and software changes), demonstrate that some of these statistics depend on external variables (e.g. time of day, temperature), highlight their strong dependence on proper handling of the MRI phantom, and show how the use of a phantom holder may balance this dependence. Site effects, however, do not only exist for the phantom data, but also for human MRI data. Using T1-weighted structural images, we show that total intracranial (TIV), grey matter (GMV), and white matter (WMV) volumes significantly differ between the MR scanners, showing large effect sizes. Voxel-based morphometry (VBM) analyses show that these structural differences observed between scanners are most pronounced in the bilateral basal ganglia, thalamus, and posterior regions. Using DTI data, we also show that fractional anisotropy (FA) differs between sites in almost all regions assessed. When pooling data from multiple centers, our data show that it is a necessity to account not only for inter-site differences but also for hardware and software changes of the scanning equipment. Also, the strong dependence of the QA statistics on the reliable placement of the MRI phantom shows that the use of a phantom holder is recommended to reduce the variance of the QA statistics and thus to increase the probability of detecting potential scanner malfunctions.

* Corresponding author. Department of Psychiatry, University of Marburg, Rudolf-Bultmann-Strasse 8, 35039 Marburg, Germany.
E-mail address: jansena2@staff.uni-marburg.de (A. Jansen).
[1] Contributed equally.

https://doi.org/10.1016/j.neuroimage.2018.01.079
Received 21 December 2017; Received in revised form 26 January 2018; Accepted 30 January 2018
Available online 1 February 2018

Introduction

Affective disorders, i.e. major depressive disorder (MDD) and bipolar disorder (BD), are common, chronic, costly and debilitating diseases. Genetic and environmental risk factors contribute to both their etiology and their longitudinal course (Meyer-Lindenberg and Tost, 2012; Tost et al., 2012). The neurobiological correlates by which these predispositions exert their influence on brain structure and function however are poorly understood. The overarching aim of the multicenter research consortium *MACS* (Marburg-Münster Affective Disorders Cohort Study, http://for2107.de/) is to decipher neurobiological mediators and pathways leading from individual configurations of genetic and environmental risk factors to the clinical presentation of symptoms and the course of illness. Within this consortium, a large cohort of subjects (n~2500) will be recruited, consisting of healthy subjects and patients suffering from either MDD or BD. All participants will be deeply phenotyped by multimodal magnetic resonance imaging (MRI), clinical assessment, neuropsychology, and biomaterial analyses. The cohort will be completely re-assessed after two years.

Large, longitudinal, multicenter MR neuroimaging studies require careful planning and coordination, making a comprehensive quality assurance (QA) protocol necessary (Glover et al., 2012). Although modern MRI systems show good technical quality (i.e. high signal-to-noise ratio, good image homogeneity, and minimal ghosting) and differentiation between tissue classes (i.e. image contrast), image characteristics may change significantly over the course of a longitudinal study and may differ between MRI scanners. This is in particular a major challenge for functional magnetic resonance imaging (fMRI) studies since functional signal changes are typically just a small fraction (~1–5%) of the raw signal intensity (Friedman and Glover, 2006a,b). Therefore in particular the temporal stability of MRI acquisitions is important, for instance to differentiate between MRI signal changes that are associated with the time course of a disease and signal changes caused by alterations in the MRI scanner environment. In a longitudinal, multicenter imaging study, there are many MRI (e.g. choice of scan parameters, selection of paradigms) and non-MRI related factors (e.g. data storage, long-term management of measurement procedures) which have to be properly controlled for in order to improve the overall quality and to reduce intersite variability (for an overview, cf (Glover et al., 2012)).

Several examples of MRI scanner QA protocols are described in the literature, mostly in the context of large-scale multicenter studies (for an overview, see (Glover et al., 2012; Van Horn and Toga, 2009)). Depending on the main neuroscientific or clinical questions, these QA protocols focused on the quality assessment for structural (e.g.Gunter et al. (2009)) or functional MRI data (e.g.Friedman and Glover (2006a, b)). In several ongoing projects, QA protocols were also developed for more specialized problems, for instance in multimodal settings as the combined acquisition of MRI with EEG (Ihalainen et al., 2015) or PET data (Kolb et al., 2012) or with regard to the development of new phantoms (Hellerbach et al., 2013; Olsrud et al., 2008; Tovar et al., 2015). The analysis of the implementation of quality assurance methods has become one important factor to look at if one is interested in evaluating the strength of large-scale neuroimaging studies. The documented adherence to QA protocols is considered a key benchmark that will help to guide both clinicians and researchers to evaluate the quality, impact, and relevance of the study to the patient-level (Van Horn and Toga, 2009).

In multicenter designs, data has to be pooled across different MR scanners. Therefore it is necessary to develop analysis techniques that properly account for intersite variability. It has, e.g., been suggested that smoothing images to an equal full-width-at-half-maximum (FWHM) level (Friedman et al., 2006) or including the signal-to-noise ratios as covariate (Friedman et al., 2006) reduces differences in BOLD effect sizes across scanners. While potentially reducing intersite differences is important when pooling data in multicenter studies, this does by no means obviate the need to account for scanner differences by dedicated covariates in the subsequent formal analysis. Here, we shall illustrate the presence and extent of between-center differences and the dependence on experimental conditions such as temperature and time of day for different imaging modalities. This illustrates that subsequent analyses performed in the MACS consortium have to account for these covariates.

The present article had two aims. The first aim was to describe the implementation of a comprehensive QA protocol for the acquisition of MRI data in the MACS consortium. This protocol aimed to monitor MR scanner performance, to assess the impact of changes in scanner hardware and software, and to serve as an early-warning system indicating potential scanner malfunctions. MR scanner characteristics were assessed by the regular measurement of a MRI phantom. Since MRI phantoms deliver more stable data than living beings, they can be used to disentangle instrumental drifts from biological variations and pathological changes. A variety of QA parameters can be calculated from phantom data, for instance geometric accuracy, contrast resolution, ghosting level, and spatial uniformity. For functional imaging studies, in particular the assessment of the temporal stability of the acquired time series is important, both within a session and between repeated measurements. In the present article, we will in particular provide a comprehensive overview of the QA statistics included in our QA protocol. The second aim was to analyze how QA data from phantom measurements was influenced by external variables. In particular, we (i) analyzed how these statistics differed between scanners, (ii) investigated the effect of changes in experimental settings (e.g. hardware changes), (iii) analyzed how QA statistics depend on time of day, temperature and helium level, and (iv) showed how the implementation of a phantom holder significantly decreased the variance of the QA statistics. We will further demonstrate that differences between MR scanners have a measurable impact on human MRI data, as exemplarily shown by standard analyses of MRI data.[2]

Methods

The *MACS* neuroimaging consortium involved two MR centers (Departments of Psychiatry at the University of Marburg and the University of Münster) with different hardware and software configurations. In Marburg, the data were acquired at a 3T MRI scanner (Tim Trio, Siemens, Erlangen, Germany) using a 12-channel head matrix Rx-coil. In Münster, data were acquired at a 3T MRI scanner (Prisma, Siemens, Erlangen, Germany) using a 20-channel head matrix Rx-coil.[3] Pulse sequence parameters were standardized across both sites to the extent permitted by each platform. Until April 30, 2017, only one major hardware change (change of a defective gradient coil, see below) took place at the University of Marburg.

The study started on September 9, 2014 at the University of Marburg, and on September 4, 2015 at the University of Münster. Re-assessment after a two-year interval started on June 21, 2016. All subjects were assessed with a large neuroimaging battery, involving both structural (high-resolution T1-weighted images, diffusion weighted imaging for DTI analyses) and functional measurements. The functional imaging

[2] At this point, it might be instructive to clarify the scope of the present article in order to guard against common misunderstandings. The focus of this article is the analysis of phantom QA data with the aim to monitor the long-term performance of the MR scanners in the MACS consortium. The phantom data, however, cannot be used to directly assess the quality of the human MRI data. Even if a MR scanner performs acceptably, human MRI data might have to be excluded for other reasons (e.g. extensive motions artefacts). For the analysis of human MRI data, a separate QA protocol has to be developed, depending on the image modality (e.g. T1-weighted image or functional image) and the analysis methods. This is, however, beyond the scope of the present article. All analyses with the human MRI data that are presented in this article were included to illustrate that differences between MR scanners used in the MACS consortium have a large impact on the human MR data.

[3] Throughout the manuscript, we will discuss the influence of differences between MR scanners used in both centers. Of note, not only the MR scanners were different, but also the head coils. Scanner differences thus comprise the combined effect of different scanner models and different head coils.

battery included a face matching task (Hariri et al., 2002), an affective priming task (Suslow et al., 2013), a face encoding task (Dietsche et al., 2014), and a 8-min resting state sequence.

In the following, we will (1) describe the QA protocol (sequence of measurements, MRI phantom, MRI parameters), (2) give an overview on the QA statistics, and (3) describe how we analyzed human MRI data.

QA study protocol

The basic idea of the QA protocol was to regularly measure a MRI phantom and perform an automated analysis of the acquired data using various QA statistics.

MRI Phantom. The phantom was a 23.5 cm long and 11.1 cm-diameter cylindrical plastic vessel (Rotilabo, Carl Roth GmbH + Co. KG, Karlsruhe, Germany) filled with a mixture of 62.5 g agar and 2000 ml distilled water. In contrast to widely used water filled phantoms, agar phantoms are more suitable for fMRI studies. On the one hand, T2 values and magnetization transfer characteristics are more similar to brain tissue (Hellerbach and Einhäuser-Treyer, 2013), on the other hand they are less vulnerable to scanner vibrations and thus avoid a long settling time prior to data acquisition (Friedman and Glover, 2006a,b).

Scanning protocol. A phantom measurement was performed after each subject. Only if two subjects were measured consecutively, it was allowed to measure the MRI phantom only once (i.e. between the two measurements). At the beginning of the study, the phantom was manually aligned in the scanner and fixated using soft foam rubber pads. Alignment of the phantom was lengthwise, parallel to the z-axis, and at the center of the head coil (Supplementary Fig. S1). The alignment of the phantom was evaluated by the radiographer performing the measurement and – if necessary – corrected using the localizer scan. The measurement volume was manually centered at the phantom with slice direction perpendicular to the phantom body. To reduce spatial variance related to different placements of the phantom in the scanner and to decrease the time-consuming alignment procedure, we developed a Styrofoam phantom holder in the course of the study (Supplementary Fig. S2). The phantom holder allowed a more time-efficient and standardized alignment of the phantom within the scanner. The measurement volume was placed automatically in the center of the phantom. Since September 17, 2015, all phantom measurements at the University of Marburg were performed with this holder. A similar holder will be used in Münster in the near future. In addition to MRI data, further scanner related parameters were collected such as the temperature of the scanner room, the helium level during each phantom measurement, MR system maintenance appointments and all MR scanner related incidents (e.g. hardware failures).

Major incidents. On June 3, 2016, service technicians detected a defective radiofrequency pulse amplifier in Marburg during a regular maintenance service performed by the manufacturer. After the replacement, it was not possible to adjust the new amplifier to the MRI system. During extensive error diagnostics, the technicians detected a water bubble around one of the gradient coils located below the body coil. After the gradient coil was replaced, the MRI system was working properly again. On August 11, 2016, the MR protocol in Münster was changed. During the analysis of human fMRI data, it was detected that activity in the amygdala, a core region activated during the face matching task, was relatively low. Therefore the *prescan normalize*[4] option was activated to increase signal-to-noise in deeper brain regions.

Table 1
MRI parameters of the imaging sequences used to measure the phantom at the sites Marburg and Münster.

Site	Marburg	Münster
Repetition time (TR)	2000 ms	2000 ms
Echo time (TE)	30 ms	29 ms
Field of View (FoV)	210 mm	210 mm
Matrix size	64 × 64	64 × 64
Slice thickness	3.8 mm	3.8 mm
Distance factor	10%	10%
Flip angle	90°	90°
Phase encoding direction	anterior >> posterior	anterior >> posterior
Bandwidth	2232 Hz/Px	2232 Hz/Px
Acquisition order	Ascending	Ascending
Number of slices	33	33
Measurements	164	164
Effective voxel size (mm^3)	3.28 × 3.28 × 4.18	3.28 × 3.28 × 4.18
Acquisition time (TA)	5:34 min	5:34 min

MRI data acquisition. We designed a QA program that focused on the temporal stability of the MRI data, necessary for fMRI studies in which MR scanners are typically highly stressed. We therefore measured the MRI phantom with a functional T2*-weighted echo planar imaging (EPI) sequence sensitive to blood oxygen level dependent (BOLD) contrast. We chose the same sequence parameters as for the resting-state measurement, albeit a lower acquisition time (Table 1). Also the same scanner specific reconstruction methods were employed, since alterations might be reflected in the resulting imaging data. The MRI parameters of the EPI sequences are listed in Table 1. 167 images were acquired. Images 1–3 were by default not recorded by the MRI system, images 4–5 were also discarded from analysis to account for equilibrium effects.

Analysis of QA data

A wide array of QA methods has been developed to describe MR scanner stability (for an overview see e.g. Glover et al. (2012)). Our QA protocol used statistics as they were previously described by Friedman et al (Friedman and Glover, 2006a,b) (the so-called "Glover parameters"). These statistics were complemented by other parameters described by Simmons et al. (1999); Stöcker et al. (2005). In the course of the project, we adapted both the algorithms used to analyze the data (e.g. by the development of air bubble detection algorithms) and the operating procedures (e.g. by the introduction of a standardized phantom holder).

The data analysis was fully automated. All algorithms were implemented in Matlab R2014a (Version 8.3.0.532, 64 Bit February 11, 2014) (The Mathworks, Inc. Natick, MA, USA) and Python 2.7.11 (https://www.python.org, 5th Dec. 2015). First, all data were converted from DICOM to the NIfTI format using the *dcm2nii* tool (https://www.nitrc.org/projects/dcm2nii/, version 7, July 2009). Second, images were binarized applying Otsu's method (Otsu, 1979), separating each data set into phantom and background. Third, an array of published and widely used QA statistics – statistic maps and summary statistics – were calculated. These statistics were summarized in a QA report file for further inspection.

In a first step, both the original phantom data and the QA statistic maps (i.e. signal image, temporal fluctuation noise image, signal-to-fluctuation noise ratio image and static spatial noise image) were visually inspected by two of the authors (C.V., M.B.) to detect potential signal abnormalities (e.g. unexpected structures, large signal intensity deviations) or artefacts (e.g. ghosting or air bubbles). In a second step, we calculated all QA values for all phantom measurements at both sites. In

[4] MR imaging is increasingly performed, as in the present case, with arrays of small surface coils placed near the body. The advantage of using small surface coils is that they produce higher signal-to-noise ratios than would be possible from a larger, more distant coil. The disadvantage is non-uniformity of the signal. The depth of penetration of coils is inversely proportional to their diameters. Signals arising superficially in the subject are thus accentuated, while those deeper in the brain (e.g. the amygdala) are attenuated. It is possible, however, to make corrections for non-uniform receiver coil profiles prior to imaging. For Siemens scanners, this method is known as "prescan normalize". The normalization process involves acquiring an additional pair of low resolution scans, one with the head coil receiving signals and the other with the body coil receiving signals instead. The body coil is used for RF transmission in both cases. Then, under the assumption that the large body coil's receive profile is homogeneous across a head-sized object, when the prescan head coil image is divided by the prescan body coil image, the resulting image is essentially an image of the receive field of the head Rx coil. This image can then be used to normalize a target image (such as an EPI), thereby removing the receive field heterogeneity.

the following, we describe both the calculation of and the motivation for these. These statistics are broadly subdivided into statistics describing either temporal or spatial characteristics of the phantom image.

Spatial characteristics and statistics

Signal image: The signal image is the voxel-wise average of the center slice of the phantom (slice-of-interest, SOI) across the time series (Friedman and Glover, 2006a,b). It can be used, by visual inspection, for a first assessment of spatial signal variability within the phantom. Inhomogeneity in the signal image might be caused, for instance, by problems of the head coils. *Static spatial noise image:* The static spatial noise image is the voxel-wise difference of the sum of all odd images and the sum of all even images in the SOI (Friedman and Glover, 2006a,b). If the signal variance at a voxel is high, this difference will consequently yield large values in magnitude, and may thus be used to visualize noise in the signal. *Signal-to-noise-ratio (SNR):* Friedman and Glover (Friedman and Glover, 2006a,b) define the SNR as the quotient of the average intensity of the signal image in a region of interest (ROI, 15 × 15 voxel), located at the center of the phantom of the SOI, and the standard deviation of the static spatial noise within the same ROI.[5]

Percent integral uniformity (PIU): PIU describes the uniformity of an image. Since the agar phantom consisted of a homogenous material, spatial inhomogeneity in the MR image may be related to scanner malfunctions. The PIU values were calculated for the SOI as follows (Mri and Program, 2005): The phantom was separated from the image background and minimum (I_{min}) and maximum intensities (I_{max}) within the phantom were detected using a moving 3 × 3 voxel ROI. The PIU was then calculated for each time point by

$$PIU = 100 \times (1 - (I_{max} - I_{min}) / (I_{max} + I_{min}))$$

From this time series, we calculated mean PIU, minimum PIU, maximum PIU, and the standard deviation of the PIU. In the results section, we will present the mean PIU.

Since PIU depends on the actual homogeneity of the agar phantom, it must be ensured that non-uniform signals, caused, e.g., by air bubbles in specific slices, are not included in the calculation of the image uniformity. We therefore implemented an air bubble detection algorithm that removed the influence of signal inhomogeneity not related to the MR scanner but the phantom itself. This algorithm significantly reduced the variance of the PIU value, making it more likely to detect image degradations caused by scanner malfunctions.

Ghosting: Ghosting is a typical artifact in MR images. In fMRI, ghosting artifacts can lead to spatially variable signals that might cause a displacement of activity or might decrease the sensitivity. In our protocol, we implemented two different methods to quantify the ghosting in the acquired imaging data. In a first approach (based on Simmons et al. (1999)), we used three 8 × 8 voxel ROIs placed in the SOI of the phantom, one at the phantom center and two in the image background, moving either in frequency or phase encoding direction across the image. Thereby, the mean intensity (I_s) within the central ROI as well as mean (I_{mean}) and maximum "mean intensity" (I_{max}) of the moving ROIs were calculated for each time point. Finally, four characteristics were defined as follows:

Signal to Maximum Ghost Ratio (Phase) = I_s/I_{max}(Phase)
Signal to Mean Ghost Ratio (Phase) = I_s/I_{mean}(Phase)
Signal to Maximum Ghost Ratio (Frequency) = I_S/I_{max}(Frequency)
Signal to Mean Ghost Ratio (Frequency) = I_S/I_{mean}(Frequency)

For all statistics, we calculated mean, maximum, minimum and standard deviation across all time points. In the results section, we will present the mean ghosting values.

In a second approach (based on the ACR protocol, (Mri and Program, 2005)), we calculated the *percent signal ghosting* (PSG). For each slice of the volume and time point, a signal ROI (8 × 8 voxels) was placed at the center of the phantom, four background ROIs (8 × 8 voxels) were located at the edges of the MR image (top, bottom, right, left) outside the phantom. The average signal intensity was calculated for each ROI. PSG was defined as

$$PSG = [(I_{top} + I_{bottom}) - (I_{left} + I_{right})] / [(2 * I_{center})]$$

In the results section, we will present PSG as time average either for the SOI (PSG_{slice}) or the total volume (PSG_{volume}).

Temporal characteristics and statistics

Temporal fluctuation noise image: The temporal fluctuation noise image was calculated, analogous to the signal image, for the SOI of the phantom. After the time series of each voxel in the SOI was detrended by a second order polynomial, we calculated the standard deviation of the residuals (Friedman and Glover, 2006a,b). It can be used, by visual inspection, to evaluate the signal variance which is left after having removed temporal drifts from the signal. *Signal-to-fluctuation-noise-ratio (SFNR) image*: The SFNR image is the voxel wise ratio of the signal image and the temporal fluctuation noise image (Friedman and Glover, 2006a, b). *SFNR:* The SFNR is the average intensity of the SFNR image in a ROI (15 × 15 voxel, placed at the center of the phantom of the SOI) (Friedman and Glover, 2006a,b). It is a signal to noise ratio, in which the denominator denotes the variability of the signal which cannot be explained purely by temporal fluctuations in the signal, and, thus, constitutes a measure of relative temporal noise.

Percent fluctuation and *drift*: First, the intensity values of all voxels in a ROI (15 × 15 voxel, placed at the center of gravity of the SOI) were averaged. Second, the mean signal intensity of the times series was calculated. Third, the standard deviation of the time series was calculated after detrending by a second order polynomial. The *drift* was defined as the ratio of the difference of highest and lowest values of the fitted polynomial and the mean signal intensity (Friedman and Glover, 2006a, b). The *percent fluctuation* was defined as the ratio of the standard deviation of the residuals (after detrending) and the mean signal intensity (Friedman and Glover, 2006a,b). Both drift and percent fluctuation are multiplied by 100.

Percent signal change (PSC): The PSC was adapted from Stöcker et al. (2005) and describes the homogeneity of the SNR over time. SNR was calculated for each slice and time point separately using five ROIs. A signal ROI (15 × 15 voxels) was defined in the center of the phantom, four background ROIs (8 × 8 voxels) to estimate the noise were placed in the corners of the image. The signal was defined as the averaged intensity in the signal ROI, the noise was defined as the Rayleigh corrected standard deviation of the signal intensity of the voxels in the four noise ROIs at one time-point. PSC was calculated as 100/SNR. It can be depicted as a time-course of the SNR either of each slice or the total volume. It thus can be used to detect deviations of the SNR for specific time-points or specific slices.

Statistical analysis of QA statistics from phantom measurements: First we studied the influence of site and experimental settings (e.g. hardware or software changes) on the normal ranges of the presented QA statistics. Time series plots were visually inspected for all QA statistics. Differences in mean, in variance, in drift, and also oscillation were, in fact, so obvious that formal statistical analysis was not needed.

Second, we studied the potential influence of external variables, such as temperature, time of day of the measurement, and helium level of the scanner, on the normal ranges of the above QA statistics. As described in the results section, differences in experimental settings (e.g. hardware and software changes) have a severe impact on the normal ranges of all

[5] Note, however, that this definition does not follow common conventions for a SNR, as the variance of the static spatial noise image is proportional to the number of images used for its calculation. When the number of images is large, also the variance in the static spatial noise image will be large. When the same number of images are used as a basis for its calculation, as it is the case in this study, it is possible, though, to use this value to compare relative spatial noise between scanners and experimental settings.

QA statistics. In order to have a homogeneous data set at hand, we opted to study the influence of these external parameters on a reduced data set which only consisted of scans measured in Marburg after the replacement of the defective coil. A linear model was fitted to each QA statistic which included covariates for temperature, time of day, helium level, and a variable which modelled the general level of the QA statistic at a specific date. The reason to include the latter is that most of the QA statistics are subject to drifts or oscillations within the almost two years of data acquisition. In order to estimate the general level of a QA statistic, a LOWESS (locally weighted scatterplot smoothing) which included 40% of the data at any given date was fitted to the respected series. Replacing the intercept of a model with this variable, makes the effect estimates for the external variables robust for shifts, drifts, or oscillations. After visual inspection of the temperature distribution in the data, temperature was dichotomized as "cold" (<20.8°) and "warm" (>20.8°). Helium level entered the model as a continuous variable. Time of day was also dichotomized into "early" (7:00–13:59) and "late" (14:00–20:00) measurements.

Assessment of human MRI data

In multicenter designs, data has to be pooled across different MR scanners. The data acquired at different scanners however can differ profoundly. To illustrate these differences, we analyzed the impact of different MR scanners on standard metrics obtained from human MRI data (e.g., total brain volume). We restricted our analysis to healthy control subjects to exclude effects related to disease status. Since in multi-site studies a prominent source of between-site bias can result from an imbalance in the distribution of subjects (e.g., differing numbers of men and women, different age distributions), we also investigated the effects of age and sex on structural and functional MRI measures. All explorative and formal statistical analysis were performed in Python (3.5.2) using packages Numpy (1.12.0), Pandas (0.19.2), and Statsmodels (0.8.0).

Analysis 1: In the first analysis, we assessed whether volumetric information from T1-weighted structural images differed between MR scanners. Total intracranial volume (TIV), total gray matter volume (GMV), and total white matter volume (WMV) were calculated using a standard processing pipeline of the CAT12 toolbox, applying a smoothing kernel of 8 mm (www.neuro.uni-jena.de/cat). An explorative data analysis assessed the general shapes and locations of the distributions of the response variables TIV, GMV, and WMV, and saw them fit for linear modelling. Linear models were fit to TIV, GMV, and WMV including the independent variables age (in years), sex, and site of the scan. For lack of influence, the variable age was dropped from the model for WMV. The validity of each model was assessed by visual inspection of the respective residual distributions. To analyze whether significant volume differences were caused by spatially localized differences between MR images acquired from both scanners, we also compared, using a voxel-based morphometry approach (VBM, (Ashburner and Friston, 2001)), as implemented in the CAT 12 toolbox, the arctan-transformed volume densities of each voxel, again using site, age and sex as covariates.

Analysis 2: In the second analysis, we assessed whether characteristics from T2*-weighted functional images obtained during the face matching task differed between MR scanners. Applying the *percent signal change* (PSC) routines for phantom data as described above on human fMRI data, we assessed the noise inherent in the functional imaging data. PSC values from fMRI data acquired in Münster and Marburg were compared using again a linear model with covariates site, age and sex.

Analysis 3: In the third analysis, we assessed whether fractional anisotropy (FA) information in selected brain regions, calculated from diffusion-weighted structural images, differed between MR scanners. DTI data analysis was performed using FSL 5.0.2 (FMRIB Software Library, Oxford, United Kingdom, http://www.fmrib.ox.ac.jk/fsl). Preprocessing of DTI data was performed according to the following protocol: First, images were corrected for head motion and eddy currents, respectively, by aligning all images onto the mean reference volume. Second, brain tissue was extracted using the FSL brain extraction tool BET. Third, the diffusion tensor and FA-maps were estimated for each voxel. Voxel-wise statistical analysis of FA-maps was performed using tract based spatial statistics (TBSS) v1.2 implemented in FSL according to the following procedure: First, all FA data sets were aligned into a common space, the standard Montreal Neurological Institute (MNI) space, using non-linear registration, and were subsequently interpolated, resulting in a spatial resolution of $1 \times 1 \times 1$ mm^3. Second, a mean FA-image was created and further thinned to generate a mean FA skeleton. Third, each subject's aligned FA data was projected onto the mean FA skeleton using a non-maximum suppression threshold of $\geq$0.3. The resulting data was then fed into the voxel-wise statistical analysis on the whole-brain mean FA skeleton using a linear regression model with site, age and sex as covariates. Each contrast was analyzed according to permutation–based non-parametric inference with 10,000 random permutations, using threshold-free cluster enhancement (TFCE), allowing for correction for multiple comparisons with a significance level of $p < 0.05$.

Results

Analysis of phantom MRI data

We set October 31, 2016, a-priori as date for a "data freeze" of the phantom data. Until this date, 1009 phantom measurements were performed in Marburg, 205 in Münster. This data set was used for the following analyses. In Marburg, 369 measurements were performed without phantom holder and 640 with holder. From the 640 phantom measurements performed with the phantom holder, 428 took place before replacement and 212 after replacement of the defective gradient coil. In Münster, 165 measurements were done without the *pre-scan normalize* option and 40 measurements with this changed routine.

First, we show that differences between the scanners, technical changes of a scanner (such as the replacement of the MRI gradient coil), and changes in the QA-protocol (such as the introduction of a phantom holder), or changes in certain sequence parameters (such as adding the *prescan normalization* option) impact many of the QA statistics in a variety of ways.

The supplementary material (Supplementary Figs. S3–S15) includes time series plots of all defined QA statistics. Three of the time series, namely *signal fluctuation noise ratio* (SFNR), PSG Signal Image (PSG-SI), and *maximal ghost frequency* (MGF), are shown in Figs. 1–3. The following properties, which are exemplarily described for SFNR, PSG-SI, and MGF, are characteristic for all implemented QA statistics (Supplementary Figs. S3–S15): normal ranges of each of the QA statistics differ between scanners and also drastically change whenever hardware or software settings are changed at a scanner – both in mean and variance. (i) The typical SFNR values for the scanner in Münster lie above the respective values in Marburg. (ii) After implementing a phantom holder in Marburg, the mean of SFNR and MGF dropped, the mean PSG-SI was raised, and the variances of SFNR and PSG-SI were considerably reduced. (iii) Whereas the coil change in Marburg did not seem to have a relevant impact on SFNR and MGF, it had a huge impact on PSG-SI. (iv) Adding the *prescan normalize* option to the protocol in Münster had an impact on the MGF. Whereas the normal range of Münster's MGF lay below the respective value in Marburg without *prescan normalize*, the typical range was later lying above. SFNR and MGF show drifts over time. Some outliers in the QA values can be identified and they are all explained by a wrong alignment of the phantom in the phantom holder in Marburg, by wrong placement of the phantom in the MR scanner in Münster or by use of a wrong phantom.

Based on the 212 phantom measurements which have been acquired in Marburg using the new phantom holder and after the coil change, we studied the dependence of QA statistics on the external variables temperature, time of day, and helium level. Helium level does not seem to have an influence on any of the QA-statistics. Measurements during the second half of the day, seem to reduce SFNR ($\beta = -1.69$, CI(0.95)=(-2.77,-

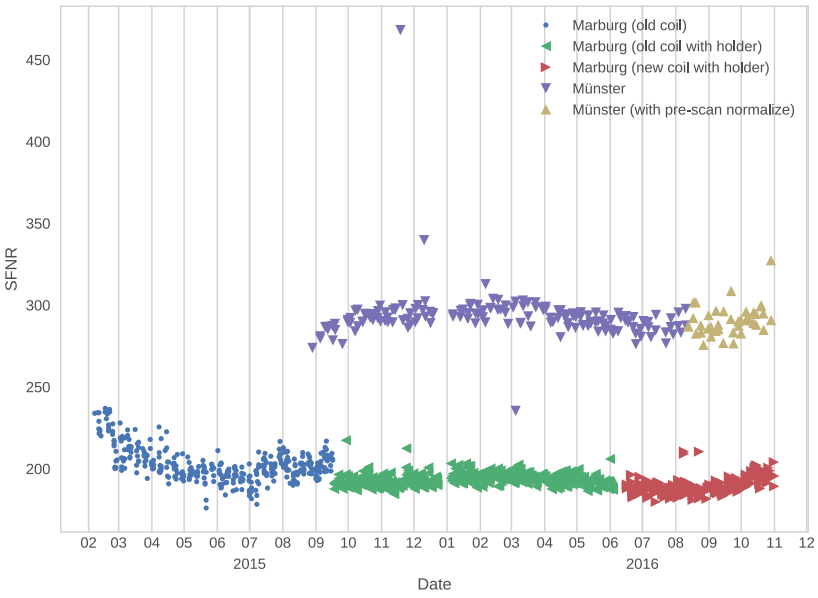

Fig. 1. SFNR values of phantom measurements in Marburg and Münster. One can see, for instance, that (i) typical SFNR values for the scanner in Münster lie above the respective values in Marburg, (ii) the implementation of a phantom holder in Marburg considerably reduced the variance of SFNR, and (iii) the coil change in Marburg did not have a relevant impact on SFNR.

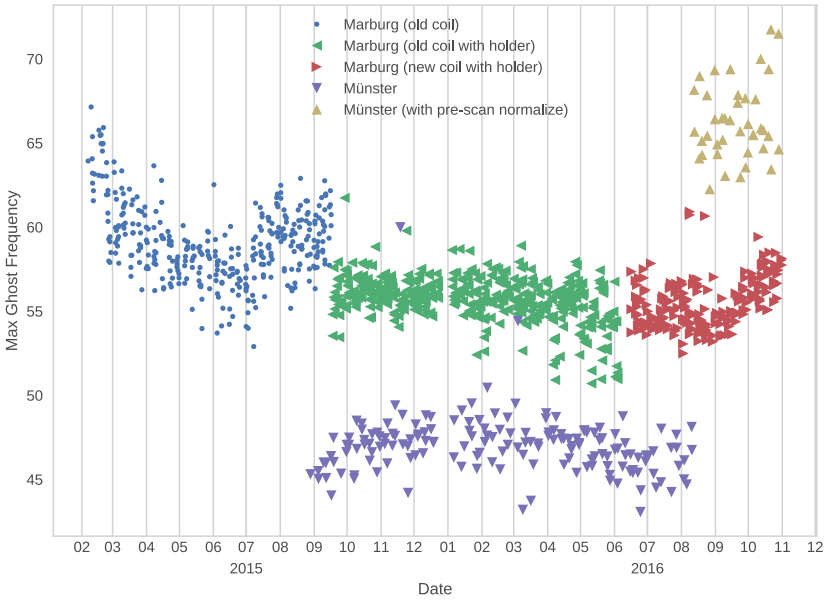

Fig. 2. MGF values of phantom measurements in Marburg and Münster. One can see, for instance, that the implementation of a phantom holder in Marburg reduced both the mean MGF and the variance. In contrast, the coil change in Marburg did not have a relevant impact on MGF. Adding the prescan normalize option to the protocol in Münster had an impact on the MGF. Whereas the normal range of Münster's MGF lay below the respective value in Marburg without prescan normalize, the typical range later was lying above.

0.62), p = 0.0021), Max Ghost Phase (β = -2.8, CI(0.95)=(-0.42, -0.14), p = 0.0001), Mean Ghost Frequency (β = -0.63, CI(0.95)=(-1.02, -0.25), p = 0.0012), Mean Ghost Phase (β = -0.42, CI(0.95)=(-0.68, -0.16), p = 0.0015), and to increase PSC (β = 0.01, CI(0.95)=(0.01, 0.02), p < 0.0001). Measurements acquired above 20.8 °C room temperature seem to reduce the SFNR (β = -2.12, CI(0.95)=(-3.24, -1), p = 0.0003).

Analysis of human MRI data

In the following, we assessed whether image characteristics from all imaging modalities differ between MR scanners and how they depend on subject covariates age and sex. In February 2016, a first data freeze was conducted after 1000 subjects (both patients and controls) were measured in the study. All 444 healthy control subjects (335 in Marburg, 109 in Münster) were used in the analysis of the volumetric information. Functional data from the face matching task and DTI data were not available from all these subjects but of a subsample of 373 (273 in Marburg, 100 Münster). Differences induced by the gradient coil change in Marburg were not assessed, as the data freeze was performed prior to this change.

Volumetric information from T1-weighted MR images

In order to assess possible differences between volumetric measurements between the scanners of Marburg and Münster, estimated brain volumes, namely TIV, GMV and WMV, were compared. Subjects' age ranged from 18 to 65 years and showed similar distributions in Marburg and Münster with almost the same range. The women to men ratio leaned towards the former with 61% women in the combined sample (66% in Münster and 59% in Marburg). Since age and sex are both known to be associated with brain volume, they need to be considered in the modelling process, as they would otherwise confound potential results.

TIVs were modelled using a linear model with age (in years), sex, and site as independent variables. As seen in Table 2, all three variables were statistically significant with p-values <0.001 for sex and site, and <0.01 for age. On average, TIV estimates in Münster were larger than in Marburg by a value of approximately 69 cm^3 or 0.58 standard deviations.

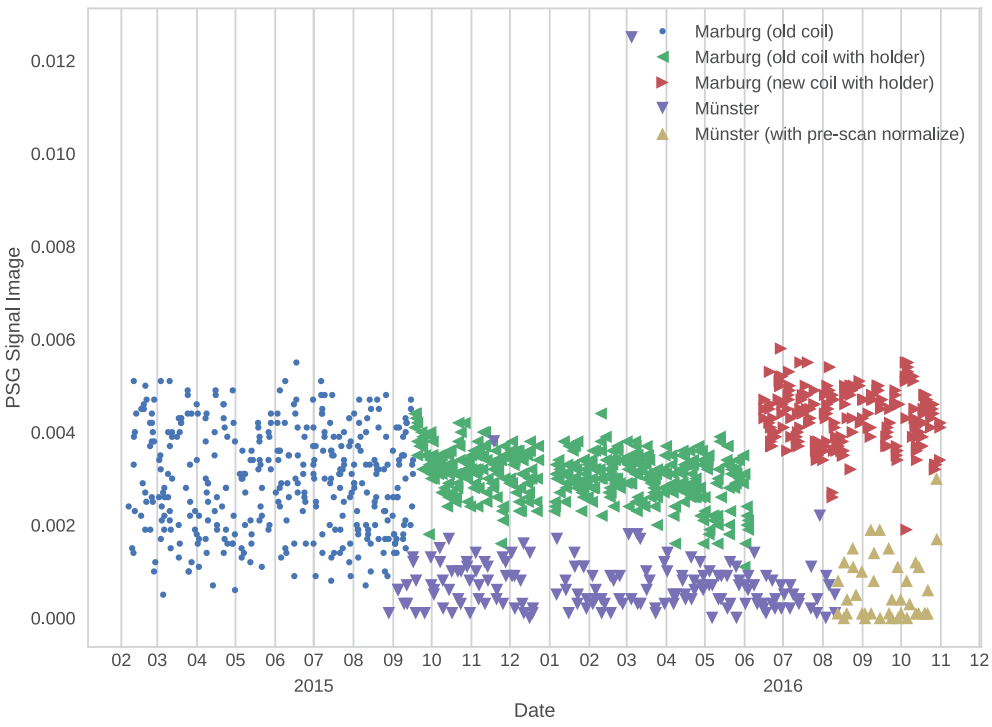

Fig. 3. PSG Signal Image values of phantom measurements in Marburg and Münster. One can see, for instance, that the mean PSG-SI was raised after implementation of a phantom holder, while the variance was considerably reduced. Interestingly, the coil change in Marburg had a huge impact on PSG-SI.

Table 2
Brain volumes (linear model coefficients with 95%-confidence intervals) calculated from T1-weighted images acquired in Marburg and Münster, respectively.

	Coefficient	Lower confidence limit (0.025)	Upper confidence limit (0.975)	p-value
TIV				
Intercept	1740.18	1703.71	1776.65	<0.0001
Sex[female]	−159.14	−181.70	−136.58	<0.0001
Site[Münster]	69.34	42.95	95.72	<0.0001
Age	−1.31	−2.26	−0.36	0.0070
GMV				
Intercept	851.00	835.06	866.93	<0.0001
Sex[female]	−69.86	−79.71	−60.00	<0.0001
Site[Münster]	30.13	18.61	41.66	<0.0001
Age	−2.88	−3.30	−2.47	<0.0001
WMV				
Intercept	575.07	567.04	583.10	<0.0001
Sex[female]	−68.67	−78.50	−58.84	<0.0001
Site[Münster]	19.09	7.95	30.23	0.0008

This corresponds to brains in Münster to be an average of 4.0% larger than in Marburg. Female TIVs lie an average of 159 cm^3 or 1.35 standard deviations below the TIVs of males, which corresponds to around 9.1% smaller TIVs of females in comparison to males. Estimated TIVs dropped by an average of 1.3 cm^3 or 0.01 standard deviations per year of age.

GMVs were also modelled by a linear model with the same independent variables age (in years), sex, and site. Table 2 shows that also for GMV all three variables were statistical significant with p-values <0.001. On average, grey matter volumes in Marburg were estimated to be 30 cm^3 or 0.59 standard deviations smaller than in Münster. This corresponds to an average of 3.5% larger grey matter estimates in Münster than in Marburg. The average grey matter volume of female lies 70 cm^3, 1.36 standard deviations, or 8.2% below the GMV of males. Grey matter decreases by 2.9 cm^3 or 0.06 standard deviations per year of age.

The first fit of a linear model to WMVs included the same independent variables as for TIV and GMV but has shown no significant impact of age to white matter. Consequently, the variable was dropped resulting in a linear model for WMV including the variables sex and site. Estimates are shown in Table 2. Both sex and site are statistical significant with p-values <0.001. Females have white matter volumes which are on average 70 cm^3 or 1.34 standard deviations below the volumes of males. This corresponds to 12% smaller volumes than males. Subjects in Münster have white matter volume estimates which lie on average 19 cm^3, 0.37 standard deviations, or 3.3% above the estimates of Marburg.

Within the age range in our sample (18–65 years), TIV estimates drop by an average of more than 61.5 cm^3. This effect is close to the effect of site on TIV, and lies well within the 95% confidence region of the site effect. If we had two subjects, one 18 and one 65 year old, both with average volumes, i.e. the 65 year old subject's TIV lies approximately 60 cm^3 below the 18 year old, their scans would yield the same TIV estimate, if the 65 year old subject was measured in Münster and the 18 year old in Marburg. GMV even drops by an average of 135.4 cm^3 within the age range from 18 to 65 years, which is larger than the effect of site and sex combined.

The sex effect (male to female), and the site effect (Marburg to Münster) go in opposite directions for TIV, GMV, and WMV respectively. As more females have been recruited in Münster, each of the two variables would act as a confounder for the estimation of the other: if sex would be dropped from the model, this would lead to an underestimation of the site effect; if site would be dropped from the model, this would lead to an underestimation of the sex effect. Even though the sex unbalance in this sample is not very severe and, dropping either of the variables only

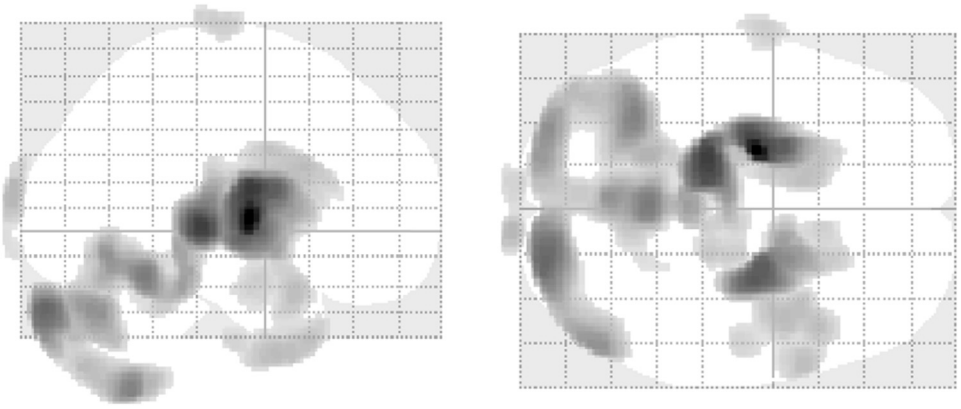

Fig. 4. Using voxel-based morphometry, volumetric differences between MRI data measured in Münster and Marburg were assessed at a local scale (contrast Münster > Marburg, $p < 0.001$ at the voxel level, $p < 0.05$ corrected at the cluster level). Brain volume differences were found in particular in the bilateral basal ganglia and thalamus and the posterior regions (occipital cortex, cerebellum).

lead to minor changes in the effect size estimates of the other (including their respected significance), and even though, as age is homogeneously distributed between Marburg and Münster, site does not act as a severe confounder, when estimating and testing for an age effect in this example, we strongly recommend to include site as a covariate to any model: Dropping either of the variables would lead to a misspecification of the same.

To assess whether the volumetric differences between MR scanners were spatially localized, we compared the data from both sites using voxel-based morphometry, again including sex and age as covariates in the model. Results are presented in Fig. 4 ($p < 0.001$ at the voxel level, $p < 0.05$ corrected at the cluster level). Brain differences were found in the bilateral basal ganglia and thalamus and the posterior regions (occipital cortex, cerebellum).

Noise in T2-weighted fMRI data*

Subjects' age range was the same as for the full sample of 444 subjects, as well as the ratio of women to men in the two centers. A linear model was fit for Percent Signal Change (PSC). Variable selection was performed by backward selection starting with the full set of available covariates: sex, age, and site, including their potential interactions. Neither sex nor site had a significant impact on PSC and were consequently dropped from the model resulting in a simple linear regression model with age as the only significant covariate ($p < 0.001$).

For each year of age, the average PSC of a person's face matching task increases by an average of 0.008 (95% confidence interval (0.005, 0.011)) or 0.5%. This corresponds to an average increase of 0.02 standard deviations. Although the covariate age is highly statistical significant, one should, however, also consider the low predictive power of the model: the adjusted coefficient of determination was estimated to 0.06, i.e., only 6% of the variation in PSC may be explained by age differences. Thus, although the effect of age is statistically significant, the statistical model also shows that its biological consequence is minor in this case.

Fractional anisotropy (FA) values from diffusion tensor imaging (DTI) data

Differences in FA maps between Marburg and Münster are depicted in Fig. 5. DTI measurements in Marburg showed significantly ($p < 0.05$, corrected for multiple comparisons) higher FA values in almost all regions assessed, while DTI data from Münster showed higher FA values specifically in the brainstem.

Discussion

For high-quality imaging studies, it is important to implement comprehensive QA protocols that assess, for instance, instabilities of the MRI system. Often malfunctions of the MR-scanner are only detected long after the study is finished and the QA data is retrospectively analyzed (see Friedman and Glover (2006a,b) for an example). In the present article, we therefore described the implementation of a comprehensive QA protocol for the acquisition of MRI data in the multicenter research consortium MACS. The protocol aimed to monitor scanner performance, to define benchmark characteristics, and to assess the impact of changes in scanner settings. We will, therefore, first discuss the impact of hardware and software changes on the normal ranges of the QA statistics, and the implications this impact has for their use in monitoring MR scanner

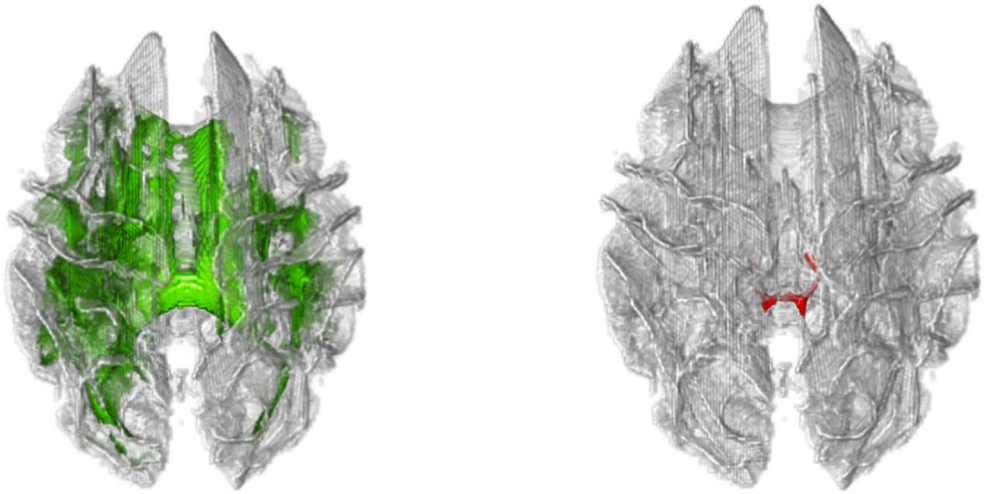

Fig. 5. Differences in FA values between DTI data measured in Münster ($p < 0.05$, corrected for multiple comparisons). For the contrast Marburg > Münster, differences (marked in green) were found in widespread regions throughout the brain (left). For the contrast Münster > Marburg, FA values were higher specifically in the brainstem (marked in red, right).

performance. Next, we will discuss the consequences that these differences have for multi-center studies or single-center studies, when changes to the scanner have occurred.

Implications of the QA data for the assessment of MR scanner performance

We have only included and implemented QA statistics which are currently published in the literature. The result section has shown that almost all of these QA statistics suffer from drastic changes in their normal ranges whenever there are changes in the MR scanning equipment. Consequently, any specific value of a QA statistic (e.g. that the PSC of a MR scanner is at a specific time 1.2%) has limited informational value for assessing or identifying sudden malfunctions of the scanning equipment. Only when embedded in the normal range of the respective QA statistic during the respective setting of the MR scanner at the time, one may be able to detect abnormal behaviors, and as scanner settings do change over time, this may only be possible in retrospect.

Another problem, which our study has revealed, is the strong dependence of all implemented QA statistics on even minor misplacements of the phantoms in the scanner. Although the agar phantom consists of homogenous material, image characteristics seem to change depending on where the slice of interest (SOI) is placed with respect to the phantom. This leads to increased variability of the QA statistics, in particular, when QA statistics are based on single slices (e.g. SNR and SFNR). This is an intrinsic problem of these statistics, and hints that it may be necessary to refine their definitions. When malfunctions in the scanning equipment produce effects in the QA statistics which are smaller than the variability of these statistics due to the day-by-day handling of the phantom, these malfunctions will remain undetected. A workaround to decrease the variability of the QA statistics, is the use of a phantom holder. This holder does not only reduce the time needed to place the phantom, it also ensures that the same volume is assessed during each measurement. This strongly decreased the variability of the QA statistics, making it more likely to detect potential scanner malfunctions. We therefore recommend, for all measurements, the use of phantom holders. This does not mean, of course, that the phantom holder improves the quality of the MRI scanner.

Some QA statistics seem to be affected by the time of day they have been acquired. This might be caused by heating up of the MR scanner due to the high amount of measurements over the day. It might be advisable to ensure that group data do not systematically differ with regard to acquisition time or temperature.

During the course of the study, both MR scanners underwent several software upgrades, but only one major hardware change. At Marburg, a defective gradient coil was replaced about one and a half years after start of the study. A comparison of phantom data before and after repair showed that some QA characteristics significantly differed. After repair, the QA statistics indicate a somewhat lower overall performance, characterized by increased noise and ghosting. We recommend to consider this event during data analysis of the human MRI data by including the scanner repair as a covariate. In particular for longitudinal aspects, e.g. when comparing data predominantly before and after the repair, special care has to be taken to disentangle for instance activation changes caused by hardware changes or by physiological effects.

Interestingly, we did not detect any hint in the QA data (acquired before the incident) that the gradient coil was defective, not even in a retrospective analysis. Instead, the defect was accidentally discovered during one of the regular maintenance services. This is much more surprising since the QA values have been proven to be sensitive with regard to many other changes in the environment. Also when artificially introducing disturbances (e.g. not fully closed door of the MR scanner room, changes of the homogeneity of the magnetic field by temporally introducing objects in MR scanner bore during phantom measurements), QA statistics strongly change. We therefore cannot say for how long the gradient coil was defective. It might be possible that the gradient coil was defective since the beginning of the study or that it occurred during or shortly before the maintenance service. In any case however, we do not have any hint that this incidence has a measurable effect on the MRI data during the time before repair.

What becomes immediate is that it is in general not possible to define universally valid normal ranges or benchmark characteristics for any of the currently published QA statistics. On the contrary, the data shown here demonstrate clearly that it is not even possible to present normal ranges for these statistics for the same scanner. Whenever there are hardware or software changes – even if the change may be considered to be minor – the change may affect the normal range of the QA statistics in mean, in variance, and drift. An interesting case example has been the introduction of the phantom holder in Marburg. The intention of monitoring QA statistics is that abnormalities in these statistics shall indicate possible malfunctions of the scanning equipment, and shall aid in the exclusion of potential faulty measurements just prior or after the respective phantom scan. In MRI experiments, we are dealing with three types of variances: biological variance (human brains are different, and even the same brain may react differently when measured repeatedly), technical variance (due to external parameters which cannot be fully controlled, e.g. temperature, voltage or magnetic fluctuations), and variance due to handling, e.g. differences in the placement of the body of interest in the scanner or placement of the measurement volume. Quality assessment aims to specifically monitor the technical variance of an experimental setting independent of handling differences. The use of a phantom reduces the biological variance and brings it close to zero - that it is not exactly zero is shown by the existence of air bubbles and other artefacts. The introduction of a phantom holder reduces the variance due to handling. The strong impact of the holder on almost all QA statistics, though, show that these statistics are not able to monitor the technical variance independent of handling. Some of the QA statistics even seem to reflect handling differences more than the technical variabilities of the scanner, e.g. MGF, PIU, and PSG-SI. Some small dependence on handling should be expected but the severity shown here is surprising. Our data show that only the use of a phantom holder currently allows to reasonably work with these QA statistics but even then: divergence from the norm are typically due to handling errors and do not reflect technical instabilities. This hints that most of the published QA statistics which are currently in use should be revised in the future.

It has been argued that the documented adherence to a QA protocol is a key benchmark in the evaluation of the quality, impact, and relevance of a study to the patient-level (Van Horn and Toga, 2009). While the implementation of a QA protocol is a straightforward procedure at the beginning of a study, the final success however largely depends on the dedication of the project teams to consistently apply the requirements of the protocol over the whole study phase. This requires an external control of all procedures, the automatic and fast analysis of all QA data, and the direct publication of QA measures via the World Wide Web. Only if the QA data is openly available and continuously documented across the whole study, one can convincingly argue that QA procedures did not only exist on paper. This will also imply that all QA data are presented, representing a departure from the prevailing practice of simply stating that, e.g., “metrics from each site complied with defined ranges and recommendations”. At present, we are working on an extension of our QA protocol to also publish phantom data openly in the Internet.

Different opinions exist on how extensive and time-consuming a QA protocol has to be. Existing QA protocols described in the literature differ depending on the main neuroscientific or clinical questions, focusing for instance on structural (e.g. Gunter et al. (2009)) or functional MRI data (e.g. Friedman and Glover (2006a,b)). In our case, we extended the standard QA protocol performed on our MR scanners, consisting of a weekly measurement of the ACR-phantom, by introducing a study-specific QA protocol focusing on the temporal stability of the MR scanner, a necessary prerequisite for the assessment of small BOLD signal changes. In the first part of the study, we performed a phantom scan after the measurement of each subject. Since both scanners showed however a stable performance, it might be feasible to only perform 1–2

measurement per week. Since the QA data depends on the time of day when the phantom is measured, these scans should be taken at fixed time points. It is advisable to document the time of day and the room temperature for the measurement of subjects since these variables might have an impact on the data quality.

For the further course of the study, it will be necessary to perform the QA assessments also with personnel that is not specifically employed for the development and implementation of QA procedures. We therefore have to develop automated warning systems giving notifications when the MR scanner is significantly changing its performance characteristics. At present, we are working on the extension of our QA protocol to also incorporate these aspects. We will use data from previous measurements to analyze whether current QA parameters are significantly changing. These changes might be operationalized, for instance, by criteria such as exhibiting values outside predefined confidence limits.

Implications of differences between MR scanners for multi-center studies

The present study was originally planned as a single-center study in Marburg. To increase the recruitment of subjects, a second center, Münster, was included, after one of the PIs (U.D.) was appointed in Münster. Although stimulus equipment and pulse sequences were standardized across both sites to the extent permitted by each platform, we expected systematic differences in image characteristics between the two MR scanners, last but not least because the MR hardware was different. Both centers used MR scanners from Siemens, the MR scanner type (Tim Trio vs. Prisma) and the head coils (12 channel vs. 20 channel) however were different. Consequently, the QA values were different across sites. This is in line with a growing body of literature that suggests that MRI scanners not only produced by different manufacturers but also different scanner models built by a single manufacturer, produce significantly different measurements (e.g. Abdulkadir et al. (2011); Bendfeldt et al, (2012); Clarkson et al. (2009); Friedman and Glover (2006a,b); Friedman and Glover (2006a,b); Reig et al. (2009); Saotome et al. (2012); Stonnington et al. (2008); Takao et al. (2012); Yendiki et al. (2010)). Although several of these studies have stated that the between-scanner differences were small compared to differences produced by, for instance, disease or aging (e.g. Abdulkadir et al. (2011); Bendfeldt et al. (2012); Evans (2006); Kruggel et al. (2010); Stonnington et al. (2008)), one has to be aware that effect sizes may depend on the respective scanner equipment and should be considered during data analysis.

We have shown that non-negligible differences exist in the MRI performance between scanners and whenever any hard- or software changes have been applied to the scanners. We have shown that the impact of using different scanners on volumetric data is comparable to the impact of age and sex of the participating subjects. Our recommendation, therefore, is to treat any change in hard- or software as an equivalent to having measured the data at a different site/scanner. A (dummy encoded) categorical variable should be part of any model used in the formal analysis that reflects that data has been measured at different sites but that also reflects changes that have been applied to the respective scanners. In our case, e.g., this variable would have four categories when analyzing human fMRI data: Marburg (old coil), Marburg (new coil), Münster (no prescan normalize), Münster (with prescan normalize). When analyzing human structural data, this variable would have three categories: Marburg (old coil), Marburg (new coil), and Münster.

In conclusion, we described the implementation of a comprehensive QA protocol for the acquisition of MRI data. This QA protocol constitutes the basis for further MRI data analysis steps in the consortium.

Funding

This study was supported by grants from the German Research Foundation (Grant No. FOR 2107, JA, 1890/7-1, DE 1614/3-1, KI 588/14-1, DA 1151/5-1, KO 4291/3-1). CV was partly funded by a research grant from the BIPOLIFE consortium (German Ministry of Research and Education, Grant No. 01EE1404F).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.01.079.

References

Abdulkadir, A., Mortamet, B., Vemuri, P., Jack, C.R., Krueger, G., Klöppel, S., 2011. Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. Neuroimage 58, 785–792. https://doi.org/10.1016/j.neuroimage.2011.06.029.

Ashburner, J., Friston, K.J., 2001. Why voxel-based morphometry should Be used. Neuroimage 14, 1238–1243. https://doi.org/10.1006/nimg.2001.0961.

Bendfeldt, K., Hofstetter, L., Kuster, P., Traud, S., Mueller-Lenke, N., Naegelin, Y., Kappos, L., Gass, A., Nichols, T.E., Barkhof, F., Vrenken, H., Roosendaal, S.D., Geurts, J.J.G., Radue, E.W., Borgwardt, S.J., 2012. Longitudinal gray matter changes in multiple sclerosis-Differential scanner and overall disease-related effects. Hum. Brain Mapp. 33, 1225–1245. https://doi.org/10.1002/hbm.21279.

Clarkson, M.J., Ourselin, S., Nielsen, C., Leung, K.K., Barnes, J., Whitwell, J.L., Gunter, J.L., Hill, D.L.G., Weiner, M.W., Jack, C.R., Fox, N.C., 2009. Comparison of phantom and registration scaling corrections using the ADNI cohort. Neuroimage 47, 1506–1513. https://doi.org/10.1016/j.neuroimage.2009.05.045.

Dietsche, B., Backes, H., Stratmann, M., Konrad, C., Kircher, T., Krug, A., 2014. Altered neural function during episodic memory encoding and retrieval in major depression. Hum. Brain Mapp. 35, 4293–4302. https://doi.org/10.1002/hbm.22475.

Evans, A.C., 2006. The NIH MRI study of normal brain development. Neuroimage 30, 184–202. https://doi.org/10.1016/j.neuroimage.2005.09.068.

Friedman, L., Glover, G.H., 2006a. Report on a multicenter fMRI quality assurance protocol. J. Magn. Reson. Imag. 23, 827–839. https://doi.org/10.1002/jmri.20583.

Friedman, L., Glover, G.H., Krenz, D., Magnotta, V., 2006. Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. Neuroimage 32, 1656–1668. https://doi.org/10.1016/j.neuroimage.2006.03.062.

Friedman, L., Glover, G.H., The FBIRN Consortium, 2006b. Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. Neuroimage 33, 471–481. https://doi.org/10.1016/j.neuroimage.2006.07.012.

Glover, G.H., Mueller, B.A., Turner, J.A., Van Erp, T.G.M.M., Liu, T.T., Greve, D.N., Voyvodic, J.T., Rasmussen, J., Brown, G.G., Keator, D.B., Calhoun, V.D., Lee, H.J., Ford, J.M., Mathalon, D.H., Diaz, M., O'Leary, D.S., Gadde, S., Preda, A., Lim, K.O., Wible, C.G., Stern, H.S., Belger, A., McCarthy, G., Ozyurt, B., Potkin, S.G., 2012. Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. J. Magn. Reson. Imag. 36, 39–54. https://doi.org/10.1002/jmri.23572.

Gunter, J.L., Bernstein, M.A., Borowski, B.J., Ward, C.P., Britson, P.J., Felmlee, J.P., Schuff, N., Weiner, M., Jack, C.R., 2009. Measurement of MRI scanner performance with the ADNI phantom. Med. Phys. 36, 2193–2205. https://doi.org/10.1118/1.3116776.

Hariri, A.R., Tessitore, A., Mattay, V.S., Fera, F., Weinberger, D.R., 2002. The amygdala response to emotional stimuli: a comparison of faces and scenes. Neuroimage 17, 317–323. https://doi.org/10.1006/nimg.2002.1179.

Hellerbach, A., Einhäuser-Treyer, W., 2013. Phantomentwicklung und Einführung einer systematischen Qualitätssicherung bei multizentrischen Magnetresonanztomographie-Untersuchungen. Philipps-Universität Marburg. https://doi.org/10.17192/z2014.0048.

Hellerbach, A., Schuster, V., Jansen, A., Sommer, J., 2013. MRI phantoms - are there alternatives to agar? PLoS One 8. https://doi.org/10.1371/journal.pone.0070343.

Ihalainen, T., Kuusela, L., Turunen, S., Heikkinen, S., Savolainen, S., Sipilä, O., 2015. Data quality in fMRI and simultaneous EEG-fMRI. MAGMA 28, 23–31. https://doi.org/10.1007/s10334-014-0443-6.

Kolb, A., Wehrl, H.F., Hofmann, M., Judenhofer, M.S., Eriksson, L., Ladebeck, R., Lichy, M.P., Byars, L., Michel, C., Schlemmer, H.P., Schmand, M., Claussen, C.D., Sossi, V., Pichler, B.J., 2012. Technical performance evaluation of a human brain PET/MRI system. Eur. Radiol. 22, 1776–1788. https://doi.org/10.1007/s00330-012-2415-4.

Kruggel, F., Turner, J., Muftuler, L.T., 2010. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. Neuroimage 49, 2123–2133. https://doi.org/10.1016/j.neuroimage.2009.11.006.

Meyer-Lindenberg, A., Tost, H., 2012. Neural mechanisms of social risk for psychiatric disorders. Nat. Neurosci. 15, 663–668. https://doi.org/10.1038/nn.3083.

Mri, A.C.R., Program, A., 2005. Phantom test guidance for the ACR MRI accreditation program. Am. Coll. Radiol. 5.

Olsrud, J., Nilsson, A., Mannfolk, P., Waites, A., Ståhlberg, F., 2008. A two-compartment gel phantom for optimization and quality assurance in clinical BOLD fMRI. Magn. Reson. Imaging 26, 279–286. https://doi.org/10.1016/j.mri.2007.06.010.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man. Cybern 9, 62–66. https://doi.org/10.1109/TSMC.1979.4310076.

Reig, S., Sánchez-González, J., Arango, C., Castro, J., González-Pinto, A., Ortuño, F., Crespo-Facorro, B., Bargalló, N., Desco, M., 2009. Assessment of the increase in variability when combining volumetric data from different scanners. Hum. Brain Mapp. 30, 355–368. https://doi.org/10.1002/hbm.20511.

Saotome, K., Ishimori, Y., Isobe, T., Satou, E., Shinoda, K., Ookubo, J., Hirano, Y., Oosuka, S., Matsushita, A., Miyamoto, K., Sankai, Y., 2012. Comparison of diffusion tensor imaging-derived fractional anisotropy in multiple centers for identical human subjects. Nihon Hoshasen Gijutsu Gakkai Zasshi 68, 1242–1249. https://doi.org/10.6009/jjrt.2012_JSRT_68.9.1242.

Simmons, A., Moore, E., Williams, S.C.R., 1999. Quality control for functional magnetic resonance imaging using automated data analysis and Shewhart charting. Magn. Reson. Med. 41, 1274–1278. https://doi.org/10.1002/(SICI)1522-2594(199906)41:6<1274::AID-MRM27>3.0.CO;2–1.

Stöcker, T., Schneider, F., Klein, M., Habel, U., Kellermann, T., Zilles, K., Shah, N.J., 2005. Automated quality assurance routines for fMRI data applied to a multicenter study. Hum. Brain Mapp. 25, 237–246. https://doi.org/10.1002/hbm.20096.

Stonnington, C.M., Tan, G., Klöppel, S., Chu, C., Draganski, B., Jack, C.R., Chen, K., Ashburner, J., Frackowiak, R.S.J., 2008. Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease. Neuroimage 39, 1180–1185. https://doi.org/10.1016/j.neuroimage.2007.09.066.

Suslow, T., Kugel, H., Ohrmann, P., Stuhrmann, A., Grotegerd, D., Redlich, R., Bauer, J., Dannlowski, U., 2013. Neural correlates of affective priming effects based on masked facial emotion: an fMRI study. Psychiatry Res. Neuroimaging. 211, 239–245. https://doi.org/10.1016/j.pscychresns.2012.09.008.

Takao, H., Hayashi, N., Kabasawa, H., Ohtomo, K., 2012. Effect of scanner in longitudinal diffusion tensor imaging studies. Hum. Brain Mapp. 33, 466–477. https://doi.org/10.1002/hbm.21225.

Tost, H., Bilek, E., Meyer-Lindenberg, A., 2012. Brain connectivity in psychiatric imaging genetics. Neuroimage. https://doi.org/10.1016/j.neuroimage.2011.11.007.

Tovar, D.A., Zhan, W., Rajan, S.S., 2015. A rotational cylindrical fMRI phantom for image quality control. PLoS One 10, e0143172. https://doi.org/10.1371/journal.pone.0143172.

Van Horn, J.D., Toga, A.W., 2009. Multisite neuroimaging trials. Curr. Opin. Neurol. 22, 370–378. https://doi.org/10.1097/WCO.0b013e32832d92de.

Yendiki, A., Greve, D.N., Wallace, S., Vangel, M., Bockholt, J., Mueller, B.A., Magnotta, V., Andreasen, N., Manoach, D.S., Gollub, R.L., 2010. Multi-site characterization of an fMRI working memory paradigm: reliability of activation indices. Neuroimage 53, 119–131. https://doi.org/10.1016/j.neuroimage.2010.02.084.