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.