AUS DER KLINIK UND POLIKLINIK FÜR PSYCHIATRIE UND PSYCHOTHERAPIE

Direktor: Univ.-Prof. Dr. T. Kircher

DES FACHBEREICHS MEDIZIN DER PHILIPPS-UNIVERSITÄT MARBURG

Thema der Dissertation:

Development of quality standards for multi-center, longitudinal magnetic resonance imaging studies in clinical neuroscience

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades der Naturwissenschaften

dem Fachbereich Medizin der Philipps-Universität Marburg

vorgelegt von

M. Sc. Christoph Vogelbacher aus Alsfeld

Marburg, 2020

Angenommen vom Fachbereich Medizin der Philipps-Universität Marburg am: 12.02.2020

Gedruckt mit Genehmigung des Fachbereichs

Dekan: Herr Prof. Dr. H. Schäfer

Referent: Herr Prof. Dr. A. Jansen

1. Korreferent: Frau PD Dr. B. Carl

SELECTED PUBLICATIONS

Accepted articles

Vogelbacher, C., Möbius, T. W., Sommer, J., Schuster, V., Dannlowski, U., Kircher, T., ... & Bopp, M. H. (2018). *The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data*. NeuroImage, 172, 450-460.

Impact factor: 5.812

Vogelbacher, C., Bopp, M. H. A., Schuster, V., Herholz, P., Jansen, A., & Sommer, J. 2019. *LAB–QA2GO: A Free, Easy-to-Use Toolbox for the Quality Assessment of Magnetic Resonance Imaging Data.* Frontiers in Neuroscience 13 (July). Frontiers: 688. doi:10.3389/fnins.2019.00688.

Impact factor: 3.648

Poster-presentations

Vogelbacher, C., Sommer, J., Jansen, A. & Bauer, M. H. A. (2016). *Quality Assurance for MACS*. Poster FOR2107 Retreat.

Vogelbacher, C., Bauer, M. H. A., Sommer, J. & Jansen, A. (2016). *Quality Assurance for functional Magnetic Resonance Imaging*. Poster OHBM 2016.

Vogelbacher, C., Bopp, M. H. A., Schuster, V., Sommer, J. & Jansen, A. (2016). *Quality Assurance for functional Magnetic Resonance Imaging for MACS*. Poster FOR2107 Retreat.

Vogelbacher, C., Bopp, M. H. A., Schuster, V., Herholz, P., Jansen, A. & Sommer, J. (2019). *LAB–QA2GO: A free, easy-to-use toolbox for the quality assessment of magnetic resonance imaging data*. Poster OHBM 2019.

TABLE OF CONTENTS

SELECTED PUBLICATIONS III

ACCEPTED ARTICLES III

POSTER-PRESENTATIONS III

TABLE OF CONTENTS IV

LIST OF ABBREVIATIONS V

LIST OF FIGURES VI

ABSTRACT VII

ZUSAMMENFASSUNG IX

1 INTRODUCTION 1

2 RESULTS OF THE STUDIES 9

2.1 STUDY QA-PROTOCOL 9

2.2 QA-TOOL 11

3 DISCUSSION 15

3.1 QUESTION 1: HOW TO SET UP AND USE A STUDY QA-PROTOCOL 15

3.2 QUESTION 2: HOW TO IMPROVE THE DISTRIBUTION OF QA-PROTOCOLS 18

3.3 CONCLUSION AND FUTURE WORK 20

4 REFERENCES 21

PUBLICATIONS 29

APPENDIX 30

MANUSCRIPT 1 30

Supplementary data 41

MANUSCRIPT 2 56

CURRICULUM VITAE 68

VERZEICHNIS DER AKADEMISCHEN LEHRER/-INNEN 72

DANKSAGUNG 73

EHRENWÖRTLICHE ERKLÄRUNG 75

EIGENER ANTEIL AN DIESER ARBEIT 76

LIST OF ABBREVIATIONS

ACR	American College of Radiology
DTI	Diffusion Tensor Imaging
EEG	Electro-encephalography
EPI	Echo Planar Imaging
FA	Fractional Anisotropy
FBIRN	Functional Bioinformatics Research Network
fMRI	functional Magnetic Resonance Imaging
GMV	Gray Matter Volume
MACS	Marburg-Münster Affective Disorders Cohort Study
MEG	Magnetoencephalography
MRI	Magnetic Resonance Imaging
MRT	Magnetresonanztomographie
PET	Positron Emission Tomography
PSC	Percent Signal Change
QS	Qualitätssicherung
QA	Quality Assurance
SNR	Signal-to-Noise Ratio / Signal-zu-Rausch Verhältnis
SOI	Slice of Interest
TE	Time of Echo
TIV	Total Intracranial Volume
TR	Time of Repetition
VBM	Voxel-Based Morphometry
WMV	White Matter Volume

LIST OF FIGURES

Figure 1: The relative spatial and temporal resolutions of common neuroimaging techniques. (Page 2)

Figure 2: Statistical Parametric Mapping (version: 2) one-sample t-test results (random effects) for the working memory contrast in a multicenter study. (Page 4*)*

ABSTRACT

Magnetic resonance imaging (MRI) data is generated by a complex procedure. Many possible sources of error exist which can lead to a worse signal. For example, hidden defective components of a MRI-scanner, changes in the static magnetic field caused by a person simply moving in the MRI scanner room as well as changes in the measurement sequences can negatively affect the signal-to-noise ratio (SNR). A comprehensive, reproducible, quality assurance (QA) procedure is necessary, to ensure reproducible results both from the MRI equipment and the human operator of the equipment. To examine the quality of the MRI data, there are two possibilities. On the one hand, water or gel-filled objects, so-called "phantoms", are regularly measured. Based on this signal, which in the best case should always be stable, the general performance of the MRI scanner can be tested. On the other hand, the actually interesting data, mostly human data, are checked directly for certain signal parameters (e.g., SNR, motion parameters).

This thesis consists of two parts. In the first part a study-specific QA-protocol was developed for a large multicenter MRI-study, FOR2107. The aim of FOR2107 is to investigate the causes and course of affective disorders, unipolar depression and bipolar disorders, taking clinical and neurobiological effects into account. The main aspect of FOR2107 is the MRI-measurement of more than 2000 subjects in a longitudinal design (currently repeated measurements after 2 years, further measurements planned after 5 years). To bring MRI-data and disease history together, MRI-data must provide stable results over the course of the study. Ensuring this stability is dealt with in this part of the work. An extensive QA, based on phantom measurements, human data analysis, protocol compliance testing, etc., was set up. In addition to the development of parameters for the characterization of MRI-data, the used QA-protocols were improved during the study. The differences between sites and the impact of these differences on human data analysis were analyzed. The comprehensive quality assurance for the FOR2107 study showed significant differences in MRI-signal (for human and phantom data) between the centers. Occurring problems could easily be recognized in time and be corrected, and must be included for current and future analyses of human data.

For the second part of this thesis, a QA-protocol (and the freely available associated software "LAB-QA2GO") has been developed and tested, and can be used for individual studies or to control the quality of an MRI-scanner. This routine was developed because at many sites and in many studies, no explicit QA is performed nevertheless suitable, freely available QA-software for MRI-measurements is available. With LAB-QA2GO, it is possible to set up a QA-protocol for an MRI-scanner or a study without much effort and IT knowledge.

Both parts of the thesis deal with the implementation of QA-procedures. High quality data and study results can be achieved only by the usage of appropriate QA-procedures, as presented in this work. Therefore, QA-measures should be implemented at all levels of a project and should be implemented permanently in project and evaluation routines.

ZUSAMMENFASSUNG

Magnetresonanztomographie (MRT)-Daten entstehen durch ein komplexes Verfahren. Es gibt dadurch viele mögliche Fehlerquellen, die zu einem schlechteren Signal führen können. Beispielsweise können defekte Bauteile des MRT-Scanners, Veränderungen des statischen Magnetfeldes (z.B. durch eine sich bewegende Person im MRT-Scannerraum) oder Veränderungen der Messsequenzen das Signal-zu-Rausch Verhältnis (SNR) negativ beeinflussen. Daher ist eine umfassende Qualitätssicherung (QS) nötig. Eine QS sollte sich neben der Qualität der MRT-Daten unter anderem mit dem Einhalten festgelegter Protokolle und der Dokumentation befassen. Um die Qualität der MRT-Daten zu untersuchen, gibt es zwei Möglichkeiten. Zum einen werden regelmäßig Wasser- oder Gel-gefüllte Behältnisse (sogenannte „Phantome") gemessen. Anhand dieses Signals, welches im besten Fall immer stabil ist, kann die generelle Performanz des MRT-Scanners getestet werden. Zum anderen werden die eigentlich interessierenden Daten, meist Humandaten, direkt auf bestimmte Signalparameter (z.B. SNR, Bewegungen) geprüft.

Die vorliegende Arbeit besteht aus zwei Teilen. Im ersten Teil wurde für eine große multizentrische MRT-Studie, FOR2107, ein studienspezifisches QS-Protokoll entwickelt. FOR2107 hat das Ziel, die Ursachen und den Verlauf von affektiven Störungen, unipolaren Depressionen und bipolaren Störungen unter der Berücksichtigung von klinischen und neurobiologischen Effekten zu untersuchen. Kern von FOR2107 ist die MRT-Messung von mehr als 2000 Probanden in einem longitudinalen Design (derzeit Wiederholungsmessung nach zwei Jahren; geplant sind weitere Messungen nach fünf Jahren). Um MRT-Daten und Krankheitsverlauf zusammenzubringen, müssen die MRT-Daten über den Verlauf der Studie stabile Ergebnisse liefern. Die Sicherstellung dieser Stabilität wird in diesem Teil der Arbeit behandelt. Hierzu wurde eine umfangreiche QS aufgesetzt, basierend auf Phantommessungen, Analyse der Humandaten, Prüfung der Einhaltung der Protokolle, usw. Neben der Entwicklung von Parametern für die Charakterisierung der MRT-Daten wurden die QS-Protokolle während der Studie verbessert. Die Unterschiede zwischen den Standorten und die Auswirkung dieser Unterschiede auf die Analyse der Humandaten wurden analysiert. Die umfassende Qualitätssicherung für die FOR2107 Studie zeigte, dass signifikante Unterschiede im

MRT-Signal (für Human- und Phantomdaten) zwischen den beteiligten Zentren bestehen. Auftretende Probleme konnten somit entweder rechtzeitig erkannt und behoben werden oder müssen für aktuelle und zukünftige Auswertungen der Humandaten beachtet werden.

Im zweiten Teil der Arbeit wurde ein QS-Protokoll (und die frei verfügbare zugehörige Software „LAB–QA2GO“) entwickelt und getestet, welches leicht für einzelne Studien oder zur Kontrolle der Qualität eines MRT-Scanners umsetzbar ist. Dies geschah vor dem Hintergrund, dass trotz der Existenz von geeigneter, frei verfügbarer QS-Software für MRT-Messungen an vielen Standorten und in vielen Studien keine explizite QS durchgeführt wird. Durch diese Software ist es möglich, ein QS-Protokoll ohne großen Aufwand und IT-Kenntnisse an einem MRT-Scanner oder in einer Studie aufzusetzen.

Beide Teile der Arbeit beschäftigen sich mit der Durchführung von QS-Maßnahmen. Erst durch den Einsatz von geeigneten QS-Maßnahmen, wie in dieser Arbeit vorgestellt, können qualitativ hochwertige Daten und Studienergebnisse erzielt werden. Daher sollten QS-Maßnahmen auf allen Ebenen eines Projekts durchgeführt werden und permanent in Projekt- und Auswerteroutinen realisiert werden.

1 INTRODUCTION

Since the 1990s, functional Magnet Resonance Imaging (fMRI) has become a common tool to investigate the human brain (Ogawa et al. (1990, 1992)), allowing the exploration of its structure and its functioning with a non-invasive procedure. Based on these features, the Magnet Resonance Imaging (MRI) technology became tremendously important within the field of neuroscience. Besides MRI, other techniques exist to investigate the human brain e.g., Electro-encephalography (EEG), Magnetoencephalography (MEG) or Positron Emission Tomography (PET). Each technique has a specific relation between spatial and temporal resolution which is highlighted in Figure 1. The EEG technology for instance has a high temporal but a relatively poor spatial resolution and measures the brain's electrical activity directly. The MEG acquisition has a good temporal and spatial resolution and uses like EEG the electrical activity of the neurons. PET has a good spatial, but poor temporal resolution, and records the metabolic activity. The fMRI has a fair temporal and spatial resolution and uses the changes in blood flow to detect the functional activity in the brain.

To analyze the function of the brain, fMRI is a prevalent method used in many neuroimaging studies. The localization of a specific brain function (functional segregation) and the investigation of connectivity between brain regions (functional integration) are hereby of major interest. Google scholar[1] lists about 870.000 “fMRI” entries. As stated above, this technique enables the investigation of the functional processes in the brain with a high spatial resolution. The current spatial resolution of an fMRI acquisition on a 3 Tesla MRI-scanner is between 2.0 and 3.2 mm (Thanh Vu et al. 2017; Jahanian et al. 2019). By using a 7 Tesla MRI-scanner in human data, the in-plane resolution results can be improved to sub-millimeter level (Murphy et al. 2019). Abe et al. (2019) improved the resolution up to 100 µm3 at a rat MRI-scanner.

In general, fMRI-studies analyze functional signal changes which are typically just a small fraction (~1-5 %) of the raw MRI-signal intensity (Friedman and Glover 2006). In the 1990s research began with studies which had a small amount of participants

[1] https://scholar.google.de/ (last visited 10/26/2019)

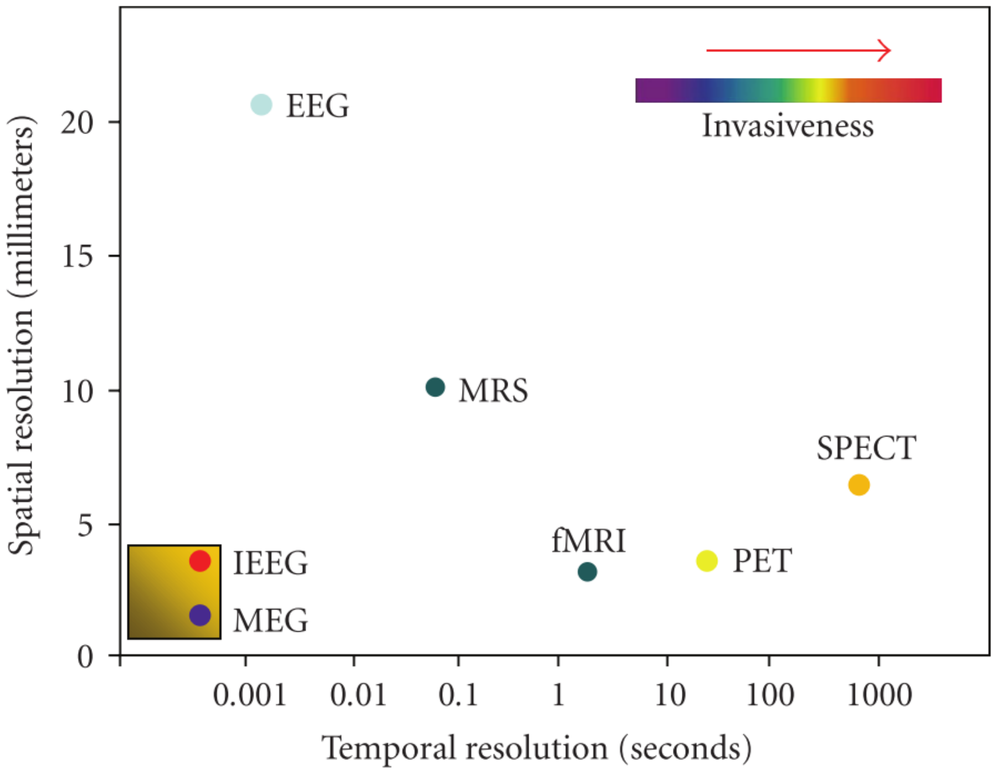


Figure 1: *Graphic showing the relative spatial and temporal resolutions of common neuroimaging techniques (EEG: Electro-encephalography, IEEG: Invasive Electroencephalography, MEG: Magnetoencephalography, MRS: Magnetic Resonance Spectroscopy, fMRI: functional MRI, SPECT: Single Photon Emission Cranial Tomography, and PET: Positron Emission Tomography) (Adapted from Zamrini et al. 2011).*

(n ~ 20), mostly performed at one center. Today, especially in psychiatric research, large cohort studies (n >> 100) are performed at multiple centers. Apart from this basic scientific research, the MRI-technology is increasingly used also in clinical context, e.g., to locate a tumor in the brain (Talos et al. 2010; Metwali et al. 2019; Zhavoronkova et al. 2019).

To obtain results which highlight active brain regions based on the experiment or differences in the brain structure (see Figure 4 in manuscript 1), many different steps have to be performed. First, the data needs to be acquired. Therefore, the right MRI-scanner parameters have to be chosen. If smaller brain regions, like the amygdala, are investigated, the MRI-scanner parameters must be adapted for the measurement (e.g., adaption of the measurement volume)(Morawetz et al. 2008). If a whole brain analysis is performed, a bigger measurement volume must be set to cover the whole brain(Yan 2010; Craddock et al. 2012). This is just one of many MRI-parameters which can be adapted. Others, such as the time of repetition (TR), the time of echo (TE), the voxel size or the matrix size, are important as well to obtain high quality data. Even if the same parameters were used for the measurements, differences in the quality of the data can be present based on the differences of participants who were measured. Because of the large variability of both equipment parameter-settings and of patients, a calibrating, measurable standard is needed to ensure the quality of the MRI-data. This quality standard is not absolute and will change during the course of a study due to (i) different external aspects and (ii) different MRI-image characteristics. An external change (i) could be the result of a MRI-protocol change during the study or due to insufficient equipment which affects the MRI-scanner. These external changes result in a different temporal stability of the MRI-signal or a change of the MRI-image contrast (ii).

The quality of the MRI-data is important for the interpretation of the human MRI-data analysis and the corresponding results. Stöcker et al. (2005) introduced the percent signal change (PSC) value, which describes the signal changes over the time course of a study for human MRI-data. This method qualifies the functional data over the time. Stöcker et al. separated the data of controls (c) and patients (p) into two quality levels (high (+) and low (-)) (Figure 2). Some regions in the low quality control group have a higher intensity than the high quality control group. In the low quality patient group some regions are not present in comparison to the high quality patient group. This highlights the differences in the quality of the data which effects the results of the analysis. To detect the changes between MRI-signal that is associated with the time course of a disease and signal changes caused by alterations in the MRI-scanner environment, a stable MRI-signal is important.

To monitor the stability of the MRI-signal, different quality assurance (QA) mechanisms are needed. These mechanisms are recorded in a QA-protocol. Besides the MRI-signal, other MRI-related (e.g., choice of scan parameters, selection of paradigms) and non-MRI-related factors (e.g., data storage, long-term management of measurement procedures) should be included in the QA-protocol to improve the overall quality and to reduce the inter-site variability of a study. Therefore, a comprehensive QA-protocol is necessary, especially in large, longitudinal, multicenter MRI-neuroimaging studies. Such a protocol also includes careful planning and coordination (Glover et al. 2012).

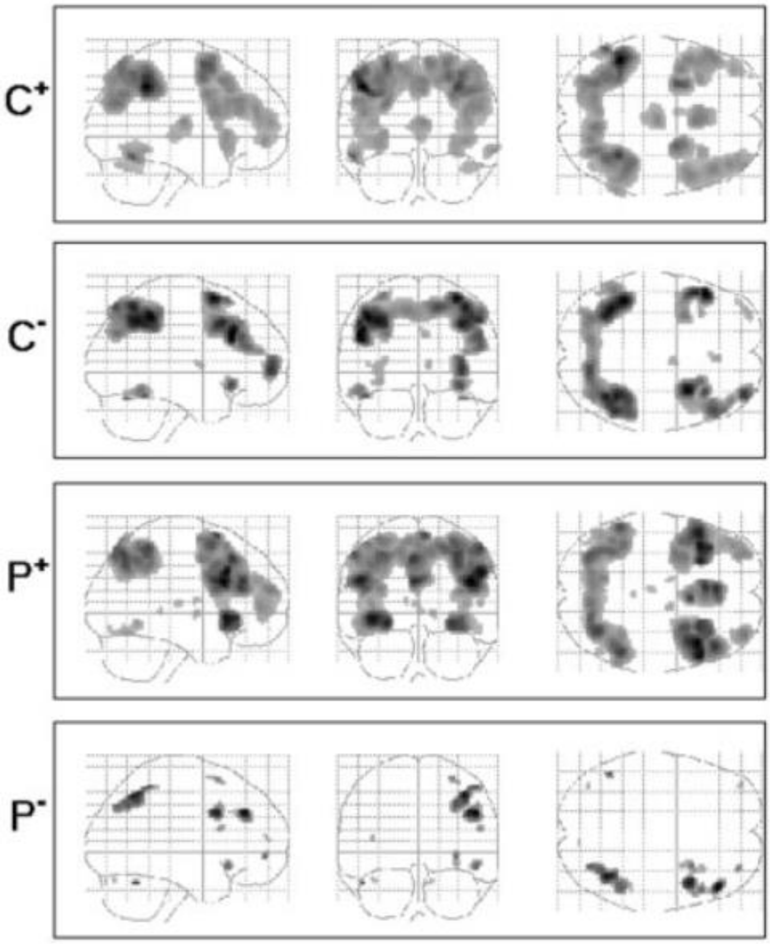


Figure 2: *Statistical Parametric Mapping (version: 2) one-sample t-test results (random effects) for the working memory contrast in the multicenter study. Groups of size n=16 were analyzed. All results are thresholded at p=0.001 (uncorrected). C+ and C- results are similar. Furthermore, the cluster size is larger in the C+ group. The P-group has extremely low data quality, which is reflected by the low activation in the statistical maps. It is the only case that does not show any activation when thresholding at p=0.05 with correction for multiple comparisons* (Stöcker et al. 2005).

The documented adherence to QA-protocols has become a key benchmark to evaluate the quality, impact and relevance of a study (Van Horn and Toga, 2009).

As important as QA-protocol documentation is, most MRI-studies do not describe any QA of their study or their MRI-data (e.g., Paret et al. (2016) or Vignali et al. (2019)). Even if a study performs QA of the MRI-data, the respective description is not detailed, but mostly refer to the Friedman and Glover study (e.g., Krystal et al. (2018)). Friedman and Glover (2006) were the first to present a QA-protocol which uses an gel filled object (so-called "phantom") in a multicenter study to investigate the stability of the MRI-signal over time. They also pointed out that modern MRI-systems show in general overall high technical quality, but image characteristics (e.g., signal-to-noise ratio (SNR)) may change over the time of a study (Friedman and Glover 2006).

Literature describes many QA-protocols now, mostly in the context of large-scale multicenter studies (Van Horn et al. 2009; Glover et al. 2012; Davids et al. 2014). Depending on the neuroscientific question at hand, some QA-protocols focus more on the quality assessment for structural (e.g., Gunter et al. (2009)) or than the functional MRI-data (e.g., Stöcker et al. (2005) or Friedman and Glover (2006)). Moreover, literature describes software tools that detect and remove movement artifacts (e.g., ARTRepair (Mazaika et al. 2009)) or investigate the temporal stability (i.e., stability of the MR-signal over the course of a measurement) of the signal (e.g., MRIQC (Esteban et al. 2017)). In human MRI-datasets, QA-protocols were also developed for more specialized problems e.g., multimodal settings such as the combined acquisition of MRI with EEG (Ihalainen et al. 2015) or PET data (Kolb et al. 2012). Other protocols were developed for a daily phantom QA-routine of MRI-data (Chen et al. n.d.; Peltonen et al. 2017).

To perform a QA-protocol analysis, these software tools need to be installed and set up for a given computer environment. The installation of these routines is often not straight-forward. It typically requires a fair level of technical experience, e.g., to install additional image processing software packages or to handle the dependence of the QA-tools on specific software versions or hardware requirements. Some QA-algorithms require the installation of standard image processing tools (e.g., Artifact Detection Tool[2] or PCP Quality Assessment Protocol (Zarrar et al. 2015)) while others are integrated in different imaging tools (Mindcontrol[3] or BXH/XCEDE (Gadde et al. 2012)). Some QA-workflows can be integrated in commercial programs, e.g., MATLAB[4] (CANlab[5] or ARTRepair), or in large image processing systems (e.g., XNat (Marcus DS, Olsen TR, Ramaratnam M et al. 2007); C-Mind (Lee et al. 2014)). Other QA-workflows

[2] http://web.mit.edu/swg/software.htm (last visited on 10/26/2019)

[3] https://github.com/akeshavan/mindcontrol (last visited on 10/26/2019)

[4] https://www.mathworks.com/ (last visited on 10/26/2019)

[5] https://canlab.github.io/ (last visited on 10/26/2019)

can only be used online, by registering with a user account and uploading data to a server (e.g., LONI (Kim et al. 2019)). Commercial software tools (e.g., BrainVoyager (Goebel 2012)) mostly have their own QA-workflow included. Also some virtualization based QA-pipeline tools exist (e.g., MRIQC (Esteban et al. 2017)).

MRI-phantoms (water or gel-filled objects) are generally used to monitor the stability of the MRI-scanner. MRI-phantoms have the advantage that they are not affected by instrumental drifts from biological variations and pathological changes, whereas human MRI-data has a lot of biological influences (Hellerbach 2013). Common MRI-phantoms are: the American College of Radiology (ACR) phantom (ACR 2005), the Eurospin test objects (Firbank et al. 2000), gel phantoms of the Functional Bioinformatics Research Network (FBIRN)-Consortium (Friedman and Glover 2006) or the Pro-MRI Agar[6] phantom. Other projects are developing new phantoms (Olsrud et al. 2008; Tovar et al. 2015; Hellerbach et al. 2013). Each of these QA-phantoms was designed for specific purposes. The ACR phantom and the Eurospin test objects were designed to test the geometry of the MRI-system, whereas the gel phantoms were developed to control for the temporal stability especially in fMRI studies. In all scenarios, an accurate alignment of the phantom in the MRI-scanner is necessary, by using a phantom holder, in order to reduce the alignment time, reduce the variance of the data, and to improve the sensitivity of the QA-parameters (Vogelbacher et al. 2016).

A specific QA-protocol to monitor the performance of an MRI-scanner would enhance the assessment of the temporal stability of the acquired time series, both within a session and between repeated measurements. To reach that aim, different types of QA-routines could be applied. On the one hand, a study related QA-protocol could be set up to monitor both the study specific MRI-settings and to the study-specific data (management). All (MRI-) parameters can be adapted to the setting used in the study (e.g., the investigation of one specific functional MRI-sequence). Measurements of a MRI-phantom could be performed at a specific time on a measurement day or subsequent to a human measurement. The advantage is a specific to the data, adapted QA-procedure which can easily be transferred to other

[6] http://pro-project.pl/pro-mri_agar (last visited on 10/26/2019)

institutes, if the study is a multicenter study. On the other hand, a QA-routine can be set up tailored to a specific center (the so-called center specific QA). This approach focuses on the monitoring of one MRI-scanner. A documented QA-protocol is used with a specific purpose (e.g., testing the cooling system of the MRI-scanner). It is executed routinely at defined time points with the goal to filter out the external influencing factors which cofounds data collection quality, i.e., controls for the same state of the MRI-scanner.

In conclusion, the quality of MRI-data is important to rephrase MRI-data, which is related to the signal that is associated with the time course of a disease and not related to signal changes caused by alterations in the MRI-scanner environment. Therefore QA-protocols are used which not only analyze (phantom) MRI-data, but extend over all parts that are related to data acquisition. Considering only phantom measurements there are many phantoms and many routines described in the literature. The main idea of these MRI-measurements is the inspection of stability of different aspects (defined in each QA-protocol).

In this work, two different questions concerning QA-procedures were investigated. First, the question how a QA-procedure must set up in a longitudinal multicenter study and second how the distribution and the usage of QA-tools can be improved. For the first question a gel phantom was used to monitor the temporal stability of the MRI-scanners in the Marburg-Münster Affective Disorders Cohort Study (http://for2107.de/, MACS). MACS is a two-center research consortium studying the neurobiological foundations of affective disorders for a large amount of participants (n>2500). To improve the implemented QA-protocol, a phantom holder was used (Vogelbacher et al. 2016). By inspecting the phantom data, differences between the centers were detected. An evaluation of the different MRI-sequences for data acquisition in humans was necessary with regard to differences found in the phantom data. Respective difference could be observed, so that this has to be considered in the human data analysis. These outcomes point to the importance of QA-protocols and QA-analysis for a (MRI) study. To investigate the partly distributed usage of QA-protocols a survey was performed, which revealed that QA is often too complex for the users. For the second question the LAB–QA2GO toolbox was implemented to minimize the inhibitions of setting up a QA-routine and to improve the distribution of QA-protocols. This toolbox provides QA-scripts for the ACR and gel phantom and can also calculate QA-parameters for structural and functional human MRI-datasets. The installation requires minimal effort and the tool is simple to use. All results are presented in a user-friendly web interface.

2 RESULTS OF THE STUDIES

In the following section, the first part describes the implementation and improvement of a QA-protocol in a longitudinal multicenter study. The first part's question results led to this thesis's second question of how to improve the distribution and the usage of QA-tools for MRI-scanners. The existing software tool, LAB–QA2GO, was developed to improve the usage of QA tools and will be described in the last section.

2.1 Study QA-protocol

For the first question how to set up and improve a MRI specific QA-protocol a comprehensive study QA for the MACS study was installed. The basic idea of this QA-protocol was to guarantee the stability of a MRI-signal across the duration of this ongoing large cohort study. About 2500 subjects will be recruited in total and two MRI-measurements for each subject will be performed. For each subject different MRI weighted measurements (e.g., a T1 weighted measurement to measure the structure of the brain) will be acquired. To investigate the neuronal activity in order to the fMRI-measurements are important, but functional signal changes are typically just a small fraction (~1-5 %) of the raw signal intensity (Friedman and Glover 2006)**.** Therefore, a stable temporal MRI-signal is important to make sure that the first and last measurements of the study are comparable to each other. To test the temporal stability of the MRI-signal, a gel phantom measurement was performed after each human measurement using a standard study fMRI protocol. Based on a prior work (Vogelbacher et al. 2016), a phantom holder was introduced during the study to align the phantom into the scanner. The structural MRI, fMRI, and diffusion tensor imaging (DTI) data of 444 healthy control subjects was also investigated with regard to the extent of between-site differences.

During the study, each center had a major incident. The Marburg site had to replace the defective gradient coil of the MRI-scanner, and at the Münster site, the MRI-protocol was changed by activating the pre-scan normalize[7] filter.

To analyze the phantom data, a set of different QA-metrics of different QA-protocols was compiled (Friedman and Glover 2006; Stöcker et al. 2005; Simmons et al. 1999; ACR 2005). These QA-metrics can be segmented into spatial (e.g., SNR or ghosting) and temporal (e.g., PSC or percent fluctuation) characteristics and statistics. For each human dataset, a related analysis method was used. For structural MRI-data the volumetric information were investigated by using the CAT12 toolbox[8] to calculate the total intracranial volume (TIV), total gray matter volume (GMV), and total white matter volume (WMV). To detect the regions where significant volume differences were caused by spatially localized differences between MRI-images of both scanners, a voxel-based morphometry approach (VBM, (Ashburner et al. 2001)) was performed. For the functional MRI-data the PSC value (Stöcker et al. 2005) was calculated for each subject and for the DTI data the fractional anisotropy (FA) information was assessed.

At both sites the phantom measurements were performed (1009 in Marburg, 205 in Münster). In Marburg, 369 measurements were performed without phantom holder and 640 with holder. Of the 640 phantom measurements performed with the phantom

[7] MRI-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 "pre-scan 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 radio frequency 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 pre-scan head coil image is divided by the pre-scan body coil image, the resulting image is essentially an image of the receive field of the head receiving coil. This image can then be used to normalize a target image, thereby removing the receive field heterogeneity.

[8] www.neuro.uni-jena.de/cat (last visited on 10/26/2019)

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.

The analysis of the phantom data showed 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) as well as changes in certain sequence parameters (such as adding the pre-scan normalization option) impacted many of the QA-statistics in a variety of ways. Based on the 212 phantom measurements which have been acquired in Marburg using the phantom holder and after the coil change, the dependence of QA-statistics on the external variables temperature, time of day, and helium level was also investigated. 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 have an effect on some QA-metrics, as well as measurements acquired above 20.8 °C room temperature.

The T1-weighted structural images analysis showed that TIV, GMV, and WMV volumes significantly differ between the MRI-scanners, showing large effect sizes. The 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, a difference of the FA between sites in almost all regions was observed. The PSC values of the fMRI data showed a significant difference between the sites as well.

In conclusion, a comprehensive QA-protocol is important to monitor a study and to detect changes in the study or a protocol. It is essential to account not only for inter-site differences but also for hardware and software changes of the MRI-scanner setting during a MRI-study. Any changes in the MRI-setting should be noted and considered for the analysis. There is also a strong dependency between the reliable placement of the phantom and the resulting QA-statistics. Therefore the usage of a phantom holder to reduce the variance of the QA-statistics and to detect potential malfunctions of the scanner is recommended.

2.2 QA-tool

Based on the findings of the first question, a comprehensive QA-protocol should be used for every study (in the neuroscience field). Even if no study related QA-protocol is used, a MRI specific QA-protocol should be applied. To underline the importance of the distribution and especially the usage of QA-protocols, a survey (in 2009) was performed in 240 university hospitals and research institutes in Germany, Austria and Switzerland (data unpublished) to investigate which kind of QA-protocols they routinely applied. The results show that some centers have a comprehensive QA-protocol established but that in practice most researchers in the cognitive and clinical neurosciences have only a vague idea to what extent QA-protocols are implemented in their studies and how to deal with potential temporal instabilities of the MRI-system (Hellerbach 2013). However, there already exist a fair amount of QA-protocols to monitor the MRI-scanner stability which could be flexibly adapted to the given QA-protocol and data by researchers (for an overview see e.g., Glover et al. (2012)). These routines are mostly publically available and need a fair level of technical experience regarding the installation. Many of these tools need additional preprocessing software (with a specific software version). This circumstance is a challenge for unexperienced researchers that deter them for performing QA.

To help all researchers getting started to perform QA on MRI-systems, an easy-to-use QA-tool (called LAB–QA2GO) was developed to minimize the inhibitions and to improve the distribution of QA-protocols. The tool was developed for users without a strong technical background or for MRI-laboratories without support of large core-facilities. Based on the virtualization approach of personal computer hardware the integration on most computer systems is given and does not require particular hardware specifications. LAB–QA2GO is available in a virtual machine with NeuroDebian (Halchenko et al. 2012) as operating system. NeuroDebian provides a large collection of neuroscience software packages and is widely used in the neuroscience community. All necessary software tools (all open source software to avoid license fees) for the analysis are installed so that the working environment is preconfigured. LAB–QA2GO provides a fully automated QA-pipeline on data of ACR phantoms and gel phantoms. These phantoms are commonly used and cover geometric and temporal stability QA-test to characterize and monitor the MRI-scanner. In addition, the movement parameters of human fMRI and the noise level of structural human MRI-data are calculated as easily interpretable QA-parameters. It is easily possible to modify these pipelines and to extend the QA-analyses by adding self-designed routines (based on a modular implementation of the source code). The results of the analysis are presented in an easy readable and easy-to-interpret web based format. The tool and these results can be accessed via a web browser so that a very user friendly usage without any specific IT knowledge is guaranteed. This also reduces the maintenance work of the tool to a minimum. The tool and all QA-scripts are available for download on GitHub[9]. Based on the virtualization approach the LAB–QA2GO can be set up in about 10 minutes and can easily be integrated into the given computer environment. The adaption to own data is necessary and can be performed in a few configuration steps.

Another aspect of a QA-protocol is the documentation of processes. Therefore most centers have their own procedure, data structure or documents. To centralize all these procedures and documents a software solution could be used to make this information easily accessible for everybody. One solution for this problem is MediaWiki[10] which is a web based system to store documents and helps organizing processes. This software is also integrated in LAB–QA2GO to give the users the possibility to document their procedures and their QA-protocol.

To give the users not only the analysis methods, an application scenario to perform MRI-scanner QA is as well given as an example QA-protocol. This routine includes runs with the ACR phantom and the gel phantom. All measurements were performed as the first measurement of the day. The ACR phantom was measured twice and the gel phantom once a week. The fix QA-protocol for the ACR phantom was used to perform the measurements. For the gel phantom a new QA MRI-acquisition protocol was installed. It consists of a localizer, a structural T1-weighted sequence, a T2*-weighted echo planar imaging (EPI) sequence, a DTI sequence, another fast T2*-weighted EPI sequence and, finally, the same T2*-weighted EPI sequence as at the beginning. This protocol is used to test the cooling system of the scanner. The QA-metrics of the first and the last EPI sequence are used to assess the impact of a highly stressed MRI-scanner on the imaging data.

[9] https://github.com/vogelbac/LAB-QA2GO (last visited on 10/26/2019)

[10] www.mediawiki.org (last visited on 10/26/2019)

In conclusion, to improve the usage and the distribution of QA-protocols for especially MRI-scanner a fully automated QA-pipeline for phantom and human MRI-datasets was developed. This tool helps unexperienced users who have no QA-routine implemented but want to assess the quality of MRI-data or to characterize the long-term performance of a MRI-scanner. By using the QA-metrics of the LAB–QA2GO tool, it is possible to detect outliers which could be an indication of insufficient data quality or a MRI-scanner malfunction. Based on the virtualization technique the LAB–QA2GO tool can easily be integrated into almost every computer environment and needs minimal maintenance costs. This tool can be used to realize either a study specific or centers specific QA-protocol. The adaption to locally used phantoms and MRI-settings can easily be realized by the usage of the user friendly web interface.

3 Discussion

This work illustrates the importance of comprehensive QA-protocols in high-quality fMRI studies, which are affected by control for MRI-scanner instabilities and protocol changes. MRI-scanner malfunctions are often detected after the study is finished (Friedman and Glover 2006), so a prompt QA-analysis based on a comprehensive QA protocol should be performed to detect these malfunctions or deviation of the QA-protocol in time.

3.1 Question 1: How to set up and use a study QA-protocol

A comprehensive QA-protocol was implemented 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. Only the current QA-statistics published in the literature were included and implemented for this analysis. Any changes in the MRI-scanner setting (equipment or protocol) would have had a major negative impact on the QA-statistics. Each QA-statistic has limited information to identify malfunctions of the MRI-scanner. To detect abnormal behavior of these QA-statistics, which lead to a possible malfunction, the QA-values must be compared continuously to those values which represent a respective setting of the MRI-scanner.

The general idea of QA-protocols is the monitoring of QA-statistics to identify possible malfunctions to aid researchers in excluding those measurements (Glover et al. 2012). Defined ranges of the QA-statistics are used to identify the outliers. Based on the results of the published QA-phantom data, a definition of normal ranges of each QA-statistics was not possible for either the whole study or especially for just one MRI-scanner. The reason is that changes in the hardware or software of the MRI-scanner may have affected the QA-statistics and consequently the ranges of the QA-statistics. This non-reproducibility of published QA-phantom data demonstrates the need and importance of QA protocols (not only in MRI-studies).

In general MRI-experiments have to deal with different types of variances (biological, technical and variances during the placement of the measurement volume). QA-protocols aim to monitor the technical variance of an experimental setting (e.g., defective MRI-coils) independent of handling differences (Glover et al. 2012). If a phantom is used, the biological variance is reduced close to zero, based on the apathy to vibrations of the MRI-scanner and the resemblance to human tissue incident to the resulting stable MRI-signal (Hellerbach 2013). The handling variance can be minimized by using a phantom holder (Vogelbacher et al. 2016), which shows a strong impact on almost all QA-statistics even though these QA-statistics are not able to monitor the technical variance independent of handling. It is also mentionable that some QA-statistics seems to monitor the handling differences more than technical variables of the MRI-scanner. This could be because the used gel phantom consists of homogenous material and the placement of the calculation slice (slice of interest (SOI)) is based on the placement of the phantom in the scanner. An inconsistent alignment of the phantom increases the variability of the QA-statistics. This leads to the fact that there is a strong dependence between the QA-statistics and the placement of the phantom. If the MRI-equipment is faulty and the effect in the resulting QA-statistics is smaller than the variability of the reference values, the malfunction remains undetected. As a workaround, to reduce the variability of the QA-statistics, the usage of a phantom holder in combination with a fix MRI-protocol is recommended and was used for this study. The advantage is not only the reduced alignment time of the phantom in the MRI-scanner, it also ensures the measurement of the same volume of the phantom over various phantom measurements (Vogelbacher et al. 2016). The decreased variability of the QA-statistics is the result of the used phantom holder and delivers reasonable values so that the easy detection of outliers is possible. Some of the detected outliers (or possible malfunctions) were caused by minor misplacements of the phantom in the scanner (handling variance) instead of technical instabilities. This does, of course, not mean that the phantom holder improves the quality of the MRI-scanner. In addition to the alignment problem, some QA-statistics seem to be sensitive to the time of day they have been acquired. This might be caused by heating up of the MRI-scanner due to the high amount of measurements over the day. An equal distribution of the measurements with regard to acquisition time or temperature is advisable. A revision of the QA-statistics should be performed in the future to detect the instabilities in the technical variances.

As an example to accentuate the need for a statistical review and an adjustment of the QA-statistics was, the MRI-manufacturer detected after about one and a half years after the start of the study a defective gradient coil in the MRI-scanner, and it was replaced at the Marburg site. An investigation of the QA-statistics before and after the replacement showed that some QA-characteristics showed significant differences in performance. Interestingly, the defect coil was not detectable in the QA-statistics, but was accidentally discovered during a regular maintenance service. This was surprising because the QA-statistics proved to be sensitive to any change in the MRI-setting. An accurate indication of the time point when the defect gradient occurs is not given. The gradient coil might have been defective since the beginning of the study or could be broken shortly before the maintenance service.

In general, a strict adherence of 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 et al. 2009). The successful execution of the QA-protocol depends on the dedication of the project teams to consistently apply the requirements of the protocol over the whole study phase. To help these teams to produce consistent results, it might be also helpful to implement the possibility of an external control (e.g., by presenting results and current working steps via the World Wide Web).

As a second aspect the differences in the MRI-performance between two sites were analyzed, too. The study was designed first for only one site and was extended to another site in Münster. The stimulus equipment and the MRI-settings were standardized across both sites. The used MRI-hardware differs between the sites (same manufacture but different scanner model), so that the QA-values were different as well. This is not surprising because different studies reported this occurrence before (e.g., Abdulkadir et al. (2011); Bendfeldt et al. (2012); Clarkson et al. (2009); Reig et al. (2009); Saotome et al. (2012); Stonnington et al. (2008); Takao et al. (2012); Yendiki et al. (2010); Friedman and Glover (2006); Friedman et al. (2006)). Other studies report that the differences between the scanners were small in comparison to the differences caused by, for instance, disease or aging (e.g., Evans (2006); Kruggel et al. (2010); Abdulkadir et al. (2011); Bendfeldt et al. (2012); Stonnington et al. (2008)). These differences could have an impact to the effect sizes so that this should be mentioned during data analysis.

The installed QA-protocol highlights the differences in the performance of a MRI-scanner if a hard- or software change has been realized. The impact for instance on volumetric data when using different scanners is comparable to the impact of age (18 vs 70 years old) and sex of the participating subjects. A recommendation for this problem is handling the data of any hard- or software changes as data that is measured at a different scanner. For any human MRI-data analysis, a categorical variable, that represents the different scanners and the changes, should be used.

3.2 Question 2: How to improve the distribution of QA-protocols

For this thesis's second question, the LAB–QA2GO tool was developed to distribute QA-procedures. The tool provides fully automated QA-routines of especially phantom MRI-data, but it can also analyze human data. The current version is able to run analysis of the ACR and gel phantoms. The ACR phantom is a widely used phantom for QA of MRI-data to test spatial properties of the MRI-scanner. The gel phantom is mainly used to assess the temporal stability of the MRI-data. In addition, QA-routines for human datasets were developed. The LAB–QA2GO tool is developed modularly to enable modifications of existing analyzes or to integrate other scripts easily. The tool is a virtual machine and has no specific hardware requirements. The approach of a virtual machine was used to have a closed environment and to preconfigure all needed software, so that the users do not have to install any software to perform the QA-analysis. LAB–QA2GO is ready-to-use in about 10 minutes and only a few configuration steps have to be performed to set it up.

The results of the LAB–QA2GO analysis are presented in tabular and graphical form in a user-friendly and easy-to-interpret web based format. The timeline graphs, presented on the overview result page, help the users to identify the outliers. An acceptance range is highlighted in each graph, as well as a warning sign if a measurement is an outlier. These outliers could indicate a malfunction of the scanner. To access the web interface, no specific IT knowledge is needed. The tool is developed in a way that only minimal maintenance work is needed by the operator.

All analysis scripts are available for download as well, if a user wants to integrate the QA-routine into an already existing environment. This requires a specific degree of technical experience though. Other tools, which are described in the literature to assess MRI-stability e.g., Glover et al. (2012), are also publically available but do not provide the configured environment as the LAB–QA2GO tool do. Most of these tools require pre-installed analysis software (e.g., MATLAB) to run their analysis scripts, so LAB–QA2GO. This installation normally needs a fair level of technical experience. So LAB–QA2GO can be a tailor-made solution for user without a strong technical background. This tool can be used to assess the quality of MRI-data in small neuroimaging studies but can also be used as monitoring tool in multicenter studies to assess the long-term stability of different MRI-scanners. It can give direct feedback to its users and can detect possible outliers or changes in the hard- or software setting. Based on the pre-configuration and the virtualization approach, this tool is easily distributable and easy to use.

A comprehensive QA-protocol should not only assess the quality of MRI-data, it needs to encompass technical issues and needs to optimize management procedures to achieve quality results (Glover et al. 2012). Especially at the beginning the study design is important. To document all these issues, a MediaWiki was integrated into this tool, to help the user realizing documentation for the study.

The current version of the LAB-QA2GO toolbox uses relatively simple QA-statistics. These techniques were developed many years ago, but still provide useful and easily accessible information for modern MRI-scanners. Modern MRI-scanners are equipped with phased array coils, a number of amplifiers and multiplexers. Parallel imaging is also available for many years now, and multiband protocols become more and more common. Small changes in the performance of the MRI-system might therefore not be detected with these parameters. The implemented QA-metrics should not be considered as “ground truth”. As mentioned before, an adjustment of these statistics is recommended. In the literature, more sophisticated QA-metrics are available, especially for the assessment of modern MRI-scanners with multi-channel coils and modern reconstruction methods (Dietrich et al. 2007, 2008; Robson et al. 2008; Goerner et al. 2011; Ogura et al. 2012). Their usage would increase the sensitivity of the QA-metrics with respect to possible hardware malfunctions. The adapted analyses workflows for the multiband protocols could be easily integrated based on the modular implementation of *LAB-QA2GO*. This tool is under further development and will be continuously updated to adapt for modern MRI-systems.

3.3 Conclusion and Future work

This work described two different QA related questions. The first question addressed an installation of a comprehensive study QA-protocol which is able to detect differences within and between scanners and any changes in the hard- and software environment. The second question dealt with the distribution and usage of automated MRI-QA-analysis using the *LAB–QA2GO* toolbox. The used analysis methods focused on monitoring the stability of an MRI-signal, which is a specific part of the wide QA-field. It must be clear that there are many other procedures which have to be controlled, to create a high quality MRI-study (e.g., careful planning (Glover et al. 2012)). Therefore a general QA-management should be included in every study to cover all parts and improve the quality of the whole study.

As mentioned before, the used QA-statistics are sufficiently sensitive to detect changes in the MRI-protocol or the MRI-hardware. These QA-statistics might, however, not be sufficient to characterize all aspects of modern MRI-scanner hardware. But they provide useful and easy accessible information also for today's MRI-scanners. As a general recommendation a revision of these parameters should be performed. Also modern reconstruction methods, which are used for multi-channel MRI-coils, and their QA-statistics should be used in the future (Dietrich et al. 2007, 2008; Robson et al. 2008; Goerner et al. 2011; Ogura et al. 2012).

4 REFERENCES

Abdulkadir, Ahmed, Bénédicte Mortamet, Prashanthi Vemuri, Clifford R. Jack, Gunnar Krueger, and Stefan Klöppel. 2011. “Effects of Hardware Heterogeneity on the Performance of SVM Alzheimer’s Disease Classifier.” *NeuroImage* 58 (3): 785–92. https://doi.org/10.1016/j.neuroimage.2011.06.029.

Abe, Yoshifumi, Tomokazu Tsurugizawa, Denis Le Bihan, and Luisa Ciobanu. 2019. “Spatial Contribution of Hippocampal BOLD Activation in High-Resolution FMRI.” *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-39614-3.

ACR. 2005. “Phantom Test Guidance for the ACR MRI Accreditation Program.” *American College of Radiology*, 5. http://www.acr.org/~/media/ACR/Documents/Accreditation/MRI/LargePhantomGuidance.pdf.

Ashburner, John, and Karl J. Friston. 2001. “Why Voxel-Based Morphometry Should Be Used.” *NeuroImage* 14 (6): 1238–43. https://doi.org/10.1006/nimg.2001.0961.

Bendfeldt, Kerstin, Louis Hofstetter, Pascal Kuster, Stefan Traud, Nicole Mueller-Lenke, Yvonne Naegelin, Ludwig Kappos, et al. 2012. “Longitudinal Gray Matter Changes in Multiple Sclerosis-Differential Scanner and Overall Disease-Related Effects.” *Human Brain Mapping* 33 (5): 1225–45. https://doi.org/10.1002/hbm.21279.

Chen, Chien-Chuan, Yung-Liang Wan, Yau-Yau Wai, and Ho-Ling Liu. n.d. “Quality Assurance of Clinical MRI Scanners Using ACR MRI Phantom: Preliminary Results.” *Journal of Digital Imaging* 17 (4): 279–84. Accessed September 6, 2017. https://doi.org/10.1007/s10278-004-1023-5.

Clarkson, Matthew J., Sébastien Ourselin, Casper Nielsen, Kelvin K. Leung, Josephine Barnes, Jennifer L. Whitwell, Jeffrey L. Gunter, et al. 2009. “Comparison of Phantom and Registration Scaling Corrections Using the ADNI Cohort.” *NeuroImage* 47 (4): 1506–13. https://doi.org/10.1016/j.neuroimage.2009.05.045.

Craddock, R. Cameron, G. Andrew James, Paul E. Holtzheimer, Xiaoping P. Hu, and Helen S. Mayberg. 2012. “A Whole Brain FMRI Atlas Generated via Spatially Constrained Spectral Clustering.” *Human Brain Mapping* 33 (8): 1914–28. https://doi.org/10.1002/hbm.21333.

Davids, Mathias, Frank G. Zöllner, Michaela Ruttorf, Frauke Nees, Herta Flor, Gunter Schumann, and Lothar R. Schad. 2014. "Fully-Automated Quality Assurance in Multi-Center Studies Using MRI Phantom Measurements." *Magnetic Resonance Imaging* 32 (6): 771–80. https://doi.org/10.1016/j.mri.2014.01.017.

Dietrich, Olaf, José G. Raya, Scott B. Reeder, Michael Ingrisch, Maximilian F. Reiser, and Stefan O. Schoenberg. 2008. "Influence of Multichannel Combination, Parallel Imaging and Other Reconstruction Techniques on MRI Noise Characteristics." *Magnetic Resonance Imaging* 26 (6): 754–62. https://doi.org/10.1016/j.mri.2008.02.001.

Dietrich, Olaf, José G. Raya, Scott B. Reeder, Maximilian F. Reiser, and Stefan O. Schoenberg. 2007. "Measurement of Signal-to-Noise Ratios in MR Images: Influence of Multichannel Coils, Parallel Imaging, and Reconstruction Filters." *Journal of Magnetic Resonance Imaging* 26 (2): 375–85. https://doi.org/10.1002/jmri.20969.

Esteban, Oscar, Daniel Birman, Marie Schaer, Oluwasanmi O. Koyejo, Russell A. Poldrack, and Krzysztof J. Gorgolewski. 2017. "MRIQC: Advancing the Automatic Prediction of Image Quality in MRI from Unseen Sites." Edited by Boris C Bernhardt. *PLOS ONE* 12 (9): e0184661. https://doi.org/10.1371/journal.pone.0184661.

Evans, Alan C. 2006. "The NIH MRI Study of Normal Brain Development." *NeuroImage* 30 (1): 184–202. https://doi.org/10.1016/j.neuroimage.2005.09.068.

Firbank, M J, R M Harrison, E D Williams, and A Coulthard. 2000. "Quality Assurance for MRI: Practical Experience 1,2." *Radiology* 73 (August 1999): 376–83. https://doi.org/10.1259/bjr.73.868.10844863.

Friedman, Lee, and Gary H. Glover. 2006. "Report on a Multicenter FMRI Quality Assurance Protocol." *Journal of Magnetic Resonance Imaging* 23 (6): 827–39. https://doi.org/10.1002/jmri.20583.

Friedman, Lee, Gary H. Glover, and The FBIRN Consortium. 2006. "Reducing Interscanner Variability of Activation in a Multicenter FMRI Study: Controlling for Signal-to-Fluctuation-Noise-Ratio (SFNR) Differences." *NeuroImage* 33 (2): 471–81. https://doi.org/10.1016/j.neuroimage.2006.07.012.

Gadde, Syam, Nicole Aucoin, Jeffrey S. Grethe, David B. Keator, Daniel S. Marcus, and Steve Pieper. 2012. "XCEDE: An Extensible Schema for Biomedical Data." *Neuroinformatics* 10 (1): 19–32. https://doi.org/10.1007/s12021-011-9119-9.

Glover, Gary H., Bryon A. Mueller, Jessica A. Turner, Theo G.M. M Van Erp, Thomas T. Liu, Douglas N. Greve, James T. Voyvodic, et al. 2012. "Function Biomedical Informatics Research Network Recommendations for Prospective Multicenter Functional MRI Studies." *Journal of Magnetic Resonance Imaging* 36 (1): 39–54. https://doi.org/10.1002/jmri.23572.

Goebel, Rainer. 2012. "BrainVoyager - Past, Present, Future." *NeuroImage* 62 (2): 748–56. https://doi.org/10.1016/j.neuroimage.2012.01.083.

Goerner, Frank L., and Geoffrey D. Clarke. 2011. "Measuring Signal-to-Noise Ratio in Partially Parallel Imaging MRI." *Medical Physics* 38 (9): 5049–57. https://doi.org/10.1118/1.3618730.

Gunter, Jeffrey L., Matt A. Bernstein, Brett J. Borowski, Chadwick P. Ward, Paula J. Britson, Joel P. Felmlee, Norbert Schuff, Michael Weiner, and Clifford R. Jack. 2009. "Measurement of MRI Scanner Performance with the ADNI Phantom." *Medical Physics* 36 (6): 2193–2205. https://doi.org/10.1118/1.3116776.

Halchenko, Yaroslav O., and Michael Hanke. 2012. "Open Is Not Enough. Let's Take the Next Step: An Integrated, Community-Driven Computing Platform for Neuroscience." *Frontiers in Neuroinformatics* 6. https://doi.org/10.3389/fninf.2012.00022.

Hellerbach, Alexandra. 2013. "Phantomentwicklung Und Einführung Einer Systematischen Qualitätssicherung Bei Multizentrischen Magnetresonanztomographie-Untersuchungen." http://archiv.ub.uni-marburg.de/diss/z2014/0048/pdf/dah.pdf.

Hellerbach, Alexandra, Verena Schuster, Andreas Jansen, and Jens Sommer. 2013. "MRI Phantoms - Are There Alternatives to Agar?" *PLoS ONE* 8 (8). https://doi.org/10.1371/journal.pone.0070343.

Horn, John Darrell Van, and Arthur W Toga. 2009. "Multisite Neuroimaging Trials." *Current Opinion in Neurology* 22 (4): 370–78. https://doi.org/10.1097/WCO.0b013e32832d92de.

Ihalainen, Toni, Linda Kuusela, Sampsa Turunen, Sami Heikkinen, Sauli Savolainen, and Outi Sipilä. 2015. “Data Quality in FMRI and Simultaneous EEG-FMRI.” *Magma (New York, N.Y.)* 28 (1): 23–31. https://doi.org/10.1007/s10334-014-0443-6.

Jahanian, Hesamoddin, Samantha Holdsworth, Thomas Christen, Hua Wu, Kangrong Zhu, Adam B. Kerr, Matthew J. Middione, Robert F. Dougherty, Michael Moseley, and Greg Zaharchuk. 2019. “Advantages of Short Repetition Time Resting-State Functional MRI Enabled by Simultaneous Multi-Slice Imaging.” *Journal of Neuroscience Methods* 311 (October 2018): 122–32. https://doi.org/10.1016/j.jneumeth.2018.09.033.

Kim, Hosung, Andrei Irimia, Samuel M. Hobel, Mher Pogosyan, Haoteng Tang, Petros Petrosyan, Rita Esquivel Castelo Blanco, et al. 2019. “The LONI QC System: A Semi-Automated, Web-Based and Freely-Available Environment for the Comprehensive Quality Control of Neuroimaging Data.” *Frontiers in Neuroinformatics* 13. https://doi.org/10.3389/fninf.2019.00060.

Kolb, Armin, Hans F. Wehrl, Matthias Hofmann, Martin S. Judenhofer, Lars Eriksson, Ralf Ladebeck, Matthias P. Lichy, et al. 2012. “Technical Performance Evaluation of a Human Brain PET/MRI System.” *European Radiology* 22 (8): 1776–88. https://doi.org/10.1007/s00330-012-2415-4.

Kruggel, Frithjof, Jessica Turner, and L. Tugan Muftuler. 2010. “Impact of Scanner Hardware and Imaging Protocol on Image Quality and Compartment Volume Precision in the ADNI Cohort.” *NeuroImage* 49 (3): 2123–33. https://doi.org/10.1016/j.neuroimage.2009.11.006.

Krystal, Andrew D., Diego A. Pizzagalli, Sanjay J. Mathew, Gerard Sanacora, Richard Keefe, Allen Song, Joseph Calabrese, et al. 2018. “The First Implementation of the NIMH FAST-FAIL Approach to Psychiatric Drug Development.” *Nature Reviews Drug Discovery* 18 (1): 82–84. https://doi.org/10.1038/nrd.2018.222.

Lee, Gregory R, Akila Rajagopal, Nicholas Felicelli, Andrew Rupert, and Michael Wagner. 2014. “Cmind-Py: A Robust Set of Processing Pipelines for Pediatric FMRI.” *Proceedings of the 20th Annual Meeting of the Organization for Human Brain Mapping*.

Marcus DS, Olsen TR, Ramaratnam M, Buckner RL, Daniel S Marcus, Timothy R Olsen, Mohana Ramaratnam, and Randy L Buckner. 2007. “The Extensible Neuroimaging Archive Toolkit: An Informatics Platform for Managing, Exploring, and Sharing Neuroimaging Data.” *Neuroinformatics* 5 (1): 11–34. https://doi.org/10.1385/NI.

Mazaika, P.K., F. Hoeft, G.H. Glover, and A.L. Reiss. 2009. “Methods and Software for FMRI Analysis of Clinical Subjects.” *NeuroImage* 47: S58. https://doi.org/10.1016/S1053-8119(09)70238-1.

Metwali, Hussam, and Amir Samii. 2019. “Seed-Based Connectivity Analysis of Resting-State FMRI in Patients with Brain Tumors: A Feasibility Study.” *World Neurosurgery*. https://doi.org/10.1016/j.wneu.2019.04.073.

Morawetz, Carmen, Petra Holz, Claudia Lange, Jürgen Baudewig, Godehard Weniger, Eva Irle, and Peter Dechent. 2008. “Improved Functional Mapping of the Human Amygdala Using a Standard Functional Magnetic Resonance Imaging Sequence with Simple Modifications.” *Magnetic Resonance Imaging* 26 (1): 45–53. https://doi.org/10.1016/j.mri.2007.04.014.

Murphy, Jerry E., Julio A. Yanes, Lauren A.J. Kirby, Meredith A. Reid, and Jennifer L. Robinson. 2019. “Left, Right, or Bilateral Amygdala Activation? How Effects of Smoothing and Motion Correction on Ultra-High Field, High-Resolution Functional Magnetic Resonance Imaging (FMRI) Data Alter Inferences.” *Neuroscience Research*. https://doi.org/10.1016/j.neures.2019.01.009.

Ogawa, S., T. M. Lee, A. R. Kay, and D. W. Tank. 1990. “Brain Magnetic Resonance Imaging with Contrast Dependent on Blood Oxygenation.” *Proceedings of the National Academy of Sciences* 87 (24): 9868–72. https://doi.org/10.1073/pnas.87.24.9868.

Ogawa, S., D. W. Tank, R. Menon, J. M. Ellermann, S. G. Kim, H. Merkle, and K. Ugurbil. 1992. “Intrinsic Signal Changes Accompanying Sensory Stimulation: Functional Brain Mapping with Magnetic Resonance Imaging.” *Proceedings of the National Academy of Sciences* 89 (13): 5951–55. https://doi.org/10.1073/pnas.89.13.5951.

Ogura, Akio, Tosiaki Miyati, Masato Kobayashi, Hiroshi Imai, Kouzou Shimizu, Toshio Tsuchihashi, Tsukasa Doi, and Yoshio Machida. 2012. “Method of SNR Determination Using Clinical Images.” *Japanese Journal of Radiological Technology* 63 (9): 1099–1104. https://doi.org/10.6009/jjrt.63.1099.

Olsrud, Johan, Anders Nilsson, Peter Mannfolk, Anthony Waites, and Freddy Ståhlberg. 2008. “A Two-Compartment Gel Phantom for Optimization and Quality Assurance in Clinical BOLD FMRI.” *Magnetic Resonance Imaging* 26 (2): 279–86. https://doi.org/10.1016/j.mri.2007.06.010.

Paret, Christian, Matthias Ruf, Martin Fungisai Gerchen, Rosemarie Kluetsch, Traute Demirakca, Martin Jungkunz, Katja Bertsch, Christian Schmahl, and Gabriele Ende. 2016. “FMRI Neurofeedback of Amygdala Response to Aversive Stimuli Enhances Prefrontal-Limbic Brain Connectivity.” *NeuroImage* 125: 182–88. https://doi.org/10.1016/j.neuroimage.2015.10.027.

Peltonen, Juha I., Teemu Mäkelä, Alexey Sofiev, and Eero Salli. 2017. “An Automatic Image Processing Workflow for Daily Magnetic Resonance Imaging Quality Assurance.” *Journal of Digital Imaging* 30 (2): 163–71. https://doi.org/10.1007/s10278-016-9919-4.

Reig, Santiago, Javier Sánchez-González, Celso Arango, Josefina Castro, Ana González-Pinto, Felipe Ortuño, Benedicto Crespo-Facorro, Nuria Bargalló, and Manuel Desco. 2009. “Assessment of the Increase in Variability When Combining Volumetric Data from Different Scanners.” *Human Brain Mapping* 30 (2): 355–68. https://doi.org/10.1002/hbm.20511.

Robson, Philip M., Aaron K. Grant, Ananth J. Madhuranthakam, Riccardo Lattanzi, Daniel K. Sodickson, and Charles A. McKenzie. 2008. “Comprehensive Quantification of Signal-to-Noise Ratio and g-Factor for Image-Based and k-Space-Based Parallel Imaging Reconstructions.” *Magnetic Resonance in Medicine* 60 (4): 895–907. https://doi.org/10.1002/mrm.21728.

Saotome, Kousaku, Yoshiyuki Ishimori, Tomonori Isobe, Eisuke Satou, Kazuya Shinoda, Jun Ookubo, Yuuji Hirano, et al. 2012. “[Comparison of Diffusion Tensor Imaging-Derived Fractional Anisotropy in Multiple Centers for Identical Human Subjects].” *Nihon Hoshasen Gijutsu Gakkai Zasshi* 68 (9): 1242–49. https://doi.org/10.6009/jjrt.2012_JSRT_68.9.1242.

Simmons, Andrew, Elizabeth Moore, and Steven C R Williams. 1999. “Quality Control for Functional Magnetic Resonance Imaging Using Automated Data Analysis and Shewhart Charting.” *Magnetic Resonance in Medicine : Official Journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine* 41 (6): 1274–78. https://doi.org/10.1002/(SICI)1522-2594(199906)41:6<1274::AID-MRM27>3.0.CO;2-1.

Stöcker, Tony, Frank Schneider, Martina Klein, Ute Habel, Thilo Kellermann, Karl Zilles, and N. Jon Shah. 2005. “Automated Quality Assurance Routines for FMRI Data Applied to a Multicenter Study.” *Human Brain Mapping* 25: 237–46. https://doi.org/10.1002/hbm.20096.

Stonnington, Cynthia M., Geoffrey Tan, Stefan Klöppel, Carlton Chu, Bogdan Draganski, Clifford R. Jack, Kewei Chen, John Ashburner, and Richard S.J. Frackowiak. 2008. “Interpreting Scan Data Acquired from Multiple Scanners: A Study with Alzheimer’s Disease.” *NeuroImage* 39 (3): 1180–85. https://doi.org/10.1016/j.neuroimage.2007.09.066.

Takao, Hidemasa, Naoto Hayashi, Hiroyuki Kabasawa, and Kuni Ohtomo. 2012. “Effect of Scanner in Longitudinal Diffusion Tensor Imaging Studies.” *Human Brain Mapping* 33 (2): 466–77. https://doi.org/10.1002/hbm.21225.

Talos, Ion-Florin, Lauren O’Donnell, Carl-Fredrick Westin, Simon K. Warfield, William Wells, Seung-Schik Yoo, Lawrence P. Panych, et al. 2010. “Diffusion Tensor and Functional MRI Fusion with Anatomical MRI for Image-Guided Neurosurgery.” In , 407–15. https://doi.org/10.1007/978-3-540-39899-8_51.

Thanh Vu, An, Keith Jamison, Matthew F. Glasser, Stephen M. Smith, Timothy Coalson, Steen Moeller, Edward J. Auerbach, Kamil Uğurbil, and Essa Yacoub. 2017. “Tradeoffs in Pushing the Spatial Resolution of FMRI for the 7T Human Connectome Project.” *NeuroImage* 154 (November 2016): 23–32. https://doi.org/10.1016/j.neuroimage.2016.11.049.

Tovar, David A., Wang Zhan, and Sunder S. Rajan. 2015. “A Rotational Cylindrical FMRI Phantom for Image Quality Control.” *Plos One* 10 (12): e0143172. https://doi.org/10.1371/journal.pone.0143172.

Vignali, Lorenzo, Stefan Hawelka, Florian Hutzler, and Fabio Richlan. 2019. "Processing of Parafoveally Presented Words. An FMRI Study." *NeuroImage* 184: 1–9. https://doi.org/10.1016/j.neuroimage.2018.08.061.

Vogelbacher, Christoph, M.H.A. Bauer, Jens Sommer, and Andreas Jansen. 2016. "Quality Assurance for Functional Magnetic Resonance Imaging." In . https://doi.org/https://doi.org/10.6084/m9.figshare.5817486.v3.

Yan. 2010. "DPARSF: A MATLAB Toolbox for 'Pipeline' Data Analysis of Resting-State FMRI." *Frontiers in System Neuroscience*. https://doi.org/10.3389/fnsys.2010.00013.

Yendiki, Anastasia, Douglas N. Greve, Stuart Wallace, Mark Vangel, Jeremy Bockholt, Bryon A. Mueller, Vince Magnotta, Nancy Andreasen, Dara S. Manoach, and Randy L. Gollub. 2010. "Multi-Site Characterization of an FMRI Working Memory Paradigm: Reliability of Activation Indices." *NeuroImage* 53 (1): 119–31. https://doi.org/10.1016/j.neuroimage.2010.02.084.

Zamrini, Edward, Fernando Maestu, Eero Pekkonen, Michael Funke, Jyrki Makela, Myles Riley, Ricardo Bajo, et al. 2011. "Magnetoencephalography as a Putative Biomarker for Alzheimer's Disease." *International Journal of Alzheimer's Disease*, no. June 2014. https://doi.org/10.4061/2011/280289.

Zarrar, Shehzad, Giavasis Steven, Li Qingyang, Benhajali Yassine, Yan Chaogan, Yang Zhen, Milham Michael, Bellec Pierre, and Craddock Cameron. 2015. "The Preprocessed Connectomes Project Quality Assessment Protocol - a Resource for Measuring the Quality of MRI Data." *Frontiers in Neuroscience* 9. https://doi.org/10.3389/conf.fnins.2015.91.00047.

Zhavoronkova, Ludmila, Sofia Moraresku, Galina Boldyreva, Elena Sharova, Svetlana Kuptsova, Alexander Smirnov, Eugen Masherov, et al. 2019. "FMRI and EEG Reactions to Hand Motor Tasks in Patients with Mild Traumatic Brain Injury: Left-Hemispheric Sensitivity to Trauma." *Journal of Behavioral and Brain Science* 09 (06): 273–87. https://doi.org/10.4236/jbbs.2019.96020.

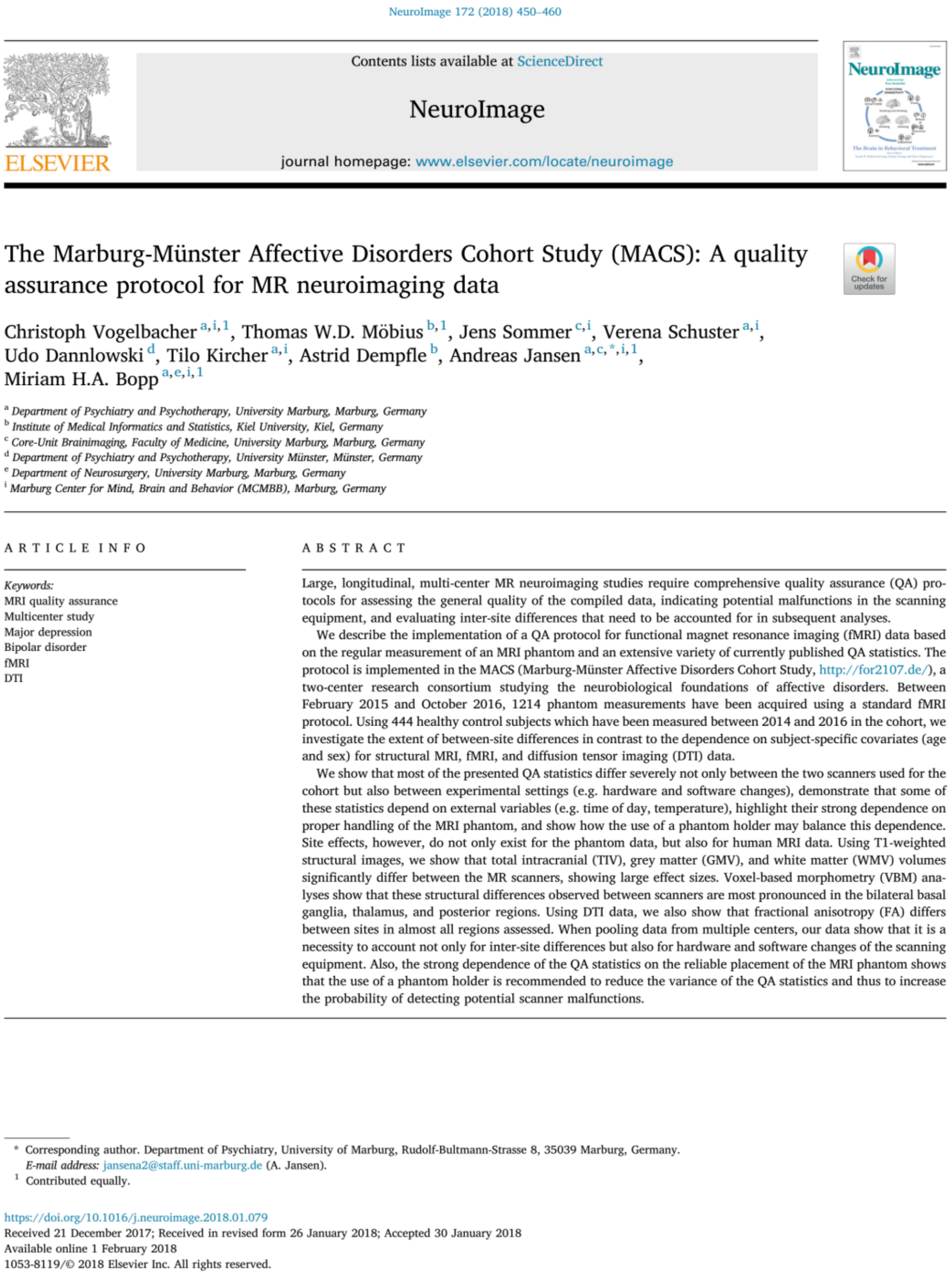
PUBLICATIONS

Vogelbacher, C., Möbius, T. W., Sommer, J., Schuster, V., Dannlowski, U., Kircher, T., ... & Bopp, M. H. (2018). *The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data*. NeuroImage, 172, 450-460.

Vogelbacher, C., Bopp, M. H. A., Schuster, V., Herholz, P., Jansen, A., & Sommer, J.. 2019. *LAB–QA2GO: A Free, Easy-to-Use Toolbox for the Quality Assessment of Magnetic Resonance Imaging Data.* Frontiers in Neuroscience 13 (July). Frontiers: 688. doi:10.3389/fnins.2019.00688.

APPENDIX

Manuscript 1



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

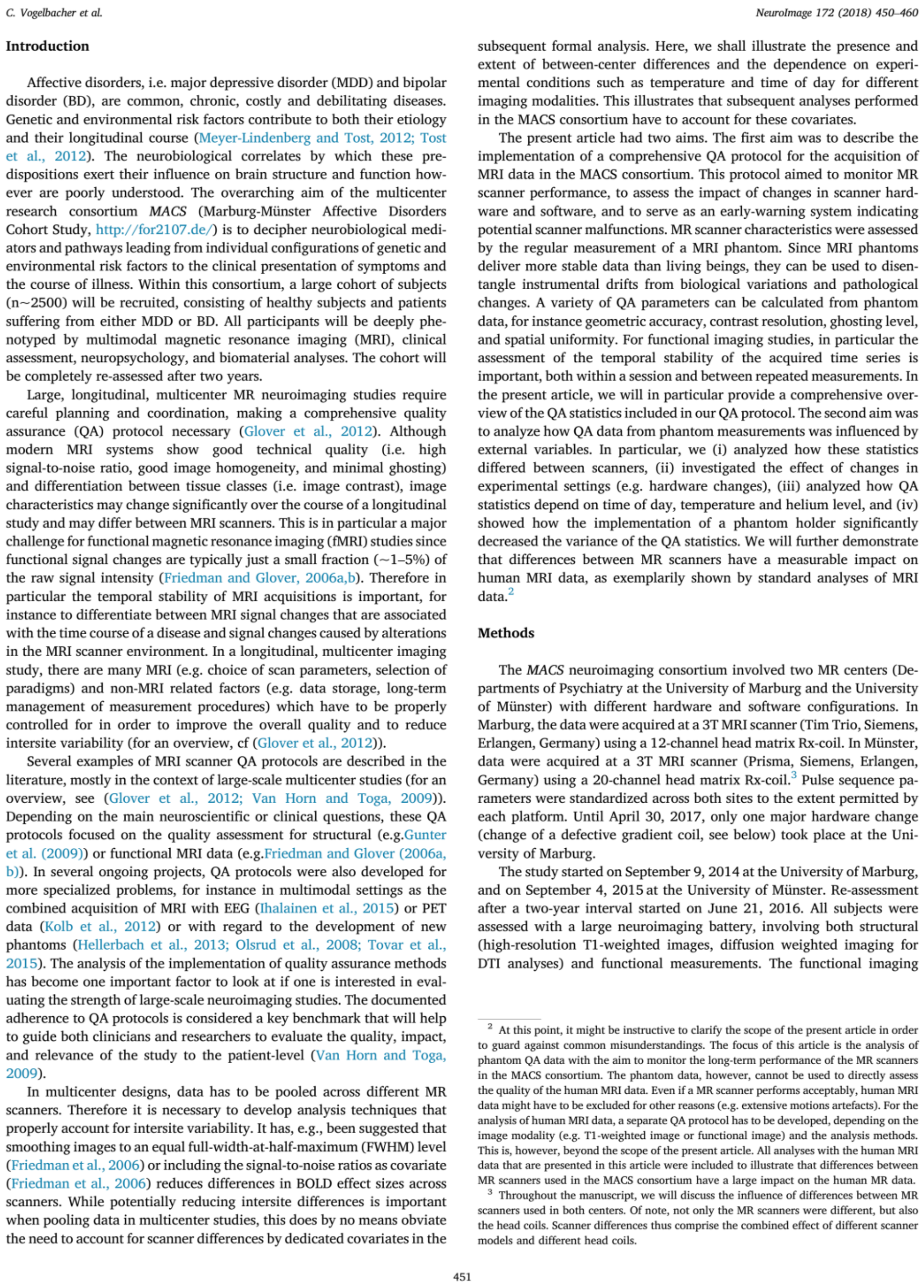
Large, longitudinal, multi-center MR neuroimaging studies require comprehensive quality assurance (QA) protocols for assessing the general quality of the compiled data, indicating potential malfunctions in the scanning equipment, and evaluating inter-site differences that need to be accounted for in 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 exemplarly shown by standard analyses of MRI data.[2]

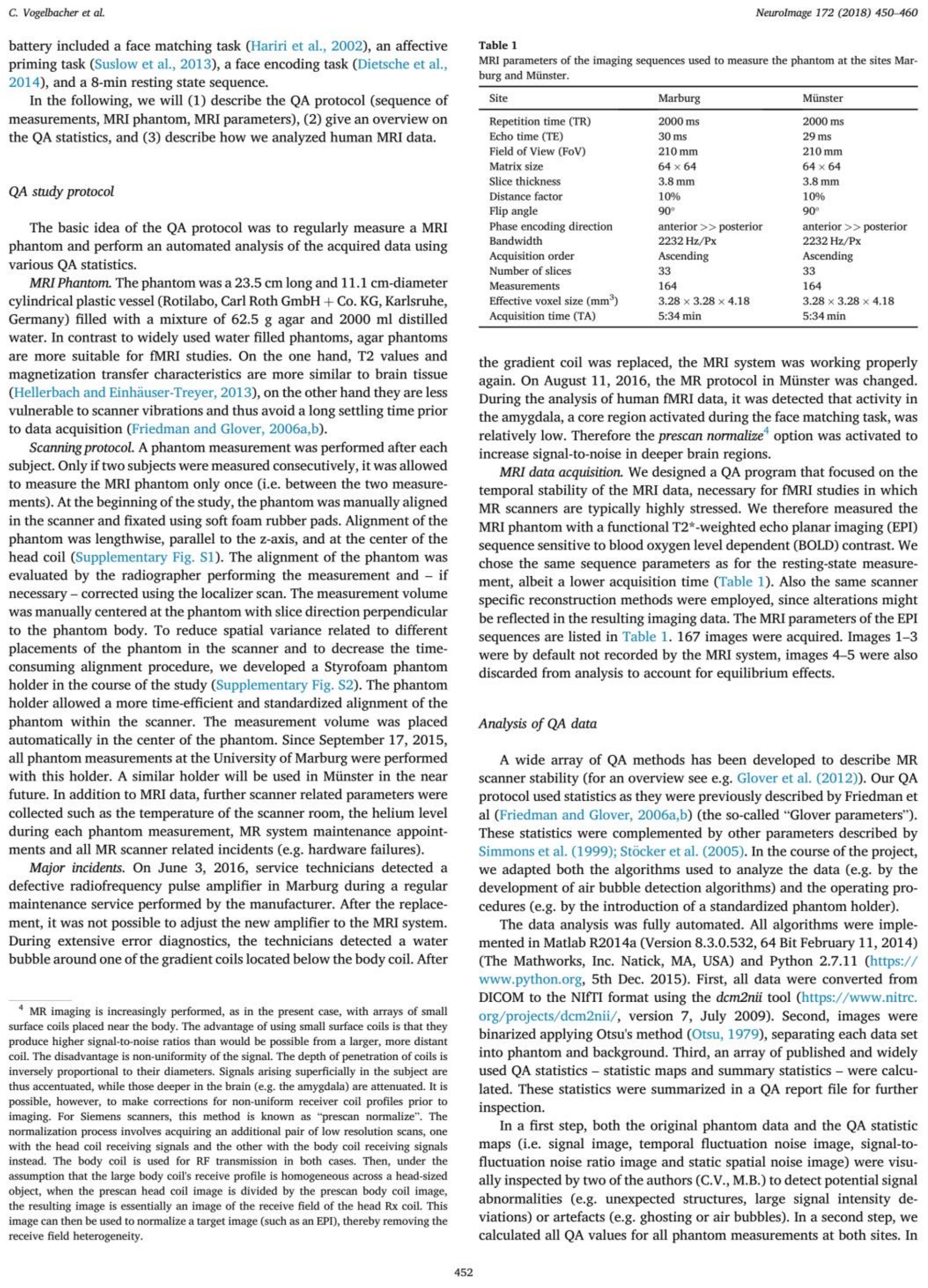
Methods

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.

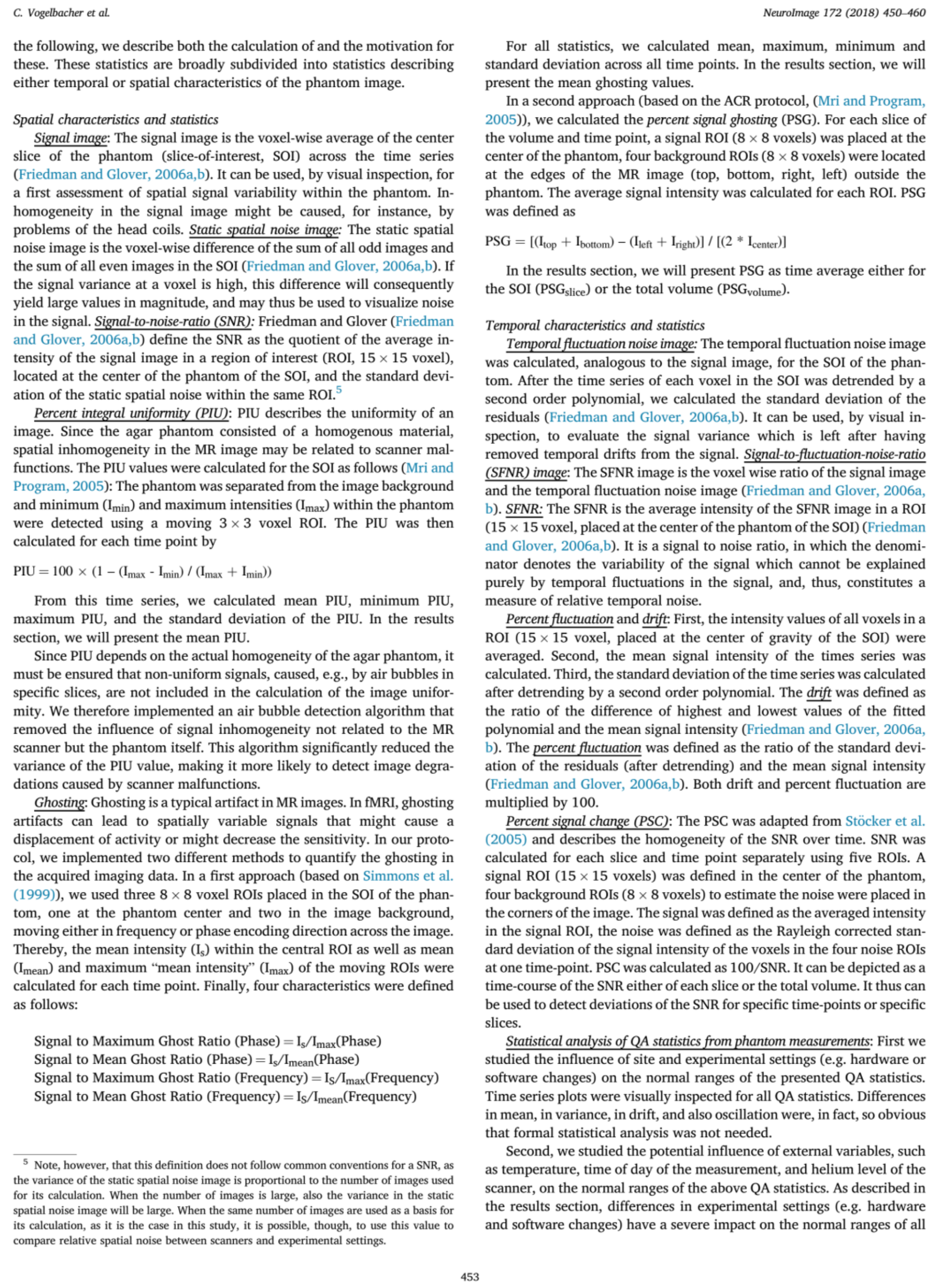
Analysis of QA data

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.

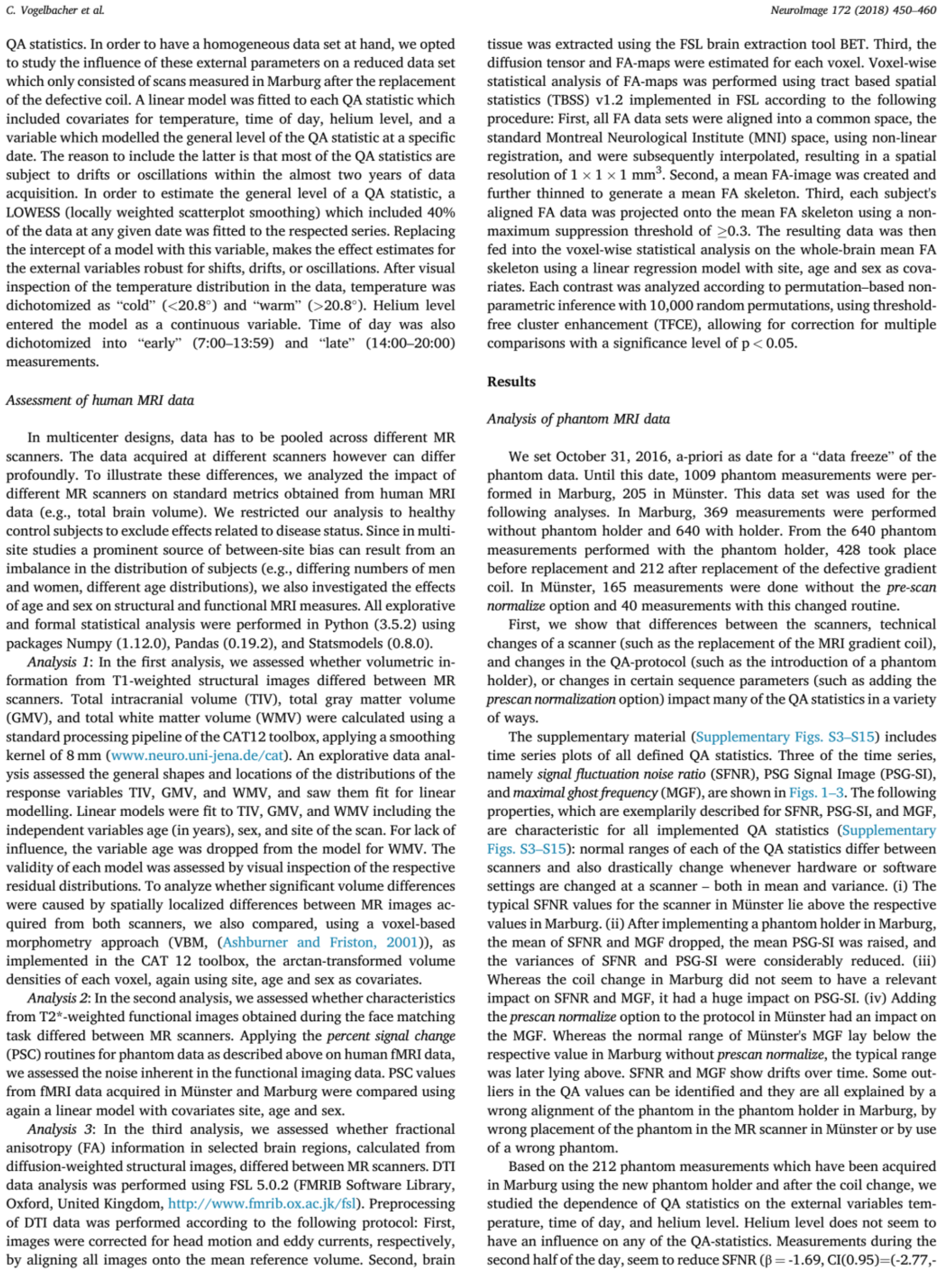
Percent fluctuation and *drift*: First, the intensity values of all voxels in a ROI (15 × 15 voxel, placed at the center of gravity of the SOI) were averaged. Second, the mean signal intensity of the times series was calculated. Third, the standard deviation of the time series was calculated after detrending 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 multisite 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 ≥ 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 nonparametric 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 (β = -1.69, CI(0.95)=(-2.77,-

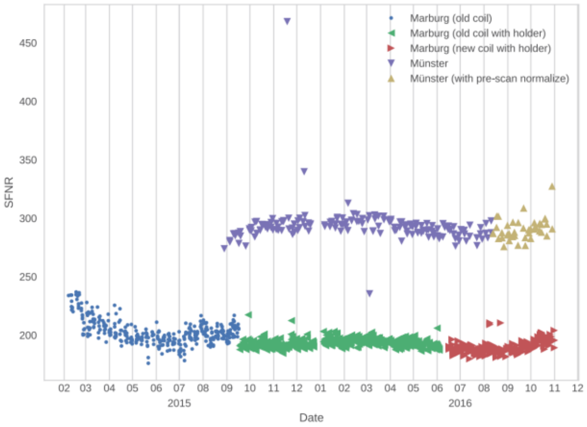


Fig. 1. SFNR values of phantom measurements in Marburg and Münster. One can see, for instance, that (i) typical SFNR values for the scanner in Münster lie above the respective values in Marburg, (ii) the implementation of a phantom holder in Marburg considerably reduced the variance of SFNR, and (iii)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

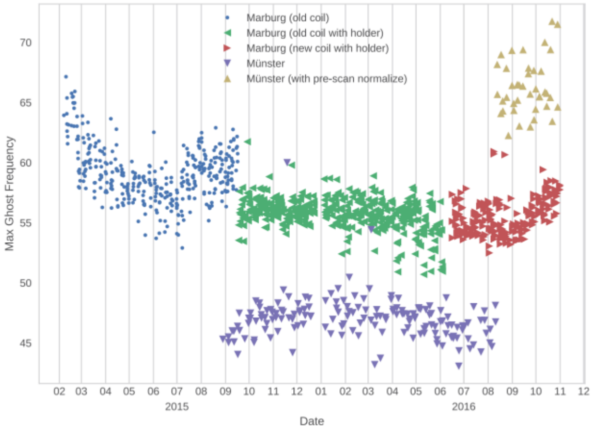


Fig. 2. MGF values of phantom measurements in Marburg and Münster. One can see, for instance, that the implementation of a phantom holder in Marburg reduced both the mean MGF and the variance. In contrast, the coil change in Marburg did not have a relevant impact on MGF. Adding the prescan normalize 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 ($\beta = -2.8$, CI(0.95)=(-0.42, -0.14), $p = 0.0001$), Mean Ghost Frequency ($\beta = -0.63$, CI(0.95)=(-1.02, -0.25), $p = 0.0012$), Mean Ghost Phase ($\beta = -0.42$, CI(0.95)=(-0.68, -0.16), $p = 0.0015$), and to increase PSC ($\beta = 0.01$, CI(0.95)=(0.01, 0.02), $p < 0.0001$). Measurements acquired above 20.8 °C room temperature seem to reduce the SFNR ($\beta = -2.12$, 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.

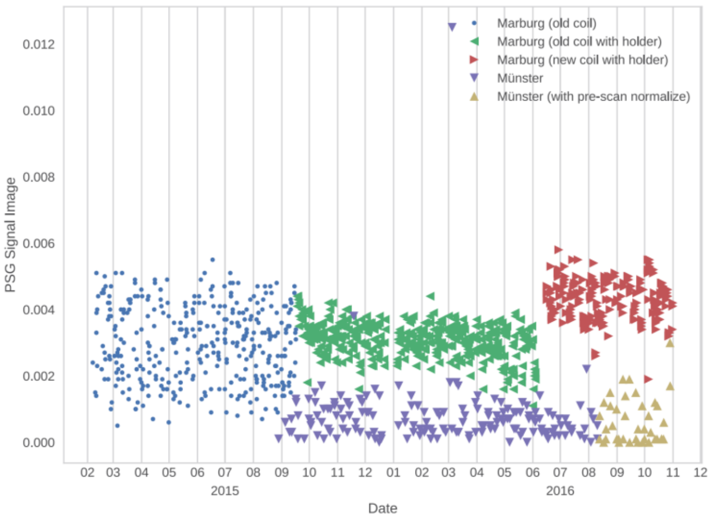


Fig. 3. PSG Signal Image values of phantom measurements in Marburg and Münster. One can see, for instance, that the mean PSG-SI was raised after implementation of a phantom holder, while the variance was considerably reduced. Interestingly, the coil change in Marburg had a huge impact on PSG-SI.

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

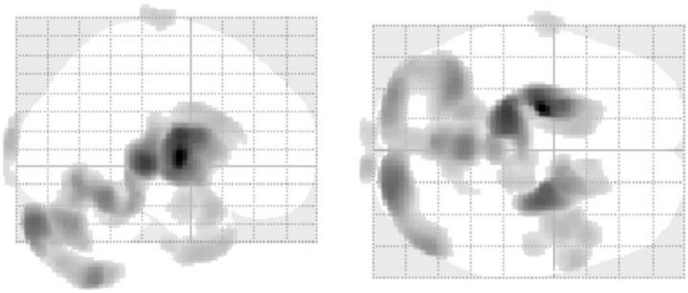


Fig. 4. Using voxel-based morphometry, volumetric differences between MRI data measured in Münster and Marburg were assessed at a local scale (contrast Münster > Marburg, $p < 0.001$ at the voxel level, $p < 0.05$ corrected at the cluster level). Brain volume differences were found in particular in the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

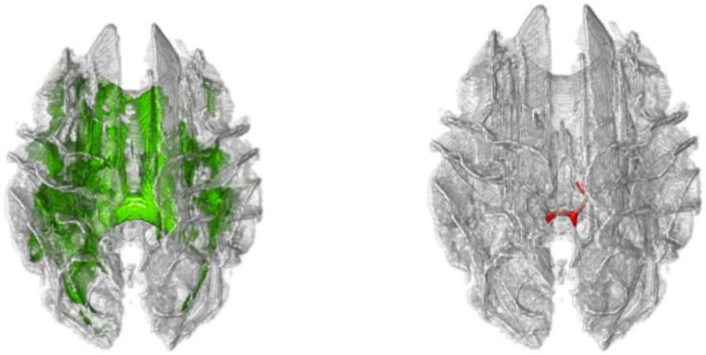
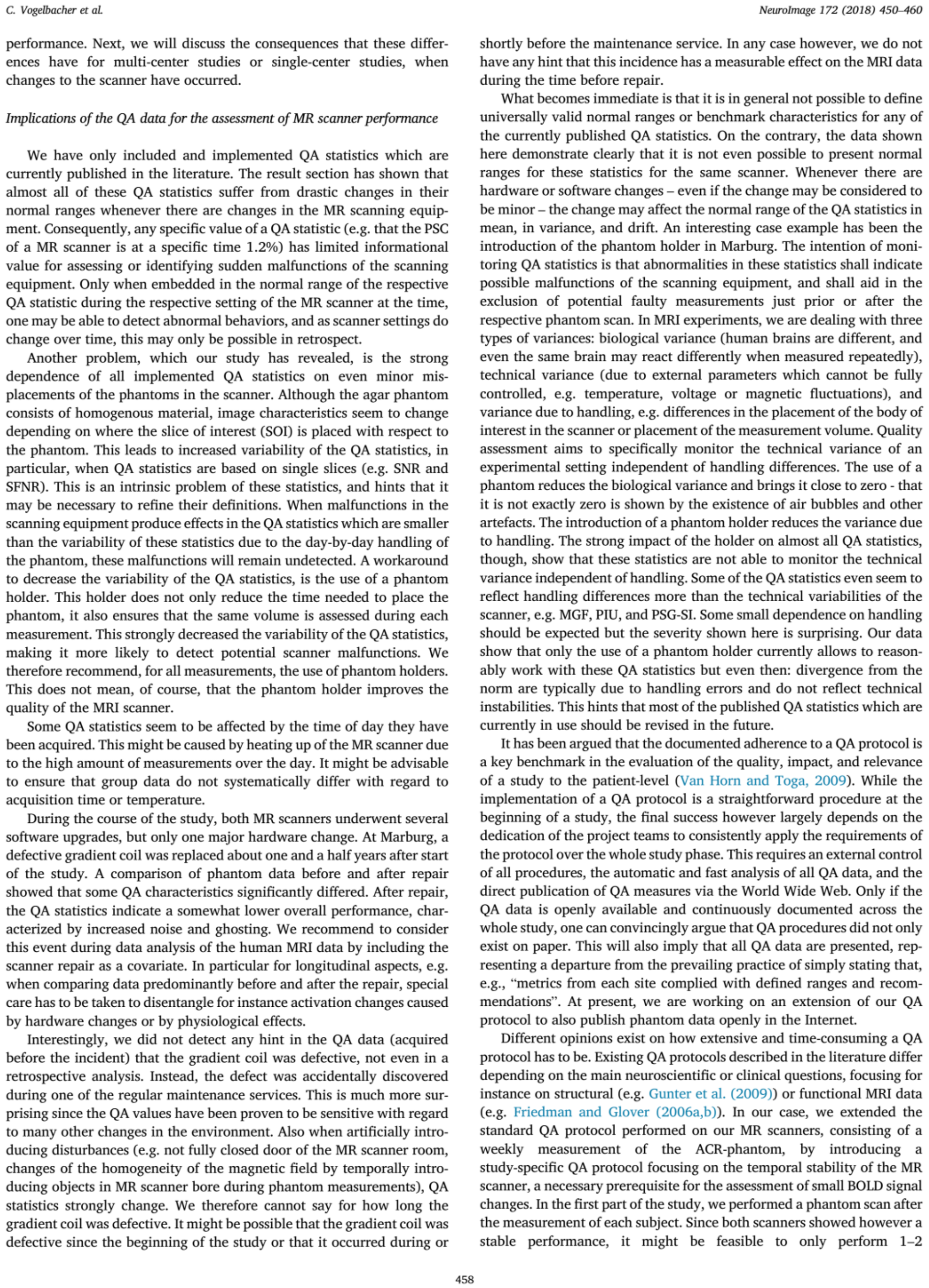


Fig. 5. Differences in FA values between DTI data measured in Münster ($p < 0.05$, corrected for multiple comparisons). For the contrast Marburg > Münster, differences (marked in green) were found in widespread regions throughout the brain (left). For the contrast Münster > Marburg, FA values were higher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

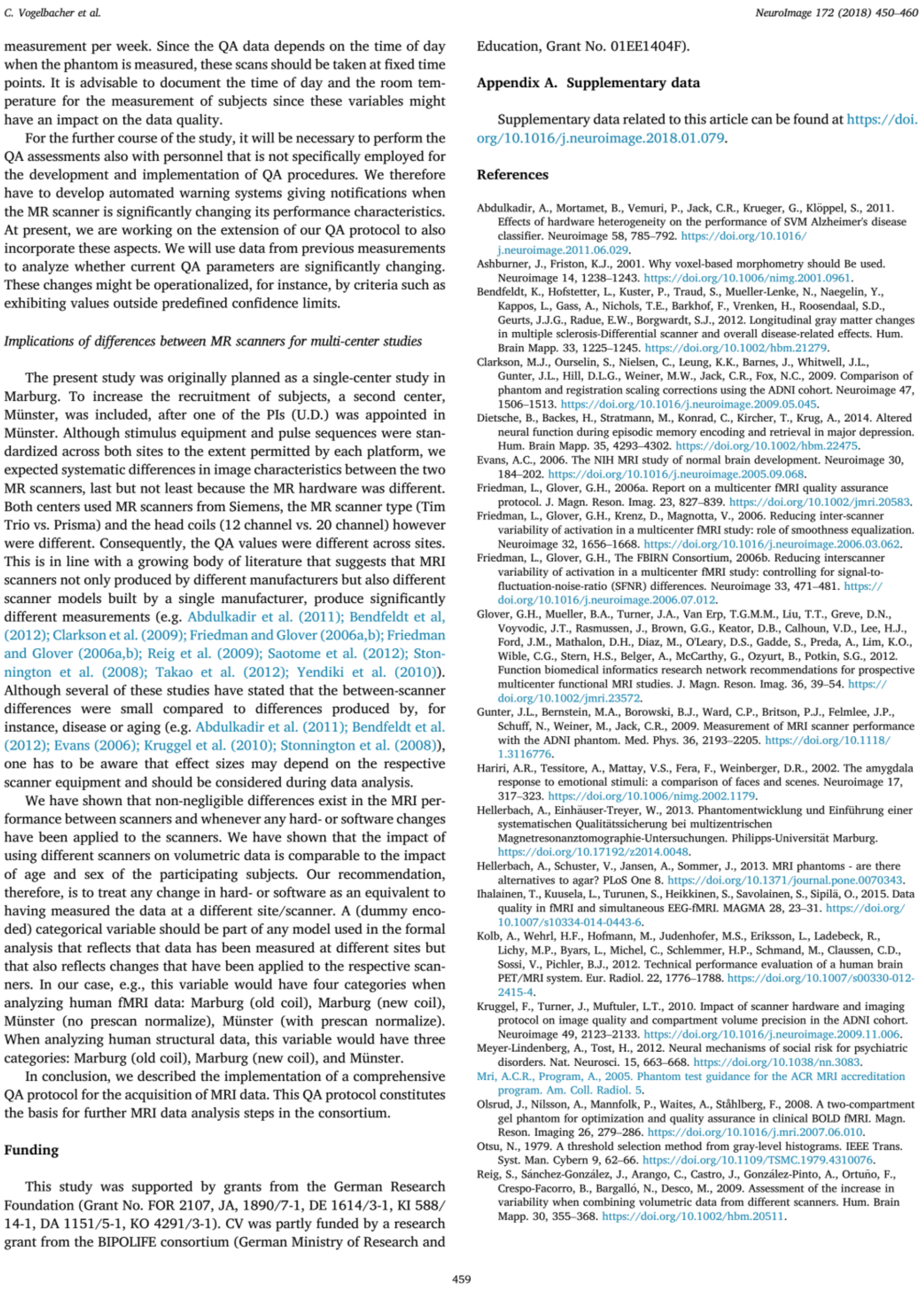
During the course of the study, both MR scanners underwent several software upgrades, but only one major hardware change. At Marburg, a defective gradient coil was replaced about one and a half years after start of the study. A comparison of phantom data before and after repair showed that some QA characteristics 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.

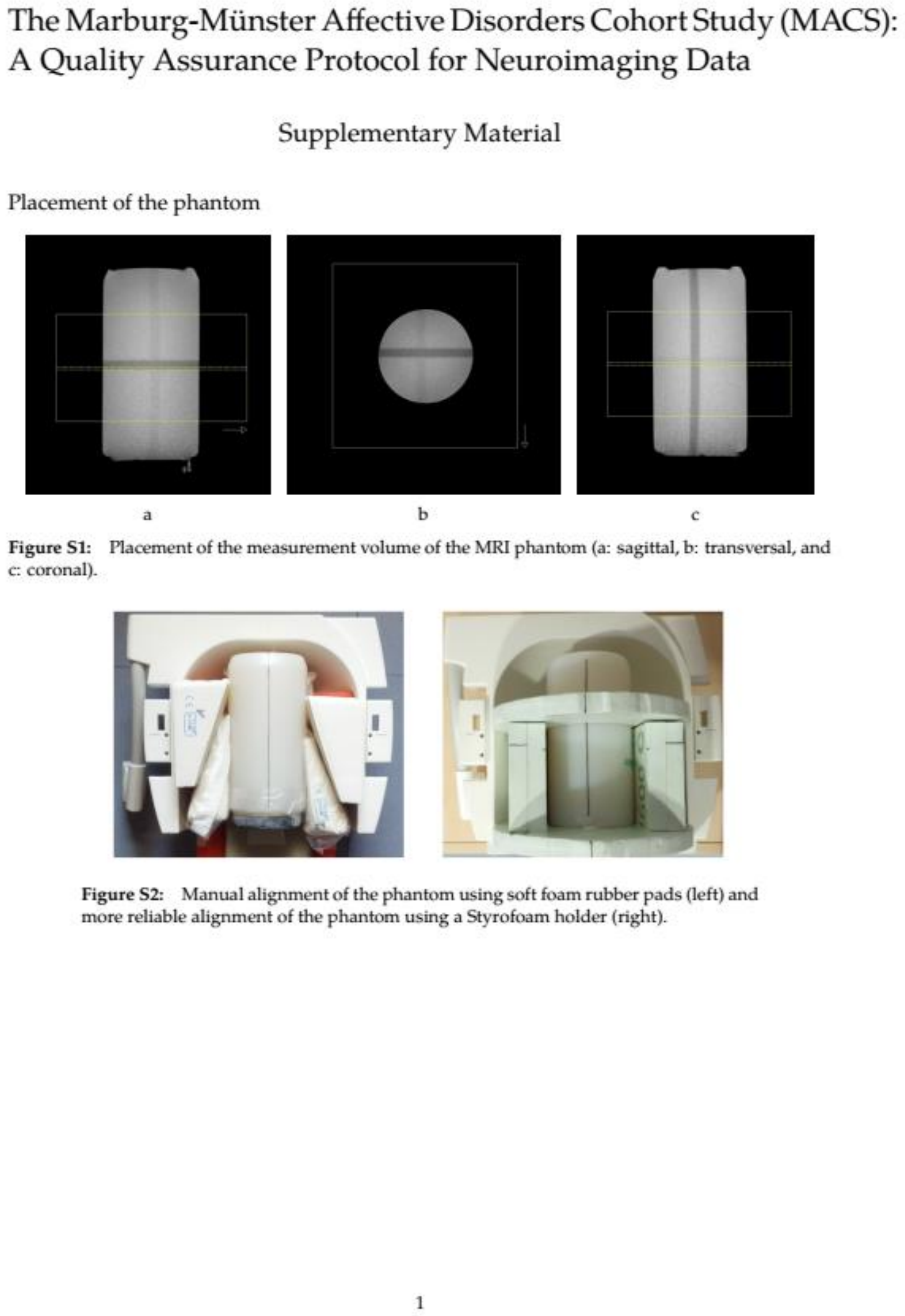
Tost, H., Bilek, E., Meyer-Lindenberg, A., 2012. Brain connectivity in psychiatric imaging genetics. Neuroimage. https://doi.org/10.1016/j.neuroimage.2011.11.007.

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.

Supplementary data



The Marburg-Münster Affective Disorders Cohort Study (MACS): A Quality Assurance Protocol for Neuroimaging Data

Supplementary Material

Placement of the phantom

a b c

Figure S1: Placement of the measurement volume of the MRI phantom (a: sagittal, b: transversal, and c: coronal).

Figure S2: Manual alignment of the phantom using soft foam rubber pads (left) and more reliable alignment of the phantom using a Styrofoam holder (right).

1

Influence of site and experimental settings on quality assurance statistics

The following figures show the course of all quality assurance (QA) statistics which have been described in the main text. The figures are based on a total of 1214 phantom measures which partitioning can be seen in Table S1.

Experimental Setting	Count
Marburg (new coil with holder)	212
Marburg (old coil with holder)	428
Marburg (old coil)	369
Münster	165
Münster (with pre-scan normalize)	40
Total	1214

Table S1: Sample sizes on which quality assurance statistic plots are based.

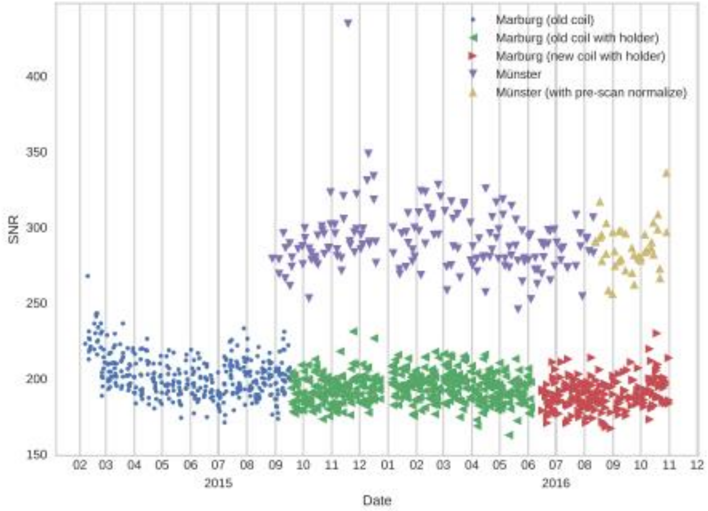


Figure S3: SNR

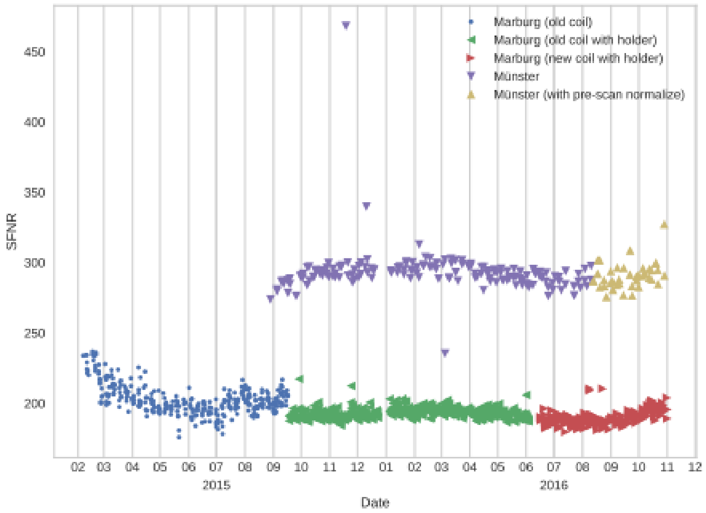


Figure S4: SFNR

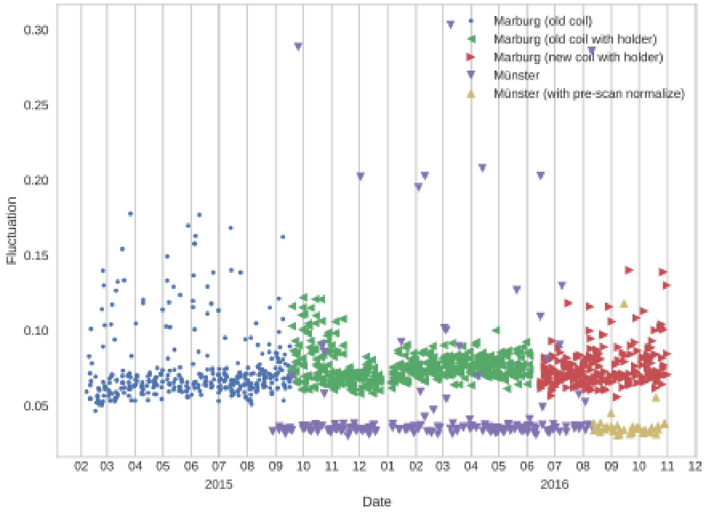


Figure S5: Fluctuation

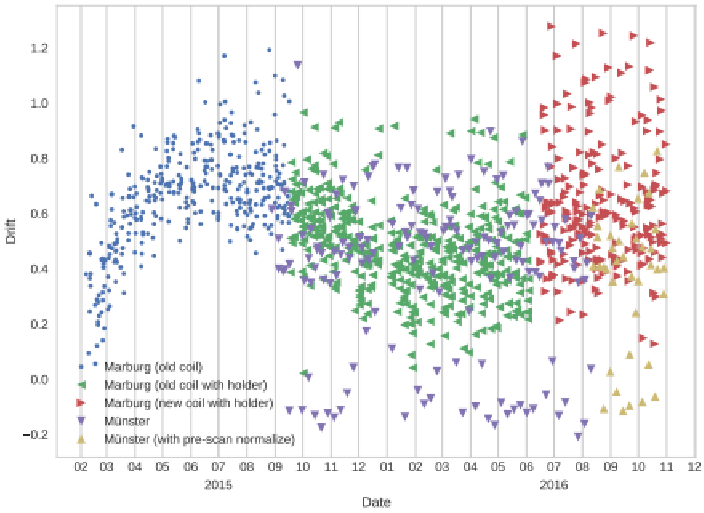


Figure S6: Drift

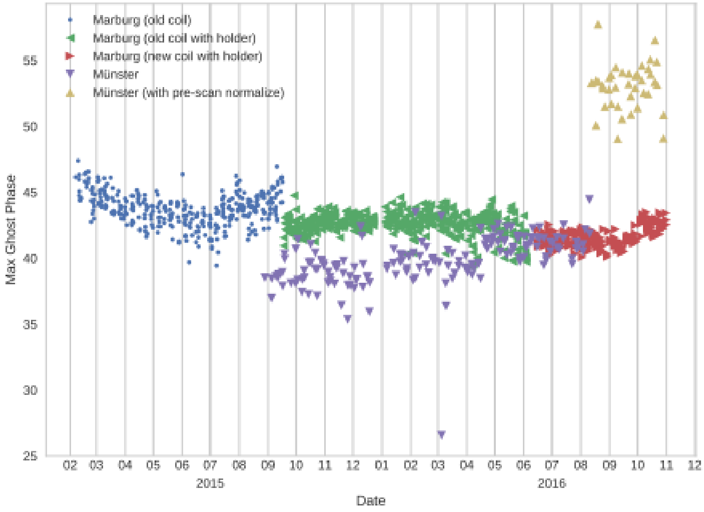


Figure S7: Max Ghost Phase

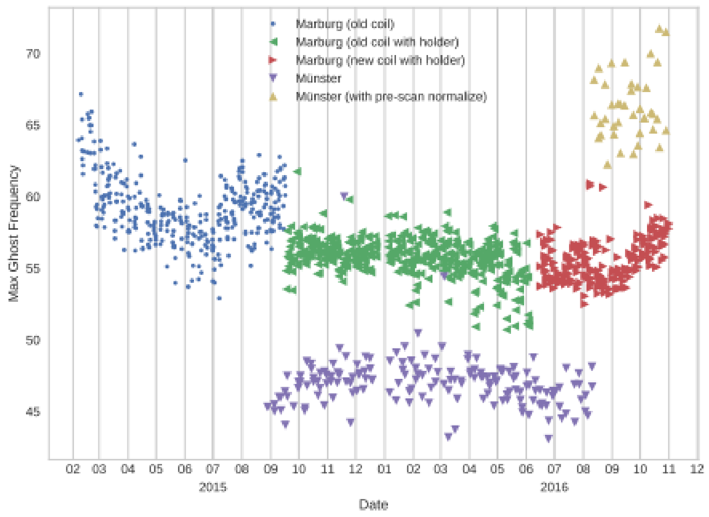


Figure S8: Max Ghost Frequency

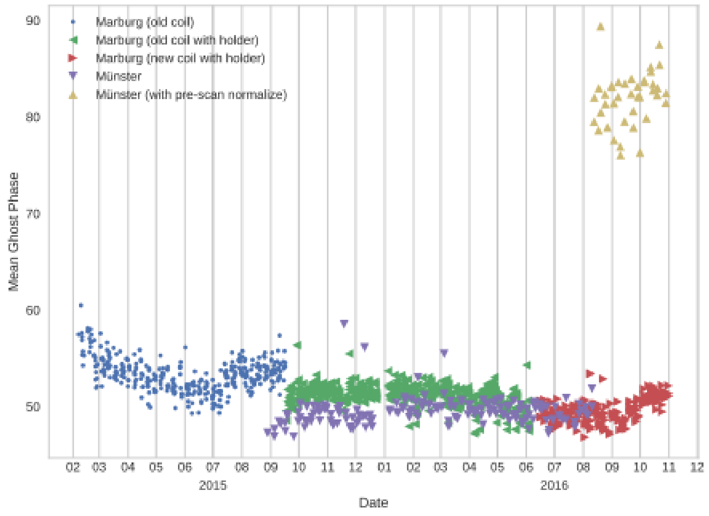


Figure S9: Mean Ghost Phase

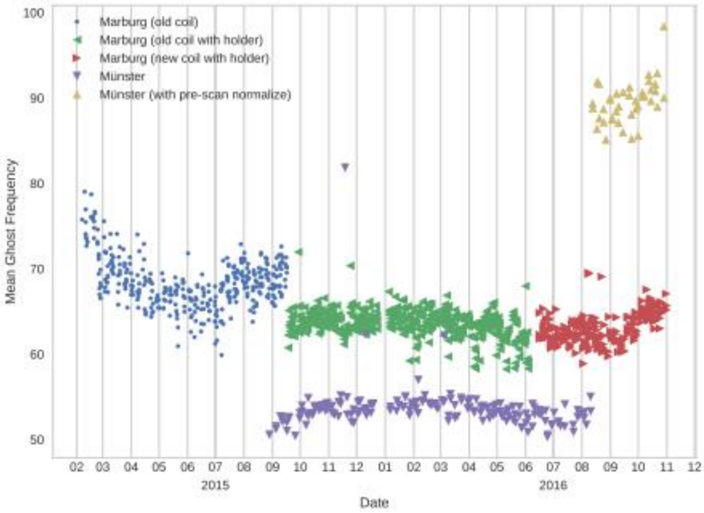


Figure S10: Mean Ghost Frequency

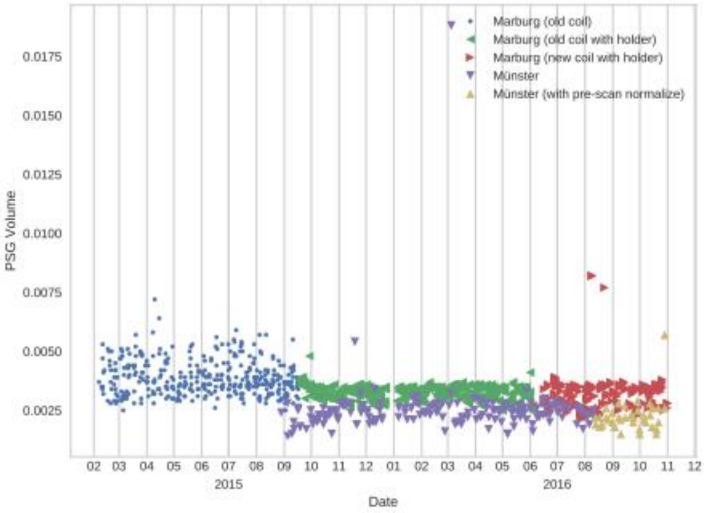


Figure S11: PSG Volume

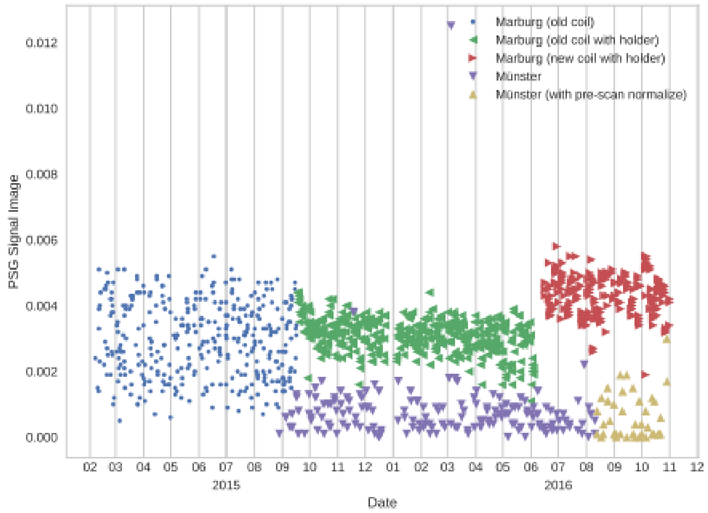


Figure S12: PSG Signal Image

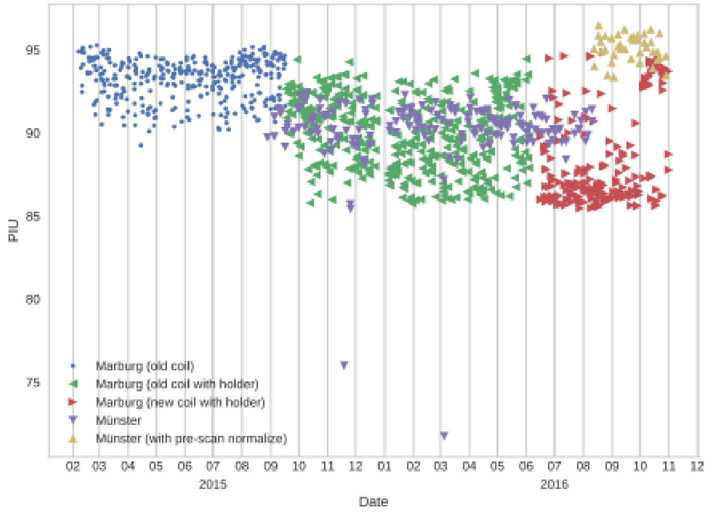


Figure S13: PIU

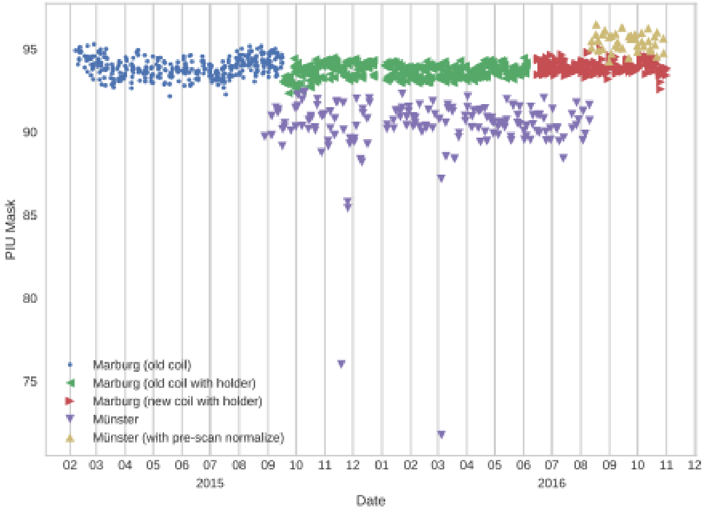


Figure S14: PIU Mask

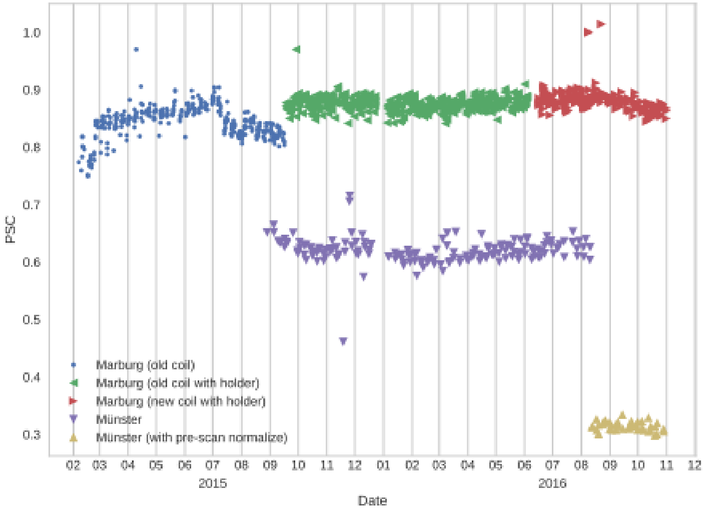
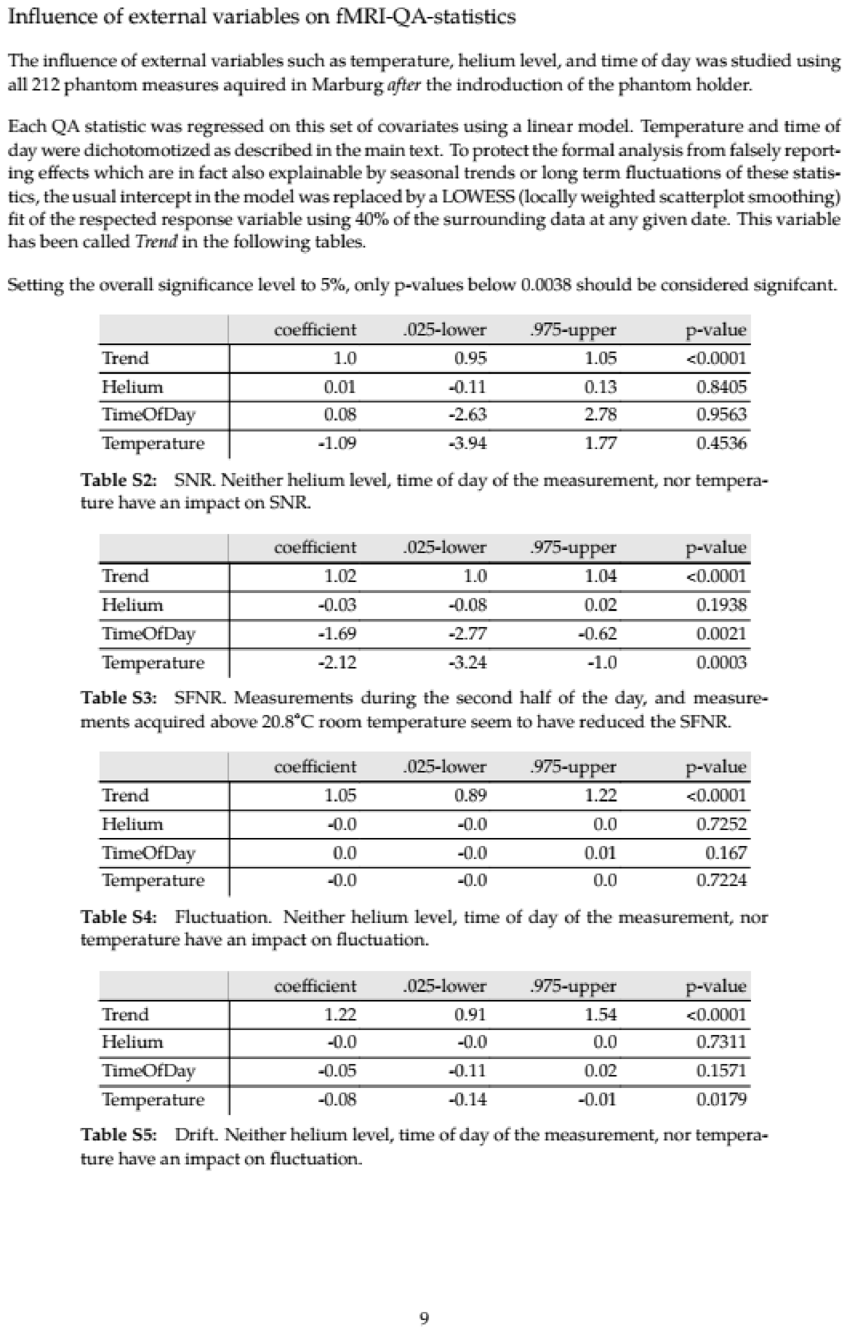


Figure S15: PSC



Influence of external variables on fMRI-QA-statistics

The influence of external variables such as temperature, helium level, and time of day was studied using all 212 phantom measures aquired in Marburg *after* the indroduction of the phantom holder.

Each QA statistic was regressed on this set of covariates using a linear model. Temperature and time of day were dichotomotized as described in the main text. To protect the formal analysis from falsely reporting effects which are in fact also explainable by seasonal trends or long term fluctuations of these statistics, the usual intercept in the model was replaced by a LOWESS (locally weighted scatterplot smoothing) fit of the respected response variable using 40% of the surrounding data at any given date. This variable has been called *Trend* in the following tables.

Setting the overall significance level to 5%, only p-values below 0.0038 should be considered signifcant.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.0	0.95	1.05	<0.0001
Helium	0.01	-0.11	0.13	0.8405
TimeOfDay	0.08	-2.63	2.78	0.9563
Temperature	-1.09	-3.94	1.77	0.4536

Table S2: SNR. Neither helium level, time of day of the measurement, nor temperature have an impact on SNR.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.02	1.0	1.04	<0.0001
Helium	-0.03	-0.08	0.02	0.1938
TimeOfDay	-1.69	-2.77	-0.62	0.0021
Temperature	-2.12	-3.24	-1.0	0.0003

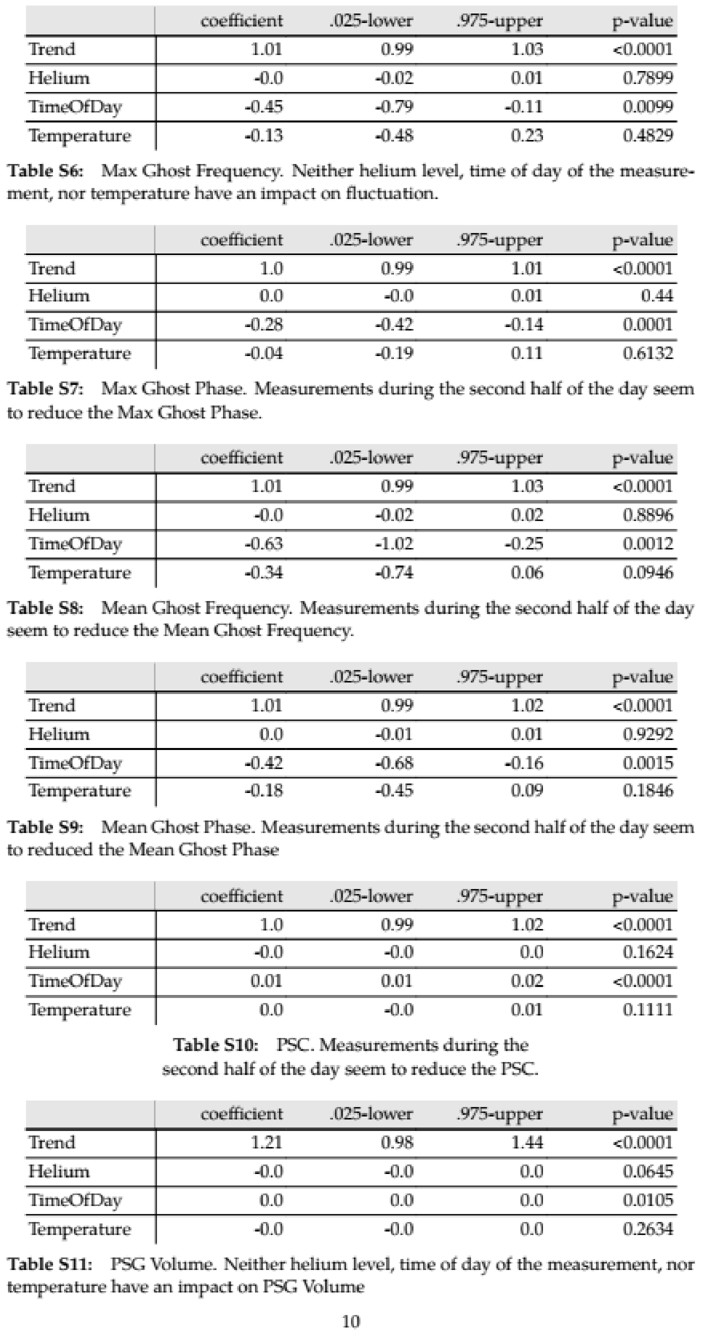
Table S3: SFNR. Measurements during the second half of the day, and measurements acquired above 20.8°C room temperature seem to have reduced the SFNR.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.05	0.89	1.22	<0.0001
Helium	-0.0	-0.0	0.0	0.7252
TimeOfDay	0.0	-0.0	0.01	0.167
Temperature	-0.0	-0.0	0.0	0.7224

Table S4: Fluctuation. Neither helium level, time of day of the measurement, nor temperature have an impact on fluctuation.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.22	0.91	1.54	<0.0001
Helium	-0.0	-0.0	0.0	0.7311
TimeOfDay	-0.05	-0.11	0.02	0.1571
Temperature	-0.08	-0.14	-0.01	0.0179

Table S5: Drift. Neither helium level, time of day of the measurement, nor temperature have an impact on fluctuation.



	coefficient	.025-lower	.975-upper	p-value
Trend	1.01	0.99	1.03	<0.0001
Helium	-0.0	-0.02	0.01	0.7899
TimeOfDay	-0.45	-0.79	-0.11	0.0099
Temperature	-0.13	-0.48	0.23	0.4829

Table S6: Max Ghost Frequency. Neither helium level, time of day of the measurement, nor temperature have an impact on fluctuation.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.0	0.99	1.01	<0.0001
Helium	0.0	-0.0	0.01	0.44
TimeOfDay	-0.28	-0.42	-0.14	0.0001
Temperature	-0.04	-0.19	0.11	0.6132

Table S7: Max Ghost Phase. Measurements during the second half of the day seem to reduce the Max Ghost Phase.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.01	0.99	1.03	<0.0001
Helium	-0.0	-0.02	0.02	0.8896
TimeOfDay	-0.63	-1.02	-0.25	0.0012
Temperature	-0.34	-0.74	0.06	0.0946

Table S8: Mean Ghost Frequency. Measurements during the second half of the day seem to reduce the Mean Ghost Frequency.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.01	0.99	1.02	<0.0001
Helium	0.0	-0.01	0.01	0.9292
TimeOfDay	-0.42	-0.68	-0.16	0.0015
Temperature	-0.18	-0.45	0.09	0.1846

Table S9: Mean Ghost Phase. Measurements during the second half of the day seem to reduced the Mean Ghost Phase

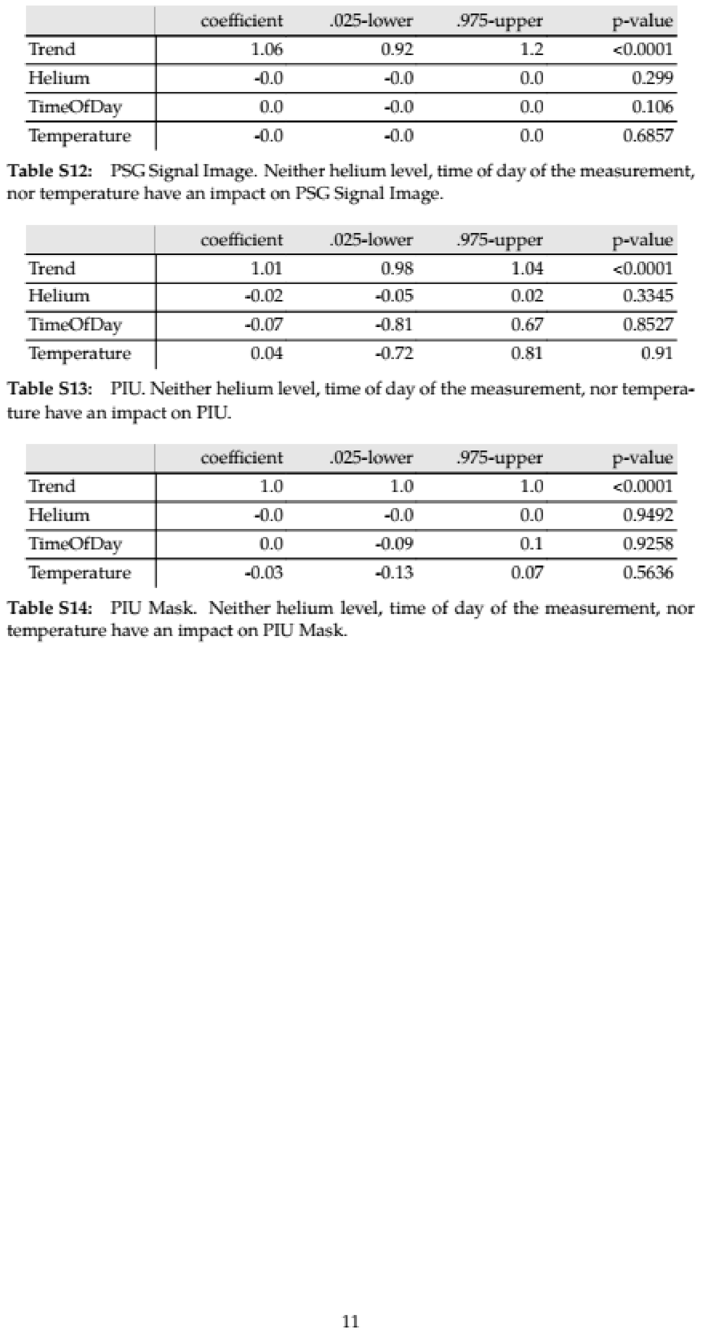
	coefficient	.025-lower	.975-upper	p-value
Trend	1.0	0.99	1.02	<0.0001
Helium	-0.0	-0.0	0.0	0.1624
TimeOfDay	0.01	0.01	0.02	<0.0001
Temperature	0.0	-0.0	0.01	0.1111

Table S10: PSC. Measurements during the second half of the day seem to reduce the PSC.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.21	0.98	1.44	<0.0001
Helium	-0.0	-0.0	0.0	0.0645
TimeOfDay	0.0	0.0	0.0	0.0105
Temperature	-0.0	-0.0	0.0	0.2634

Table S11: PSG Volume. Neither helium level, time of day of the measurement, nor temperature have an impact on PSG Volume

10



	coefficient	.025-lower	.975-upper	p-value
Trend	1.06	0.92	1.2	<0.0001
Helium	-0.0	-0.0	0.0	0.299
TimeOfDay	0.0	-0.0	0.0	0.106
Temperature	-0.0	-0.0	0.0	0.6857

Table S12: PSG Signal Image. Neither helium level, time of day of the measurement, nor temperature have an impact on PSG Signal Image.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.01	0.98	1.04	<0.0001
Helium	-0.02	-0.05	0.02	0.3345
TimeOfDay	-0.07	-0.81	0.67	0.8527
Temperature	0.04	-0.72	0.81	0.91

Table S13: PIU. Neither helium level, time of day of the measurement, nor temperature have an impact on PIU.

	coefficient	.025-lower	.975-upper	p-value
Trend	1.0	1.0	1.0	<0.0001
Helium	-0.0	-0.0	0.0	0.9492
TimeOfDay	0.0	-0.09	0.1	0.9258
Temperature	-0.03	-0.13	0.07	0.5636

Table S14: PIU Mask. Neither helium level, time of day of the measurement, nor temperature have an impact on PIU Mask.

Assessment of PSC using human FMRI data

Analysis of percent signal change (PSC) is based on all 373 healthy subjects measured in Marburg (273) and Münster (100).

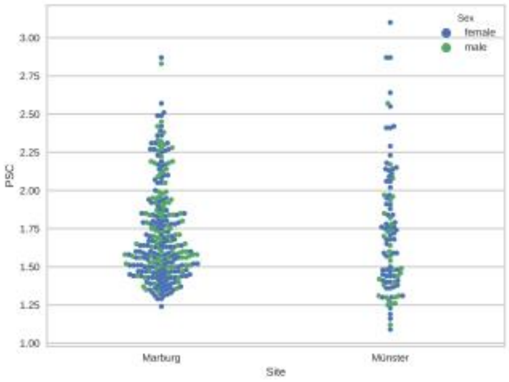


Figure S16: Swarm plot of the empirical PSC distribution stratified by site (Marburg/Münster) and sex. No site nor sex impact is visible in the empirical densities which is also in compliance with the formal analysis shown in the main text.

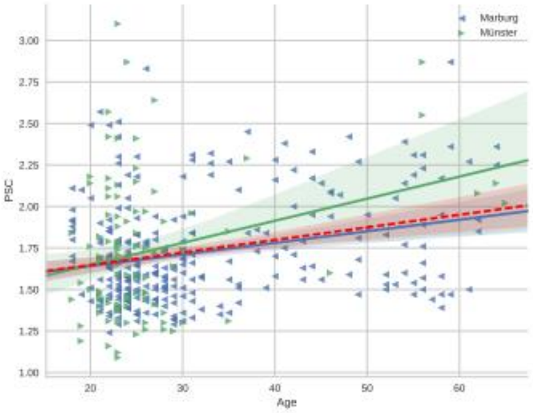


Figure S17: Scatter plot of PSC on the y-axis and age on the x-axis together with seperate estimates of a regression line for Münster and Marburg. Shaded are 95% confidence bands for the respected regression lines. As one can see, the areas strongly overlap and indeed the differences are non-significant. The red line shows the common regression line for both sites which slope, as the main text shows, is significantly different from zero.

Assessment of TIV, GMV, and WMV using human FMRI data

The analysis of total intracranial volume (TIV), gray matter volume (GMV), and white matter volume (WMV) is based on 444 healthy subjects measured in Marburg (445) and Münster (109).

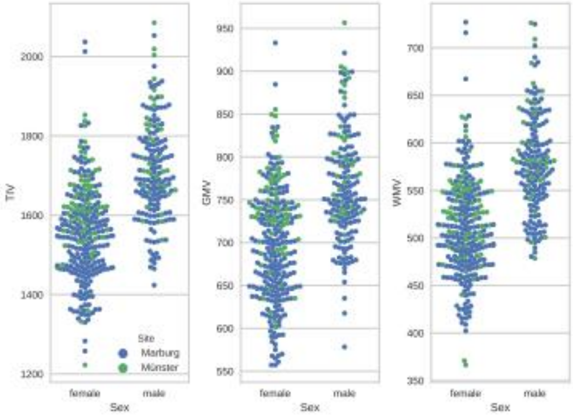


Figure S18: Swarm plots of the empirical TIV, GMV, and WMV distributions stratified by sex and coloured by site. Female subjects appear to produce smaller volumes than males. Since more females have been recruited in Münster than in Marburg, it is important to consider sex as a covariate when evaluating site effects on TIV, GMV, and WMV.

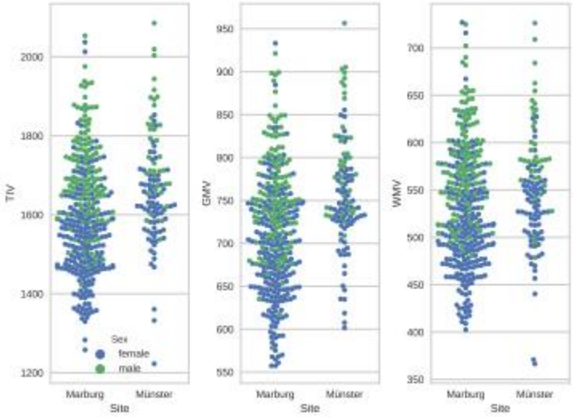


Figure S19: Swarm plots of the empirical TIV, GMV, and WMV distributions stratified by site and coloured by sex. Even though more females have been recruited in Münster than in Marburg, estimated volumes seem to be larger in Münster.

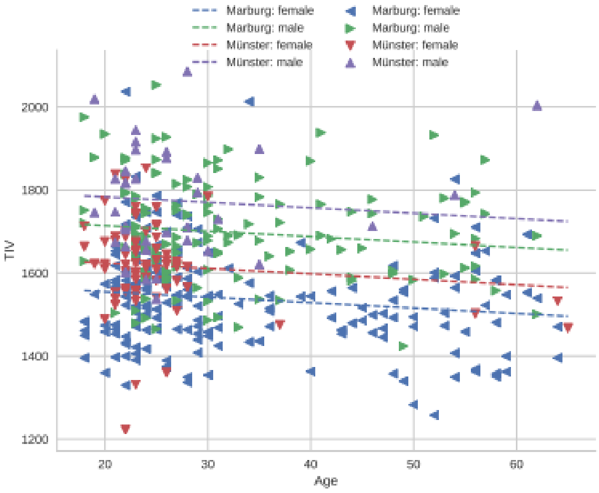


Figure S20: Scatter plot with Total Intracranial Volmune (TIV) on the y-axses and age on the x-axes. Regression lines through these points depend on site and sex and are depicted as dashed lines respectively. Parameter estimates and their significance are found in the main text.

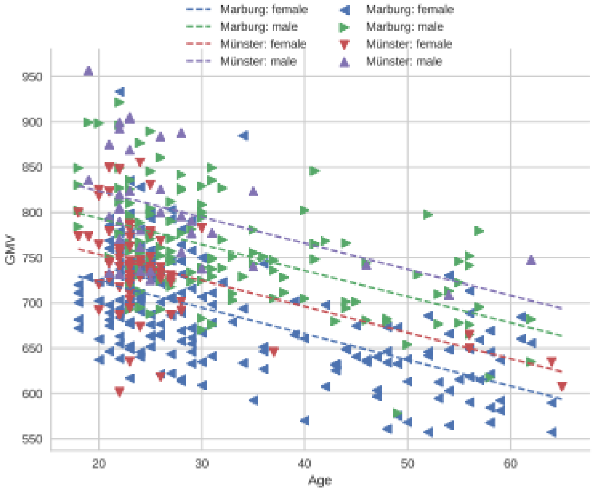


Figure S21: Scatter plot with Grey Matter Volume (GMV) on the y-axses and age on the x-axes. The intercept of a regression of GMV on age shows a significant dependence on site and sex. Estimates are depicted as dashed lines respectively. Parameter estimates and their significance are found in the main text.

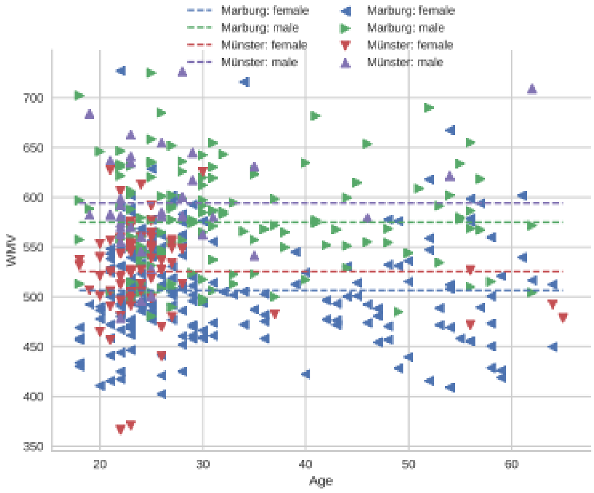
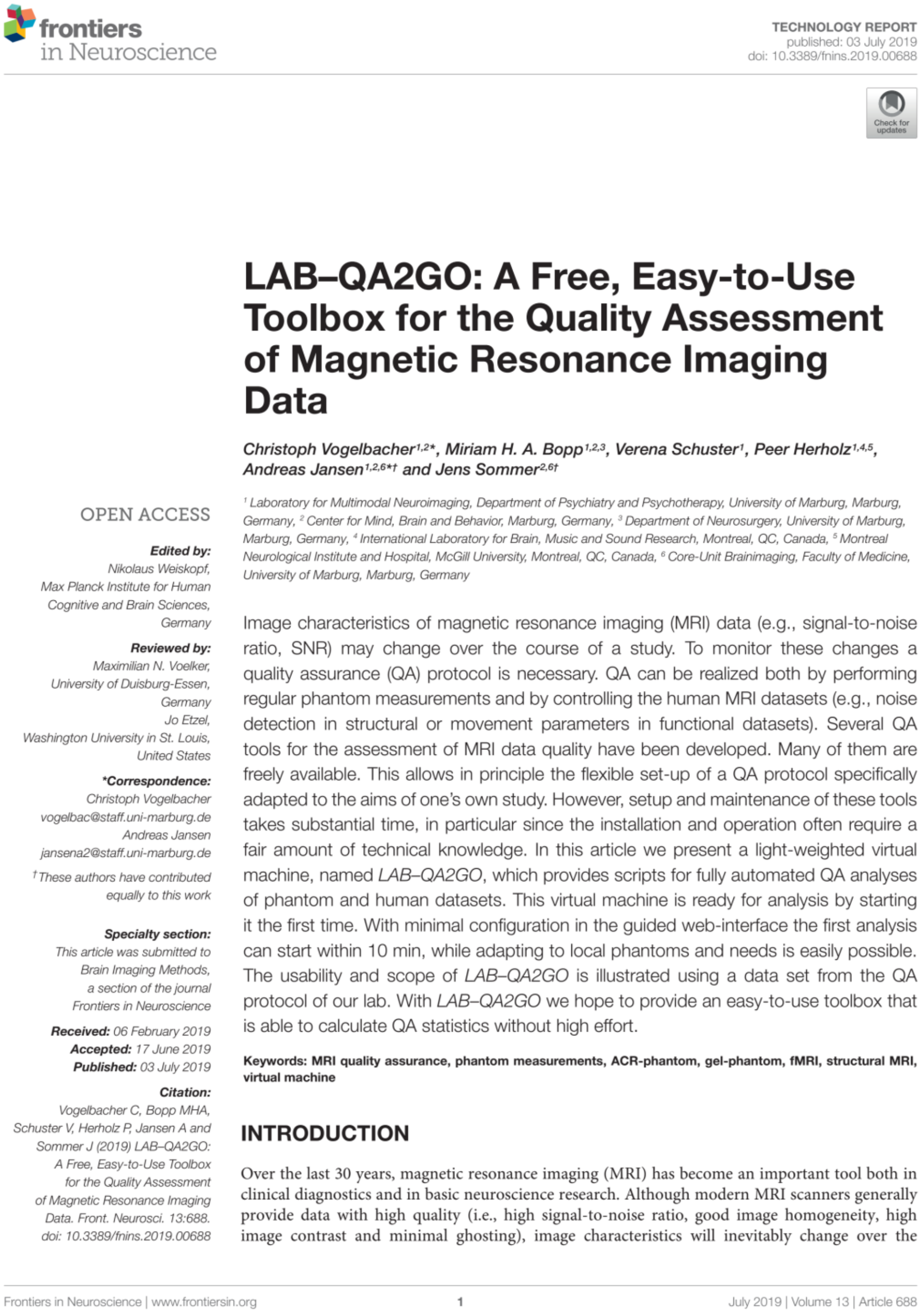


Figure S22: Scatter plot with White Matter Volume (WMV) on the y-axses and age on the x-axes. The intercept of a regression of WMV on age shows a significant dependence on site and sex. Age had a non-significant impart on WMV and has consequently been set to 0 resulting in constant regression line estimates. Parameter estimates and their significance are found in the main text.

Manuscript 2



frontiers in Neuroscience

TECHNOLOGY REPORT
published: 03 July 2019
doi: 10.3389/fnins.2019.00688

LAB–QA2GO: A Free, Easy-to-Use Toolbox for the Quality Assessment of Magnetic Resonance Imaging Data

Christoph Vogelbacher[1,2], Miriam H. A. Bopp[1,2,3], Verena Schuster[1], Peer Herholz[1,4,5], Andreas Jansen[1,2,6]*† and Jens Sommer[2,6†]*

[1] Laboratory for Multimodal Neuroimaging, Department of Psychiatry and Psychotherapy, University of Marburg, Marburg, Germany, [2] Center for Mind, Brain and Behavior, Marburg, Germany, [3] Department of Neurosurgery, University of Marburg, Marburg, Germany, [4] International Laboratory for Brain, Music and Sound Research, Montreal, QC, Canada, [5] Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada, [6] Core-Unit Brainimaging, Faculty of Medicine, University of Marburg, Marburg, Germany

OPEN ACCESS

Edited by:
Nikolaus Weiskopf,
Max Planck Institute for Human
Cognitive and Brain Sciences,
Germany

Reviewed by:
Maximilian N. Voelker,
University of Duisburg-Essen,
Germany
Jo Etzel,
Washington University in St. Louis,
United States

****Correspondence:***
Christoph Vogelbacher
vogelbac@staff.uni-marburg.de
Andreas Jansen
jansena2@staff.uni-marburg.de

† These authors have contributed equally to this work

Specialty section:
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

Received: *06 February 2019*
Accepted: *17 June 2019*
Published: *03 July 2019*

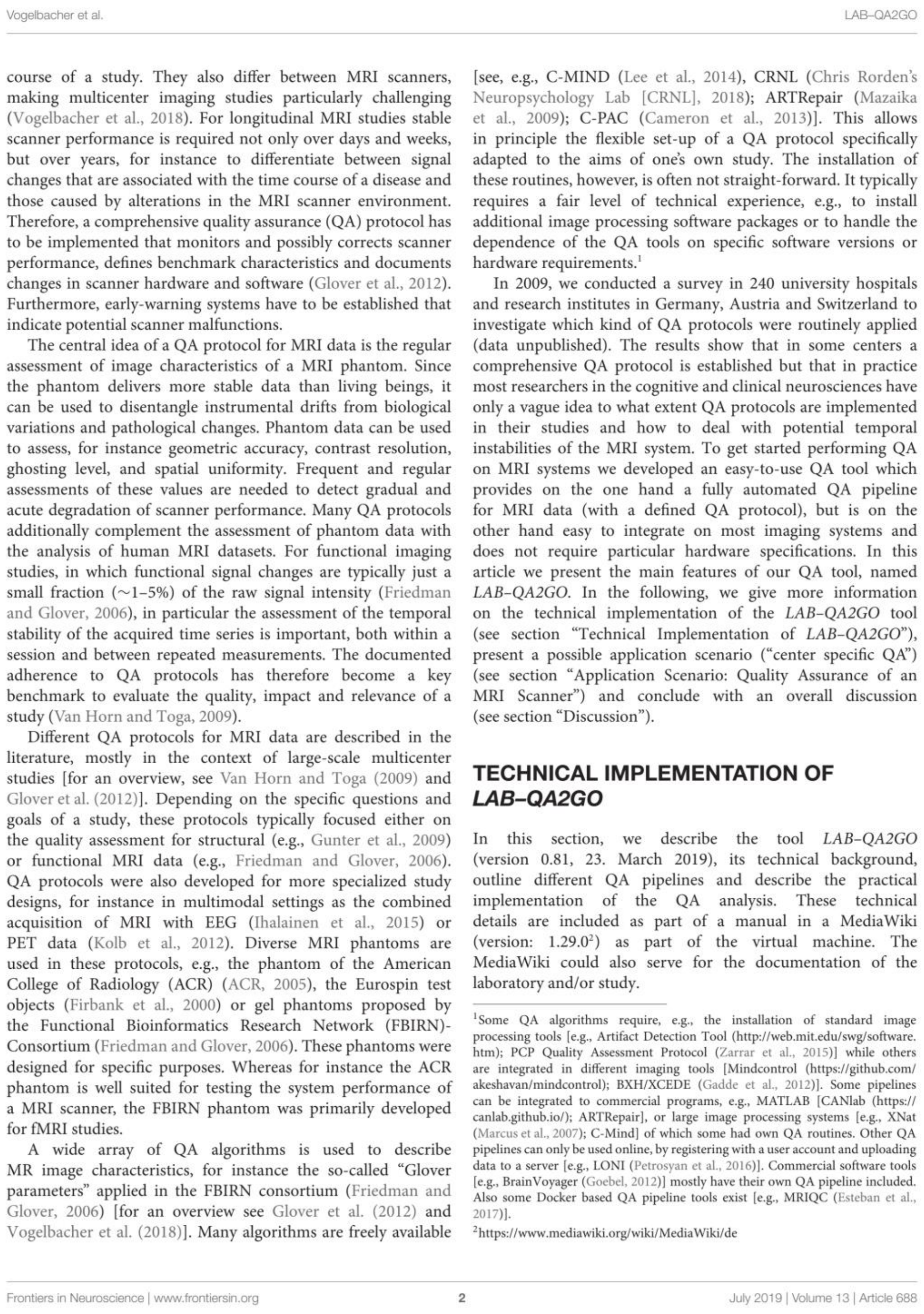
Citation:
Vogelbacher C, Bopp MHA, Schuster V, Herholz P, Jansen A and Sommer J (2019) LAB–QA2GO: A Free, Easy-to-Use Toolbox for the Quality Assessment of Magnetic Resonance Imaging Data. Front. Neurosci. 13:688. doi: 10.3389/fnins.2019.00688

Image characteristics of magnetic resonance imaging (MRI) data (e.g., signal-to-noise ratio, SNR) may change over the course of a study. To monitor these changes a quality assurance (QA) protocol is necessary. QA can be realized both by performing regular phantom measurements and by controlling the human MRI datasets (e.g., noise detection in structural or movement parameters in functional datasets). Several QA tools for the assessment of MRI data quality have been developed. Many of them are freely available. This allows in principle the flexible set-up of a QA protocol specifically adapted to the aims of one's own study. However, setup and maintenance of these tools takes substantial time, in particular since the installation and operation often require a fair amount of technical knowledge. In this article we present a light-weighted virtual machine, named *LAB–QA2GO*, which provides scripts for fully automated QA analyses of phantom and human datasets. This virtual machine is ready for analysis by starting it the first time. With minimal configuration in the guided web-interface the first analysis can start within 10 min, while adapting to local phantoms and needs is easily possible. The usability and scope of *LAB–QA2GO* is illustrated using a data set from the QA protocol of our lab. With *LAB–QA2GO* we hope to provide an easy-to-use toolbox that is able to calculate QA statistics without high effort.

Keywords: MRI quality assurance, phantom measurements, ACR-phantom, gel-phantom, fMRI, structural MRI, virtual machine

INTRODUCTION

Over the last 30 years, magnetic resonance imaging (MRI) has become an important tool both in clinical diagnostics and in basic neuroscience research. Although modern MRI scanners generally provide data with high quality (i.e., high signal-to-noise ratio, good image homogeneity, high image contrast and minimal ghosting), image characteristics will inevitably change over the



course of a study. They also differ between MRI scanners, making multicenter imaging studies particularly challenging (Vogelbacher et al., 2018). For longitudinal MRI studies stable scanner performance is required not only over days and weeks, but over years, for instance to differentiate between signal changes that are associated with the time course of a disease and those caused by alterations in the MRI scanner environment. Therefore, a comprehensive quality assurance (QA) protocol has to be implemented that monitors and possibly corrects scanner performance, defines benchmark characteristics and documents changes in scanner hardware and software (Glover et al., 2012). Furthermore, early-warning systems have to be established that indicate potential scanner malfunctions.

The central idea of a QA protocol for MRI data is the regular assessment of image characteristics of a MRI phantom. Since the phantom delivers more stable data than living beings, it can be used to disentangle instrumental drifts from biological variations and pathological changes. Phantom data can be used to assess, for instance geometric accuracy, contrast resolution, ghosting level, and spatial uniformity. Frequent and regular assessments of these values are needed to detect gradual and acute degradation of scanner performance. Many QA protocols additionally complement the assessment of phantom data with the analysis of human MRI datasets. For functional imaging studies, in which functional signal changes are typically just a small fraction (∼1–5%) of the raw signal intensity (Friedman and Glover, 2006), in particular the assessment of the temporal stability of the acquired time series is important, both within a session and between repeated measurements. The documented adherence to QA protocols has therefore become a key benchmark to evaluate the quality, impact and relevance of a study (Van Horn and Toga, 2009).

Different QA protocols for MRI data are described in the literature, mostly in the context of large-scale multicenter studies [for an overview, see Van Horn and Toga (2009) and Glover et al. (2012)]. Depending on the specific questions and goals of a study, these protocols typically focused either on the quality assessment for structural (e.g., Gunter et al., 2009) or functional MRI data (e.g., Friedman and Glover, 2006). QA protocols were also developed for more specialized study designs, for instance in multimodal settings as the combined acquisition of MRI with EEG (Ihalainen et al., 2015) or PET data (Kolb et al., 2012). Diverse MRI phantoms are used in these protocols, e.g., the phantom of the American College of Radiology (ACR) (ACR, 2005), the Eurospin test objects (Firbank et al., 2000) or gel phantoms proposed by the Functional Bioinformatics Research Network (FBIRN)-Consortium (Friedman and Glover, 2006). These phantoms were designed for specific purposes. Whereas for instance the ACR phantom is well suited for testing the system performance of a MRI scanner, the FBIRN phantom was primarily developed for fMRI studies.

A wide array of QA algorithms is used to describe MR image characteristics, for instance the so-called "Glover parameters" applied in the FBIRN consortium (Friedman and Glover, 2006) [for an overview see Glover et al. (2012) and Vogelbacher et al. (2018)]. Many algorithms are freely available [see, e.g., C-MIND (Lee et al., 2014), CRNL (Chris Rorden's Neuropsychology Lab [CRNL], 2018); ARTRepair (Mazaika et al., 2009); C-PAC (Cameron et al., 2013)]. This allows in principle the flexible set-up of a QA protocol specifically adapted to the aims of one's own study. The installation of these routines, however, is often not straight-forward. It typically requires a fair level of technical experience, e.g., to install additional image processing software packages or to handle the dependence of the QA tools on specific software versions or hardware requirements.[1]

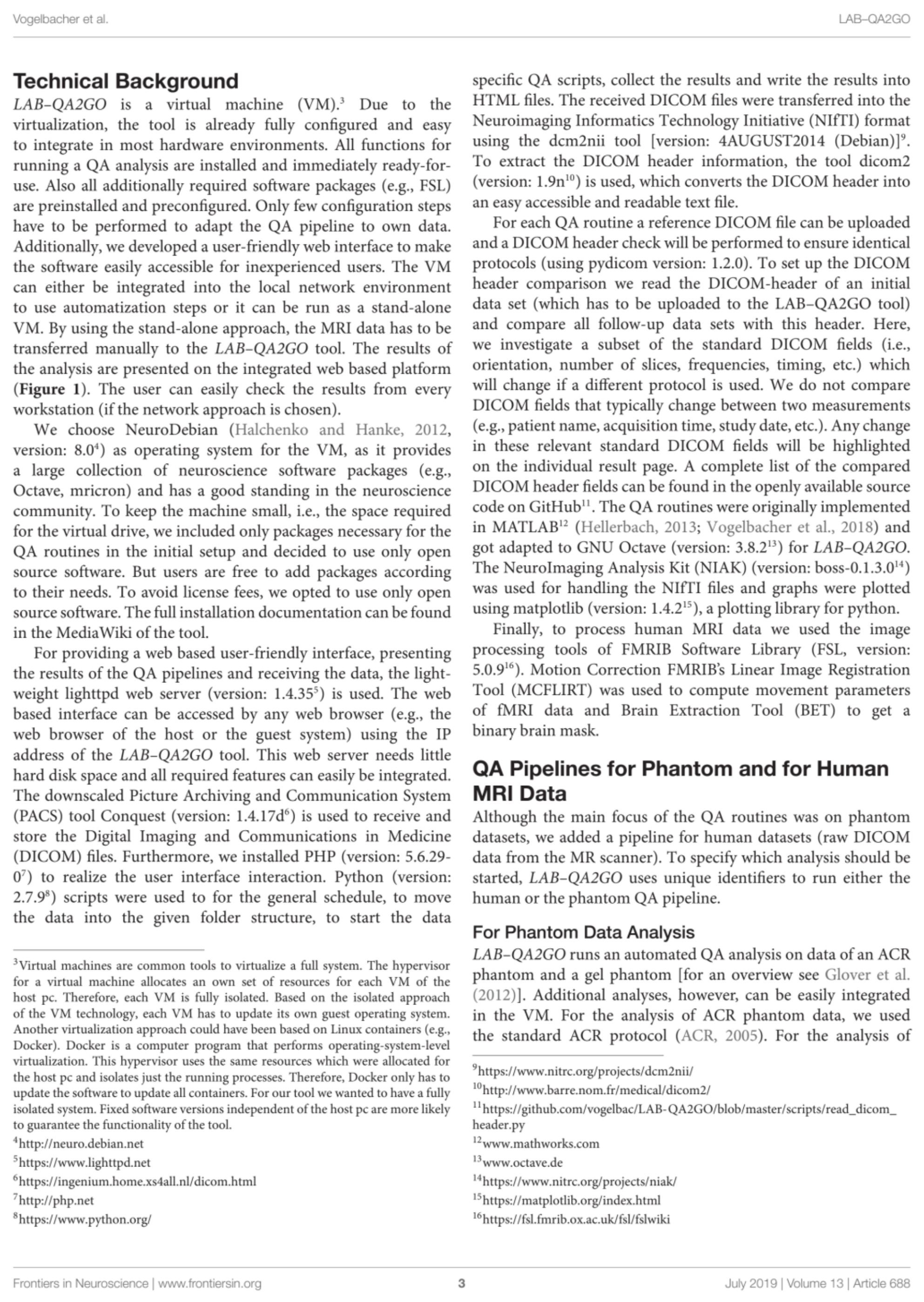
In 2009, we conducted a survey in 240 university hospitals and research institutes in Germany, Austria and Switzerland to investigate which kind of QA protocols were routinely applied (data unpublished). The results show that in some centers a comprehensive QA protocol is established but that in practice most researchers in the cognitive and clinical neurosciences have only a vague idea to what extent QA protocols are implemented in their studies and how to deal with potential temporal instabilities of the MRI system. To get started performing QA on MRI systems we developed an easy-to-use QA tool which provides on the one hand a fully automated QA pipeline for MRI data (with a defined QA protocol), but is on the other hand easy to integrate on most imaging systems and does not require particular hardware specifications. In this article we present the main features of our QA tool, named *LAB–QA2GO*. In the following, we give more information on the technical implementation of the *LAB–QA2GO* tool (see section "Technical Implementation of *LAB–QA2GO*"), present a possible application scenario ("center specific QA") (see section "Application Scenario: Quality Assurance of an MRI Scanner") and conclude with an overall discussion (see section "Discussion").

TECHNICAL IMPLEMENTATION OF *LAB–QA2GO*

In this section, we describe the tool *LAB–QA2GO* (version 0.81, 23. March 2019), its technical background, outline different QA pipelines and describe the practical implementation of the QA analysis. These technical details are included as part of a manual in a MediaWiki (version: 1.29.0[2]) as part of the virtual machine. The MediaWiki could also serve for the documentation of the laboratory and/or study.

[1] Some QA algorithms require, e.g., the installation of standard image processing tools [e.g., Artifact Detection Tool (http://web.mit.edu/swg/software.htm); PCP Quality Assessment Protocol (Zarrar et al., 2015)] while others are integrated in different imaging tools [Mindcontrol (https://github.com/akeshavan/mindcontrol); BXH/XCEDE (Gadde et al., 2012)]. Some pipelines can be integrated to commercial programs, e.g., MATLAB [CANlab (https://canlab.github.io/); ARTRepair], or large image processing systems [e.g., XNat (Marcus et al., 2007); C-Mind] of which some had own QA routines. Other QA pipelines can only be used online, by registering with a user account and uploading data to a server [e.g., LONI (Petrosyan et al., 2016)]. Commercial software tools [e.g., BrainVoyager (Goebel, 2012)] mostly have their own QA pipeline included. Also some Docker based QA pipeline tools exist [e.g., MRIQC (Esteban et al., 2017)].

[2] https://www.mediawiki.org/wiki/MediaWiki/de



Technical Background

LAB–QA2GO is a virtual machine (VM).[3] Due to the virtualization, the tool is already fully configured and easy to integrate in most hardware environments. All functions for running a QA analysis are installed and immediately ready-for-use. Also all additionally required software packages (e.g., FSL) are preinstalled and preconfigured. Only few configuration steps have to be performed to adapt the QA pipeline to own data. Additionally, we developed a user-friendly web interface to make the software easily accessible for inexperienced users. The VM can either be integrated into the local network environment to use automatization steps or it can be run as a stand-alone VM. By using the stand-alone approach, the MRI data has to be transferred manually to the *LAB–QA2GO* tool. The results of the analysis are presented on the integrated web based platform (**Figure 1**). The user can easily check the results from every workstation (if the network approach is chosen).

We choose NeuroDebian (Halchenko and Hanke, 2012, version: 8.0[4]) as operating system for the VM, as it provides a large collection of neuroscience software packages (e.g., Octave, mricron) and has a good standing in the neuroscience community. To keep the machine small, i.e., the space required for the virtual drive, we included only packages necessary for the QA routines in the initial setup and decided to use only open source software. But users are free to add packages according to their needs. To avoid license fees, we opted to use only open source software. The full installation documentation can be found in the MediaWiki of the tool.

For providing a web based user-friendly interface, presenting the results of the QA pipelines and receiving the data, the light-weight lighttpd web server (version: 1.4.35[5]) is used. The web based interface can be accessed by any web browser (e.g., the web browser of the host or the guest system) using the IP address of the *LAB–QA2GO* tool. This web server needs little hard disk space and all required features can easily be integrated. The downscaled Picture Archiving and Communication System (PACS) tool Conquest (version: 1.4.17d[6]) is used to receive and store the Digital Imaging and Communications in Medicine (DICOM) files. Furthermore, we installed PHP (version: 5.6.29-0[7]) to realize the user interface interaction. Python (version: 2.7.9[8]) scripts were used to for the general schedule, to move the data into the given folder structure, to start the data specific QA scripts, collect the results and write the results into HTML files. The received DICOM files were transferred into the Neuroimaging Informatics Technology Initiative (NIfTI) format using the dcm2nii tool [version: 4AUGUST2014 (Debian)][9]. To extract the DICOM header information, the tool dicom2 (version: 1.9n[10]) is used, which converts the DICOM header into an easy accessible and readable text file.

For each QA routine a reference DICOM file can be uploaded and a DICOM header check will be performed to ensure identical protocols (using pydicom version: 1.2.0). To set up the DICOM header comparison we read the DICOM-header of an initial data set (which has to be uploaded to the LAB–QA2GO tool) and compare all follow-up data sets with this header. Here, we investigate a subset of the standard DICOM fields (i.e., orientation, number of slices, frequencies, timing, etc.) which will change if a different protocol is used. We do not compare DICOM fields that typically change between two measurements (e.g., patient name, acquisition time, study date, etc.). Any change in these relevant standard DICOM fields will be highlighted on the individual result page. A complete list of the compared DICOM header fields can be found in the openly available source code on GitHub[11]. The QA routines were originally implemented in MATLAB[12] (Hellerbach, 2013; Vogelbacher et al., 2018) and got adapted to GNU Octave (version: 3.8.2[13]) for *LAB–QA2GO*. The NeuroImaging Analysis Kit (NIAK) (version: boss-0.1.3.0[14]) was used for handling the NIfTI files and graphs were plotted using matplotlib (version: 1.4.2[15]), a plotting library for python.

Finally, to process human MRI data we used the image processing tools of FMRIB Software Library (FSL, version: 5.0.9[16]). Motion Correction FMRIB's Linear Image Registration Tool (MCFLIRT) was used to compute movement parameters of fMRI data and Brain Extraction Tool (BET) to get a binary brain mask.

QA Pipelines for Phantom and for Human MRI Data

Although the main focus of the QA routines was on phantom datasets, we added a pipeline for human datasets (raw DICOM data from the MR scanner). To specify which analysis should be started, *LAB–QA2GO* uses unique identifiers to run either the human or the phantom QA pipeline.

For Phantom Data Analysis

LAB–QA2GO runs an automated QA analysis on data of an ACR phantom and a gel phantom [for an overview see Glover et al. (2012)]. Additional analyses, however, can be easily integrated in the VM. For the analysis of ACR phantom data, we used the standard ACR protocol (ACR, 2005). For the analysis of

[3] Virtual machines are common tools to virtualize a full system. The hypervisor for a virtual machine allocates an own set of resources for each VM of the host pc. Therefore, each VM is fully isolated. Based on the isolated approach of the VM technology, each VM has to update its own guest operating system. Another virtualization approach could have been based on Linux containers (e.g., Docker). Docker is a computer program that performs operating-system-level virtualization. This hypervisor uses the same resources which were allocated for the host pc and isolates just the running processes. Therefore, Docker only has to update the software to update all containers. For our tool we wanted to have a fully isolated system. Fixed software versions independent of the host pc are more likely to guarantee the functionality of the tool.

[4] http://neuro.debian.net

[5] https://www.lighttpd.net

[6] https://ingenium.home.xs4all.nl/dicom.html

[7] http://php.net

[8] https://www.python.org/

[9] https://www.nitrc.org/projects/dcm2nii/

[10] http://www.barre.nom.fr/medical/dicom2/

[11] https://github.com/vogelbac/LAB-QA2GO/blob/master/scripts/read_dicom_header.py

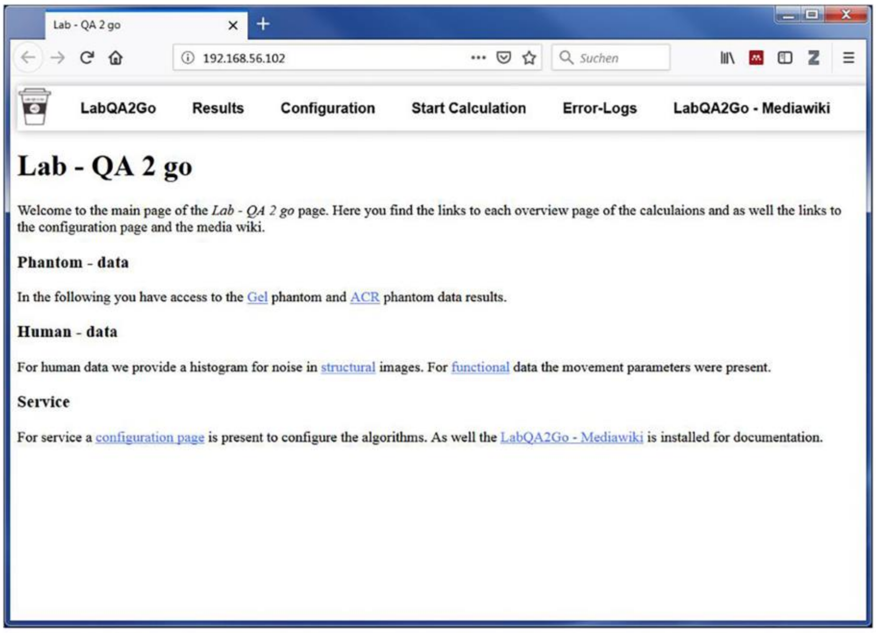
[12] www.mathworks.com

[13] www.octave.de

[14] https://www.nitrc.org/projects/niak/

[15] https://matplotlib.org/index.html

[16] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki



Lab - QA 2 go

Welcome to the main page of the *Lab - QA 2 go* page. Here you find the links to each overview page of the calculaions and as well the links to the configuration page and the media wiki.

Phantom - data

In the following you have access to the Gel phantom and ACR phantom data results.

Human - data

For human data we provide a histogram for noise in structural images. For functional data the movement parameters were present.

Service

For service a configuration page is present to configure the algorithms. As well the LabQA2Go - Mediawiki is installed for documentation.

FIGURE 1 | The web based graphical user interface of the LAB–QA2GO tool presented in a web browser.

gel phantom data, we used statistics previously described by Friedman et al. (2006) (the so-called "Glover parameters"), Simmons et al. (1999) and Stöcker et al. (2005). These statistics assess, e.g., the signal-to-noise ratio, the uniformity of an image or the temporal fluctuation. Detailed information on these statistics can be found elsewhere (Vogelbacher et al., 2018). In **Figure 2** the calculation of the signal-to-noise-ratio (SNR) based on the "Glover parameters" is shown exemplarily.

For Human Data Analysis

We use movement parameters from fMRI and noise level from structural MRI as easily interpretable QA parameters (**Figure 3**). The head movement parameters (translation and rotation) are calculated using FSL MCFLIRT and FSL FSLINFO with default settings, i.e., motion parameters relative to the middle image of the time series. Each parameter (pitch, roll, yaw, movement in x, y, and z direction) is plotted for each time-point in a graph (**Figures 3A,B**). Additionally, a histogram is generated of the step width between two consecutive time points to detect large movements between two time points (**Figures 3C,D**). For structural MRI data, a brain mask is calculated by FSL'S BET (using the default values) first. Subsequently, the basal noise of the image background (i.e., the area around the head) is detected. First, a region of interest (ROI) is defined in the corner of the three-dimensional image. Second, the mean of this ROI aggregated by an initial user defined threshold multiplier is used to mask the head in a first step. Third, for every axial and sagittal slice the edges of the scalp were detected by using a differentiation algorithm between two images to create a binary mask of the head (**Figure 3G**). Fourth, this binary mask is multiplied with the original image to get the background of the head image. Fifth, for this background a histogram (**Figure 3**) of the containing intensity values is generated. The calculated mask is saved to create images for the report. Also a basal SNR (bSNR) value is calculated by the quotient of the mean intensity in the brain mask and the standard deviation of the background signal. Each value is presented individually in the report to easily see by which parameter the SNR value was influenced. These two methods should give the user an overview of existing noise in the image. Both methods can be independently activated or deactivated by the user to individually run the QA routines.

Practical Implementation of QA Analyses in *LAB–QA2GO*

The *LAB–QA2GO* pipelines (phantom data, human data) are preconfigured, but require unique identifiers as part of the dicom

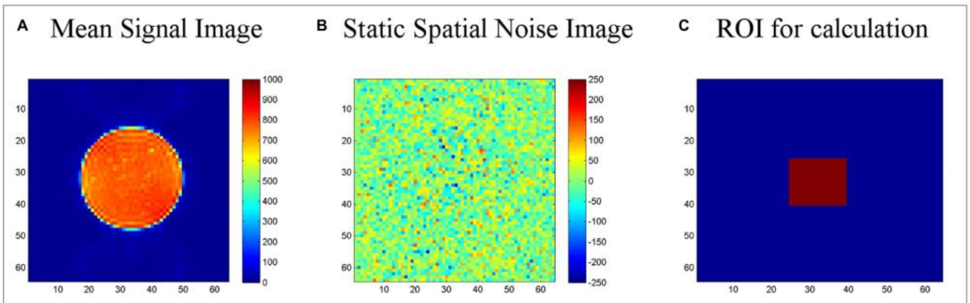


FIGURE 2 | Calculation of the signal-to-noise-ratio (SNR) for the gel phantom. **(A)** First, a "signal image" is calculated as the voxel-wise average of the center slice of the gel phantom (slice-of-interest, SOI) across the time series. **(B)** Second, a "static spatial noise image" is calculated as the voxel-wise difference of the sum of all odd images and the sum of all even images in the SOI. **(C)** Third, the SNR is defined as the quotient of the average intensity of the mean signal image in a region of interest (ROI, 15 × 15 voxel), located in the center of the phantom of the SOI, and the standard deviation of the static spatial noise within the same ROI.

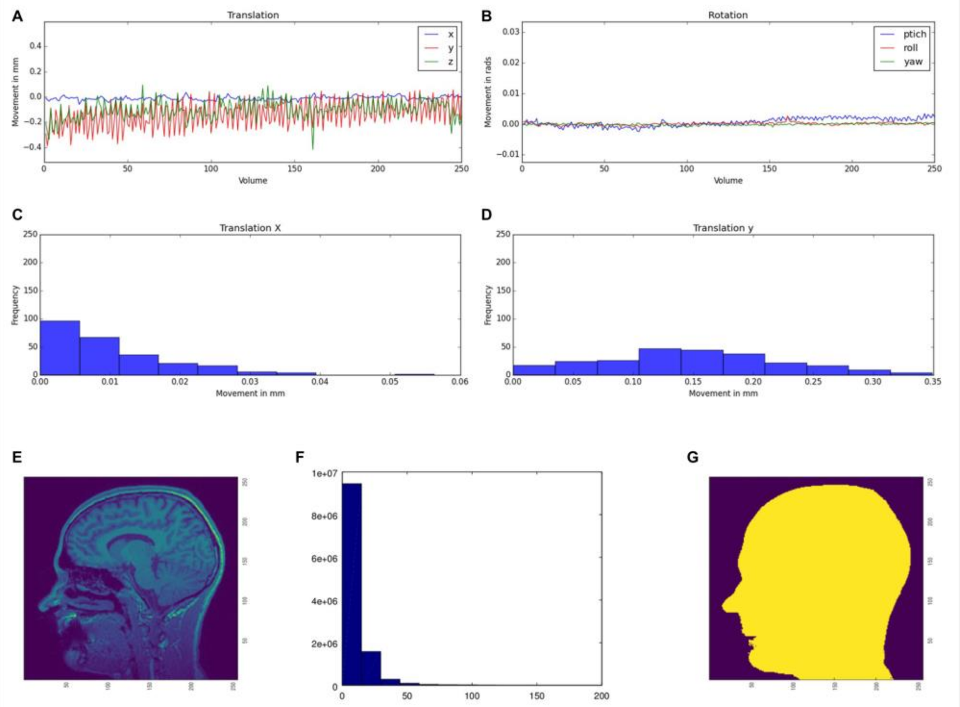
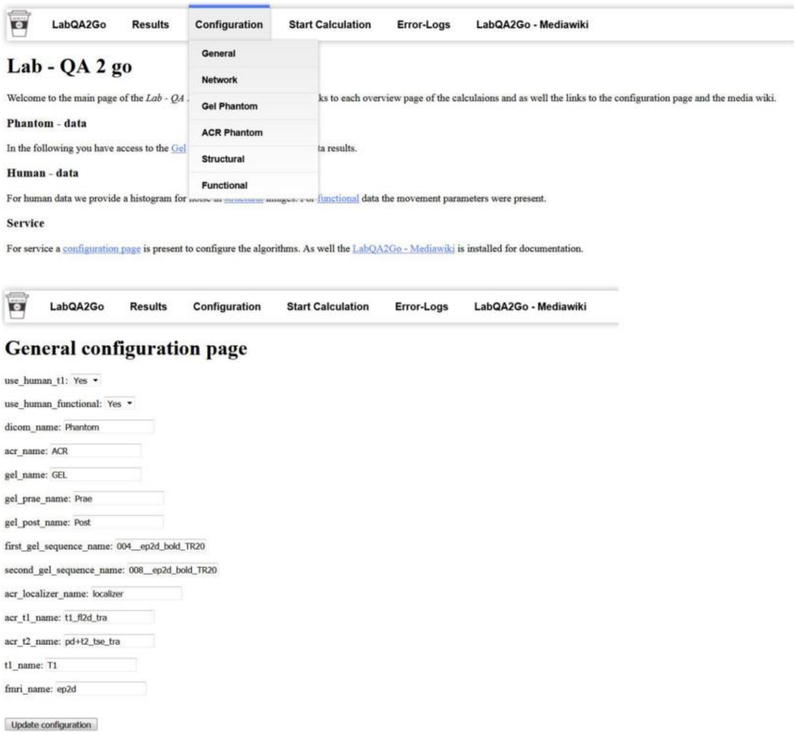


FIGURE 3 | Head movement parameters [translation **(A)** and rotation **(B)**] for functional MRI data. The movement parameters are transferred into a histogram, illustrating the amount of movement between two consecutive time points (exemplarily shown for the x **(C)** and y **(D)** translation parameters). Original structural image **(E)**, the calculated mask **(F)**, and the histogram of the basal noise **(G)**.



Lab - QA 2 go

Welcome to the main page of the *Lab - QA* … ks to each overview page of the calculaions and as well the links to the configuration page and the media wiki.

Phantom - data

In the following you have access to the Gel … ta results.

Human - data

For human data we provide a histogram for … functional data the movement parameters were present.

Service

For service a configuration page is present to configure the algorithms. As well the LabQA2Go - Mediawiki is installed for documentation.

General configuration page

use_human_t1: Yes
use_human_functional: Yes
dicom_name: Phantom
acr_name: ACR
gel_name: GEL
gel_prae_name: Prae
gel_post_name: Post
first_gel_sequence_name: 004__ep2d_bold_TR20
second_gel_sequence_name: 008__ep2d_bold_TR20
acr_localizer_name: localizer
acr_t1_name: t1_fl2d_tra
acr_t2_name: pd+t2_tse_tra
t1_name: T1
fmri_name: ep2d

FIGURE 4 | The general configuration page used to set the unique identifier and to activate or deactivate the human dataset QA pipeline.

field "patient name" to distinguish between data sets, i.e., which pipeline should be used for the analysis of the specific data set. Predefined are "Phantom," "ACR," and "GEL" in the field "patient name," but can be adopted to the local needs. These unique identifiers have to be inserted into the configuration page (a web based form) on the VM (**Figure 4**). The algorithm checks for the field "patient name" of the DICOM header so that the unique identifier has to be part of the "patient name" and has to be set during the registration of the patient at the MR scanner.

The MRI data are integrated into the VM either by sending them ("dicom send," network configuration) or providing them manually (directory browsing, stand-alone configuration). Using the network configuration, the user has to integrate the IP address of the VM as a DICOM receiver in the PACS first. *LAB–QA2GO* runs the Conquest tool as receiving process to receive the data from the local setup, i.e., either the MRI camera, the PACS, etc., and stores them in the VM. Using the stand-alone configuration, the user has to copy the data manually to the VM. This can be done using, e.g., a USB-Stick or a shared folder with the host system (provided by the virtualization software). In the stand-alone configuration, the VM can handle both DICOM and NIfTI format data. The user has to specify the path to the data in the provided web interface and then just press start. If the data is present as DICOM files, then the DICOM data send process is started to transfer the DICOM files to the conquest tool, to run the same routine as described above. If the data is present in NIfTI format, the data is copied into the temporal folder and the same routine is started without converting the data.

After the data is available in the *LAB–QA2GO* tool, the main script for analysis is either started automatically at a chosen time point or can be started manually by pressing a button in the web interface. The data processing is visualized in **Figure 5**. First, the data is copied into a temporal folder. Data processing is performed on NIfTI formatted data. If the data is in DICOM format, it will be converted into NIfTI format using the dcm2nii tool. Second, the names of the NIfTI files are compared to the predefined unique identifiers. If the name of the NIfTI data partly matches with a predefined identifier, then the corresponding QA

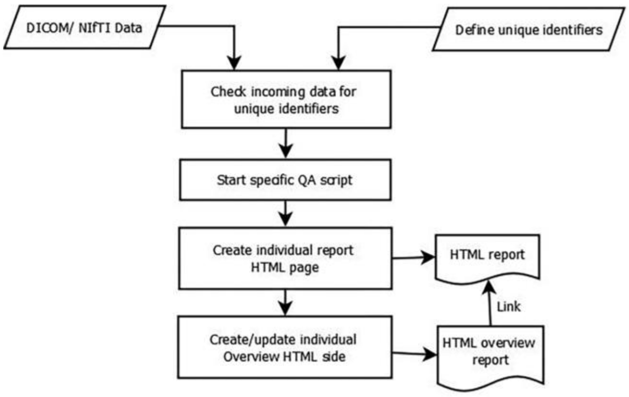
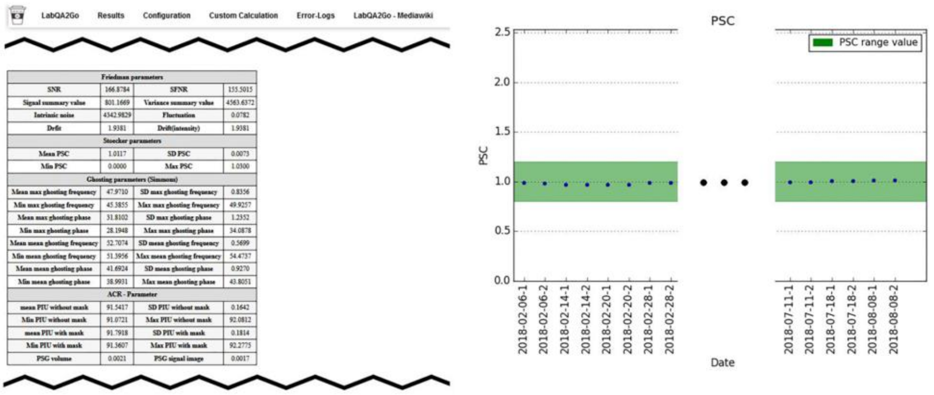

DICOM/ NIfTI Data
Define unique identifiers
Check incoming data for unique identifiers
Start specific QA script
Create individual report HTML page
HTML report
Link
Create/update individual Overview HTML side
HTML overview report

FIGURE 5 | Data flow diagram of the QA analysis.



Friedman parameters			
SNR	166.8784	SFNR	155.5015
Signal summary value	801.1669	Variance summary value	4563.6372
Intrinsic noise	4342.9829	Fluctuation	0.0782
Drfit	1.9381	Drift(intensity)	1.9381
Stoecker parameters			
Mean PSC	1.0117	SD PSC	0.0073
Min PSC	0.0000	Max PSC	1.0300
Ghosting parameters (Simmons)			
Mean max ghosting frequency	47.9710	SD max ghosting frequency	0.8356
Min max ghosting frequency	45.3855	Max max ghosting frequency	49.9257
Mean max ghosting phase	31.8102	SD max ghosting phase	1.2352
Min max ghosting phase	28.1948	Max max ghosting phase	34.0878
Mean mean ghosting frequency	52.7074	SD mean ghosting frequency	0.5699
Min mean ghosting frequency	51.3956	Max mean ghosting frequency	54.4737
Mean mean ghosting phase	41.6924	SD mean ghosting phase	0.9270
Min mean ghosting phase	38.9931	Max mean ghosting phase	43.8051
ACR - Parameter			
mean PIU without mask	91.5417	SD PIU without mask	0.1642
Min PIU without mask	91.0721	Max PIU without mask	92.0812
mean PIU with mask	91.7918	SD PIU with mask	0.1814
Min PIU with mask	91.3607	Max PIU with mask	92.2775
PSG volume	0.0021	PSG signal image	0.0017

FIGURE 6 | Excerpt from the results page summarizing the results of a QA analysis of a gel phantom. **Left**: summary of all QA statistics measured at a specific time point. **Right**: overview of the percent signal change (PSC) of the MRI signal over a period of 6 months. This graphic shows stable MRI scanner performance.

routine is started (e.g., gel phantom analysis; see section "QA Pipelines for Phantom and for Human MRI Data").

Third, after each calculation step, a HTML file for the analyzed dataset is generated. In this file, the results of the analysis are presented (e.g., the movement graphs for functional human datasets). In **Figure 6**, we show an exemplary file for the analysis of gel phantom data. Furthermore, an overview page for each analysis type is generated or updated. On this overview page, the calculated parameters of all measurements of one data type are presented as a graph. An individual acceptance range can be defined using the configuration page, which is visible in the graph. Additionally, all individual measurement result pages are linked at the bottom of the page for a detailed overview. Outliers (defined by either an automatically calculated or self-defined acceptance range) are highlighted to detect them easily.

APPLICATION SCENARIO: QUALITY ASSURANCE OF AN MRI SCANNER

There are many possible application scenarios for the *LAB–QA2GO* tool. It can be used, for instance, to assess the quality of MRI data sets acquired in specific neuroimaging studies (e.g., Frässle et al., 2016) or to compare MRI scanners in multicenter

imaging studies (e.g., Vogelbacher et al., 2018). In this section we will describe another application scenario in which the *LAB–QA2GO* tool is used to assess the long-term performance of one MRI scanner ("center-specific QA"). We will illustrate this scenario using data from our MRI lab at the University of Marburg. The aim of this QA is not to assess the quality of MRI data collected in a specific study, but to provide continuously information on the stability of the MRI scanner across studies.

Center-Specific QA Protocol

The assessment of MRI scanner stability at our MRI lab is based on regular measurements of both the ACR phantom and a gel phantom. The phantoms are measured at fix time points. The ACR phantom is measured every Monday and Friday, the gel phantom each Wednesday. All measurements are performed at 8 a.m., as first measurement of the day. For calculating the QA statistics, the *LAB–QA2GO* tool is used in the network configuration. As unique identifiers (see section "Technical Implementation of *LAB–QA2GO*"), we determined that all phantom measurements must contain the keywords "phantom" and either "GEL" or "ACR" in the "patient name." If these keywords are detected by *LAB–QA2GO*, the processing pipelines for the gel phantom analysis and the ACR phantom analysis, respectively, are started automatically. In the following, we describe the phantoms and the MRI protocol in detail. We also present examples how the QA protocol can be used to assess the stability of the MRI scanner.

Gel Phantom

The gel phantom is 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, 2013). Furthermore, gel phantoms are less vulnerable to scanner vibrations and thus avoid a long settling time prior to data acquisition (Friedman and Glover, 2006). For the gel phantom, we chose MR sequences that allowed to assess the temporal stability of the MRI data. This stability is in particular important for fMRI studies in which MRI scanners are typically operated close to their load limits. The MRI acquisition protocol consists of a localizer, a structural T1-weighted sequence, a T2*-weighted echo planar imaging (EPI) sequence, a diffusion tensor imaging (DTI) sequence, another fast T2*-weighted EPI sequence and, finally, the same T2*-weighted EPI sequence as at the beginning. The comparison of the quality of the first and the last EPI sequence allows in particular to assess the impact of a highly stressed MRI scanner on the imaging data. The MRI parameters of all sequences are listed in **Table 1**.

ACR Phantom

The ACR phantom is a commonly used phantom for QA. It uses a standardized imaging protocol with standardized MRI parameters (for an overview, see ACR, 2005, 2008). The protocol tests geometric accuracy, high-contrast spatial resolution, slice thickness accuracy, slice position accuracy, image intensity uniformity, percent-signal ghosting, and low-contrast object detectability.

Phantom Holder

At the beginning, both phantoms were manually aligned in the scanner and fixated using soft foam rubber pads. The alignment of the phantoms was evaluated by the radiographer performing the measurement and – if necessary – corrected using the localizer scan. 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^{TM} phantom holder (**Figure 7**). The phantom holder allowed a more time-efficient and standardized alignment of the phantoms within the scanner on the one hand. The measurement volumes of subsequent MR sequences could be placed automatically in the center of the phantom. On the other hand, the variability of QA statistics, related to different phantom mountings, was strongly reduced. This allowed a more sensitive assessment of MRI scanner stability (see **Figure 8**, left).

In **Figure 8**, we present selected QA data (from the gel phantom) collected over a duration of 22 months (February 2015–December 2016) during the set-up of a large longitudinal imaging study (FOR, 2107, Kircher et al., 2018). The analysis of phantom data is able to show that changes in the QA-protocol (such as the introduction of a phantom holder, **Figure 8A**), technical changes of a scanner (such as the replacement of the MRI gradient coil, **Figure 8B**) or changes in certain sequence parameters (such as adding the prescan normalization option, **Figure 8C**), impact many of the QA statistics in a variety of ways. It is also possible to use QA statistics to quantify the data quality of different MRI scanners (**Figure 8D**). In summary, this exemplary selection of data shows the importance of QA analyses to assess the impact external events on the MRI data. The normal ranges of many QA statistics drastically change whenever hardware or software settings are changed at a scanner – both in mean and variance.

DISCUSSION

In this article, we described a tool, *LAB–QA2GO*, for the fully automatic quality assessment of MRI data. We developed two different types of QA analyses, a phantom and a human data QA pipeline. In its present implementation, *LAB–QA2GO* is able to run an automated QA analysis on data of ACR phantoms and gel phantoms. The ACR phantom is a widely used phantom for QA of MRI data. It tests in particular spatial properties, e.g., geometric accuracy, high-contrast spatial resolution or slice thickness accuracy. The gel phantom is mainly used to assess the temporal stability of the MRI data. For phantom data analysis, we used a wide array of previously described QA statistics (for an overview see, e.g., Glover et al., 2012). Although the main focus of the QA routines was the analysis of the phantom datasets, we additionally developed routines to analyze the quality of human datasets (without any pre-processing steps). *LAB–QA2GO* was developed in a modular fashion, making it easily possible to

TABLE 1 | Magnetic resonance imaging parameters for the gel phantom measurements.

Sequence (position in QA protocol)	Localizer (1)	T1 (2)	Bold sensitive EPI (3, 6)	Diffusion sensitive EPI (4)	Bold sensitive EPI (5)
Repetition time (TR)	8.6 ms	1900 ms	2000 ms	3900 ms	177 ms
Echo time (TE)	4 ms	2.26 ms	30 ms	90 ms	30 ms
Field of view (FoV)	250 mm	256 mm	210 mm	256 mm	210 mm
Matrix size	256 × 256	256 × 256	64 × 64	128 × 128	64 × 64
Slice thickness	7.0 mm	1.0 mm	3.0 mm	2.0 mm	3.0 mm
Distance factor		50%	20%	0%	20%
Flip angle	20°	9°	70°		70°
Phase encoding direction	Anterior > > posterior, anterior > > posterior, right > > left	Anterior > > posterior	Anterior > > posterior	Anterior > > posterior	Anterior > > posterior
Bandwidth (Hz/Px)	320	200	2894	1502	2894
Acquisition order (series)	Sequential (interleaved)	Single shot (ascending)	Interleaved (interleaved)	Interleaved (interleaved)	Interleaved (interleaved)
Number of slices	Two in each direction	176	34	30	3
Measurements	1	1	200		322
Effective voxel size (mm)	1.0 × 1.0 × 7.0	1.0 × 1.0 × 1.0	3.3 × 3.3 × 3.0	2.0 × 2.0 × 2.0	3.3 × 3.3 × 3.0
Acquisition time (TA)	0.25	4:26	6:44	2:14	1:00

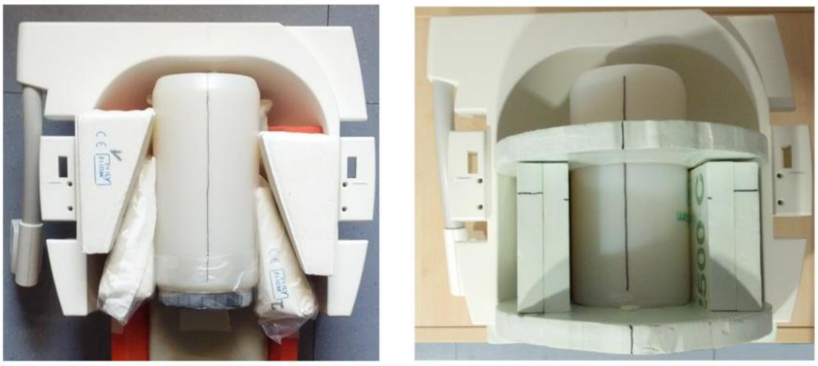


FIGURE 7 | Manual alignment of the gel phantom using soft foam rubber pads **(left)** and more reliable alignment of the phantom using a Styrofoam holder **(right)**.

modify existing algorithms and to extend the QA analyses by adding self-designed routines. The tool is available for download on github[17]. License fees were avoided by using only open source software that was exempt from charges.

LAB–QA2GO is ready-to-use in about 10 min. Only a few configuration steps have to be performed. The tool does not need any further software or hardware requirements. *LAB–QA2GO* can receive MRI data either automatically ("network approach") or manually ("stand-alone approach"). After sending data to the *LAB–QA2GO* tool, analysis of MRI data is performed automatically. All results are presented in an easy readable and easy-to-interpret web based format. The simple access via web-browser guarantees a user friendly usage without any specific IT knowledge as well as the minimalistic maintenance work of the tool. Results are presented both tabular and in graphical form. By inspecting the graphics on the overview page, the user is able to detect outliers easily. Potential outliers are highlighted by a warning sign. In each overview graph, an acceptance range

[17] Github: https://github.com/vogelbac.

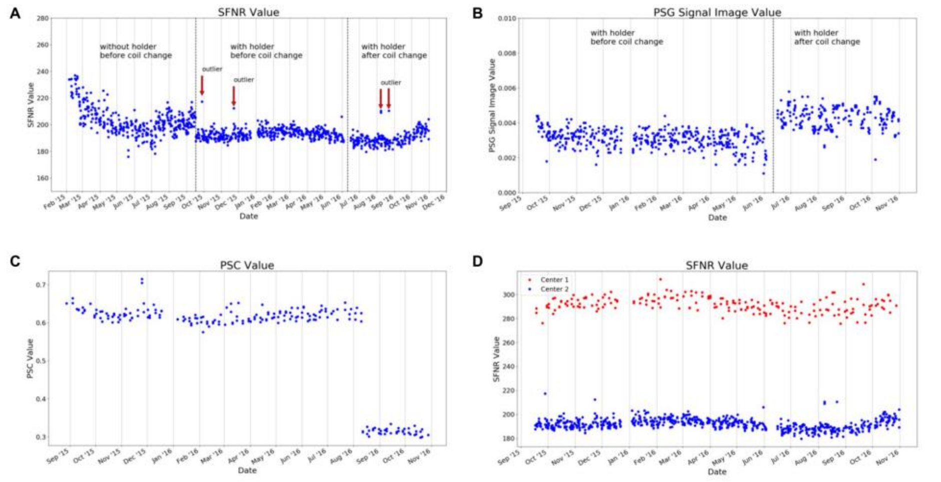


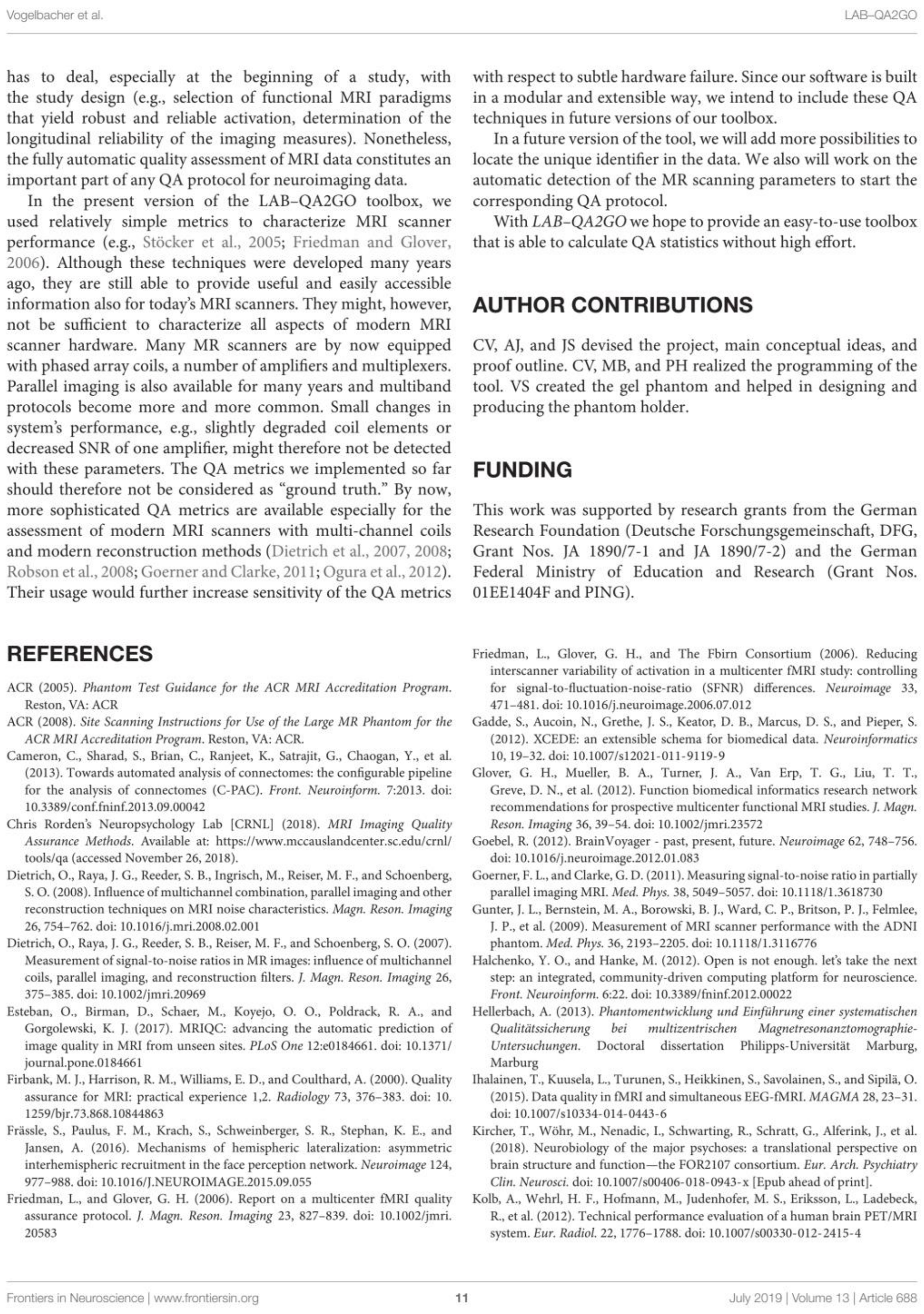
FIGURE 8 | Selected QA statistics from gel phantom measurements. The data was collected over a duration of >1.5 years (February 2015–December 2016) during the set-up of a large longitudinal imaging study (FOR, 2107, Kircher et al., 2018). **(A)** After the implementation of a phantom holder (October 2015), the variance of many QA statistics was considerably reduced, as exemplarily shown for the signal-to-fluctuation-noise ratio (SFNR). This made it possible to detect outliers in future measurements (defined as four times SD of the mean; red arrows). **(B)** In June 2016, the gradient coil had to be replaced. This had a major impact on the percent signal ghosting (PSG). **(C)** Changes in the MRI sequence, such as the introduction of the "prescan normalization" option that is used to make corrections for non-uniform receiver coil profiles prior to imaging, has a significant impact on the MRI data. This can be quantified using phantom data, as seen in the PSC. **(D)** Imaging data in the study was collected at two different scanners. The scanner characteristics can also be determined using QA statistics, as shown for the SFNR [for a detailed description of QA statistics, see Vogelbacher et al. (2018)].

(green area) is viable. This area can be defined for each graph individually (except for the ACR phantom because of the fixed acceptance values defined by the ACR protocol). To set up the acceptance range for a specific MRI scanner, we recommend some initial measurements to define the acceptance range. If a measurement is not in this range this might indicate performance problems of the MRI scanner.

Different QA protocols that assess MRI scanner stability are described in the literature, mostly designed for large-scale multicenter studies (for an overview see, e.g., Glover et al., 2012). Many of these protocols and the corresponding software tools are openly available. This allows in principle the flexible set-up of a QA protocol adapted for specific studies. The installation of these routines, however, is often not easy. The installation therefore often requires a fair level of technical experience, e.g., to install additional image processing software or to deal with specific software versions or hardware requirements. *LAB–QA2GO* was therefore developed with the aim to create an easily applicable QA tool. It provides on the one hand a fully automated QA pipeline, but is on the other hand easy to install on most imaging systems. Therefore, we envision that the tool might be a tailor-made solution for users without a strong technical background or for MRI laboratories without support of large core-facilities. Moreover, it also gives experienced users a minimalistic tool to easily calculate QA statistics for specific studies.

We outlined several possible application scenarios for the *LAB–QA2GO* tool. It can be used to assess the quality of MRI data sets acquired in small (with regard to sample size and study duration) neuroimaging studies, to standardize MRI scanners in multicenter imaging studies or to assess the long-term performance of MRI scanners. We outlined the use of the tool presenting data from center-specific QA protocol. These data showed that it was possible to detect outliers (i.e., bad data quality at some time points), to standardize MRI scanner performance and to evaluate the impact of hardware and software adaptation (e.g., the installation of a new gradient coil).

In the long run, the successful implementation of a QA protocol for imaging data does not only comprise the assessment of MRI data quality. QA has to be implemented on many different levels. A comprehensive QA protocol also has to encompass technical issues (e.g., monitoring of the temporal stability of the MRI signal in particular after hardware and software upgrades, use of secure database infrastructure that can store, retrieve, and monitor all collected data, documentation of changes on the MRI environment for instance with regard to scanner hardware, software updates) and should optimize management procedures (e.g., the careful coordination and division of labor, the actual data management, the long-term monitoring of measurement procedures, the compliance with regulations on data anonymity, the standardization of MRI measurement procedures). It also



has to deal, especially at the beginning of a study, with the study design (e.g., selection of functional MRI paradigms that yield robust and reliable activation, determination of the longitudinal reliability of the imaging measures). Nonetheless, the fully automatic quality assessment of MRI data constitutes an important part of any QA protocol for neuroimaging data.

In the present version of the LAB–QA2GO toolbox, we used relatively simple metrics to characterize MRI scanner performance (e.g., Stöcker et al., 2005; Friedman and Glover, 2006). Although these techniques were developed many years ago, they are still able to provide useful and easily accessible information also for today's MRI scanners. They might, however, not be sufficient to characterize all aspects of modern MRI scanner hardware. Many MR scanners are by now equipped with phased array coils, a number of amplifiers and multiplexers. Parallel imaging is also available for many years and multiband protocols become more and more common. Small changes in system's performance, e.g., slightly degraded coil elements or decreased SNR of one amplifier, might therefore not be detected with these parameters. The QA metrics we implemented so far should therefore not be considered as "ground truth." By now, more sophisticated QA metrics are available especially for the assessment of modern MRI scanners with multi-channel coils and modern reconstruction methods (Dietrich et al., 2007, 2008; Robson et al., 2008; Goerner and Clarke, 2011; Ogura et al., 2012). Their usage would further increase sensitivity of the QA metrics with respect to subtle hardware failure. Since our software is built in a modular and extensible way, we intend to include these QA techniques in future versions of our toolbox.

In a future version of the tool, we will add more possibilities to locate the unique identifier in the data. We also will work on the automatic detection of the MR scanning parameters to start the corresponding QA protocol.

With *LAB–QA2GO* we hope to provide an easy-to-use toolbox that is able to calculate QA statistics without high effort.

AUTHOR CONTRIBUTIONS

CV, AJ, and JS devised the project, main conceptual ideas, and proof outline. CV, MB, and PH realized the programming of the tool. VS created the gel phantom and helped in designing and producing the phantom holder.

FUNDING

This work was supported by research grants from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, Grant Nos. JA 1890/7-1 and JA 1890/7-2) and the German Federal Ministry of Education and Research (Grant Nos. 01EE1404F and PING).

REFERENCES

ACR (2005). *Phantom Test Guidance for the ACR MRI Accreditation Program*. Reston, VA: ACR

ACR (2008). *Site Scanning Instructions for Use of the Large MR Phantom for the ACR MRI Accreditation Program*. Reston, VA: ACR.

Cameron, C., Sharad, S., Brian, C., Ranjeet, K., Satrajit, G., Chaogan, Y., et al. (2013). Towards automated analysis of connectomes: the configurable pipeline for the analysis of connectomes (C-PAC). *Front. Neuroinform.* 7:2013. doi: 10.3389/conf.fninf.2013.09.00042

Chris Rorden's Neuropsychology Lab [CRNL] (2018). *MRI Imaging Quality Assurance Methods*. Available at: https://www.mccauslandcenter.sc.edu/crnl/tools/qa (accessed November 26, 2018).

Dietrich, O., Raya, J. G., Reeder, S. B., Ingrisch, M., Reiser, M. F., and Schoenberg, S. O. (2008). Influence of multichannel combination, parallel imaging and other reconstruction techniques on MRI noise characteristics. *Magn. Reson. Imaging* 26, 754–762. doi: 10.1016/j.mri.2008.02.001

Dietrich, O., Raya, J. G., Reeder, S. B., Reiser, M. F., and Schoenberg, S. O. (2007). Measurement of signal-to-noise ratios in MR images: influence of multichannel coils, parallel imaging, and reconstruction filters. *J. Magn. Reson. Imaging* 26, 375–385. doi: 10.1002/jmri.20969

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One* 12:e0184661. doi: 10.1371/journal.pone.0184661

Firbank, M. J., Harrison, R. M., Williams, E. D., and Coulthard, A. (2000). Quality assurance for MRI: practical experience 1,2. *Radiology* 73, 376–383. doi: 10.1259/bjr.73.868.10844863

Frässle, S., Paulus, F. M., Krach, S., Schweinberger, S. R., Stephan, K. E., and Jansen, A. (2016). Mechanisms of hemispheric lateralization: asymmetric interhemispheric recruitment in the face perception network. *Neuroimage* 124, 977–988. doi: 10.1016/J.NEUROIMAGE.2015.09.055

Friedman, L., and Glover, G. H. (2006). Report on a multicenter fMRI quality assurance protocol. *J. Magn. Reson. Imaging* 23, 827–839. doi: 10.1002/jmri.20583

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

Gadde, S., Aucoin, N., Grethe, J. S., Keator, D. B., Marcus, D. S., and Pieper, S. (2012). XCEDE: an extensible schema for biomedical data. *Neuroinformatics* 10, 19–32. doi: 10.1007/s12021-011-9119-9

Glover, G. H., Mueller, B. A., Turner, J. A., Van Erp, T. G., Liu, T. T., Greve, D. N., et al. (2012). Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. *J. Magn. Reson. Imaging* 36, 39–54. doi: 10.1002/jmri.23572

Goebel, R. (2012). BrainVoyager - past, present, future. *Neuroimage* 62, 748–756. doi: 10.1016/j.neuroimage.2012.01.083

Goerner, F. L., and Clarke, G. D. (2011). Measuring signal-to-noise ratio in partially parallel imaging MRI. *Med. Phys.* 38, 5049–5057. doi: 10.1118/1.3618730

Gunter, J. L., Bernstein, M. A., Borowski, B. J., Ward, C. P., Britson, P. J., Felmlee, J. P., et al. (2009). Measurement of MRI scanner performance with the ADNI phantom. *Med. Phys.* 36, 2193–2205. doi: 10.1118/1.3116776

Halchenko, Y. O., and Hanke, M. (2012). Open is not enough. let's take the next step: an integrated, community-driven computing platform for neuroscience. *Front. Neuroinform.* 6:22. doi: 10.3389/fninf.2012.00022

Hellerbach, A. (2013). *Phantomentwicklung und Einführung einer systematischen Qualitätssicherung bei multizentrischen Magnetresonanztomographie-Untersuchungen*. Doctoral dissertation Philipps-Universität Marburg, Marburg

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

Kircher, T., Wöhr, M., Nenadic, I., Schwarting, R., Schratt, G., Alferink, J., et al. (2018). Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *Eur. Arch. Psychiatry Clin. Neurosci.* doi: 10.1007/s00406-018-0943-x [Epub ahead of print].

Kolb, A., Wehrl, H. F., Hofmann, M., Judenhofer, M. S., Eriksson, L., Ladebeck, R., et al. (2012). Technical performance evaluation of a human brain PET/MRI system. *Eur. Radiol.* 22, 1776–1788. doi: 10.1007/s00330-012-2415-4

Lee, G. R., Rajagopal, A., Felicelli, N., Rupert, A., Wagner, M., et al. (2014). cmind-py: a robust set of processing pipelines for pediatric fMRI in *Proceedings of the 20th Annual Meeting of the Organization for Human Brain Mapping*. Hamburg

Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007). The extensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 5, 11–34. doi: 10.1385/NI

Mazaika, P. K., Hoeft, F., Glover, G. H., and Reiss, A. L. (2009). Methods and software for fMRI analysis of clinical subjects. *Neuroimage* 47:S58.

Ogura,A., Miyati, T., Kobayashi, M., Imai, H., Shimizu, K., Tsuchihashi, T., et al. (2012). Method of SNR determination using clinical images. *Japanese J. Radiol. Technol.* 63, 1099–1104. doi: 10.6009/jjrt.63.1099

Petrosyan, P., Hobel, S., Irimia, A., and John Van Horn, A. T. (2016). *LONI QC: a System for the Quality Control of Structural, Functional and Diffusion Brain Images*. Available at: https://qc.loni.usc.edu/

Robson, P. M., Grant, A. K., Madhuranthakam, A. J., Lattanzi, R., Sodickson, D. K., and McKenzie, C. A. (2008). Comprehensive quantification of signal-to-noise ratio and g-factor for image-based and k-space-based parallel imaging reconstructions. *Magn. Reson. Med.* 60, 895–907. doi: 10.1002/mrm.21728

Simmons, A., Moore, E., and 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. doi: 10.1002/(sici)1522-2594(199906)41:6<1274::aid-mrm27>3.3.co;2-t

Stöcker, T., Schneider, F., Klein, M., Habel, U., Kellermann, T., Zilles, K., et al. (2005). Automated quality assurance routines for fMRI data applied to a multicenter study. *Hum. Brain Mapp.* 25, 237–246. doi: 10.1002/hbm.20096

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

Vogelbacher, C., Möbius, T. W. D., Sommer, J., Schuster, V., Dannlowski, U., Kircher, T., et al. (2018). The marburg-münster affective disorders cohort study (MACS): a quality assurance protocol for MR neuroimaging data. *Neuroimage* 172, 450–460. doi: 10.1016/j.neuroimage.2018.01.079

Zarrar, S., Steven, G., Qingyang, L., Yassine, B., Chaogan, Y., Zhen, Y., et al. (2015). The preprocessed connectomes project quality assessment protocol - a resource for measuring the quality of MRI data. *Front. Neurosci.* 9:47. doi: 10.3389/conf.fnins.2015.91.00047

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CURRICULUM VITAE

Die Seiten 68-71 (Curriculum Vitae) enthalten persönliche Daten. Sie sind deshalb nicht Bestandteil der Online-Veröffentlichung.

Pages 68-71 (Curriculum Vitae) contain personal information. They are therefore not part of the online publication.

VERZEICHNIS DER AKADEMISCHEN LEHRER/-INNEN

Meine akademischen Lehrenden an der Philipps-Universität in Marburg waren:

Fachbereich Informatik:		
Dohmann	Freisleben	Gumm
Guthe	Hesse	Loogen
Ostermann	Seeger	Sommer
Taenzer	Ultsch	
Fachbereich Mathematik:		
Hinz	Hüllermeier	Upmeier
Fachbereich Biologie:		
Bremer	Hassel	Homberg
Mösch		
Fachbereich Wirtschaftswissenschaften:		
Lingenfelder	Stephan	

DANKSAGUNG

Die Seiten 73-74 (Danksagung) enthalten persönliche Daten. Sie sind deshalb nicht Bestandteil der Online-Veröffentlichung.

Pages 73-74 (acknowledgement) contain personal information. They are therefore not part of the online publication.

EHRENWÖRTLICHE ERKLÄRUNG

Ich erkläre ehrenwörtlich, dass ich die dem Fachbereich Medizin Marburg zur Promotionsprüfung eingereichte Arbeit mit dem Titel *„Development of quality standards for multi-center, longitudinal magnetic resonance imaging studies in clinical neuroscience“* in der Klinik für Psychiatrie und Psychotherapie unter Leitung von Prof. Dr. Tilo Kircher mit Unterstützung durch Prof. Dr. Andreas Jansen und PD Dr. Sommer ohne sonstige Hilfe selbst durchgeführt und bei der Abfassung der Arbeit keine anderen als die in der Dissertation aufgeführten Hilfsmittel benutzt habe. Ich habe bisher an keinem in- oder ausländischen Medizinischen Fachbereich ein Gesuch um Zulassung zur Promotion eingereicht, noch die vorliegende oder eine andere Arbeit als Dissertation vorgelegt.

Ich versichere, dass ich sämtliche wörtlichen oder sinngemäßen Übernahmen und Zitate kenntlich gemacht habe.

Mit dem Einsatz von Software zur Erkennung von Plagiaten bin ich einverstanden.

Vorliegende Arbeit wurde in folgenden Publikationsorganen NeuroImage mit dem Titel *„The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data*“und Frontiers in Neuroscience mit dem Titel *„A Free, Easy-to-Use Toolbox for the Quality Assessment of Magnetic Resonance Imaging Data*“ veröffentlicht.

Marburg, 29.10.2019,

Ort, Datum, Unterschrift Christoph Vogelbacher

„Die Hinweise zur Erkennung von Plagiaten habe ich zur Kenntnis genommen.“

Marburg, 29.10.2019,

Ort, Datum, Unterschrift Prof. Dr. Andreas Jansen

EIGENER ANTEIL AN DIESER ARBEIT

Laut §8, Absatz 3 der Promotionsordnung der Philipps-Universität Marburg (Fassung vom 15.07.2009) müssen bei den Teilen der Dissertation, die aus gemeinsamer Forschungsarbeit entstanden sind, „die individuellen Leistungen des Doktoranden deutlich abgrenzbar und bewertbar sein“. Dies betrifft die Manuskripte 1 und 2, und wird im Folgenden detailliert erläutert.

Manuskript 1: ***The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data***

Folgende Punkte wurden als eigener Anteil zum Manuskript beigetragen:

- Konzept für Qualitätssicherungsprotokoll erstellt
- Entwicklung der Phantomhalterung
- Analyse der Human- und Phantomdaten
- Skripte zur Auswertung geschrieben
- Anfertigung aller Abbildungen und Tabellen
- Anfertigung des Manuskriptes (Korrektur durch Dr. Möbius, Prof. Dr. Jansen und Dr. Bopp)

Anteil gesamt: 70%

Dieses Manuskript wurde in der vorliegenden Form im Journal *NeuroImage* veröffentlicht:

Vogelbacher, C., Möbius, T. W., Sommer, J., Schuster, V., Dannlowski, U., Kircher, T., ... & Bopp, M. H. (2018). *The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data*. NeuroImage, 172, 450-460.

Manuskript 2: ***LAB–QA2GO: A Free, Easy-to-Use Toolbox for the Quality Assessment of Magnetic Resonance Imaging Data***

Folgende Punkte wurden als eigener Anteil zum Manuskript beigetragen:

- Entwicklung des Konzepts (Unterstützung durch Dr. Sommer)
- Vollständige Implementierung des Tools
- Entwicklung der Zentrumsspezifischen Qualitätssicherung
- Anfertigung aller Abbildungen und Tabellen
- Anfertigung des Manuskriptes (Korrektur durch DR. Sommer und Prof. Dr. Jansen)

Anteil gesamt: 80%

Dieses Manuskript wurde in vorliegender Form im Journal *Frontiers in Neuroscience* veröffentlicht:

Vogelbacher, C., Bopp, M. H. A., Schuster, V., Herholz, P., Jansen, A., & Sommer, J.. 2019. *LAB–QA2GO: A Free, Easy-to-Use Toolbox for the Quality Assessment of Magnetic Resonance Imaging Data.* Frontiers in Neuroscience 13 (July). Frontiers: 688. doi:10.3389/fnins.2019.00688.

Unterschrift
Prof. Dr. Andreas Jansen

Unterschrift
Christoph Vogelbacher