RESEARCH ARTICLE

A new virtue of phantom MRI data: explaining variance in human participant data [version 1; peer review: 1 approved, 2 approved with reservations, 1 not approved]

Christopher P. Cheng [1], Yaroslav O. Halchenko [2]

[1]Dartmouth College, Hanover, NH, 03755, USA
[2]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, 03755, USA

v1 **First published:** 14 Sep 2020, **9**:1131
https://doi.org/10.12688/f1000research.24544.1
Latest published: 14 Sep 2020, **9**:1131
https://doi.org/10.12688/f1000research.24544.1

Abstract

Background: Magnetic resonance imaging (MRI) is an important yet complex data acquisition technology for studying the brain. MRI signals can be affected by many factors and many sources of variance are often simply attributed to "noise". Unexplained variance in MRI data hinders the statistical power of MRI studies and affects their reproducibility. We hypothesized that it would be possible to use phantom data as a proxy of scanner characteristics with a simplistic model of seasonal variation to explain some variance in human MRI data.

Methods: We used MRI data from human participants collected in several studies, as well as phantom data collected weekly for scanner quality assurance (QA) purposes. From phantom data we identified the variables most likely to explain variance in acquired data and assessed their statistical significance by using them to model signal-to-noise ratio (SNR), a fundamental MRI QA metric. We then included phantom data SNR in the models of morphometric measures obtained from human anatomical MRI data from the same scanner.

Results: Phantom SNR and seasonal variation, after multiple comparisons correction, were statistically significant predictors of the volume of gray brain matter. However, a sweep over 16 other brain matter areas and types revealed no statistically significant predictors among phantom SNR or seasonal variables after multiple comparison correction.

Conclusions: Seasonal variation and phantom SNR may be important factors to account for in MRI studies. Our results show weak support that seasonal variations are primarily caused by biological human factors instead of scanner performance variation. The phantom QA metric and scanning parameters are useful for more than just QA. Using QA metrics, scanning parameters, and seasonal variation data can help account for some variance in MRI studies, thus making them more powerful and reproducible.

Open Peer Review

Reviewer Status ? ✓ ✗ ?

	Invited Reviewers			
	1	**2**	**3**	**4**
version 1 14 Sep 2020	? report	✓ report	✗ report	? report

1. **Xiangrui Li**, The Ohio State University, Columbus, USA
2. **Blaise Frederick**, McLean Hospital, Belmont, USA
Harvard Medical School, Boston, USA
3. **Simon Duchesne**, Université Laval, Quebec City, Canada
4. **Jens Sommer**, Philipps-University Marburg, Marburg, Germany
Philipps-University Marburg, Marburg, Germany
Philipps-University Marburg, Marburg, Germany

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
Neuroimaging, seasonal variation, MRI QA, MRI, reproducibility

This article is included in the INCF gateway.

Corresponding authors: Christopher P. Cheng (cheng1928c@gmail.com), Yaroslav O. Halchenko (yoh@dartmouth.edu)

Author roles: Cheng CP: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Halchenko YO**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: YOH was supported by NIH-NIBIB [P41EB019936]. CPC was supported by Dartmouth College's Sophomore Research Scholarship. The datasets from Dartmouth College's Psychological and Brain Sciences Department Chang and Haxby Labs were acquired as part of the work supported by NIH [R01MH116026, R56MH080716, and R01MH116026]. Gobbini's lab's data acquisition was supported by NSF [1835200].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Cheng CP and Halchenko YO. **A new virtue of phantom MRI data: explaining variance in human participant data [version 1; peer review: 1 approved, 2 approved with reservations, 1 not approved]** F1000Research 2020, **9**:1131 https://doi.org/10.12688/f1000research.24544.1

First published: 14 Sep 2020, **9**:1131 https://doi.org/10.12688/f1000research.24544.1

Introduction

Magnetic resonance imaging (MRI) is an important data acquisition technology used to unravel the mysteries of the brain, and it is very complex. The exact constitution of MRI signals is not entirely known since it could also potentially be affected by factors such as temperature and humidity variations across seasons[1]. Notably, a study by Meyer *et al.* noted that seasonal variations may not even correspond to the four seasons, indicating a complicated relationship between environmental factors and brain functions[2]; thus, whether or not this variance in MRI is due to scanner effects or biological causes is unclear. This unexplained variance in MRI data hinders the statistical power of MRI studies and affects their reproducibility.

MRI quality assurance (QA) metrics are indicators of the condition of the scanner at the time of a given scan, and are used for quality control in MRI centers[3]. In cases of significant deviation from the norm, MRI personnel look into resolving underlying hardware or software issues. Otherwise, QA results are not used for anything else, and not shared alongside large shared datasets, such as Human Connectome Project (HCP)[4] or the ABCD study where data is acquired across different scanners and potentially affected by scanner idiosyncrasies. It is typically unknown how seasonal and operational factors affect different types of scanning (on phantom and real subjects). We hypothesize that there may be a relationship between the QA metrics of a scanner (obtained on a phantom) and the characteristics of the MRI scans (on human participants), which affect the consecutive data analysis results and drawn conclusions.

The purpose of this study was to evaluate if phantom data could be used as a useful proxy for overall scanner operational characteristics that can help explain variance in real human subject data acquired using the same MRI scanner on dates nearby phantom QA scans. The first stage of this study analyzed the influence of phantom scanning parameters on signal-to-noise ratio (SNR), which is known to be a fundamental QA metric for MRI. The second stage of this study used SNRs from stage one from phantom data to model morphometric measures obtained using human MRI data.

Methods

Ethical statement

This study was approved by the Dartmouth Committee for the Protection of Human Subjects (CPHS 31408). Data collection in the individual studies was approved by the same committee (CPHS 17763, 28486, 29780, 21200 and 30389). All participants gave written informed consent for participation and re-analysis of data.

Human participants

We used the data from 206 participants (261 scans) collected from October 30, 2017 to August 28, 2018 who participated in five studies[17–20] of three labs at the Dartmouth Brain Imaging Center (DBIC). Participants ranged from 18–64 years of age. There were 78 male and 128 female participants. The MRI scanner used was the 3.0 Tesla Siemens MAGNETOM Prisma whole-body MRI system from Siemens Medical Solutions. Human participant and phantom data were collected using a 32-channel head coil.

Phantom

The DBIC collects MRI QA data weekly (typically each Monday) on an agar phantom. For the purposes of this study we did not use DBIC QA estimates but carried out QA using MRIQC BIDS-App[7].

QA data, converted to a BIDS dataset (including original DICOM data under sourcedata/), is available as a ///dbic/QA DataLad dataset[8]. Subject "qa" within that dataset contains data for the agar phantom used in weekly QA scans. QA scans contain a single T1 weighted anatomical image (192×256×256 matrix at 0.90×0.94×0.94mm) and two functional T2* weighted echo planar imaging (EPI) scans (80×80×30 matrix at 3.00×3.00×3.99mm with 200 volumes acquired with time of repetition (TR) of two seconds; not used in this study).

Data preparation

All data at the DBIC is collected following the ReproIn convention on organizing and naming scanning protocols[9]. To guarantee that the data would not contain variance caused by different conversion software versions through time, data from all phantom and human subjects was reconverted from raw DICOMs into BIDS datasets using consistent versions of ReproIn/HeuDiConv and dcm2niix[10]. All phantom QA data was re-converted using ReproIn/HeuDiConv with dcm2niix (v1.0.20171215 (OpenJPEG build) GCC6.3.0), and human data from different studies re-converted using ReproIn/HeuDiConv (v0.5.3) with dcm2niix (v1.0.20181125 GCC6.3.0). HeuDiConv is programmed to automatically extract many acquisition and scanner operation parameters from DICOMs and place them alongside neuroimaging files in the BIDS dataset. For the purposes of this study, a subset of those parameters was selected as variables of interest for analysis based on prior knowledge regarding which variables could potentially affect the collected data (see Table 1). Furthermore, we added seasonal effects by using NumPy 1.18.4 inserting sine and cosine waves into the model with a period of one year to roughly estimate the four seasons; arguably, this was a very simplistic model due to our data's short duration of under two years. Our data's limited time range precluded us from using more elaborate seasonal models, and as noted in the introduction, seasonal effects may not exactly correspond to the four seasons. Still, we felt a simplistic representation of seasonal effects could help indicate the possibility of further investigation.

We used MRIQC (v0.14.2)[7] on both the QA phantom and the human data from October 30, 2017 to August 28, 2018. MRIQC provided us with proxy measures of scanner operation characteristics, such as total SNR for anatomicals. Figure 1 presents a correlation structure between all variables of interest for phantom MRI data visualized using Seaborn 0.10.1. DataLad[11] with the datalad-container extension[12] was used for

Table 1. Quality assurance metrics of interest in this study, categorically divided by interest.

Category	Variable description (variable name)
MR scanner operation characteristics (outside of operator control)	Transmission amplifier reference amplitude (TxRefAmp)
	Specific absorption rate (SAR)
	Scanner software version (SoftwareVersions)
Acquisition specifics (affected directly or indirectly by operator for any given acquisition)	Day time of acquisition (AcquisitionTime)
	Patient position in the scanner (ImageOrientationPatientDICOM, abbreviated as IOPD)
Proxy measures of scanner operation characteristics (possibly affected by all other variables)	Total signal-to-noise ratio (snr_total)

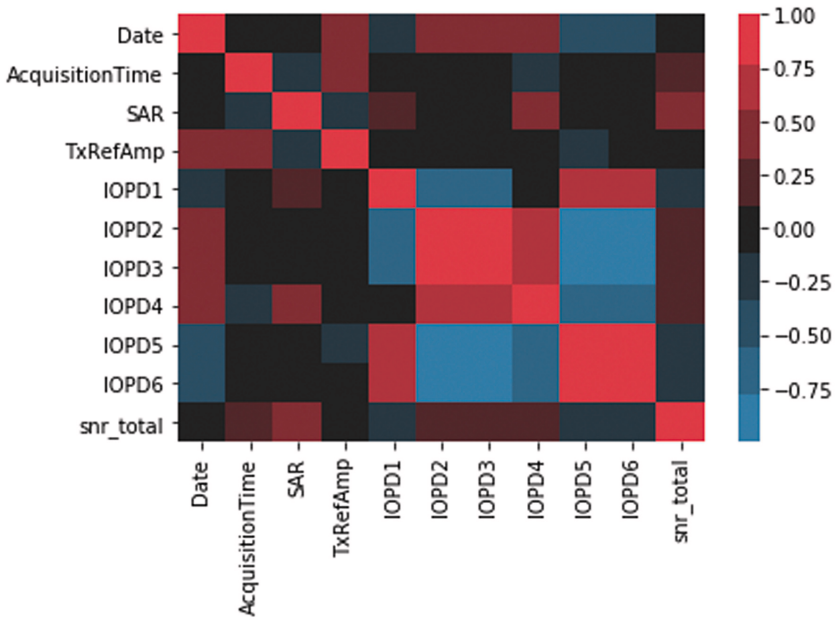

Figure 1. Pearson correlations between different variables on phantom MRI phantom data. Refer to Table 1 for explanation of abbreviated variables.

version control of all digital objects (data, code, singularity containerized environments), and all code and shareable data were made available on GitHub (see *Data availability*[22] and *Code availability*[21]) with containers and data available from the ///con/nuisance DataLad dataset. At the moment we have concentrated on analysis of anatomical data, so only T1w images from phantom and human participants were used.

We used DataLad to run a modified version of the simple_workflow container and script[13], which extracts certain segmentation statistics of the brain from real human MRI data. These include metrics relating to the accumbens area, amygdala, caudate, hippocampus, pallidum, putamen, thalamus proper, cerebrospinal fluid, and the gray and white matter in the brain. The original simple_workflow container is fully reproducible (frozen to the state of NeuroDebian as of 20170410 using nd_freeze) and uses FSL 5.0.9-3~nd80+1.

The free open source software facilitating our data preparation was Pandas 1.0.4.

Data modeling

Ordinary least squares (OLS) regression, as implemented in StatsModels Python package (v 0.9.0)[15], was used to model target variables of interest. As part of the modeling, certain interdependent variables were orthogonalized to account for possible covariance (in the order presented in Figure 1; seasonal variation data was orthogonalized last) using NumPy 1.18.4.

The extent to which an independent variable affects the dependent variable was assessed using a t-test for single-valued variables, and using an F-test for arrayed values (such as patient position in the scanner). Subsequently, the explanatory power of the scanning parameters and characteristics on the phantom QA metric (snr_total) as the dependent variable was evaluated.

Next, a segmentation statistic - gray brain matter on human participant data - was modeled using a proxy QA measure from phantom data (snr_total) and a scanner characteristic of the human participant scanning session (IOPD), demographics (age, gender), and seasonal effects. Gray brain matter was chosen because either gray or white brain matter were deemed likely to yield a statistically significant relationship. Because phantom QA data was acquired typically only each Monday, its value was interpolated in time to obtain values for the dates of human participants scanning. After modeling gray brain matter, other structures (such as white matter, cerebrospinal fluid, and subcortical regions) were analyzed. Our reasoning was that if gray brain matter yielded a significant result, then other brain segmentation statistics could also yield significant results, which could subsequently be investigated.

The free open source software facilitating the visualization of our model was Matplotlib 3.2.1.

Results

Statistical significance of variables

We found that we could describe the total SNR of phantom data well with just a limited set of scanner operational characteristics. The R^2 value of the model shown in Figure 2 was 0.533. Multiple variables (day time of acquisition, subject position, and SAR) were statistically significant and all survived false discovery rate (FDR) correction, as shown in Table 2.

Given that certain scanning parameters and a QA metric were determined to have significant explanatory power in the

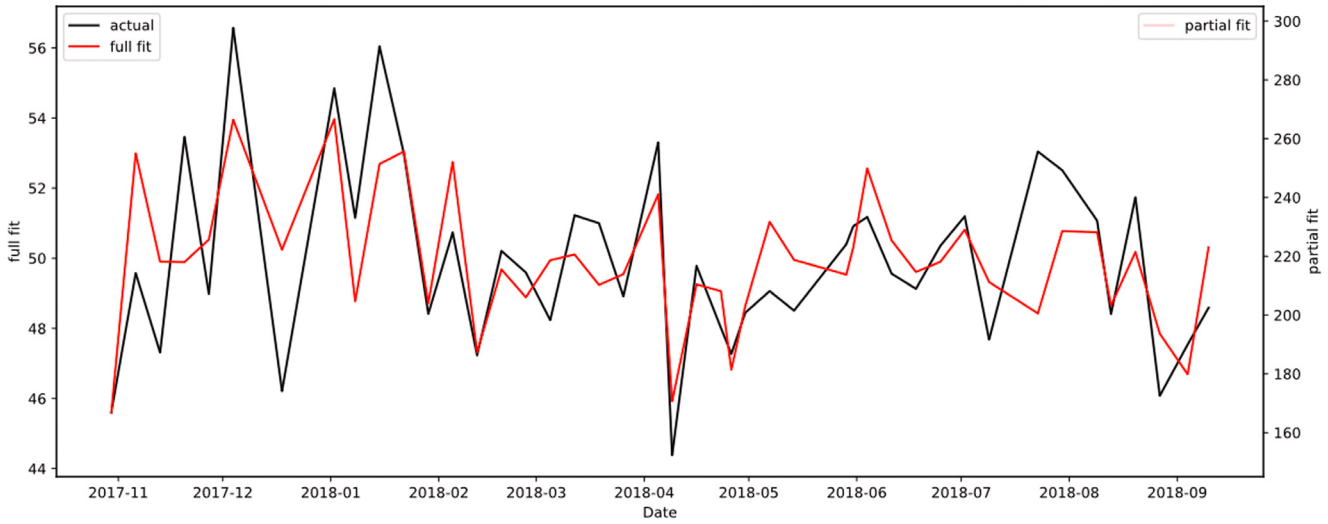

Figure 2. Model of total signal-to-noise ratio using scanning parameters from phantom data. R^2 value was 0.533, indicating scanning parameters have explanatory power.

Table 2. Independent variables' original and corrected p-values from the model in Figure 2. FDR, false discovery rate.

Independent variable	p-value	
	Original	FDR-corrected
Day time of acquisition	**0.019**	**0.031**
Specific absorption rate (SAR)	**0.003**	**0.017**
Transmission amplifier reference amplitude (TxRefAmp)	0.382	0.477
Patient position in the scanner (IOPD)	**0.015**	**0.031**
Seasonal variation	0.669	0.669

phantom scans, we decided to proceed to model human participants' gray brain matter volume using participants' basic demographic data, variables we found significant from phantom data (IOPD), and total SNR. Gray (or white) brain matter was deemed likely to be affected as they are two large "structures".

After correction, the phantom's total SNR was found to be a statistically significant (p=0.002 and FDR-corrected=0.003) predictor of gray brain matter, as shown in Table 3. Other scanning parameters and subject characteristics found significant were subject age, sex, weight, as well as seasonal variation in data. Note that the orthogonalization carried out for the model in Table 3 was carried out in the order shown from top-to-bottom.

After modeling gray brain matter, we modeled the other structures (background, subcortical regions, etc.), and assessed statistical significance for two independent variables of interest in each model: phantom's total SNR, and seasonal variables. Results in Table 4 show no statistically significant results after FDR correction across structures except phantom total SNR in gray brain matter; notably, seasonal variables in gray brain matter were not significant.

Model selection: investigation of QA metric vs seasonal effects

Given that scanning characteristics, a QA metric and seasonal effects seem to have a statistically significant effect on gray brain matter volume estimates (see Table 3), we decided to compare the fit of the model with and without those independent variables. We did so by removing one of either the QA metric or seasonal effects from the list of independent variables in the model for gray brain matter and observing the fit of the resulting model. From the initial R^2 value of 0.242 (Akaike Information Criterion, AIC = 7044; Bayesian Information Criterion, BIC = 7091) with both QA metric and seasonal effects shown in Figure 3, removing the QA metric dropped the model's R^2 value to 0.208 (AIC = 7054, BIC = 7097), and removing seasonal effects resulted in an R^2 value of 0.207 (AIC = 7052, BIC = 7092), thus showing that the QA metric and seasonal effects both have similar effects on the fit of the model. However, after removing both the QA metric and seasonal effects, the R^2 value drops to 0.185 (AIC = 7058, BIC = 7093), which suggests that they are complementary in terms of explanatory power.

Table 3. Independent variables' original and FDR-corrected p-values from the model for gray brain matter volume (Figure 3). FDR, false discovery rate; SNR, signal-to-noise ratio.

Independent variable	p-value	
	Original	FDR-corrected
Subject age	**0.0008**	**0.002**
Subject sex	**0.003**	**0.005**
Subject weight	**0.0000001**	**0.0000009**
Phantom's total SNR	**0.002**	**0.003**
Subject position in scanner (IOPD)	0.056	0.056
Seasonal	**0.029**	**0.035**

Discussion

Our results have indicated that the QA metric of the phantom data can be useful beyond routine monitoring of MRI scanner health. Specifically, our use case demonstrates the viability of using a QA metric to predict variance in estimates derived from human MRI scan data. Our results show also that the following scanning parameters: "patient" (in this case, phantom) position in scanner, day time of acquisition, and specific absorption rate; were statistically significant predictors of the phantom QA metric: total SNR. In turn, the phantom total SNR ratio was a statistically significant predictor of gray brain matter volume, even in the presence of actual data parameters relating to the patient such as age, sex, and weight. It seems that effects depend on the scale of data; in our data sample the uncorrected phantom QA metric provided an explanation for coarse "structures" (such as all of the gray brain matter) but failed to significantly explain any subcortical structure estimate.

Furthermore, we provide further support for the idea that seasonal variations affect human data. The initial R^2 value of the QA metric-human data model was 0.242; yet, by removing seasonal effects from the model, the R^2 value dropped to 0.207, suggesting that including seasonal effect data is useful for attributing variance in MRI data. It should be noted that the effect of the proxy scanner health metric seems to have a similar magnitude as that of seasonal variation effects, given that the removal of the QA metric from the model also results in a similar decline in the R^2 value to 0.208. However, we determined that this effect of seasonal variation is distinct from that from the QA metric, as can be seen by the fact that after removing both the QA metric and seasonal variation from the model, the R^2 value declines further to 0.185. This result indicates that both the QA metric and seasonal variations may be important variables to account for when seeking to explain variance in MRI data, given that using the MRI scanner's phantom QA metric was not sufficient to account for all seasonal variance (at least when gray brain matter was the dependent variable). This result weakly supports the idea that seasonal variation in human data is caused by biological, rather than scanner, effects due to the fact that seasonal variation was not significant for phantom data (i.e. scanner-only) but was significant for gray brain matter in human data.

To supplement our findings on seasonal variation effects, it should be noted that our seasonal model was very simplistic and was composed of sine and cosine waves. Given that we

Table 4. Brain segmentation statistics results where bolded values are statistically significant.
FDR, false discovery rate; SNR, signal-to-noise ratio; CSF, cerebrospinal fluid.

	Structure\Model	Phantom total SNR p value		Seasonal variables p value		Model R^2 value
		Original	FDR-corrected	Original	FDR-corrected	
0	Background	0.269	0.722	0.182	0.328	0.467
1	Left-Accumbens-area	0.981	0.981	0.879	0.973	0.104
2	Left-Amygdala	0.462	0.722	0.639	0.820	0.328
3	Left-Caudate	0.088	0.396	0.955	0.973	0.216
4	Left-Hippocampus	0.464	0.722	0.168	0.328	0.218
5	Left-Pallidum	0.563	0.722	0.160	0.328	0.520
6	Left-Putamen	0.464	0.722	0.087	0.225	0.262
7	Left-Thalamus-Proper	0.301	0.722	0.214	0.351	0.428
8	Right-Accumbens-area	0.569	0.722	**0.044**	0.211	0.146
9	Right-Amygdala	0.748	0.842	0.973	0.973	0.334
10	Right-Caudate	0.079	0.396	0.893	0.973	0.165
11	Right-Hippocampus	0.277	0.722	0.055	0.211	0.313
12	Right-Pallidum	0.453	0.722	0.081	0.225	0.544
13	Right-Putamen	0.602	0.722	0.362	0.543	0.289
14	Right-Thalamus-Proper	0.533	0.722	0.480	0.665	0.434
15	CSF	**0.031**	0.278	0.059	0.211	0.148
16	gray	**0.002**	**0.028**	**0.029**	0.211	0.242
17	white	0.862	0.912	**0.004**	0.067	0.276

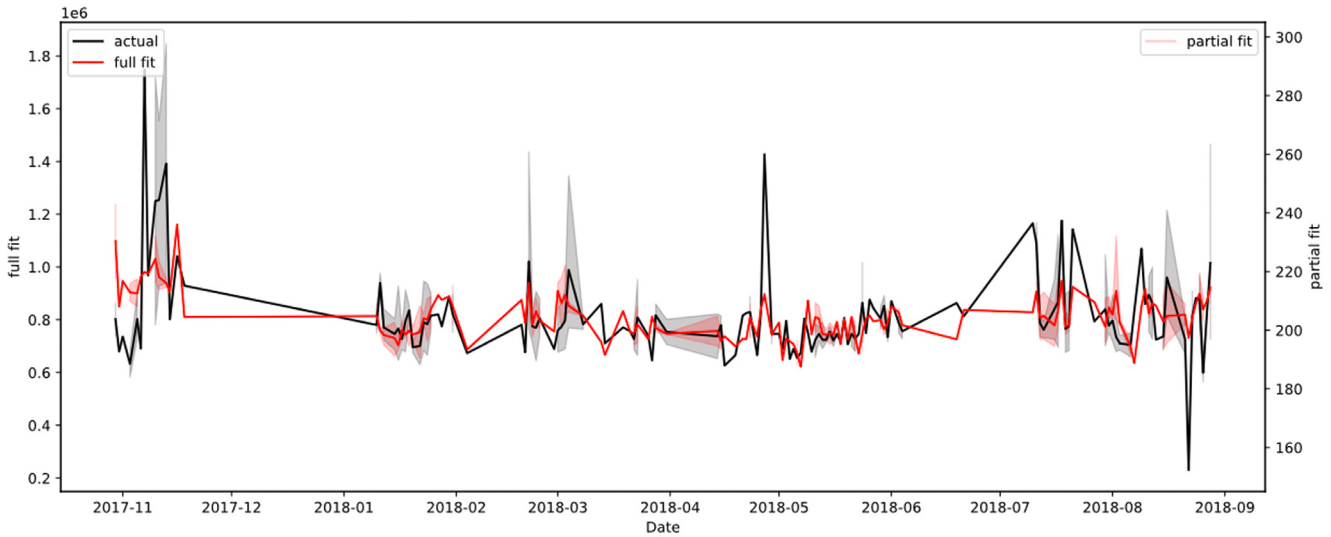

Figure 3. Model of gray brain matter using patient age, patient position in scanner, patient sex, patient weight, total signal-to-noise ratio, and seasonal variation. R^2 value was 0.242. The full fit plot (in red) shows the plot of all independent variables, whereas the partial fit plot (in pink) shows the plot of only the statistically significant variables. The black plot shows the actual fit of the real data.

noticed that it affected our models to some extent, we anticipate developing a more sophisticated model of seasonal effects for a more accurate model.

An unsurprising observation we made is that positioning of the patient (or phantom) in the scanner accounted for some variance. In the case of the phantom, its significance did not pass the significance threshold after FDR correction, but was very significant for human participants. Meanwhile, patient weight was a very statistically significant predictor for gray matter volume (Table 2); conversely, the size of the phantom remained constant while SNR was affected by position in the scanner. However, prior studies have indicated that brain size strongly correlates with patient height and thus weight[16]. This result suggests that it is still beneficial to add patient position in the scanner into the models to account for position-specific variance in addition to patient size (weight).

One interesting negative side-effect of establishing a fully reproducible pipeline, as we have done, is that we cannot share even highly compressed derivatives of the human data, such as morphometric estimates, with the subjects' participation dates. This information could potentially be used to cross-reference with datasets where such anonymized MRI data is fully shared, albeit with their dates stripped, and thereby used to violate the confidentiality of these subjects' data. Unfortunately, to our knowledge, no large public datasets are accompanied with phantom QA data scans from the participating sites, which made it impossible to reuse publicly available datasets.

Our use of ReproIn for "turnkey" collection of MRI data into BIDS datasets at DBIC was a highly beneficial methodology shown in our approach. An example of such a dataset is the phantom QA dataset we used in our study. The standardized structure of our dataset collection, from filenames to data format, facilitated our establishment of "meta-datasets" comprising data from multiple studies.

Future directions

Our investigation has used phantom and human data for the period from October 30, 2017 to August 28, 2018. We are going to compare the model's predictions on additional data (from other studies and later dates) with actual data to check the generalizability of our established models on future data and feed new data into the model to make it more robust.

As mentioned in the Methods/Phantom section, we will consider using DBIC QA estimates such as T2* weighted EPI scans instead of MRIQC estimates to evaluate the significance of other sources of QA metrics in reducing variance.

In this study we used only anatomical (T1 weighted) data. We will investigate temporal SNR, which is a QA metric only available for functional data. Functional QA metrics are an interesting area of investigation in the future, as there is some notion of functional connectivity in resting state data, and statistical estimates from GLM on task data. Functional phantom QA and other scanner characteristics could provide explanatory power to analyses.

The software we used to derive morphometric estimates of the brain could have been affected by the software used, and an investigation into the effects of conversion software (e.g., FSL) and their versions on morphometric estimates could yield valuable insights.

Conclusions

We showed that the scanning parameters and QA metric of phantom data are useful for more than just QA. To maximize the statistical power of MRI studies, we propose using scanning parameters and a QA metric, total SNR, from an MRI scanner's phantom data to reduce the unexplained variance that exists in MRI data.

Furthermore, we have found that our simple representation of seasonal variation can help explain gray brain matter volume in human MRI data, and deserves further investigation to determine if this effect is truly of a biological origin, as our results weakly suggest. The incorporation of seasonal variables can also help reduce variance in MRI data.

Data availability

The human participants' data used in this study cannot be shared, either in its anonymized form or in the form of derivative estimates of the brain structure volumes, due to the nature of the study which relies on the dates of scanning. As outlined in HIPAA, "The following identifiers of the individual or of relatives, employers, or household members of the individual must be removed to achieve the "safe harbor" method of de-identification: […](C) All elements of dates (except year) for dates directly related to the individual, including […] **admission date, discharge date**." To apply for access to the human participant data, you may contact either of the authors (cheng1928c@gmail.com and yoh@dartmouth.edu) and request a Data User Agreement (DUA) to fill out; upon approval, you will be granted access subject to the conditions of the DUA.

Harvard Database: Phantom MRI (Quality Assurance) Data From October 2017 to August 2018 at DBIC. https://doi.org/10.7910/DVN/PFH4FL[22].

This project contains the following underlying data:

- JSON files contain quality assurance metrics to be used as independent variables (AcquisitionTime, SAR, TxRefAmp, & ImageOrientationPatientDICOM) and as the dependent variable (snr_total) for logistic regression analysis
- NII.GZ files that contain a compressed version of an MRI scan in the NIfTI-1 Data Format, and is the basis from which the quality assurance metrics of the JSON file were extracted. These nii.gz files were obtained from

the original DICOM files using the HeuDiConv conversion program, which you may use to validate the JSON file's metrics.

Code availability

Code available from: https://github.com/proj-nuisance/nuisance

Archived code at time of publication: https://doi.org/10.5281/zenodo.3865441[21].

License: Apache License 2.0

The aforementioned git repositories are also DataLad datasets that provide the complete computing environments used in the study (via git-annex), and contain full history of the analyses recorded in git commits history.

Acknowledgements

We would like to express our thanks to Professors Chang, Gobbini and Haxby and Dr. Jolly of Dartmouth College's Psychological and Brain Sciences Department for making their data available for re-analysis, and to Chandana Kodiweera and Terry Sackett for their weekly collection of the phantom QA data, consultation on the details of the MRI hardware characteristics, and support in collecting all other datasets used in the study.

Besides the tools we have already mentioned, we would like to acknowledge numpy, matplotlib, pandas, seaborn, statsmodels, and the other free open source software we used in our study.

References

1. Di X, Wolfer M, Kühn S, *et al.*: **Estimations of the weather effects on brain functions using functional MRI – a cautionary tale.** *bioRxiv.* 2019; **443**.
Publisher Full Text
2. Meyer C, Muto V, Jaspar M, *et al.*: **Seasonality in human cognitive brain responses.** *Proc Natl Acad Sci U S A.* 2016; **113**(11): 3066–3071.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Lu W, Dong K, Cui D, *et al.*: **Quality assurance of human functional magnetic resonance imaging: a literature review.** *Quant Imaging Med Surg.* 2019; **9**(6): 1147–1162.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Glasser MF, Smith SM, Marcus DS, *et al.*: **The Human Connectome Project's neuroimaging approach.** *Nat Neurosci.* 2016; **19**(9): 1175–1187.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Esteban O, Birman D, Schaer M, *et al.*: **MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites.** *PLoS One.* 2017; **12**(9): e0184661.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Halchenko Y, Dartmouth Brain Imaging Center: **dbic/QA.** *Zenodo.* 2020.
Publisher Full Text
9. Visconti di Oleggio Castello M, Dobson JE, Sackett T, *et al.*: **ReproNim/reproin 0.6.0.** 2020.
Publisher Full Text
10. Li X, Morgan PS, Ashburner J, *et al.*: **The first step for neuroimaging data analysis: DICOM to NIfTI conversion.** *J Neurosci Methods.* 2016; **264**: 47–56.
PubMed Abstract | Publisher Full Text
11. Halchenko YO, Hanke M, Poldrack B, *et al.*: **datalad/datalad 0.11.6.** 2019.
Publisher Full Text
12. Hanke M, Meyer K, Halchenko YO, *et al.*: **datalad/datalad-container 1.0.0 (Version 1.0.0).** *Zenodo.* 2020.
Publisher Full Text
13. Ghosh SS, Poline JB, Keator DB, *et al.*: **A very simple, re-executable neuroimaging publication [version 2; peer review: 1 approved, 3 approved with reservations].** *F1000Res.* 2017; **6**: 124.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Seabold S, Perktold J, Fulton C, *et al.*: **statsmodels/statsmodels: Version 0.8.0 Release (Version v0.8.0).** *Zenodo.* 2017.
Publisher Full Text
16. Skullerud K: **Variations in the size of the human brain. Influence of age, sex, body length, body mass index, alcoholism, Alzheimer changes, and cerebral atherosclerosis.** *Acta Neurol Scand Suppl.* 1985; **102**: 1–94.
PubMed Abstract
17. Chang LJ, Jolly E, Cheong JH, *et al.*: **Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience.** *bioRxiv.* 2018; **487892**.
Publisher Full Text
18. Jolly E, Sadhukha S, Chang LJ: **Custom-molded headcases have limited efficacy in reducing head motion during naturalistic fMRI experiments.** *Neuroimage.* 2020; **222**: 117207.
PubMed Abstract | Publisher Full Text
19. Jiahui G, Feilong M, di Oleggio Castello MV, *et al.*: **Predicting individual face-selective topography using naturalistic stimuli.** *Neuroimage.* 2020; **216**: 116458.
PubMed Abstract | Publisher Full Text
20. Haxby JV: **Functional Anatomic Studies of Self-Affect: A Multimodal Approach.** NIMH Data Archive.
Reference Source
21. Cheng C, Halchenko Y: **proj-nuisance/nuisance 0.20200520.0 (Version 0.20200520.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3865441
22. Cheng C, Halchenko Y: **Phantom MRI (Quality Assurance) Data From October 2017 to August 2018 at DBIC.** Harvard Dataverse, V1. 2020.
http://www.doi.org/10.7910/DVN/PFH4FL

Open Peer Review

Current Peer Review Status: ? ✓ ✗ ?

Version 1

Reviewer Report 04 January 2021

https://doi.org/10.5256/f1000research.27077.r75970

?

Jens Sommer
[1] Center for Mind, Brain and Behavior, Marburg and Justus-Liebig University Giessen, Philipps-University Marburg, Marburg, Germany
[2] Core Facility Brainimaging, Philipps-University Marburg, Marburg, Germany
[3] Psychiatry Department, Philipps-University Marburg, Marburg, Germany

The article describes in a very reasonable way how to estimate previously unexplained variance of MRI studies using data from the regular QA procedure of the radiology department. As the QA procedure is only run once a week the authors interpolate QA data or QA parameters to fit the dates of subject measurements.
The original subject data is not available as the combination of individual measurement dates and mri data could disclose details about the subjects. But all of the procedures are available and described in sufficient detail, so it is possible to reproduce the study with local mri data.

Right now I expect that the article could benefit if you would include more details about the QA procedure of the radiology department. Is the procedure described somewhere? If not, what gel phantom is used (FBIRN)? Do you use a mount for reproducible positioning? Has the phantom been replaced during the study period?

Have there been any modifications or replacement of parts of the scanner hardware or software?

Did all subjects undergo the same structural scan protocol, i.e. MPRAGE with identical TR, TI, TE, bandwidth? What kind of filters were applied? Was the volume of interest automatically or manually aligned?
As you use some of the positional data for your analysis, it would also be nice to get an idea of the value ranges. Could you add a table or histogram to display the positional data?

For the segmentation you used FAST from FSL 5.0.9 (described in Gosh *et al.*, your reference #13). My former colleague L. Eggert found FAST to be more sensitive to varying image quality than other algorithms (Eggert *et al.*[1]). In his study he used FAST 4.1 from FSL 4.1.6. I did not check the FAST version used in the container you used but according the FSL version history FSL 5.0.9 should

have also included FAST 4.1.

I would be glad if in a further analysis the statistics would still remain significant when grey matter volume is based on a different segmentation algorithm. But even without this analysis your idea of QA data usage is interesting.

References

1. Eggert LD, Sommer J, Jansen A, Kircher T, et al.: Accuracy and reliability of automated gray matter segmentation pathways on real and simulated structural magnetic resonance images of the human brain.*PLoS One*. 2012; **7** (9): e45081 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: QA procedures and improvement of MRI data in general

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 21 December 2020

https://doi.org/10.5256/f1000research.27077.r75971

Simon Duchesne

Department of Radiology and Nuclear Medicine, Faculty of Medicine, Université Laval, Quebec City, QC, Canada

This is a great study addressing the subject of image quality and its influence on brain morphometric measurements. While most investigators assume that these measurements are indeed near infaillible, work of this kind serves to improve our collective ability at making these measurements ever more robust.

The authors have used a longitudinal series of AGAR phantom scans as well as human participants being recruited and scanned over the course of one year on a single (Siemens Healthcare Magnetom Prisma) machine. The statistical analysis is strong and demonstrate a possible effect of seasonal variations on the main outcome, namely grey matter segmentation.

The paper has a lot of merit, and hopefully these comments will serve to further improve it.

Major comments:

- The choice of the segmentation technique matters a lot. The inherent reliability of the technique may be the biggest inherent contributor to the variability in the results; further, this can be nonlinearly influenced by the scan characteristics. Speaking to the first point, the authors have chosen to use FSL, by which it is assumed they used the FAST technique. The authors are encouraged to read (and cite!) our recent publication on this topic (Dadar et al., NeuroImage 2020[1] in which we compared various segmentation algorithms (FAST amongst them) on a human phantom dataset composed of repeat scans on multiple scanners of the same individual (albeit without correcting for time of day, SNR, and other metrics discussed here). In this work FAST was shown to be at the lower range of reliability, when compared to other publicly available techniques (an expedited review can be promised if the authors use our own technique!)(these open reviews allow for such clarity in reviewer motivations...). To speak to the second point, it becomes hard to dissociate without further experiments how a software segmentation tool may be more able (i.e. robust) than another at glossing over image quality variations in order to produce results with high reliability. Thus it may be that all of the quality metrics measured do influence the image; but the segmentation technique is able to remove these differences. Thus the agar phantom data is of greater importance here.

- Time of day seems indeed to matter a lot, possibly more than shown here. It would appear that most of the phantom scans were taken at the beginning of the week. It is rather typical to do these scans early on the first day of scans (i.e. Monday AM). Yet, multiple reports have shown that there is quite a lot of variability in image quality due to the scanner “warming up” after repeated activity, i.e. over the course of the day. Trefler and colleagues (NeuroImage 2016;[2]) have done a nice study of this (you could cite them too). Thus in this study this effect may not have shown up in the phantom data, but could possibly affect the human results. Scattering phantom scans could address this issue.

- More generally, it is hard to draw conclusions on these effects with the choice of population. Studying GM in such a large number of different individuals on such a large age range is bound to be highly variable. It would be much better to focus on a subset of individuals (at the very least, only the young adults) and preferably only those with repeated measures, so that we can have a reasonable assumption that the scan-rescan variability should be very

low.

Minor comments:

- It is unclear what kind of interpolation between time points was done on the phantom data QC metric to estimate its value at the time of human scanning;

- The notion of IOPD is not quite clear. Does this mean some individuals were scanned 6 times? Otherwise, what goes into this metric? Patient position in the scanner is mentioned, but the position should always be the same (not like other techniques (e.g. chest x-ray) where you could have supine or prone, etc.)

- It should be discussed that the choice of "seasonal effect" as a variable on participants being recruited over a year masks realities that may bias the study irrespective of the seasonal effect on the MRI scanner itself. For example, a competing hypothesis could be that seasons generated a recruitment bias, with young participants being more available in the summer, while older participants in the winter. One would need to better define seasonal variations and what they entail in terms of signal characteristics that are related to the scanner - yet as independent as possible to the SNR. For example, seasonal effects could serve as a proxy of room temperature, which can be measured (and may even be retrospectively available, depending on the building information system). The latter would affect cooling of the main field - which in turn would affect the main B0 but possibly other sources of thermal noise in the RF apparati. Thus a metric such as room temperature would be better suited than a sinusoidal model of season. Unpublished data from a colleague (J. Doyon, Montreal Neurological Institute) have showed a distinct relationship between room temperature and phantom SNR.

References

1. Dadar M, Duchesne S: Reliability assessment of tissue classification algorithms for multi-center and multi-scanner data. *NeuroImage*. 2020; **217**. Publisher Full Text
2. Trefler A, Sadeghi N, Thomas AG, Pierpaoli C, et al.: Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging.*Neuroimage*. **133** : 41-52 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neuroimaging, imaging protocols, image processing, biomarkers, Alzheimer's disease

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 22 October 2020

https://doi.org/10.5256/f1000research.27077.r71693

Blaise Frederick
[1] McLean Imaging Center, McLean Hospital, Belmont, MA, USA
[2] Harvard Medical School, Boston, MA, USA

This article explores the relationship between various factors (SNR on a QA phantom, patient positioning, time of year, etc) and the results of quantitation of anatomic MRI images.

This is a solid piece of work, well executed and described, which demonstrates a useful method for reducing unexplained variance in MR data, in order to reveal underlying biological phenomena. The methods are very clearly described and reproducible, and all data is available for replication. I look forward to the companion paper on fMRI data. I would recommend accepting the paper as is (other than correcting Figure 2).

Specific comments:
It would be worth explaining why different versions of dcm2niix were used for the phantom and human data. This is not a major issue - each dataset was processed consistently. It is a little odd however.

Figure 2 shows a legend and y scale for a partial model fit, however no partial model fit appears in the graph, nor is there any mention of one in the text.

While the correlation between gray matter volume and QA SNR is shown to be significant, it would be useful to mention the sign of the interaction - does apparent GM volume increase or decrease with increasing SNR? This would be an interesting thing to know.

It would be a nice addition to the paper to see if there were any correlations between day of the

week and any of the patient measures (not possible for the phantom data, since it was done once a week) to complement the Time-of-day and season metrics.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My area of expertise is fMRI and NIRS technique development, in particular methods for characterizing and interpreting physiological "noise" effects, and using them to quantify hemodynamics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 06 October 2020

https://doi.org/10.5256/f1000research.27077.r71690

?

Xiangrui Li
Department of Psychology, The Center for Cognitive and Behavioral Brain Imaging, The Ohio State University, Columbus, OH, USA

This is an interesting topic and has great practical application for the MRI studies.
My major concern is the conclusion about effect of variables, such as day time of acquisition, subject position, and SAR, on the brain tissue volume. It is easy to understand the potential effect of these variables on the phantom SNR, while it sounds a big jump to claim the effect on tissue

volume. Even if other biological variables are included as confounds, the variability among participants is too large to be accounted by the potential minor effect of the phantom SNR.
A better approach may be to test the effect of the phantom SNR on the volume of the same brain, but it seems there is not enough repeats for a single brain in the dataset.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: MRI

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.