Convolutional neural networks for the automatic quality control of brain T1-weighted MRI from a clinical data warehouse

Simona Bottani, Ninon Burgos, Aurélien Maire, Adam Wild, Sébastian Ströer, Didier Dormont, Olivier Colliot

Convolutional neural networks for the automatic quality control of brain T1-weighted MRI from a clinical data warehouse

Simona Bottani[1] SIMONABOTTANI92@GMAIL.COM
Ninon Burgos[1] NINON.BURGOS@ICM-INSTITUTE.ORG
Aurélien Maire[2] AURELIEN.MAIRE@APHP.FR
Adam Wild[1] ADAM.WILD@ICM-INSTITUTE.ORG
Sébastian Ströer[3] SEBASTIAN.STROER@APHP.FR
Didier Dormont[1,3] DIDIER.DORMONT@APHP.FR
Olivier Colliot[1] OLIVIER.COLLIOT@SORBONNE-UNIVERSITE.FR
APPRIMAGE Study Group

[1] *Paris Brain Institute (ICM), Inserm U 1127, CNRS UMR 7225, Sorbonne Université, Inria, Aramis project-team, F-75013, Paris, France*

[2] *AP-HP, WIND department, F-75012, Paris, France*

[3] *AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France*

Editors: Under Review for MIDL 2021

Abstract

Many studies on machine learning (ML) for computer-aided diagnosis are restricted to high-quality research data. Clinical data warehouses, gathering routine examinations from hospitals, offer great promises for training and validation of ML models in a realistic setting. However, the use of such clinical data warehouses requires quality control (QC) tools. Visual QC by experts is time-consuming and does not scale to large datasets. The aim of this work is to develop a convolutional neural network (CNN) for the automatic QC of 3D T1w brain MRI for a large heterogeneous clinical data warehouse. Specifically, the objectives were: 1) to identify images which are not proper T1w brain MRIs; 2) to identify acquisitions for which gadolinium was injected; 3) to rate the overall image quality. We used 5000 images for training and validation and a separate set of 500 images for testing. In order to train/validate the CNN, the data were annotated by two trained raters according to a visual QC protocol that we specifically designed for application in the setting of a data warehouse. For objectives 1 and 2, our approach achieved excellent accuracy, similar to the human raters. For objective 3, the performance was good but substantially lower to that of human raters.

Keywords: Quality control, Neuroimaging, MRI, Brain, Deep learning

1. Introduction

Structural T1-weighted (T1w) magnetic resonance imaging (MRI) is useful for diagnosis of various brain disorders, in particular neurodegenerative diseases (Frisoni et al., 2010; Harper et al., 2016). They have thus often been used as inputs of machine learning (ML) algorithms for computer-aided diagnosis (CAD) (Koikkalainen et al., 2016).

Most ML methods are trained and validated on high-quality research data: protocols for image acquisition are standardized and a strict quality control is applied (Jack Jr et al.,

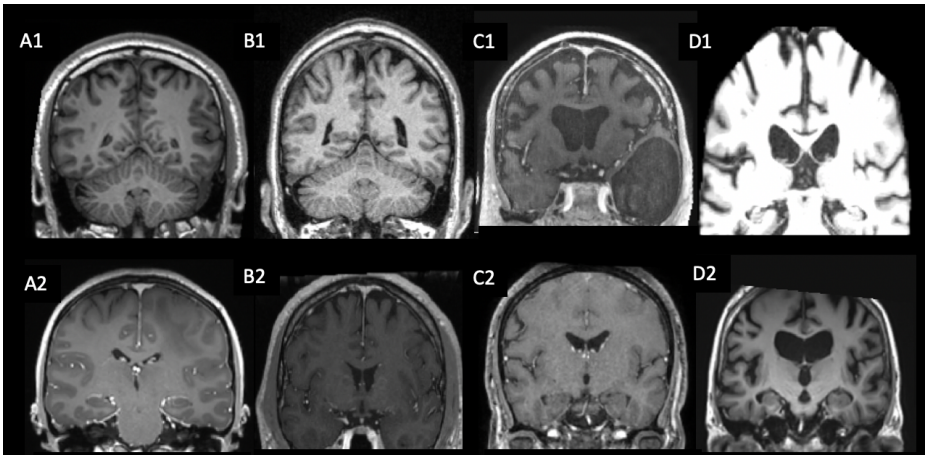

Figure 1: Examples of T1w brain images from the clinical data warehouse and the corresponding labels. A1: Image of good quality (tier 1), without gadolinium; A2: Good quality (tier 1), with gadolinium; B1: Medium quality (tier 2), without gadolinium, noise grade 1; B2: Medium quality (tier 2), with gadolinium, contrast grade 1; C1: Bad quality (tier 3), without gadolinium, contrast grade 2, motion grade 2; C2: Bad quality (tier 3), with gadolinium, contrast grade 2, motion grade 1; D1: Straight rejection, segmented; D2: Straight rejection, cropped.

2008; Littlejohns et al., 2020). However, to be applied in the clinic, ML methods need to be validated on clinical routine images. The quality of such images can greatly vary (see Figure 1), since the acquisition protocols are not standardized, scanners may not be recent and patients may have moved during the acquisition. All these factors can prevent algorithms from working properly (Reuter et al., 2015; Gilmore et al., 2019). Quality control (QC) is thus a fundamental step before training and evaluating ML approaches on clinical routine data.

Manual QC takes time and is thus not always doable, especially in the context of ML-based CAD, where a large number of training samples is needed. Typically, clinical data warehouses can contain hundreds of thousands of samples. Even if web-based systems facilitate annotation (Kim et al., 2019; Keshavan et al., 2018), the task remains unfeasible for very large datasets. In this context, automatic QC is needed. The works of (Alfaro-Almagro et al., 2018; Esteban et al., 2017) propose a set of QC metrics automatically extracted from T1w brain MRI data, such as the signal-to-noise ratio or the volume of the gray and white matters, to use as input for a classifier. The pipelines proposed by these works are very extensive: registration and segmentation steps are used for the feature extraction. It is not possible to assume a priori that these steps will perform well with a new unseen clinical dataset. On the contrary, it is likely that the segmentation will fail for the lowest quality images, thus making it impossible to apply the QC tool. Moreover, the extracted features may not be representative of the problems affecting clinical routine data. As proposed by (Sujit et al., 2019), convolutional neural networks (CNNs) are a good option

for automatic QC because they can learn features without a priori knowledge on which are the most adapted. A further limitation of these works is that they rely on images acquired following a well-defined research protocol. The pipeline presented in (Alfaro-Almagro et al., 2018) was developed for the large, but well-standardized, UK Biobank dataset containing mostly healthy volunteers. (Esteban et al., 2017) and (Sujit et al., 2019) trained their algorithms on ABIDE, a research multicenter study including patients with autism and control subjects and used another research dataset for testing.

Our work was done using a clinical data warehouse. It assembles all MRI data from all hospitals of the greater Paris area. Images come from different sites and different machines with no homogenization on the parameters. Their acquisition cover several decades. The patient may have any disease for which a brain MRI exam is required. All these factors are not present in the approaches already proposed in the literature: even when images come from different sites, the acquisition protocol is harmonized, the number of machines is limited and they are usually acquired within a few years, avoiding intrinsic problems of quality due to the progress in the technology. Additionally, the presence of different diseases such as neurodegenerative diseases, stroke, multiple sclerosis, brain tumours, or metastases, is typical of clinical dataset: they can strongly alter the structure of the brain and it may be difficult to use a specific set of features to characterize the quality of the images independently of the disease. In addition, due to security reasons, images from the data warehouse cannot be uploaded to a web server and we had to work in a restricted IT environment.

The objective of our work was to develop a method for the automatic QC of T1w brain MRI in large clinical data warehouses. The specific objectives were: 1) to discard images which are not proper T1w brain MRI; 2) identify images with gadolinium; 3) to recognise images of bad, medium and good quality. We used 5000 images for training/validation and 500 for testing. To train/validate the models, the data was annotated by two trained raters. To that purpose, we introduced an original visual QC protocol that is applicable to clinical data warehouses.

2. Materials and methods

2.1. Dataset description

This work relies on a large clinical routine dataset containing all the T1w brain MR images of adult patients scanned in hospitals of the Greater Paris area (Assistance Publique-Hopitaux de Paris [AP-HP]). The images were selected according to DICOM fields. A first query on the PACS was performed to list the DICOM fields (series, modality and body part description) corresponding to MRI. A neuroradiologist then selected the fields referring to 3D T1w brain images. These fields were used to automatically select the images. The data were made available by the data warehouse of the AP-HP and the study was approved by the Ethical and Scientific Board of the AP-HP. According to French regulation, consent was waived as these images were acquired as part of the routine clinical care of the patients. For the present study, we randomly selected 5500 images. These corresponded to 4177 patients. The images were acquired on various scanners from four manufacturers: Siemens Healthineers ($n = 3752$), GE Healthcare ($n = 1710$), Philips ($n = 33$) and Toshiba ($n = 5$).

2.2. Image preprocessing

The T1w MR images were converted from DICOM to NIfTI using the software dicom2niix (Li et al., 2016) and organized using the Brain Imaging Data Structure (BIDS) standard (Gorgolewski et al., 2016). Images with a voxel dimension smaller than 0.9 mm were resampled using a 3rd-order spline interpolation to obtain 1 mm isotropic voxels. To facilitate annotations, we applied the following pre-processing using the 't1-linear' pipeline of Clinica (Routier et al., 2019), which is a wrapper of the ANTs software (Avants et al., 2014). Bias field correction was applied using the N4ITK method (Tustison et al., 2010). An affine registration to MNI space was performed using the SyN algorithm (Avants et al., 2008). The registered images were further rescaled based on the min and max intensity values, and cropped to remove background resulting in images of size 169×208×179, with 1 mm isotropic voxels (Wen et al., 2020). One should note that we only aimed to obtain a rough alignment and intensity rescaling to facilitate annotation.

2.3. Manual labeling of the dataset

In this section, we introduce the visual QC protocol. We describe the different characteristics noted on the images and how we created the final label for the automatic QC.

2.3.1. Quality criteria

Five characteristics were manually annotated. The first two (straight rejection and gadolinium) are binary flags, while the other three (motion, contrast and noise) are assessed with a three-level grade.

- **Straight rejection (SR)**: images not containing a T1w MRI of the whole brain (for instance images of segmented tissues or truncated images). Note that these images still have DICOM fields corresponding to T1w brain MRI and thus were not removed through the selection step based on DICOM fields.
- **Gadolinium**: presence of gadolinium-based contrast agent.
- **Motion** 0: no motion, 1: some motion but the structures of the brain are still distinguishable, 2: severe motion, the cortical and subcortical structures are difficult to distinguish.
- **Contrast** 0: good contrast, 1: medium contrast (gray matter and white matter are difficult to distinguish in some parts of the image), 2: bad contrast (gray matter and white matter are difficult to distinguish everywhere in the brain).
- **Noise** 0: no noise, 1: presence of noise that does not prevent identifying structures, 2: severe noise that does prevent identifying structures.

Gadolinium injection, motion, contrast and noise were noted for all the images which were not defined as SR. According to the grades given to the motion, contrast and noise characteristics, we determined three tiers corresponding to images of good, medium and bad quality. The tiers, along with the rules used to defined them, are described in Table 1.

Table 1: Description and determination rules of the proposed quality control tiers.

Tier	Description	Determination rule
Tier 1	3D T1w brain MRI of good quality	Grade 0 for motion, contrast and noise
Tier 2	3D T1w brain MRI of medium quality	At least one characteristic among motion, contrast and noise with grade 1 and none with grade 2
Tier 3	3D T1w brain MRI of bad quality	At least one characteristic among motion, contrast and noise with grade 2

2.3.2. Annotation set-up

Our aim was to annotate the largest possible number of images in an efficient manner. Moreover, we were restricted to the environment of the data warehouse which only included a Jupyter notebook and a command-line interface. We thus implemented a graphical interface in a Jupyter notebook displaying only the central axial, sagittal and coronal slices of the brain. Loading the whole 3D volume or inspecting all the slices would have been too time consuming. Each image was labeled by two trained raters.

2.3.3. Consensus label

The final label used to train and validate the automatic QC is a consensus between the two raters. If the users labeled different image characteristics, we determined a procedure to define a consensus label. We distinguished two types of disagreement: one regarding the SR status and the other one regarding the other characteristics based on which the tiers are assigned. When the two raters disagreed on the SR status, we manually set the consensus label: the two raters reviewed the images and decided together to keep the SR label or assign the alternative label. In case of disagreement regarding the other characteristics, the consensus was chosen as follows. The objective was to be as conservative as possible: we wanted to retain all the imperfections that may have been seen by one annotator and not by the other. For a given characteristic, the consensus grade was chosen as the maximum of the two grades of the observers. The tier was recomputed accordingly.

2.4. Automatic quality control method

We developed an automatic QC method based on CNNs trained to perform several classification tasks: 1) discard images which were not proper T1w brain MRI (SR: yes vs no)); 2) identify images with gadolinium (gadolinium: yes vs no); 3) differentiate images of bad quality from images of medium and good quality (tier 3 vs tiers 2-1); 4) differentiate images of medium quality from images of good quality (tier 2 vs tier 1).

2.4.1. Network architecture

The network was composed of five convolutional and max pooling layers and of three fully connected layers. The models were trained using the cross entropy loss, which was weighted according to the proportion of images per class for each task. We used the Adam optimizer

with a learning rate of 1e-4. We implemented early stopping and all the models were evaluated with a maximum of 50 epochs. The batch size was set to 2. The model with the lowest loss was saved as final model. Implementation was done using Pytorch.

2.4.2. Experiments

Before starting the experiments, we defined a test set by randomly selecting 500 images which respected the same distribution of tiers as the images in the training/validation set. The remaining 5000 images were split into training and validation using a 5-fold cross validation (CV). The separation between training, validation and test sets was made at the patient level in order to avoid data leakage. For each of the four tasks considered (SR, gadolinium, tier 3 vs 2-1, tier 2 vs 1), the five models trained in the CV were evaluated on the test set. We also studied the influence of the size of the training set on the performance by computing learning curves. We compared the output of each classifier with the consensus label. To set the automatic QC results in perspective, we computed the balanced accuracy (BA) for the raters (defined as the average of the BAs between each rater and the consensus).

3. Results

3.1. Manual quality control

The inter-rater agreement was evaluated using the weighted Cohen's kappa (Watson and Petrie, 2010) between the two annotators for each of the characteristics. Results are presented in Table 2. The agreement is strong for the SR label and the gadolinium injection (0.88 and 0.89) and moderate for the other characteristics (from 0.68 to 0.79).

The distribution of the consensus labels for the 5500 patients is shown in Figure 2. 26% of the images are labeled as SR, 16% as tier 1, 28% as tier 2, and 30% as tier 3. As expected, the proportion of images with gadolinium increased when the quality decreased. A vast majority of tier 3 images had a contrast of 2 and were with gadolinium. Figure 1 shows some representative examples of T1w brain images with the corresponding labels.

3.2. Automatic quality control

The BAs obtained for the four tasks of interest by the CNN classifiers and by the annotators are presented in Table 3. For the recognition of SR images, we used all the images available in the training/validation set ($n = 5000$); for the gadolinium and tier 3 vs tiers 2-1 tasks,

Table 2: Weighted Cohen's kappa between the two annotators

Characteristics	**Weighted Cohen's kappa**
SR (yes vs no)	0.88
Gadolinium injection (yes vs no)	0.89
Contrast (0 vs 1 vs 2)	0.79
Motion (0 vs 1 vs 2)	0.68
Noise (0 vs 1 vs 2)	0.70

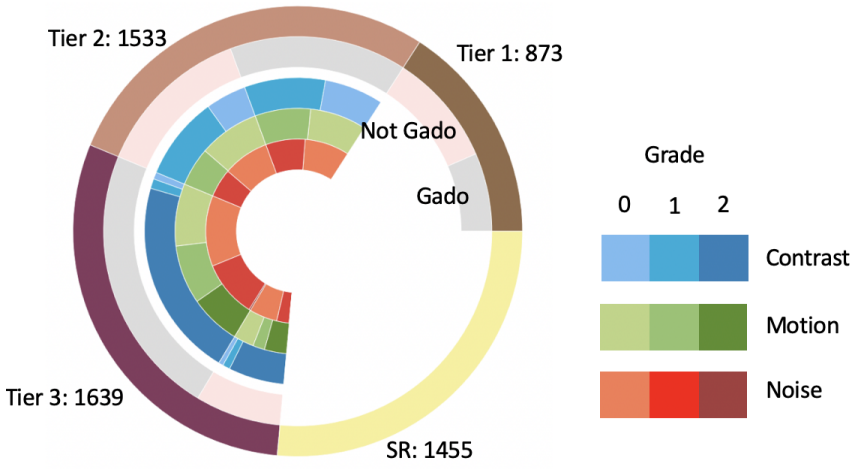

Figure 2: Distribution of the consensus labels for the whole dataset of 5500 images. Outermost circle: images in SR and in the different tiers. For every tier, we divide between images with and without gadolinium injection. For each injection status we see the grade distribution of the contrast, motion and noise characteristics.

the training/validation set does not include SR images ($n = 3770$); and for the tier 2 vs tier 1 task, the training/validation set does not include SR and tier 3 images ($n = 2182$).

Balanced accuracies for SR and gadolinium are excellent (94% and 97%). For SR, the CNN is slightly less good than the annotators. For gadolinium, the CNN is as good as the raters. For tier 3 vs 2-1, the classifier BA is good but lower than that of the annotators. For tier 2 vs 1, CNN BA is low (71%) and much lower than that of the raters (88%).

The influence of the size of the training set on the performance is shown in Figure Figure 3. For SR, the performance increases with sample size, even if it is also good with few examples (90% for 500 images) because of the easiness of the task. For gadolinium,

Table 3: Results of the CNN classifier for all the tasks. For the balanced accuracy of the classifier, we report the mean and the empirical standard deviation across the five folds.

Task	**BA classifiers**	**BA annotators**
SR (yes vs no)	93.76 ± 0.57	97.13
Gadolinium injection (yes vs no)	97.14 ± 0.34	96.10
Tier 3 vs tiers 2-1	83.51 ± 0.93	91.56
Tier 2 vs tier 1	71.65 ± 2.15	88.27

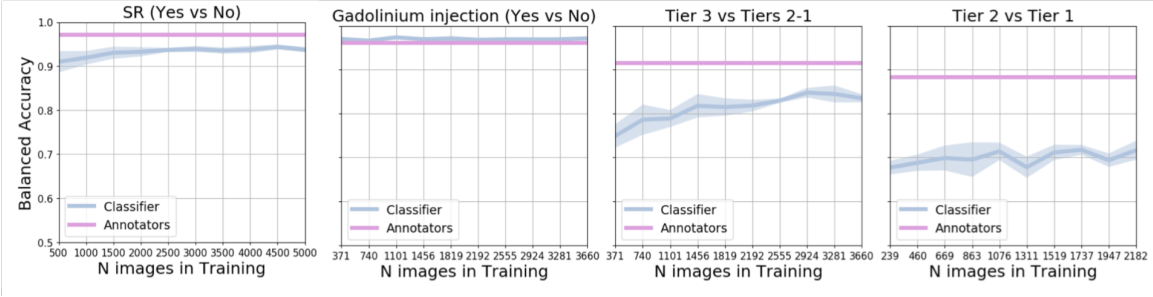

Figure 3: Learning curves for the SR (yes vs no), gadolinium injection (yes vs no), tier 3 vs tier 2-1 and tier 2 vs tier 1 tasks. Blue: balanced accuracy of the classifier across the five folds. Violet: balanced accuracy of the annotators on the testing set.

performance is very high regardless of the sample size. For tier 3 vs tiers 2-1, adding more training samples helps the classifier while this is not the case for tier 2 vs 1.

4. Discussion and conclusion

In this work, we developed a method for the automatic QC of T1w brain MRI for a large clinical data warehouse. In order to achieve this goal, we devised a manual QC protocol to build the training/validation/test sets of 5500 images in total.

Manual annotation results showed that our protocol is reproducible across all tasks, even though agreement was less for more challenging characteristics. They also provide interesting information on the variability of image quality in a clinical routine data warehouse. As much as 25% are totally unusable (SR), and almost a third has a low quality (tier 3). We also confirmed that gadolinium has a strong impact on image quality, hence the critical importance of detecting it accurately, the DICOM fields being unreliable in that regard.

For detecting straight reject, our CNN had excellent performance. Even though the task is relatively easy, this is very important in order to automatically discard images in a very large scale study. This was also the case for gadolinium, an important characteristic that strongly impacts the behavior of many image analysis methods. We thus believe that these tools can be reliably used on the rest of this large data warehouse and already have an important practical impact for researchers in deep learning for medical imaging.

For detecting low quality data (tier 3), the performance was good even though lower than that of manual raters. On the other hand, it was substantially lower for differentiating between high and medium quality images. Nevertheless, such tools still seem useful for analysing the failure modes of CAD systems or other ML approaches, as such correlative work is still doable with an imperfect tool. More work is nevertheless needed in order to use them for a strict rating of MRI quality.

Acknowledgments

The research was done using the Clinical Data Warehouse of the Greater Paris University Hospitals. The authors are grateful to the members of the AP-HP WIND and URC teams, and in particular Stéphane Bréant, Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret, Christel Daniel, Martin Hilka, Yannick Jacob, Julien Dubiel and Cyrina Saussol. They would also like to thank the "Collégiale de Radiologie of AP-HP" as well as, more generally, all the radiology departments from AP-HP hospitals. Finally, the authors are very appreciative of the support and guidance they have received from Quentin Vanderbecq when setting up the visual quality control protocol.
The research leading to these results has received funding from the Abeona Foundation (project Brain@Scale), from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6). The authors have no relevant financial or non-financial interests to disclose.

APPRIMAGE Study Group

Olivier Colliot, Ninon Burgos, Simona Bottani [1]
Didier Dormont [1,2], Samia Si Smail Belkacem, Sebastian Ströer [2]
Nathalie Boddaert [3]
Farida Benoudiba, Ghaida Nasser, Claire Ancelet, Laurent Spelle [4]
Hubert Ducou-Le-Pointe[5]
Catherine Adamsbaum[6]
Marianne Alison[7]
Emmanuel Houdart[8]
Robert Carlier, Myriam Edjlali[9]
Betty Marro[10,11]
Lionel Arrive[10]
Alain Luciani[12]
Aurélien Maire, Stéphane Bréant, Christel Daniel, Martin Hilka, Yannick Jacob, Julien Dubiel, Cyrina Saussol [13]
Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret [14]
[1] Paris Brain Institute (ICM), Inserm U 1127, CNRS UMR 7225, Sorbonne Université, Inria, Aramis project-team, F-75013, Paris, France
[2] AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France
[3] AP-HP, Hôpital Necker, Department of Radiology, F-75015, Paris, France
[4] AP-HP, Hôpital Bicêtre, Department of Radiology, F-94270, Le Kremlin-Bicêtre, France
[5] AP-HP, Hôpital Armand-Trousseau, Department of Radiology, F-75012, Paris, France
[6] AP-HP, Hôpital Bicêtre, Department of Pediatric Radiology, F-94270, Le Kremlin-Bicêtre, France
[7] AP-HP, Hôpital Robert-Debré, Department of Radiology, F-75019, Paris, France
[8] AP-HP, Hôpital Lariboisière , Department of Neuroradiology, F-75010, Paris, France
[9] AP-HP, Hôpital Raymond-Poincaré, Department of Radiology, F-92380, Garches, France

[10] AP-HP, Hôpital Saint-Antoine, Department of Radiology, F-75012, Paris, France
[11] AP-HP, Hôpital Tenon, Department of Radiology, F-75020, Paris, France
[12] AP-HP, Hôpital Henri-Mondor , Department of Radiology, F-94000, Créteil, France
[13] AP-HP, WIND department, F-75012, Paris, France
[14] AP-HP, Unité de Recherche Clinique, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France

References

Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*, 166:400–424, 2018.

Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

Brian B Avants, Nicholas J Tustison, Michael Stauffer, Gang Song, Baohua Wu, and James C Gee. The Insight ToolKit image registration framework. *Frontiers in Neuroinformatics*, 8:44, 2014.

Oscar Esteban, Daniel Birman, Marie Schaer, Oluwasanmi O Koyejo, Russell A Poldrack, and Krzysztof J Gorgolewski. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS one*, 12(9):e0184661, 2017.

Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2): 67–77, 2010. ISSN 1759-4758, 1759-4766. doi: 10.1038/nrneurol.2009.215. 01394.

Alysha Gilmore, Nicholas Buser, and Jamie L Hanson. Variations in structural MRI quality impact measures of brain anatomy: Relations with age and other sociodemographic variables. *Biorxiv*, page 581876, 2019.

Krzysztof J Gorgolewski, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, Satrajit S Ghosh, Tristan Glatard, Yaroslav O Halchenko, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):1–9, 2016.

Lorna Harper, Giorgio G Fumagalli, Frederik Barkhof, Philip Scheltens, John T O'Brien, Femke Bouwman, Emma J Burton, Jonathan D Rohrer, Nick C Fox, Gerard R Ridgway, et al. MRI visual rating scales in the diagnosis of dementia: evaluation in 184 post-mortem confirmed cases. *Brain*, 139(4):1211–1225, 2016.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

Anisha Keshavan, Esha Datta, Ian M McDonough, Christopher R Madan, Kesshi Jordan, and Roland G Henry. Mindcontrol: A web application for brain segmentation quality control. *NeuroImage*, 170:365–372, 2018.

Hosung Kim, Andrei Irimia, Samuel M Hobel, Rita I Esquivel Castelo-Blanco, Ben Duffy, Lu Zhao, Karen L Crawford, Sook-Lei Liew, Kristi Clark, Meng Law, et al. LONI QC system: a semi-automated, web-based and freely-available environment for the comprehensive quality control of neuroimaging data. *Frontiers in Neuroinformatics*, 13:60, 2019.

Juha Koikkalainen, Hanneke Rhodius-Meester, Antti Tolonen, Frederik Barkhof, Betty Tijms, Afina W Lemstra, Tong Tong, Ricardo Guerrero, Andreas Schuh, Christian Ledig, et al. Differential diagnosis of neurodegenerative diseases using structural MRI data. *NeuroImage: Clinical*, 11:435–449, 2016.

Xiangrui Li, Paul S Morgan, John Ashburner, Jolinda Smith, and Christopher Rorden. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, 264:47–56, 2016.

Thomas J Littlejohns, Jo Holliday, Lorna M Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D Bell, Chris Boultwood, Rory Collins, Megan C Conroy, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications*, 11(1):1–12, 2020.

Martin Reuter, M Dylan Tisdall, Abid Qureshi, Randy L Buckner, André JW van der Kouwe, and Bruce Fischl. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage*, 107:107–115, 2015.

Alexandre Routier, Ninon Burgos, Jérémy Guillon, Jorge Samper-González, Junhao Wen, Simona Bottani, Arnaud Marcoux, Michael Bacci, Sabrina Fontanella, Thomas Jacquemont, Adam Wild, Pietro Gori, Alexis Guyot, Pascal Lu, Mauricio Díaz, Elina Thibeau-Sutre, Tristan Moreau, Marc Teichmann, Marie-Odile Habert, Stanley Durrleman, and Olivier Colliot. Clinica: an open source software platform for reproducible clinical neuroscience studies. *hal-02308126*, 2019. URL https://hal.inria.fr/hal-02308126.

Sheeba J Sujit, Ivan Coronado, Arash Kamali, Ponnada A Narayana, and Refaat E Gabr. Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. *Journal of Magnetic Resonance Imaging*, 50(4):1260–1267, 2019.

Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.

PF Watson and A Petrie. Method agreement analysis: a review of correct methodology. *Theriogenology*, 73(9):1167–1179, 2010.

Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation. *Medical Image Analysis*, page 101694, 2020.