Exploration of How MRI Acquisition parameter leads to Acquisition Shift

Fateme Sadat Haghpanah & Beiqin Zeng & Zhongkang Guo
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
{fateme.haghpanah,beiqin.zeng,zhongkang.guo}@mail.utoronto.ca

Abstract

Machine learning has played a key role in medical imaging analysis, especially for MRI, a commonly used medical imaging technique. The previous works show that researchers can tell the potential bias based on images and annotations. In this project, we explore the causal analysis, especially on the acquisition shift. Given the MRI scans from PPMI dataset, we implemented skull stripping, bias field correction, and tissue segmentation for the preprocessing. Then we used ResNet to implement several different experiments. In this study, we explored the affect of different MRI acquisition parameters on the performance of severity of Parkinson's Disease classification models. We looked for any causal relationship leading to acquisition shift.

1 Introduction

There are a lot of works in the application of machine learning in medical imaging, e.g., detecting COVID from chest X-Ray images(Oh et al. (2020)). Since MRI is one of the most common modalities in medical imaging, the machine learning for healthcare community did a lot of research using MRI scans to detect diseases. For example, Moradi et al. (2015) developed a machine learning framework to make Alzheimer's conversion prediction in Mild Cognitive Impairment (MCI) subjects based on MRI, and it achieved high accuracy. Salvatore et al. (2014) developed a machine learning algorithm allowing individual differential diagnosis of Parkinson's Disease and Progressive Supranuclear Palsy based on MRI scans.

However, the medical imaging field is still suffering from generalization problems caused by different reasons, such as lack of labeled data, or shift distribution. Castro et al. (2020) claims that establishing a causal relationship between images and annotations will help researchers to identify potential biases and issues in advance. It also offers step-by-step recommendations to consider potential underlying biases more thoroughly.

Inspired by this paper, we would like to further investigate to what extent the causal analysis and the suggested step-by-step recommendation will make difference. Especially, we mainly focus on the acquisition shift, which is discussed in Castro et al. (2020). We focus on data mismatch due to different MRI acquisition parameters based on MRI dataset of Parkinson's Disease and investigate whether different MRI acquisition parameters can be detected by machine learning approaches.

It matters because if machine learning methods can easily detect the difference due to different acquisition parameters, it's likely that if a group of subjects with the same medical condition are under similar acquisition parameters, the classifier may learn a criterion based on the acquisition parameters rather than the medical condition. In this case, the high accuracy of the machine learning model is not reliable and the model can not be generalized to other datasets.

In this project, we follow the step-by-step recommendations and do experiments on whether several MRI acquisition parameters can be detected by convolutional neural networks.

2 Related Work

Our original inspiration comes from Castro et al. (2020). In this paper, the author proposes a framework for how causality can benefit medical imaging using the concepts from causal inference. And it discusses different kinds of dataset shift including acquisition shift, which is what we would like to investigate further in this project.

There are some previous papers related to acquisition shift in medical imaging. Ferrari et al. (2018) shows that the differences due to the acquisition protocol can have a strong impact on machine learning models. Wachinger et al. (2019) shows MRI scans from different datasets can be correctly assigned to their respective dataset with 73.3% accuracy. Although these previous papers prove the existence of the acquisition shift in MR images to some extent, they focus on the effect of different sites or different sources of data instead of acquisition parameters. In addition, none of them uses deep learning methods rather than machine learning algorithms.

In terms of our problem setting - MRI scans of Parkinson's disease dataset, there are some related papers that apply machine learning and deep learning algorithms to detect Parkinson's disease based on the dataset we use. Sivaranjini & Sujatha (2020) uses AlexNet and transfer learning to classify Parkinson's disease and achieves 88.9% accuracy. Esmaeilzadeh et al. (2018) develops a 3D-CNN deep learning framework based on MR-Images combined with age and gender features even achieves 100% accuracy. Although our goal in this project is not to build a highly accurate model to predict disease, these papers provides experience of building a model on PPMI MRI data.

3 Dataset

3.1 Introduction of PPMI

As for the dataset, we choose PPMI (Parkinson Progression Marker Initiative), a five-year observational, international, multi-center study designed to identify PD progression biomarkers(Marek et al. (2011)). It includes 797 diagnosed subjects for the PD group and 234 healthy subjects for the control group. The data includes clinical information, medical imaging, biospecimen biomarker assessment, and metadata, such as sex, age, weight, acquisition settings, etc.

3.2 Introduction of MRI

Magnetic resonance imaging (MRI) is a very commonly used medical imaging technique, which uses a magnetic field and computer-generated radio waves to create detailed images of the organs and tissues in the body. Compared with Computed Tomography (CT), MRI does not use the damaging ionizing radiation of X-rays. As a result, MRI is preferred while frequent diagnosis is required, especially for the brain.

There are three main parameters of MRI imaging:

- TE (Echo Time): the time from the center of the radiofrequency pulse to the center of the gradient echo sequences.
- TR (Repetition Time): the length of time between corresponding consecutive points on a repeating series of pulses and echoes.
- TI (Inversion Time): the time between the 180-degree inverting pulse and the 90-degree pulse.

The two basic types of MRI images are T1-weighted and T2-weighted scans, produced by using different configurations of TE, TR, and TI.

- T1-weighted images are produced by short TE and TR time, and they can highlight the fat within the body.
- T2-weighted images are produced by long TE and TR time, and they can highlight the fat and water within the body.

3.3 Schwab and England ADL Scale

The PPMI dataset contains several scales to assess the levels of severity of Parkinson's disease, including Schwab and England ADL scale we use in this project. The Schwab and England ADL (Activities of Daily Living) scale is a method of assessing the capabilities of people with impaired mobility. The scale uses percentages to represent how much effort and dependence on other people need to complete daily chores. It has a range from 0% to 100%. For example, 100% means the person is able to do all chores without slowness, difficulty, or impairment while 0% the person is bedridden and helpless.

4 Preprocessing

4.1 Skull Stripping

First of all, we need to isolate the brain part from extra-cranial or other non-brain tissues in the MRI original scans, which is referred to as skull stripping. Because the brain images preprocessed by skull stripping could get better segmentation of different brain regions, the brain regions must be skull-stripped before the application of other preprocessing algorithms. In our study, we implemented the skull stripping with ROBEX v 1.2 (Iglesias et al. (2011)).

4.2 Bias Field Correction

Bias field signal is a low-frequency and very smooth signal which would corrupt MRI images, especially the images produced by old MRI machines. Affected by bias field signal, the image processing algorithms based on the intensity value of image pixels, such as segmentation, texture analysis, or classification, will not have satisfactory results. As a result, the correction for the bias field signal is necessary here, before applying the MRI images to other preprocessing steps. In our study, we implemented bias field correction with N4ITK (Tustison et al. (2010)).

4.3 Tissue Probability Maps

Image segmentation plays a key role in subsequent quantitative analysis of brain images. Tissue probability maps are probability maps for different tissue classes. In our study, we obtain the tissue probability maps with SPM12 (Penny et al. (2011); Ashburner (2012)). The algorithm can generate 3 probability maps for grey matter, white matter, and CSF (cerebrospinal fluid) respectively. We provided the probability maps as the feature to train the random forest classification.

Figure 1 shows the original MRI scan and the results of preprocessing steps.

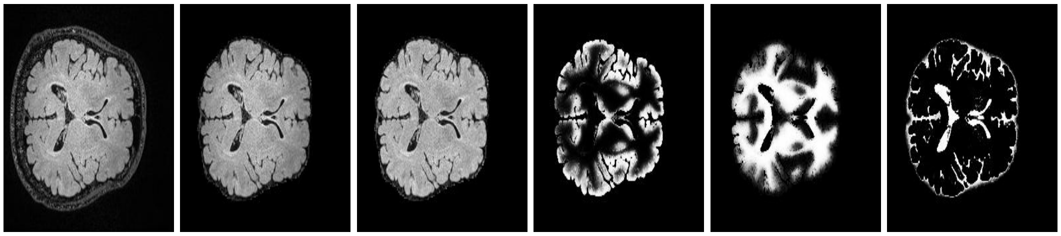

Figure 1: From left to right, the original MRI image, the image after skull stripping, the image after bias field correction, and the tissue probability maps of grey matter, white matter, and CSF.

5 Causal Diagram

According to step-by-step recommendations by Castro et al. (2020), we gathered meta-information of the PPMI dataset, established the predictive causal direction and identified potential dataset mismatch and drew the full casual diagram of our project.

According to the inclusion criteria mentioned in the paper of PPMI(Marek et al. (2011), PD subjects are required to have an asymmetric resting tremor or asymmetric bradykinesia or two of bradykinesia, resting tremor and rigidity with diagnosis within two years and to be untreated for PD and all subjects should undergo dopamine transporter (DAT) imaging and DAT deficit will be required for PD subject eligibility.

This is to say, MRI scans are not one of the diagnosis criterias in the PPMI dataset. Instead, the differences in MRI scans are caused by the disease. Therefore, if a machine learning or deep learning model predicts diagnosis based on MRI scans like the works we mentioned in related work section(Sivaranjini & Sujatha (2020), Esmaeilzadeh et al. (2018)), the task is anticausal, which means it predicts cause from effect. In this case, the researchers working on the PPMI MRI scans are supposed to pay attention to prevalence shift, manifestation and acquisition shift.

Figure 2 shows a full causal diagram and a formal description of acquisition shift.

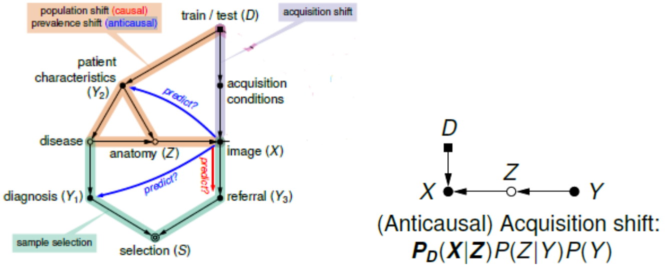

Figure 2: Left: a causal diagram summarizing the typical medical imaging workflows that may apply in the PPMI MRI scans and the potential dataset shift. Right: diagrams for acquisition shift. X is the acquired image, Y is the prediction target, Z is the unobserved true anatomy and D is the domain indicator(in this case it means the acquisition parameters).

After identifying the potential possible acquisition shift in the PPMI MRI dataset, we used deep learning models to verify whether the acquisition shift exists in the dataset.

6 Model

ResNet (He et al. (2016)) is the model we use for most of the experiments. ResNet uses a residual learning framework to make it possible to train deeper neural networks. As a classic and powerful CNN model, it works well in most imaging tasks and requires a relatively reasonable amount of computational resources compared with other state-of-the-art deep learning models. Since the goal of this project is to detect the differences due to different acquisition parameters instead of training the most accurate disease prediction model, ResNet18 is a reasonable choice for this project considering the computational resources limit.

We trained severity score classification models using different settings of ResNet18. 1) Train all the weights and layers. 2) Unfreeze the last two convolution layers and only train weights for those layers. 3) Unfreeze only the last convolution layer and use the pre-trained weights for the other layers. 4) Use pre-trained weights and only train the classifier layers. Since there was no significant difference in result for different settings, we used the third one due to time and resource imitation to evaluate the performance of the model after combining MRI acquisition features into the model, as you can see in Figure 3. Since the MRI acquisition features are categorical ones, we added them to the model using one-hot embedding vectors.

Model has been trained with cross entropy loss function, gradient descent optimizer with learning rate of 0.001 and momentum of 0.9. Also, we have calculated the precision and recall to evaluate and compare the performance of model in different settings. We have reported the f1 score, which is reasonable combination of precision and recall metrics and suitable for our work.

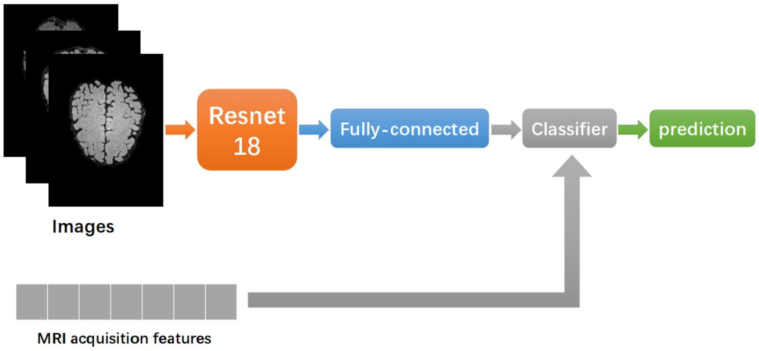

Figure 3: The structure of the model. MRI acquisition features is added to schwab severity score classification model.

7 Experiments and Results

As mentioned in the related work section, the previous study focused on finding the location of the image acquisition and classifying the images based on that information mostly using the machine learning algorithm like random forest. Our original goal was to focus on classifying the images based on their acquisition locations using the deep learning model and see how they can differentiate images. The PPMI dataset was acquired from 20 different locations, and it was a great choice of dataset for our study. However, after plotting the distribution of "siteKey" metadata included in the files, we found that the number of images and subjects per site is not the same or even comparable (Figure 4). After contacting the PPMI data coordinator, we were told that the "siteKey" was not what we were looking for and site information was not disclosed in the PPMI dataset.

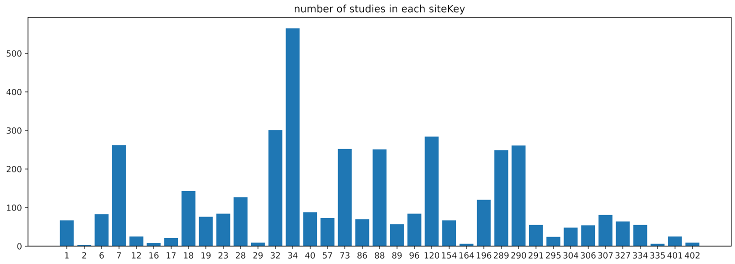

Figure 4: The distribution of siteKey.

We started looking for other metadata and features that can be used as a proxy of siteKey to pursue the study. In the previous study, images were acquired from different locations with pretty similar MRI acquisition parameters. We have looked into the parameters and found out the TE, TR, and TI are more important than the other, and the combination of these three parameters leads to different pulse sequences. Fig 5 is showing the distribution of combinations of TE, TR, and TI in T1 image modalities. As you can see, most of the images have been acquired with the combination of 2.95, 2300, and 900 ms. We have looked into that category more and found out that the images belong to four different manufacturing models ("Mfg model" field in metadata), Biograph_mMR, Prisma_fit, TrioTim, and Verio.

Our first experiment was classifying the images to these four different manufacturing models using the deep learning based classification models.

As you can see in table 1, the dataset is unbalanced in case of distribution in different manufacturing models categories. So, we have trained a model in the original dataset and a random sampling version of that in order to make it a balanced one. In table 2 you can see the results. Neither CNN classification models nor the random forest algorithm was not able to learn features and weights that can classify the manufacturing model of these images. In order to validate the model we are using,

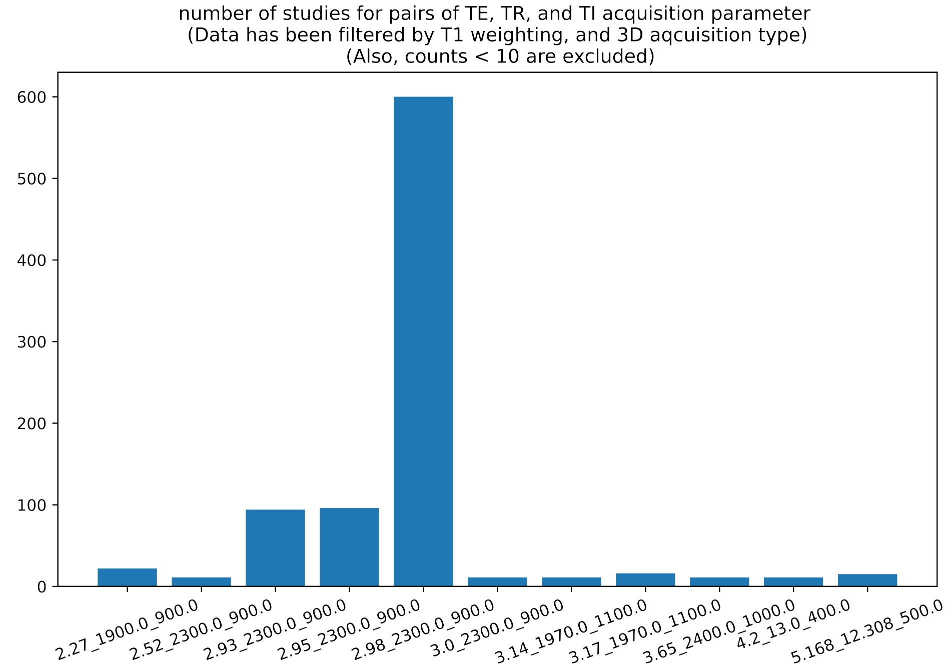

Figure 5: The distribution of TE, TR and TI.

we trained the CNN classification on T1 modality vs T2 modality of MRI images, both in unbalanced and balanced settings. This time the model was able to do classification with considerable performance.

Table 1: The counts of different Mfg Model

Mfg Model	Counts
Biograph_mMR	26
Prisma_fit	17
TrioTim	491
Verio	66

Table 2: First two rows are results of manufacture model classification of most frequent combination of TE, TR, and TI in T1 images, both balanced and unbalanced. Row 3 and 4 are result of balanced and unbalanced T1 vs T2 classification. ”Random” means the model cannot differentiate the acquisition parameter in the experiment.

Experiment	dataset	Result (RF)	Result (DL)
Most frequent TE, TR, TI classification	unbalanced	random	random
Most frequent TE, TR, TI classification	balanced	random	random
T2 vs T1 modality	unbalanced	-	0.94
T2 vs T1 modality	balanced	-	0.96

Based on the results of the first experiments and after consulting with the instructing team, we pursued in a different direction. We added acquisition parameters to the model as features and wanted to see if the features could change the performance of the model. If the model with acquisition parameter features has better performance than the one without these features, it means the acquisition parameters we use leads to acquisition shift. To do it, first, we trained a different ResNet18 based model in case of freezing and trained the weights of different layers of the architecture to classify the T1 images with different schwab severity scores for parkinson disease. For T1 modalities, we have focused on only classifying the images in one of the four most frequent categories of schwab severity scores.

In the table 3 you can see all different experiments for this part. First we trained different models to increase the performance of severity classification. And in the following, in each experiment we added a different MRI acquisition parameter as a feature to the classifier to investigate which one of these parameters will change the performance of severity score model. As you can see, there is not any significant difference in f1 score of these models for different severity scores. Also, for the score of 80 we got zero for all different versions of models. It is worth noting that the distribution of images for each score is not balanced in these experiments.

Table 3: This table includes different experiments on severity score classification without and with adding different MRI acquisition features. The f1 score for test set of epoch 10 is reported for the most four frequent schwab score labels. The "+" in the feature column means that the models' inputs are "preprocessed images + features".

Version	Model	Features	80	90	95	100
v3	ResNet18, training all weights	Preprocessed Images	0	0.61	0.36	0.34
v4	Unfreeze two last Conv layer	Preprocessed Images	0	0.61	0.34	0.34
v5	Unfreeze one last Conv layer	Preprocessed images	0	0.58	0.32	0.35
v6	Freeze all layers of ResNet18	Preprocessed images	0	0.60	0.36	0.34
v1_2	Freeze all layers of ResNet18	+ all MRI acquisition parameter	0	0.60	0.39	0.35
v3_2	Freeze all layers of ResNet18	+ pulse Sequence	0	0.59	0.33	0.33
v4_2	Freeze all layers of ResNet18	+ TE	0	0.62	0.35	0.33
v5_2	Freeze all layers of ResNet18	+ TR	0	0.60	0.38	0.38
v6_2	Freeze all layers of ResNet18	+ TI	0	0.60	0.36	0.36
v7_2	Freeze all layers of ResNet18	+ Manufacturer	0	0.57	0.40	0.35
v8_2	Freeze all layers of ResNet18	+ Mfg Model	0	0.61	0.38	0.35
v9_2	Freeze all layers of ResNet18	+ TE, TR, TI	0	0.59	0.35	0.33

In order to prevent zero f1 score, we have randomly selected the images for each schwab score to create a balanced dataset and trained v6, v3_2, and v7_2 from Table 3. The results are in Table 4.

Table 4: This table include results of selected experiments from Table 3 and train them on balanced dataset. The "+" in the feature column means that the models' inputs are "preprocessed images + features".

Version	Model	Features	80	90	95	100
v6_B	Freeze all layers of ResNet18	Preprocessed images	0.47	0.40	0.50	0.41
v3_2_B	Freeze all layers of ResNet18	+ Pulse Sequence	0.48	0.43	0.42	0.40
v7_2_B	Freeze all layers of ResNet18	+ Manufacturer	0.50	0.43	0.48	0.40

In Figure 6 you can see the confusion matrix of model v7_2 and v7_2_B for both train and test set. Obviously the models are over fitting and for future work, we need to explore more classification model with better performance and then investigate which of the MRI acquisition affects the performance of the model. Also, by training a models on balanced data set, the accuracy of schwab score of 80 is not zero anymore.

8 LIMITATIONS

Our work has several limitations. These limitations may be part of the reasons why we do not get clear conclusions in the experiments.

1. Because the PPMI dataset decides not to disclose the site of MRI scans where the scans were acquired, we lack the site data. Since several previous studies focused on the differences due to sites and were able to detect the differences using machine learning methods,

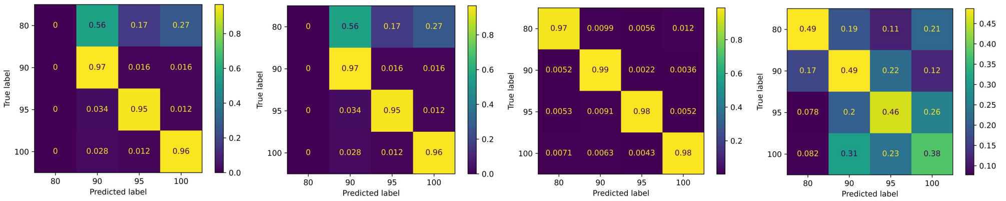

Figure 6: Confusion matrices (CM) for train and test set of model v7_2 and v7_2_B. From left to right, train CM of v7_2, test CM of v7_2, train CM of v7_2_B, and test CM of v7_2_B. Training the model on balanced dataset leads to model without zero accuracy. However, comparing the train and test CM of both model it is obvious both are over fitting. The results are from 10th epoch.

the results may be different if we work on another dataset that includes the site information in the metadata. If so, we could first do the same classification experiment we did with site information. Then if there is difference, we could check if a specific acquisition parameter has a strong correlation with sites and do the same classification experiment. In this way, it is more likely to find the acquisition parameters that may lead to acquisition shift.

2. We used Schwab and England ADL scale to assess the severity of patients in this project. It would be better if we use a more comprehensive and complicated scale called MDS-UPDRS which is also included in the PPMI dataset. The MDS-UPDRS was developed to evaluate various aspects of Parkinson's disease including non-motor and motor experiences of daily living and motor complications(Goetz et al. (2008)). Due to the time limit, we were unable to figure it out and use it in our experiments for now.

3. It would help if we explore images and parameters more by clustering control MR images using the features based on auto-encoder. We failed to do so due to lack of time and GPU ram for now.

4. The PPMI dataset contains the longitudinal information but we did not consider the longitudinal images in our study cohort.

Therefore, if we continue working on this project, there are some directions we can try in the future. We can include other data included in the PPMI dataset, like medical history for PD Severity classification, or use MDS-UPDRS score to improve the PD classification and then investigate if the MRI acquisition parameters are introducing any bias leading to acquisition shift in the dataset. Or we can use another dataset that site information has been disclosed and our model can differentiate the site of MRI acquisition and then investigate if the MRI acquisition parameters introduce any bias. In addition, we can also consult with MRI experts, MRI physicists, or radiologists on the importance and differences of different MRI parameters.

9 Conclusions

It is necessary to draw the causal diagram and be aware of different biases, such as data mismatch or acquisition shift in medical imaging studies.

Machine learning and deep learning model can focus and detect other difference in the images besides medical conditions and that is important to investigate more and find out what are the possibilities.

Contributions

All of us three made equal contributions to this project. Listing order is random. Beiqin explored the PPMI dataset and implemented the skull stripping and bias field correction of the preprocessing part. Also, she explored and read the previous works. Zhongkang integrated the metadata and implemented the tissue class segmentation to obtain the tissue probability maps. Fateme explored metadata information for experiments design possibilities and built the models and trained the models on the dataset. All of us three contributed to the presentation and report writing.

REFERENCES

John Ashburner. Spm: a history. *Neuroimage*, 62(2):791–800, 2012.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

Soheil Esmaeilzadeh, Yao Yang, and Ehsan Adeli. End-to-end parkinson disease diagnosis using brain mr-images by 3d-cnn. *arXiv preprint arXiv:1806.05233*, 2018.

Elisa Ferrari, Paolo Bosco, Giovanna Spera, Maria Evelina Fantacci, and Alessandra Retico. Common pitfalls in machine learning applications to multi-center data: tests on the abide i and abide ii collections. In *Joint Annual Meeting ISMRM-ESMRMB*, 2018.

Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M Thompson, and Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging*, 30(9):1617–1634, 2011.

Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.

Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Alzheimer's Disease Neuroimaging Initiative, et al. Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. *Neuroimage*, 104:398–412, 2015.

Yujin Oh, Sangjoon Park, and Jong Chul Ye. Deep learning covid-19 features on cxr using limited training data sets. *IEEE transactions on medical imaging*, 39(8):2688–2700, 2020.

William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

Christian Salvatore, Antonio Cerasa, Isabella Castiglioni, F Gallivanone, A Augimeri, M Lopez, G Arabia, M Morelli, MC Gilardi, and A Quattrone. Machine learning on brain mri data for differential diagnosis of parkinson's disease and progressive supranuclear palsy. *Journal of neuroscience methods*, 222:230–237, 2014.

S Sivaranjini and CM Sujatha. Deep learning based diagnosis of parkinson's disease using convolutional neural network. *Multimedia tools and applications*, 79(21):15467–15479, 2020.

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.

Christian Wachinger, Benjamin Gutierrez Becker, Anna Rieckmann, and Sebastian Pölsterl. Quantifying confounding bias in neuroimaging datasets with causal inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 484–492. Springer, 2019.