



Predicting Brain Age at Slice Level: Convolutional Neural Networks and Consequences for Interpretability

Pedro L. Ballester^{1*}, Laura Tomaz da Silva², Matheus Marcon^{2,3},
Nathalia Bianchini Esper^{3,4}, Benicio N. Frey^{5,6}, Augusto Buchweitz^{3,5,7} and
Felipe Meneguzzi²

¹ Neuroscience Graduate Program, McMaster University, Hamilton, ON, Canada, ² School of Technology, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, Brazil, ³ BRAINS - Brain Institute of Rio Grande do Sul, Porto Alegre, Brazil, ⁴ Graduate School of Medicine, School of Medicine, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, Brazil, ⁵ Mood Disorders Program, Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada, ⁶ Women's Health Concerns Clinic, St. Joseph's Healthcare, Hamilton, ON, Canada, ⁷ Graduate School of Psychology, School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, Brazil

OPEN ACCESS

Edited by:

James H. Cole,
University College London,
United Kingdom

Reviewed by:

Ryan Muetzel,
Erasmus Medical Center, Netherlands
Sebastian Gabriel Popescu,
Imperial College London,
United Kingdom

*Correspondence:

Pedro L. Ballester
ballestp@mcmaster.ca

Specialty section:

This article was submitted to
Neuroimaging and Stimulation,
a section of the journal
Frontiers in Psychiatry

Received: 24 August 2020

Accepted: 22 January 2021

Published: 25 February 2021

Citation:

Ballester PL, da Silva LT, Marcon M,
Esper NB, Frey BN, Buchweitz A and
Meneguzzi F (2021) Predicting Brain
Age at Slice Level: Convolutional
Neural Networks and Consequences
for Interpretability.
Front. Psychiatry 12:598518.
doi: 10.3389/fpsy.2021.598518

Problem: Chronological aging in later life is associated with brain degeneration processes and increased risk for disease such as stroke and dementia. With a worldwide tendency of aging populations and increased longevity, mental health, and psychiatric research have paid increasing attention to understanding brain-related changes of aging. Recent findings suggest there is a *brain age gap* (a difference between chronological age and brain age predicted by brain imaging indices); the magnitude of the gap may indicate early onset of brain aging processes and disease. Artificial intelligence has allowed for a narrowing of the gap in chronological and predicted brain age. However, the factors that drive model predictions of brain age are still unknown, and there is not much about these factors that can be gleaned from the black-box nature of machine learning models. The goal of the present study was to test a brain age regression approach that is more amenable to interpretation by researchers and clinicians.

Methods: Using convolutional neural networks we trained multiple regressor models to predict brain age based on single slices of magnetic resonance imaging, which included gray matter- or white matter-segmented inputs. We evaluated the trained models in all brain image slices to generate a final prediction of brain age. Unlike whole-brain approaches to classification, the slice-level predictions allows for the identification of which brain slices and associated regions have the largest difference between chronological and neuroimaging-derived brain age. We also evaluated how model predictions were influenced by slice index and plane, participant age and sex, and MRI data collection site.

Results: The results show, first, that the specific slice used for prediction affects prediction error (i.e., difference between chronological age and neuroimaging-derived brain age); second, the MRI site-stratified separation of training and test sets removed site effects and also minimized sex effects; third, the choice of MRI slice plane influences the overall error of the model.

Conclusion: Compared to whole brain-based predictive models of neuroimaging-derived brain age, slice-based approach improves the interpretability and therefore the reliability of the prediction of brain age using MRI data.

Keywords: brain age, deep learning, neuroimaging, convolutional neural networks, model interpretability

1. INTRODUCTION

Brain age prediction involves estimating chronological age based on information typically gleaned from neuroimaging data. The prediction may be referred to as the biological or neuroanatomical age of the brain. Although brain age can be computed from other approaches, such as the epigenetic clock from brain tissue (1), in this paper we use brain age as a synonym for neuroimaging-derived brain age. The difference between the predicted age and the actual chronological age is called *brain age gap*, which has been associated with a number of lifestyle factors (2) [e.g., tobacco and alcohol consumption (3), obesity (4), diabetes, schooling, physical activity (5), higher mortality risk (6), lower fluid intelligence, psychiatric disorders (7), and neurological diseases (8)].

Recent advances in machine learning, specifically on deep convolutional neural networks, have gradually improved brain age prediction by lowering prediction error (9). However, brain age prediction methods still receive criticism due to the lack of interpretability (10). The criticism stems from the limited information about what the model uses to predict brain age, and which regions might bias findings. Hidden biases and poor generalization are a recurrent theme in machine learning and deep learning research (11), including its medical imaging applications (12). Thus to fulfill the promise of translational research, AI needs to establish reliable and reproducible prediction methods, and to generate models that are more amenable to clinical interpretation (10). Identification of clinical neural markers and association with clinical and behavioral data may render AI applications more meaningful (10, 13, 14).

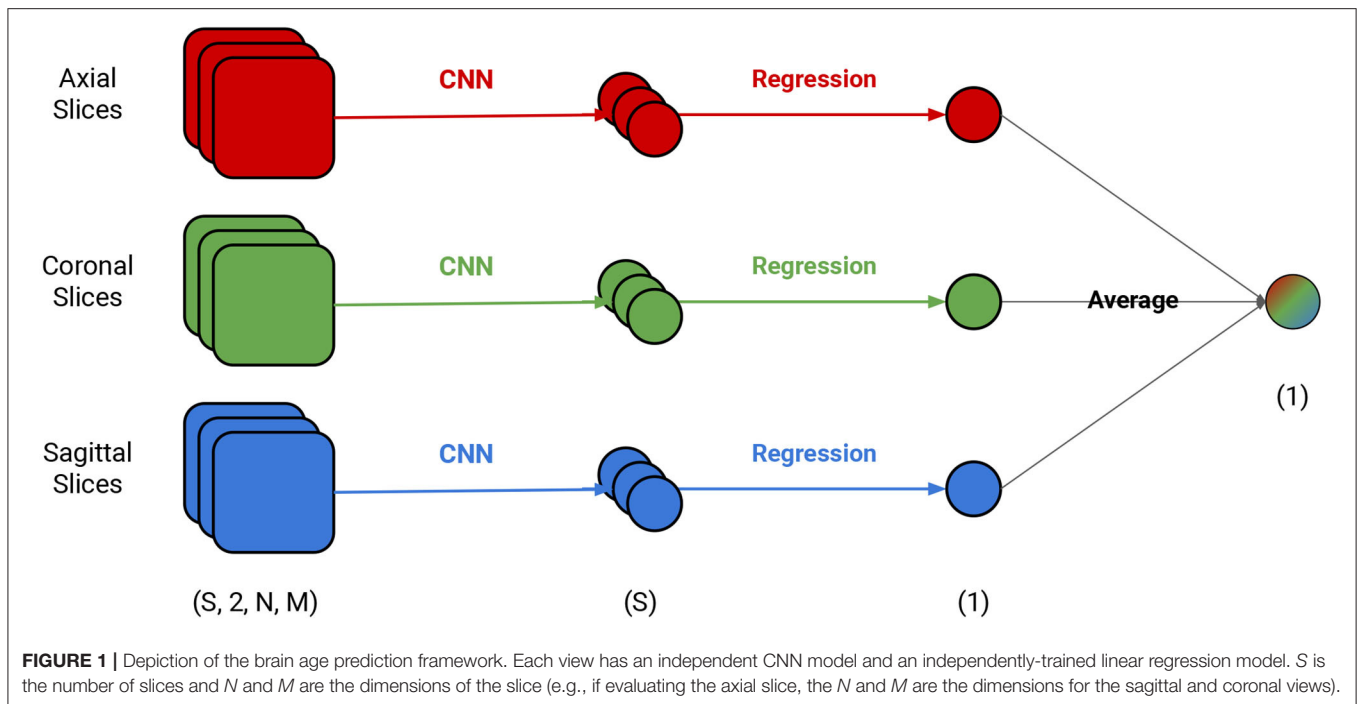
In this article, we report on a model developed for the PAC-2019 brain age prediction competition. Our goal was to generate competitive predictions using meaningful neuroanatomical information. We developed a deep learning framework whose predictions draw on features from every single slice of brain imaging combined with average or linear regression models. The resulting model associates each slice with an independent age prediction for the same patient, allowing researchers to scrutinize the areas of the brain responsible for the overall brain-age gap. Our hypothesis was that our approach would help understand the behavior of brain age prediction at each part of the brain. We also believe that this method, alongside other approaches that try to move away from single predictions of brain age (16), may help us get a comprehensive picture of the parts and characteristics of the aging brain that inform prediction. Such picture should allow for identification of diverse, slice-level, and eventually voxel-level, neuroanatomical traits of age-related diseases.

2. BACKGROUND

The known patterns of brain development associated with aging, such as a decline in gray matter volume (17), are readily identifiable by magnetic resonance imaging (MRI). These images are extensively used for diagnostic and research of disorders associated with brain tissue loss, such as Alzheimer's disease, Parkinsonian dementias, and Frontotemporal lobe degeneration (18). More recently, machine learning techniques have been used to draw on the rich MR images to predict the brain age of healthy people (19), and the aging processes of neurodegenerative disorders (20). The mismatch in chronological and brain age has been investigated in schizophrenia (8), bipolar disorder (21), and in association with factors associated with mortality risk, physical and mental fitness, and biological health (6).

Recent advances in deep learning models, specifically Convolutional Neural Networks (CNNs) achieve state-of-the-art performance in computer vision tasks (22), while requiring little to no prior hand-engineering of data. CNN architectures using 3D convolutions have been used to predict brain age with segmented GM and white matter (WM), and raw T1-weighted MRI scans (10, 23). The use of 3D convolutions allows the model to take in whole-volume information for convolutional filtering operations, which, given enough data, learn feature detection and extraction. CNN models provide highly accurate predictions for regression and classification tasks on multiple medical imaging datasets (22, 24).

CNN models have remarkable predictive power, but the results are typically difficult to interpret. Whereas manual feature selection in classic machine learning simplifies interpretation of the model's results, CNNs require further processing steps to interpret the model's decision processes due to the use of less processed data (15). Examples of interpretation-seeking mechanisms include saliency maps (25) and activation mappings (26), which aim to identify the regions in an image that are responsible for assisting model predictions, thus allowing for some visualization of key input features. These maps trace network outputs back to the input image voxels through the computation of their partial derivatives. For example, regression activation mapping applied to age prediction models on newborn structural MRI generated brain maps of rapid growth during early development (27). Saliency maps have three key limitations. First, they depend on human validation, a time consuming task that also entails the potential for confirmation bias. Second, these methods can produce results that are independent of model and data, and thus inadequate for model debugging and inspection (28). Finally, additional techniques must be used for



combining individual subject saliency maps of into population-level visualizations (29).

3. METHOD

We tested several models and found that the ones with a RESNET18 architecture had a good trade-off between size and prediction error (30). In order to use it in our context with the dimensions of our input, we modified it in three simple ways: (1) the input size had one or two channels, depending on the experiment, (2) the kernel size from the average pool was changed from 7 to 4, (3) the final fully connected layer was changed from 512 to 1,024. The code to build this architecture as well as the steps to reproduce all experiments are available on GitHub (see data availability statement). The input was one brain slice with segmented GM in the first channel and white matter in the second channel. The segmentation was provided by PHOTON-AI¹ and we made no adjustments to it apart from scaling. The output of each model was a brain age estimation for a single segmented slice (GM, WM, or both) from the structural MRI. We illustrate our framework in **Figure 1**. The framework relies on three different, simultaneously-run models that are combined by three linear regression models and a final average. Each model is trained independently to predict brain age from a single MRI slice and draws on different MRI slice orientation (coronal, sagittal, or axial) as input. Each of these models used the validation set to generate error estimates for each slice and each volume. We then used the error estimates to determine the importance of each slice for the model. All three views were

combined to locate specific regions in the brain responsible for a particular classification based on the contribution of each slice for the brain age prediction, which provides an additional source of interpretation.

The final age prediction for a single individual was calculated as follows:

$$a_x = M_x(s^i) \quad i \in [0, S_x] \quad (1)$$

$$age = \frac{\frac{1}{e_a} * L_a(a_a) + \frac{1}{e_s} * L_s(a_s) + \frac{1}{e_c} * L_c(a_c)}{\frac{1}{e_a} + \frac{1}{e_s} + \frac{1}{e_c}} \quad (2)$$

where $x \in a, s, c$ represents the axial, sagittal, and coronal views, a_x is the age vector for each slice for the subject, M_x is one of the CNN models, L_x is the linear regression model, e_x is the error of model M_x in the validation set, and S_x is the total number of slices for an orientation x . Our rationale was that each model's contribution was inversely proportional to the error in the validation set with a weighted average. The influence of each slice for the final prediction is weighted by the linear regression model. This slice-level rationale can also be applied to understand the independent contributions of gray and white matter to the brain age estimate. While using gray or white matter alone causes the model to lose predictive power, it improves interpretability by estimating the independent gray and white matter contributions to the age prediction.

All segmented MRI scans for the dataset provided by the competition were shaped (121, 145, 121). To reduce the amount of empty space on the corners of the input, we removed 20% of the image corners, resulting in a (72, 88, 72) image. The input for the model is thus shaped ($batch, c, 72, 88$) where $c = 1$ when using either gray or white matter alone or $c = 2$ when

¹<https://photon-ai.com/>

using both types of brain tissue. For the coronal view, we zero-pad the image so that the (72, 72) slice also becomes (72, 88) to keep the consistency across all models. The participants from the PHOTON-AI dataset included healthy individuals from a wide age range, males and females, and from 17 different centers. We included basic demographic information about the sample in **Table 1**.

We trained the CNN models using an Adam Optimizer set with the learning rate at $6e - 4$ and weight decay of $6e - 4$. The training also used a sigmoid learning rate rampup for 20 epochs followed by a cosine rampdown until a total of 100 epochs. The batch size was set to 64. We conducted a data augmentation with Elastic Transform (31) with an α range between [28, 30], σ with a range of [3.5, 4.0], and $p = 0.3$,

TABLE 1 | Participants information.

Center	Age mean (std)	Sex (F-M)
0	34.24 (12.67)	197-133
1	26.76 (9.23)	79-55
2	35.51 (12.18)	331-244
3	25.76 (6.62)	18-129
4	21.24 (2.01)	85-58
5	31.25 (7.46)	21-18
6	62.70 (6.75)	3-7
7	43.44 (11.27)	15-10
8	24.82 (5.16)	121-137
9	49.13 (16.62)	255-194
10	33.19 (11.34)	23-51
11	69.92 (7.97)	9-9
12	28.77 (7.77)	16-15
13	41.00 (17.80)	52-76
14	44.41 (22.81)	142-88
15	23.27 (1.27)	16-3
16	22.76 (2.80)	20-9
Total	35.88 (16.21)	1,403-1,236

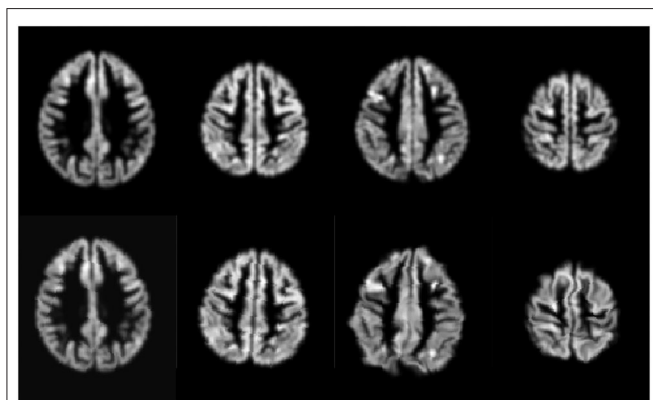


FIGURE 2 | Examples of the augmentation procedure. First row are gray matter segmented images before augmentation; the second row are their augmented counterparts.

representing the scaling factors, the Gaussian spatial smoothing of the deformation field, and the probability of the augmentation being applied.; Random Affine transformations with 4.6 degrees, [0.98, 1.02] scale, and translation of 0.03; finally, we used a Random Tensor Channel Shift with the range of $[-0.1, 0.1]$. Some examples of the augmentation procedure can be seen in **Figure 2**. All data augmentation procedures were implemented in the `medicaltorch` framework².

4. RESULTS

For the competition, we achieved a mean absolute error of 4.44 years on the test set, with a Spearman correlation of -0.25 between the age estimates and chronological age. The model that won the competition achieved 2.90. Due to time restrictions, we employed axial slice predictions only, combining gray and white matter. In what follows, we present the results for the combined gray and white matter models for each separate orientation, the combined predictions; we also present how predictions improve interpretability and decrease model errors. The results of this article are based on predictions made on the competition's validation set. We did not have access to the test set's ground truth at any point during our experiments, prior to nor after the competition.

Instead of using the validation set to train the linear regression, we applied the regression to the training set, in order to avoid circular analysis. However, we believe this can limit the accuracy that otherwise would be achievable with the linear regression model. We leave the comparison of using the validation dataset and reusing the training set to train the linear regression model for future work with more data available, as we restricted ourselves to use data exclusively from the challenge for this study.

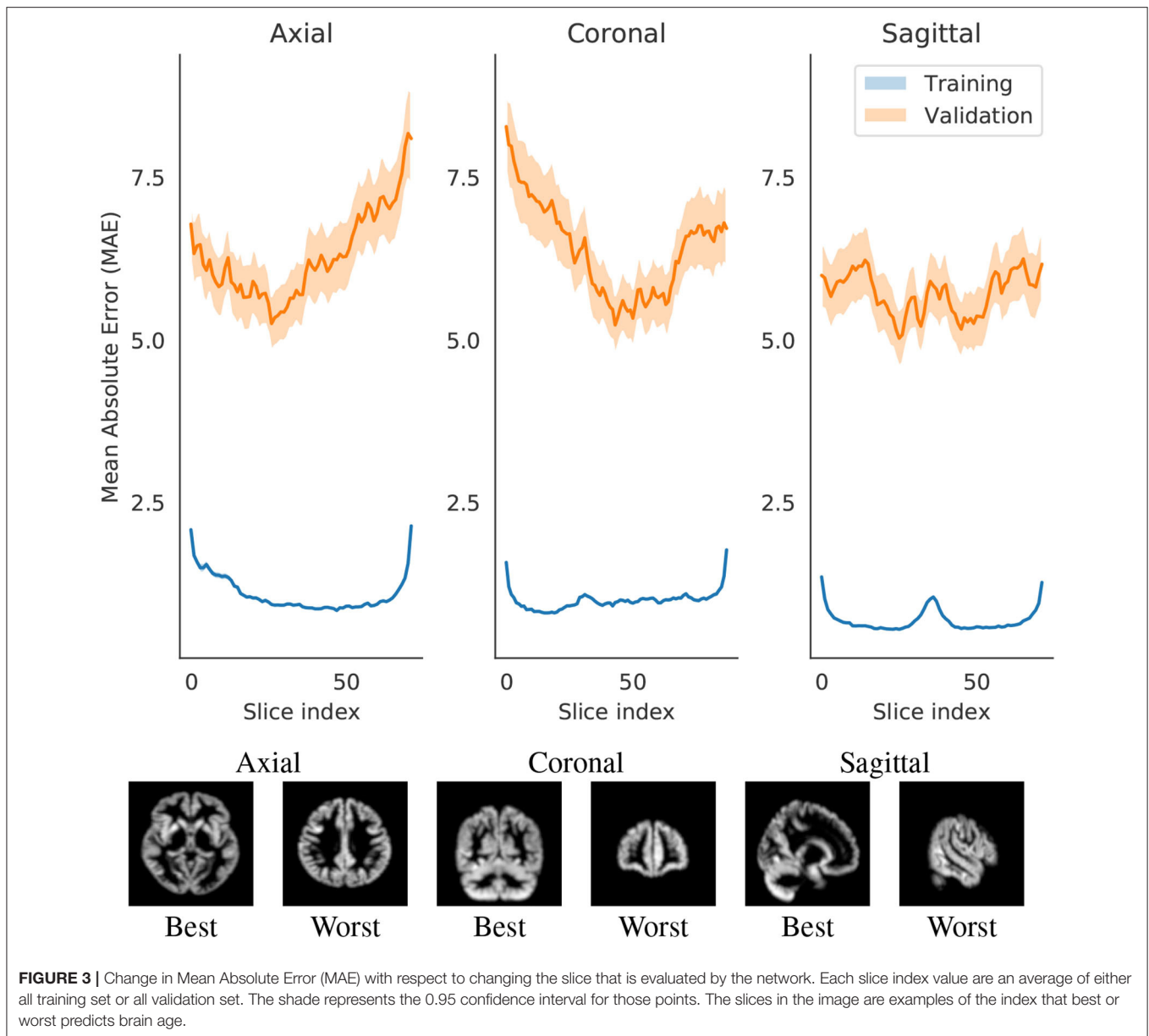
4.1. Combined Gray and White Matter

Our approach used slices for both gray and white matter in individual channels. The use of the two tissues simultaneously may have sacrificed obtaining more fine-grained information about brain aging from the each independent tissue, but it was done in favor of feeding additional data to the model. Gray and white matter were concatenated on the first channel, resulting in an input of (2, 72, 88). We then trained and evaluated the model, and then assessed the effects of age, sex, and site on its predictions. In the sections, we explain the key findings from our experimental analysis.

4.1.1. Models for Different Views of the Brain Have Different Errors

We trained three independent models, taking the input from either axial, coronal, or sagittal views (Note: for the competition, in the interest of time, we used only the axial orientation information). After training, we estimated the error for each slice. The estimate is used to gauge how much each slice contributes to the prediction of brain age.

²<https://github.com/perone/medicaltorch>



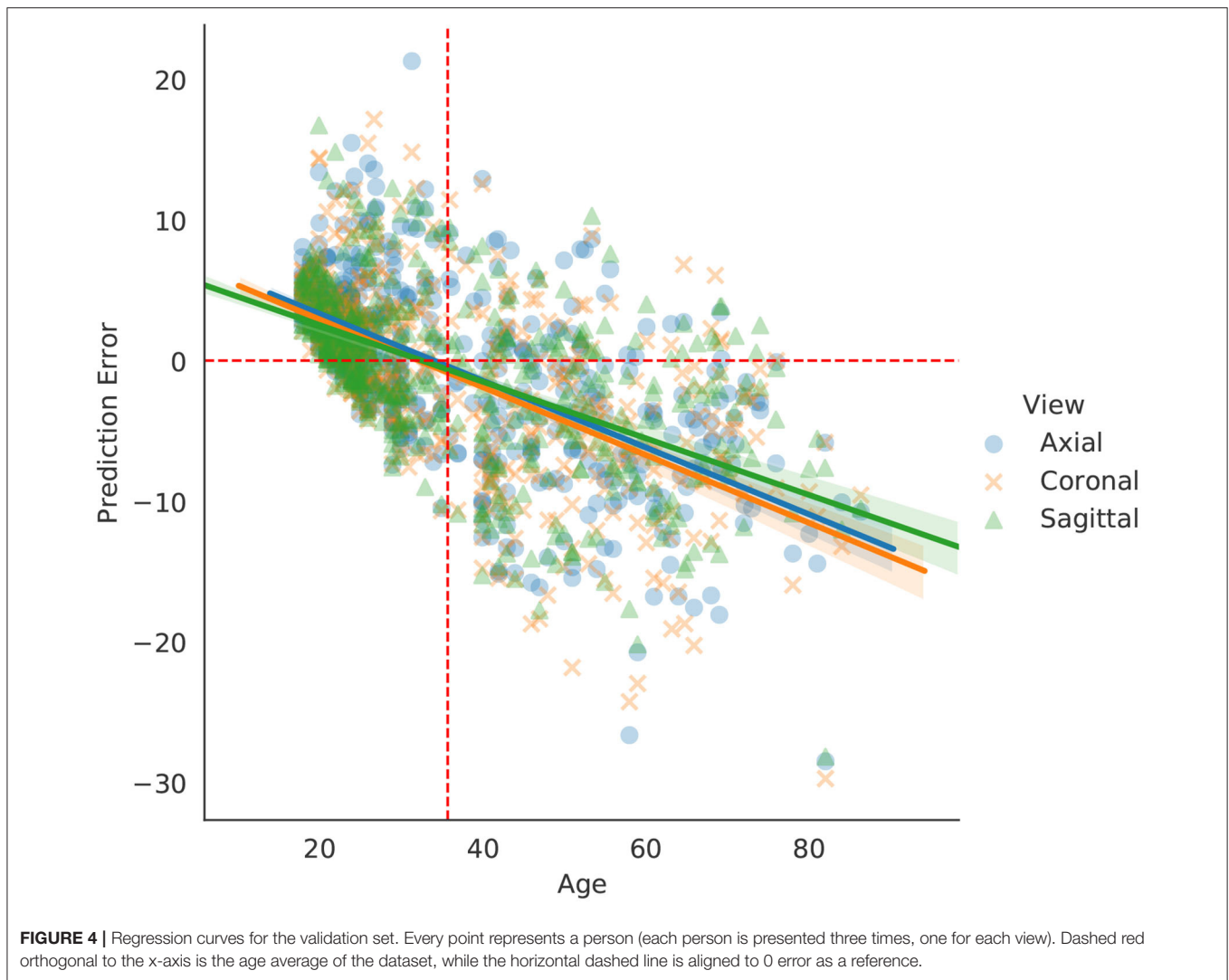
We identified a pattern, most present in axial and coronal models, but to some extent also visible in the sagittal. With more distal slices, the average prediction error seems to be higher. This could arguably be attributed to several reasons: (1) differences of brain matter across regions of the brain, (2) more age-related changes in some regions than others, and (3) the tendency of noise from the scanner to be concentrated on the extremities of the image (and this tendency is fairly visible in **Figure 3**).

The three models afforded different final prediction errors (**Table 2**, which we will further discuss in section 4.1.2). The models also resulted in different estimates for the contribution of slices for predicting brain age. **Figure 3** shows the error variation for slices for the axial, coronal and sagittal slice models. We hypothesize that the differences in mean error may stem from: intrinsic and extrinsic factors that contributed to a

TABLE 2 | Final validation results for all views.

View	Average slice error	Average error	Regression error	R^2
Axial	6.28	4.88	5.09	0.82
Coronal	6.38	4.91	5.04	0.83
Sagittal	5.71	4.45	4.52	0.86
Combined			4.62	0.86

The Average Slice Error column evaluates the final error of our model by doing a simple average across all slices of the volume. The Average Error column uses the same information for the slice column, but instead of calculating the average error for each slice, calculates the mean prediction of all slices, and compare with the actual age. The Regression Error column is the error of using linear regression instead of an average to classify the whole volume. All values are mean absolute errors.



poorer segmentation (and therefore an input of lesser quality); the randomness of the modeling process; and from sample heterogeneity for each of the regions with respect to age.

4.1.2. Final Predictions

Using Equations (1) and (2), we generated the predictions for the final dataset. After pre-training each of three models (one for each view), we evaluated every slice in the dataset and generated a dataset of predictions with a row for each individual and a column for model predictions of every slice. We trained the linear regression for the generated dataset using `scikit-learn`³ library.

Table 2 shows the results for the three models. The sagittal slice prediction showed the lowest error, and it outperformed the outputs combined using Equation (2). We argue that this discrepancy in the error is due to the lack of another validation set for proper out-of-sample error estimation. Additionally, the

differences present in a model trained with this dataset may not accurately translate to other data sources. To ensure that the sagittal slice prediction is actually superior to the axial and coronal slice predictions, the model needs to be further validated and trained using yet another dataset.

4.1.3. Age Effects

Brain age prediction methods usually perform better around the mean chronological age of the dataset. Research shows that models tend to overestimate brain age for participants younger than the mean, and underestimate it for participants older than the mean (32, 33).

The behavior of our model's prediction error changed with respect to age, as shown in **Figure 4**. Not surprisingly, we found the same pattern of over and underestimation of age reported in the literature. One major implication of this type of error is the possibility of inserting biases when trained models are applied to external test sets that have a different age distribution than the training set. Thus, brain age prediction models should

³<https://scikit-learn.org/>

TABLE 3 | Site effects for each orientation and each site. $p < 0.03$ in bold.

Site	View	t-Statistic	p-value
0	Axial	-2.825	0.006
	Coronal	-1.581	0.119
	Sagittal	-1.985	0.051
1	Axial	0.829	0.415
	Coronal	1.904	0.068
	Sagittal	1.788	0.085
2	Axial	3.248	0.002
	Coronal	3.248	0.002
	Sagittal	3.097	0.002
3	Axial	-3.006	0.005
	Coronal	-2.335	0.027
	Sagittal	-2.845	0.008
4	Axial	-5.293	0.000
	Coronal	-6.040	0.000
	Sagittal	-5.188	0.000
5	Axial	-1.945	0.093
	Coronal	-1.111	0.303
	Sagittal	-1.323	0.227
6	Axial	5.178	0.121
	Coronal	2.203	0.271
	Sagittal	22.157	0.029
7	Axial	-0.010	0.992
	Coronal	1.080	0.341
	Sagittal	0.529	0.625
8	Axial	-1.436	0.157
	Coronal	-1.921	0.060
	Sagittal	-0.957	0.343
9	Axial	0.698	0.487
	Coronal	1.845	0.068
	Sagittal	0.914	0.363
10	Axial	0.801	0.437
	Coronal	0.961	0.353
	Sagittal	1.675	0.116
11	Axial	3.146	0.051
	Coronal	3.574	0.037
	Sagittal	1.575	0.213
12	Axial	0.873	0.416
	Coronal	1.203	0.274
	Sagittal	1.387	0.215
13	Axial	-1.973	0.060
	Coronal	-1.293	0.208
	Sagittal	-1.296	0.207

(Continued)

TABLE 3 | Continued

Site	View	t-Statistic	p-value
14	Axial	4.478	0.000
	Coronal	3.960	0.000
	Sagittal	4.114	0.000
15	Axial	-2.058	0.132
	Coronal	-0.393	0.721
	Sagittal	-0.336	0.759
16	Axial	-0.527	0.621
	Coronal	-0.858	0.430
	Sagittal	-0.418	0.693

consider or aim for age-matched test sets (if possible), even if the model has been validated in external datasets. That way, if any underestimation or overestimation of age is happening with the target set, it can be identified and properly dealt with before any conclusions are drawn. Other means of mitigating the effect of testing models on datasets with different age ranges has been addressed elsewhere (33, 34).

4.1.4. Site Effects

The dataset included contributions from 17 different sites. To investigate any biases associated with the different sites, we extracted the age and predicted age for each of our models and each of the 17 sites. We ran a two-tailed dependent t -test to compare age and predicted age among sites and found statistically significant differences for just 6 out of the 17 centers. The data are presented in **Table 3** and **Figure 5**. In four out of the six significantly different centers, all views had significant differences simultaneously, thus suggesting that the models tend to operate in similar manner, and that an actual superiority of the sagittal slice model needs to be further investigated.

4.1.5. Sex Effects

Previous papers suggest that sex can play a role in biasing brain age prediction models. For that reason, we assessed how sex influenced our predictions. We executed a two-tailed paired t -test for age and model predictions for both males and females independently and found no significant differences ($p < 0.03$). We ran a two-tailed unpaired t -test for males and females to see whether predictions or age was significantly different between groups, also with negative findings ($p < 0.03$). **Table 4** summarizes our tests. We applied a two-tailed dependent t -test to compare group means, the test showed no significant differences.

4.1.6. Voxel-Wise Level Brain Age Predictions

We investigated whether a voxel-level model could benefit from the three slice predictions. For each axial slice, information was gathered from each intersecting coronal and sagittal slice, and the combined with the axial prediction to generate an average for each voxel. We show an example of its use in **Figure 6**. We believe there is a multitude of applications and benefits that come with this sort of approach, such as: (1) the exploration of distinct brain

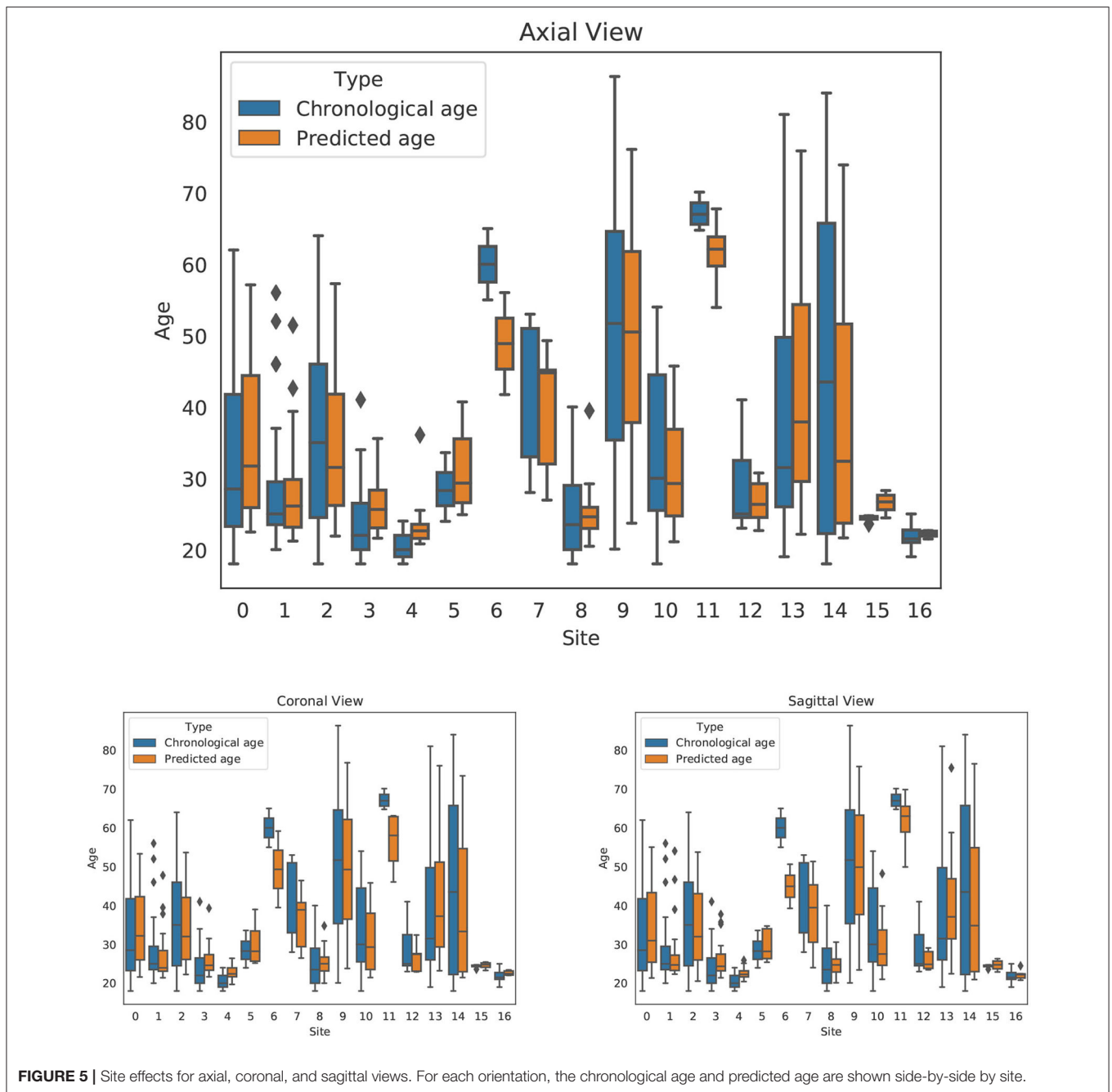
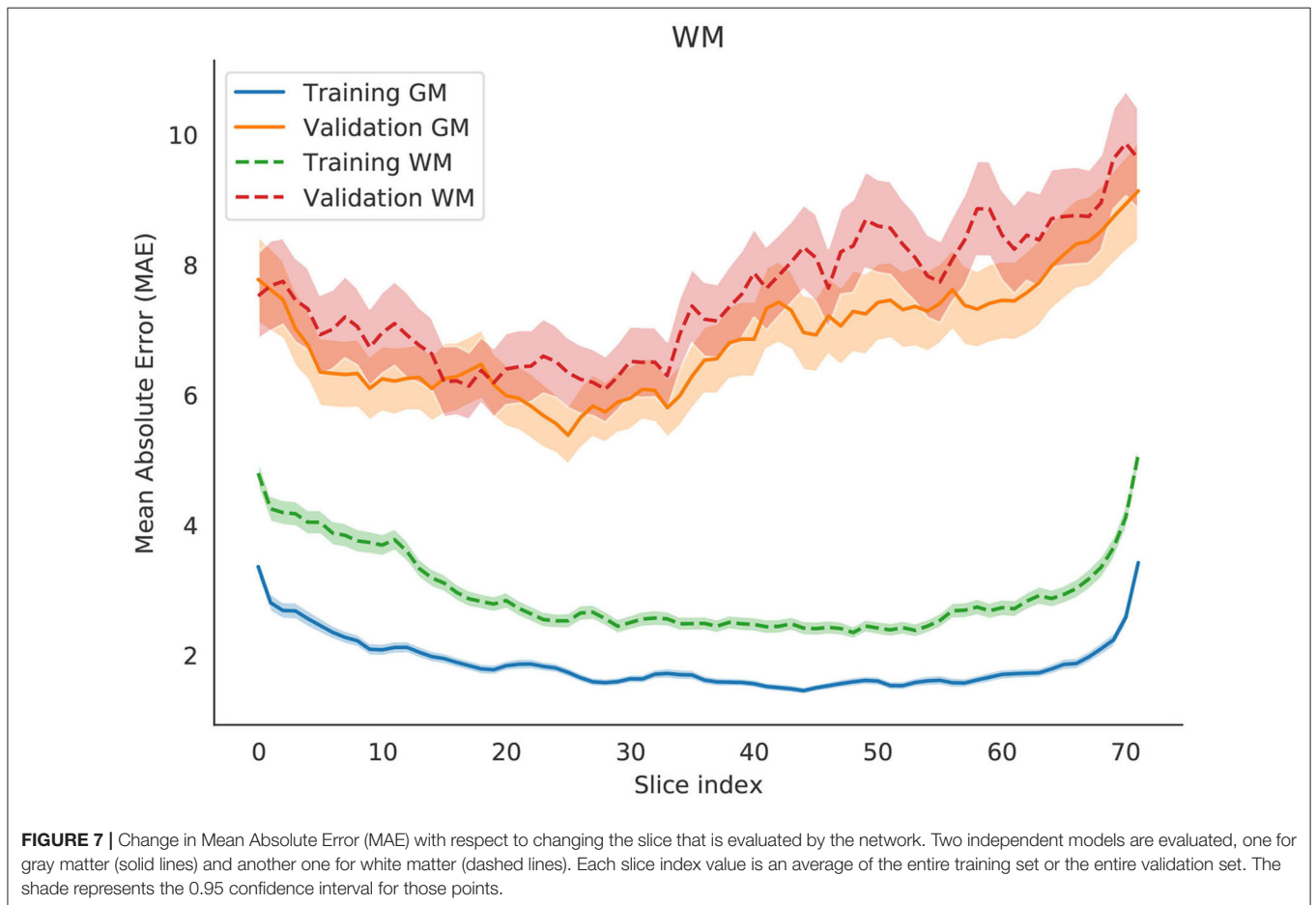
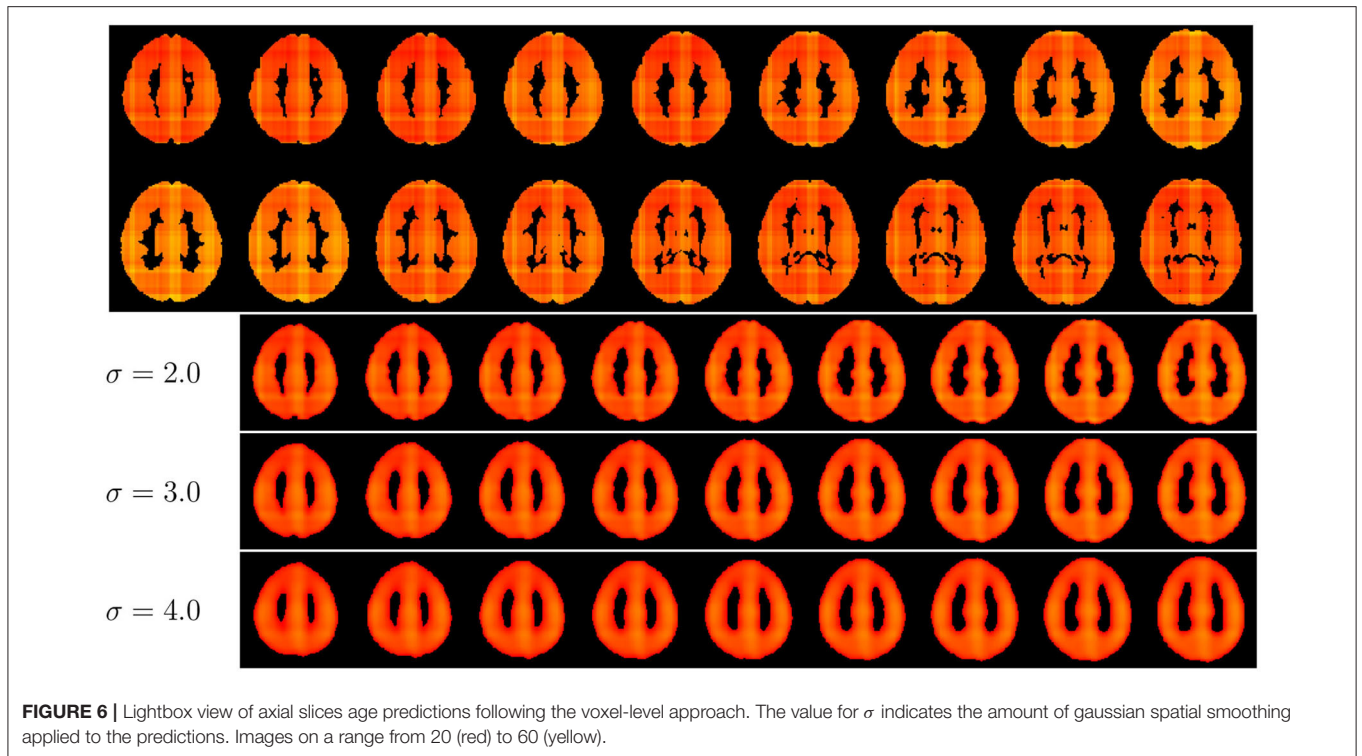


TABLE 4 | Sex effects for each combination of age and prediction and males and females.

Group 1	Group 2	Variable 1	Variable 2	Test type	t	p-value
Sex M	Sex F	Age	Age	I	$t(531) = -1.957$	0.051
Sex M	Sex F	Pred	Pred	I	$t(531) = -2.088$	0.037
Sex M	Sex M	Age	Pred	D	$t(237) = 0.558$	0.578
Sex F	Sex F	Age	Pred	D	$t(293) = 1.423$	0.156

All tests are two-tailed. Results presented here are for the axial orientation model in the validation set. Values in parentheses are the degrees of freedom for each test. I, independent (or unpaired); D, dependent (or paired); Pred, prediction.

aging patterns in different regions of the brain; these patterns do not need to be bounded by anatomically or functionally defined regions, (2) the investigation of region-specific brain aging in neuropsychiatric disorders that present a brain age gap, and (3) the identification of voxel- or region-specific prediction biases (e.g., regions that consistently present higher error). But this approach led to unstable predictions, possibly because the model was not being trained using voxel-level information. There were high frequency changes in predictions for neighboring slices; it is expected that neighboring slices be the opposite. This problem has been documented in deep learning patch-based prediction methods (35). A possible simple but sub-optimal



solution is to use gaussian spatial smoothing to remove the high-frequency changes, as we demonstrate in **Figure 6**. We contend that future approaches may aim to develop means to mitigate the artifacts that come “stitching” slice-level predictions into voxel-level ones. Such a successful approach could improve the current model.

4.2. Independent Gray and White Matter

To estimate the contribution of gray and white matter tissue to the prediction, we created another model (for axial slices) to test age predictions using gray matter and white matter separately. Based on the results in **Figure 7**, the regions that are best predictors of brain age are similar between gray and white matter. The model using solely white matter presented higher errors in most slices. Previous studies have shown that gray matter is a better predictor of brain age than white matter (9).

5. DISCUSSION AND CONCLUSION

Identifying the most predictive regions of the brain for brain age models may help interpreting results, but may also introduce model biases that are unrelated to a neurological condition. Our study presents a model that attempted to balance accuracy and interpretability of the results. Our model provides a level of scrutability for the decision-making process, and can thus help researchers and clinicians understand its limitations.

Given the number of perspectives on interpretability for machine learning (36, 37), we must clarify exactly what we mean by interpretability. Although the interpretability or explainability are commonly used to refer to strategies that explain model predictions, we use it more broadly to define that the final predicted brain age can be attributed to specific slices or regions of the brain. In case additional explanation is needed for a single slice, the usual strategies, carrying the limitations we discussed in section 2, should be applied. We believe that for research purposes, knowing the influence of each part of the brain in the final prediction is imperative to guide and interpret findings of research using the brain age gap.

As a whole, neuropsychiatric research on the brain age gap has mostly focused on associating the difference between the chronological and predicted age to clinical populations. Moreover, most of this research was conducted on the assumption that the brain age gap actually encompasses a larger brain-wide phenomenon responsible for accelerated brain aging.

The body of work on the prediction of brain age at specific regions is growing, two contributions of which are of note in comparison to our contribution. First, a preprint paper (37) uses a similar approach to ours, but instead of using slices, they rely on 3d patches that are later combined with averages or linear regression. Second, a recently published approach (38) uses slice-level predictions, but instead of combining it with linear regression, uses a recurrent neural network for the job.

Attention-based, especially Transformers, models, originally purposed for text-based modeling (39), have recently shown to offer substantial improvement on classification accuracy for

image-based problems (40). However, at the time we developed our approach, attention models were more popular for text-based than image-based tasks. More importantly, we were not aware of any work with neuroimaging data that had shown substantial improvements using those approaches. Indeed, future work should focus on comparing the explanations we provide in terms of slides to the attention maps generated by attention-based models.

Our experiments corroborate some findings from the field. First, we showed that age effects are significant and should need to be accounted for in predictive models of brain age. Second, our results suggest that proper training and test splits that keep site data proportional may mitigate site effects. Third, gray matter seems to be more predictive of age than white matter. Interestingly, our model had similar performance in both male and female sex, although sex is not explicitly used by the model and no separate models were trained.

Since the competition dataset was preprocessed by segmenting gray and white matter, future work should look at less processed data to try to replicate these results. Differences between the unprocessed and segmented inputs might help us understand the extent that possible segmentation errors may influence the behavior of models of brain age.

6. LIMITATIONS

The convolutional neural network fails to combine information from different regions of the brain due to its 2D nature. Although aggregated at later stages of our framework, 3D CNNs might be able to capture patterns that our proposed method misses.

We also found that the aggregated information from each slice prediction to form a voxel-level age prediction to be noisy enough to be unusable. Adjacent voxels had usually the same error but occasionally, while changing the slice index, the prediction had a drastic change. This behavior can probably be attributable to several issues, such as lack of regularization for more stable predictions or possible unidentified problems with the segmentation maps.

DATA AVAILABILITY STATEMENT

The code to reproduce our experiments is available at GitHub (<https://github.com/lisa-pucrs/pac-2019>). Data was provided by PHOTON-AI (<https://www.photon-ai.com/explorer>) during the PAC-2019 challenge.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

PB, LT, MM, and NE designed the study. PB implemented the framework and ran experiments. FM and AB supervised the implementation and engineering of the work. AB and BF helped interpreting the findings and provided neuroimaging-related insights. All authors contributed to writing the manuscript.

FUNDING

NE was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. MM was financed in part by the Conselho Nacional de Pesquisa—Brasil (CNPq).

REFERENCES

- Voisey J, Lawford BR, Morris CP, Wockner LF, Noble EP, Young RM, et al. Epigenetic analysis confirms no accelerated brain aging in schizophrenia. *NPJ Schizophr.* (2017) 3:1–3. doi: 10.1038/s41537-017-0026-4
- Cole JH. Multi-modality neuroimaging brain-age in UK Biobank: relationship to biomedical, lifestyle and cognitive factors. *Neurobiol Aging.* (2020) 92:34–42. doi: 10.1101/812982
- Ning K, Zhao L, Matloff W, Sun F, Toga AW. Association of relative brain age with tobacco smoking, alcohol consumption, and genetic variants. *Sci Rep.* (2020) 10:1–10. doi: 10.1038/s41598-019-56089-4
- Kolenic M, Franke K, Hlinka J, Matejka M, Capkova J, Pausova Z, et al. Obesity, dyslipidemia and brain age in first-episode psychosis. *J Psychiatr Res.* (2018) 99:151–8. doi: 10.1016/j.jpsychires.2018.02.012
- Steffener J, Habeck C, O'Shea D, Razlighi Q, Bherer L, Stern Y. Differences between chronological and brain age are related to education and self-reported physical activity. *Neurobiol Aging.* (2016) 40:138–44. doi: 10.1016/j.neurobiolaging.2016.01.014
- Cole JH, Ritchie SJ, Bastin ME, Hernández MCV, Maniega SM, Royle N, et al. Brain age predicts mortality. *Mol Psychiatry.* (2018) 23:1385–92. doi: 10.1038/mp.2017.62
- Kaufmann T, van der Meer D, Doan NT, Schwarz E, Lund MJ, Agartz I, et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat Neurosci.* (2019) 22:1617–23. doi: 10.1038/s41593-019-0471-7
- Schnack HG, Van Haren NE, Nieuwenhuis M, Hulshoff Pol HE, Cahn W, Kahn RS. Accelerated brain aging in schizophrenia: a longitudinal pattern recognition study. *Am J Psychiatry.* (2016) 173:607–16. doi: 10.1176/appi.ajp.2015.15070922
- Cole JH, Poudel RP, Tsagkrasoulis D, Caan MW, Steves C, Spector TD, et al. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage.* (2017) 163:115–24. doi: 10.1016/j.neuroimage.2017.07.059
- Cole JH, Franke K. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends Neurosci.* (2017) 40:681–90. doi: 10.1016/j.tins.2017.10.001
- Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: *International Conference on Learning Representations*. New Orleans, LA (2019). p. 22. Available online at: <https://openreview.net/forum?id=Bygh9j09KX>
- Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* (2018) 15:e1002683. doi: 10.1371/journal.pmed.1002683
- Just MA, Pan L, Cherkassky VL, McMakin DL, Cha C, Nock MK, et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat Hum Behav.* (2017) 1:911. doi: 10.1038/s41562-017-0234-y
- Heinsfeld, A., Sólón F, Alexandre R, Craddock RC, Buchweitz A, and Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage.* (2018) 17:16–23. doi: 10.1016/j.nicl.2017.08.017
- Da Silva LT, Esper NB, Ruiz DD, Meneguzzi F, Buchweitz A. Visual explanation for identification of the brain bases for dyslexia on fMRI data. *CoRR.* (2020).
- Popescu SG, Cole JH, Sharp DJ, Glocker B. Deep learning methods for estimating "brain age" from structural MRI scans (2018).
- Good CD, Johnsrude IS, Ashburner J, Henson RNA, Friston KJ, Frackowiak RSJ. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage.* (2001) 14:21–36. doi: 10.1006/nimg.2001.0786
- Wattjes MP. Structural MRI. *Int Psychogeriatr.* (2011) 23:S13–24. doi: 10.1017/S1041610211000913
- Franke K, Ziegler G, Klöppel S, Gaser C. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *NeuroImage.* (2010) 50:883–92. Available online at: doi: 10.1016/j.neuroimage.2010.01.005
- Franke K, Gaser C, Initiative f. Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. *GeroPsych.* (2012) 25:235–45. doi: 10.1024/1662-9647/a000074
- Hajek T, Franke K, Kolenic M, Capkova J, Matejka M, Propper L, et al. Brain age in early stages of bipolar disorders or schizophrenia. *Schizophr Bull.* (2019) 45:190–8. doi: 10.1093/schbul/sbx172
- Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12*. Red Hook, NY: Curran Associates Inc. (2012). p. 1097–105.
- Jónsson BA, Bjornsdottir G, Thorgeirsson T, Ellingsen LM, Walters GB, Gudbjartsson D, et al. Brain age prediction using deep learning uncovers associated sequence variants. *Nat Commun.* (2019) 10:1–10. doi: 10.1038/s41467-019-13163-9
- Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005
- Zeiler D, Rob F. (2013). Visualizing and understanding convolutional networks. *arXiv:1311.2901*. (2013).
- Arslan S, Ktena SI, Glocker B, Rueckert D. Graph saliency maps through spectral convolutional networks: application to sex classification with brain connectivity. In: *International Workshop on Integrating Medical Imaging and Non-Imaging Modalities for Healthcare Challenges*. Granada (2018). doi: 10.1007/978-3-030-00689-1_1
- Duffy BA, Liu M, Flynn T, Toga AW, Barkovich AJ, Xu D, et al. Regression activation mapping on the cortical surface using graph convolutional networks. In: *International Conference on Medical Imaging with Deep Learning-Extended Abstract Track*. London (2019). p. 4.
- Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems*. Montreal, QC (2018). p. 9505–15.
- Levakov G, Rosenthal G, Shelef I, Raviv TR, Avidan G. From a deep learning model back to the brain-identifying regional predictors and their relation to aging. *Hum Brain Mapp.* (2020) 41:3235–52. doi: 10.1101/803742
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV (2016). p. 770–8. doi: 10.1109/CVPR.2016.90
- Simard PY, Steinkraus D, Platt JC, et al. Best practices for convolutional neural networks applied to visual document analysis. In: *ICDAR. Vol. 3*. Edinburgh (2003). p. 6.

32. Aycheh HM, Seong JK, Shin JH, Na DL, Kang B, Seo SW, et al. Biological brain age prediction using cortical thickness data: a large scale cohort study. *Front Aging Neurosci.* (2018) 10:252. doi: 10.3389/fnagi.2018.00252
33. Liang H, Zhang F, Niu X. Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. *Hum Brain Mapp.* (2019) 40:3143–52. doi: 10.1002/hbm.24588
34. Le TT, Kuplicki RT, McKinney BA, Yeh HW, Thompson WK, Paulus MP, et al. A nonlinear simulation framework supports adjusting for age when analyzing BrainAGE. *Front Aging Neurosci.* (2018) 10:317. doi: 10.3389/fnagi.2018.00317
35. Pielawski N, Wählby C. Introducing Hann windows for reducing edge-effects in patch-based image segmentation. *PLoS ONE.* (2020) 15:e0229839. doi: 10.1371/journal.pone.0229839
36. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. Long Beach, CA (2017). p. 4765–74.
37. Bintsi KM, Baltatzis V, Kolbeinsson A, Hammers A, Rueckert D. Patch-based brain age estimation from MR images. *arXiv preprint arXiv:200812965.* (2020). doi: 10.1007/978-3-030-66843-3_10
38. Lam PK, Santhalingam V, Suresh P, Baboota R, Zhu AH, Thomopoulos SI, et al. Accurate brain age prediction using recurrent slice-based networks. In: *16th International Symposium on Medical Information Processing and Analysis, Vol. 11583*. Lima: International Society for Optics and Photonics (2020). p. 1158303. doi: 10.1117/12.2579630
39. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inform Process Syst.* (2017) 30:5998–6008.
40. Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A, et al. Image Transformer. In: Dy J, Krause A, editors. *Proceedings of the 35th International Conference on Machine Learning. vol. 80 of Proceedings of Machine Learning Research*. Stockholm (2018). p. 4055–64. Available online at: <http://proceedings.mlr.press/v80/parmar18a.html>

Conflict of Interest: BF had a research grant from Pfizer outside of this study.

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

Copyright © 2021 Ballester, da Silva, Marcon, Esper, Frey, Buchweitz and Meneguzzi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.