Assessing Generalizability of Predictions of Age from Anatomical Images with 3D CNN

Patrick Sadil, Farzad V. Farahani, Tanmay Nath, Martin Lindquist

Johns Hopkins University | Bloomberg School of Public Health | Baltimore, MD

Introduction

Neuroimaging remains an untapped but potentially rich source of biomarkers. However, predictive models trained with neuroimaging data can exhibit substantial bias when they make predictions from scans of participants whose demographics differ from the demographics of the population used to develop the marker (Greene et al., 2022; Li et al., 2022). Although generalizability may improve as neuroimaging datasets become larger, similar biases persist in models that are trained with hundreds of millions of observations (e.g., Buolamwini & Gebru, 2018), and so simply increasing the size of neuroimaging datasets will likely not be sufficient to prevent algorithmic bias. Instead, minimally, models need to be audited to understand the scale of biases before deployment in the clinic.

Here, we explored issues of generalizability for one putative biomarker: the brain age gap. The gap refers to the difference between a participant's chronological age and a model's prediction of that age. Participants whose ages are predicted to be higher tend to have worse health along several axes (e.g., Cole, Marioni, Harris, & Deary, 2019; Cruz-Almeida et al., 2019; Gaser et al., 2013; Smith et al., 2019), and so the gap could serve as a screening test. Therefore, poor generalizability could lead to misallocation of medical resources across demographics. But the extent to which models that predict age are robust to demographic shifts remains unclear.

We focus on two participant characteristics: sex and body-mass index. Sex was chosen due to its possible relationships with features that may be learned by the model. BMI was chosen due to its associations with motion-driven imaging artifacts.

Objectives

- Test generalizability of age predictions by 3D CNN across demographics
- Explore relevancy of anatomical features for age predictions

Materials and Methods

We built a neural network (tensorflow) using a modified version of a previously published architecture (Jonsson et al., 2019). The network was trained, validated, and tested on a 40000-participant subset of the anatomical images from the UK Biobank that had been normalized to MNI152NLin6Sym space and rescaled to 0-1 (age field: 21022).

Training (20k) and validation (2k) datasets were created with varied proportions (100%, 75%, 50%, 25%, 0%) of the provided field for participant sex (field 31; split on male/female) and Body Mass Index (field 21001; split on median: 25.91; cf. normal/overweight boundary of 25). Test sets comprised 2k participants (equal proportions of demographic).

Models were trained with batched gradient descent (256 samples per batch) and Adam (learning rate: 0.001; decay: 10-6; β1: 0.9; β2: 0.999). After visiting the entire training sample, the mean squared error on the validation set was stored. This was repeated until 50 validation evaluations passed without improvement. Final performance was then evaluated on a test split and was measured with both the product-moment correlation and mean absolute error.

References

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Cole, J. H., Marioni, R. E., Harris, S. E., & Deary, I. J. (2019). Brain age and other bodily 'ages': implications for neuropsychiatry. Molecular psychiatry, 24(2), 266-281.

Cruz-Almeida, Y., Fillingim, R. B., Riley III, J. L., Woods, A. J., Porges, E., Cohen, R., & Cole, J. (2019). Chronic pain is associated with a brain aging biomarker in community-dwelling older adults. Pain, 160(5), 1119.

Greene, A. S., Shen, X., Noble, S., Horien, C., Hahn, C. A., Arora, J., ... & Constable, R. T. (2022). Brain–phenotype models fail for individuals who defy sample stereotypes. Nature, 609(7925), 109-118.

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., ... & Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. Science advances, 8(11), eabj1812.

Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., & Miller, K. L. (2019). Estimation of brain age delta from brain imaging. Neuroimage, 200, 528-539.

Results

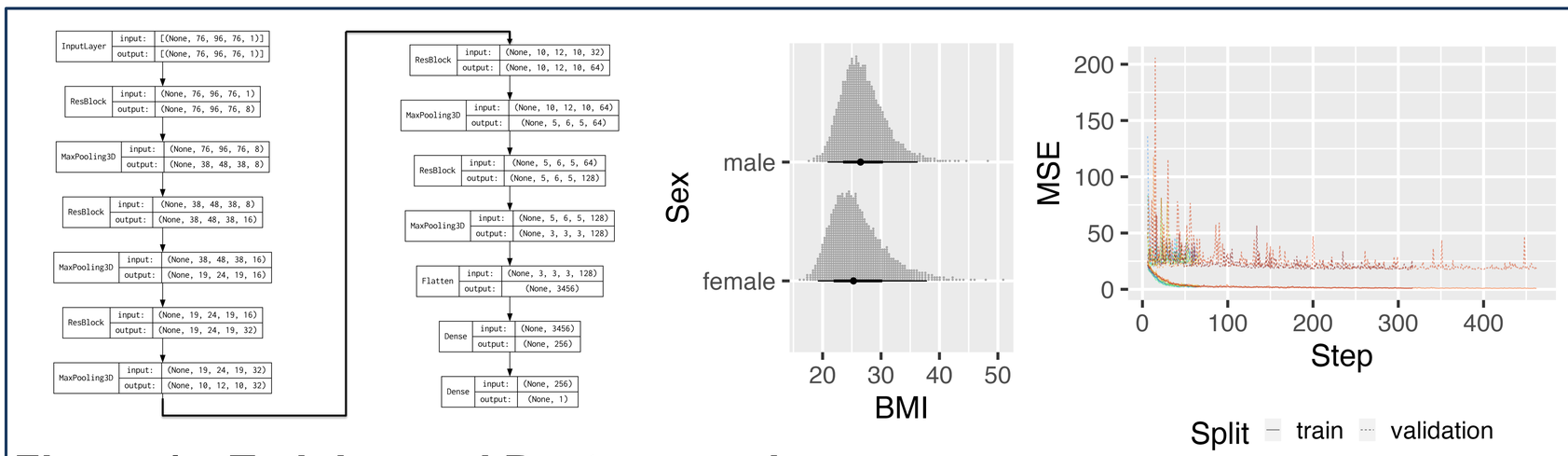

Figure 1—Training and Postprocessing

Left) Model Diagram. **Middle)** BMI. **Right)** Mean Squared Error During Training, All Models.

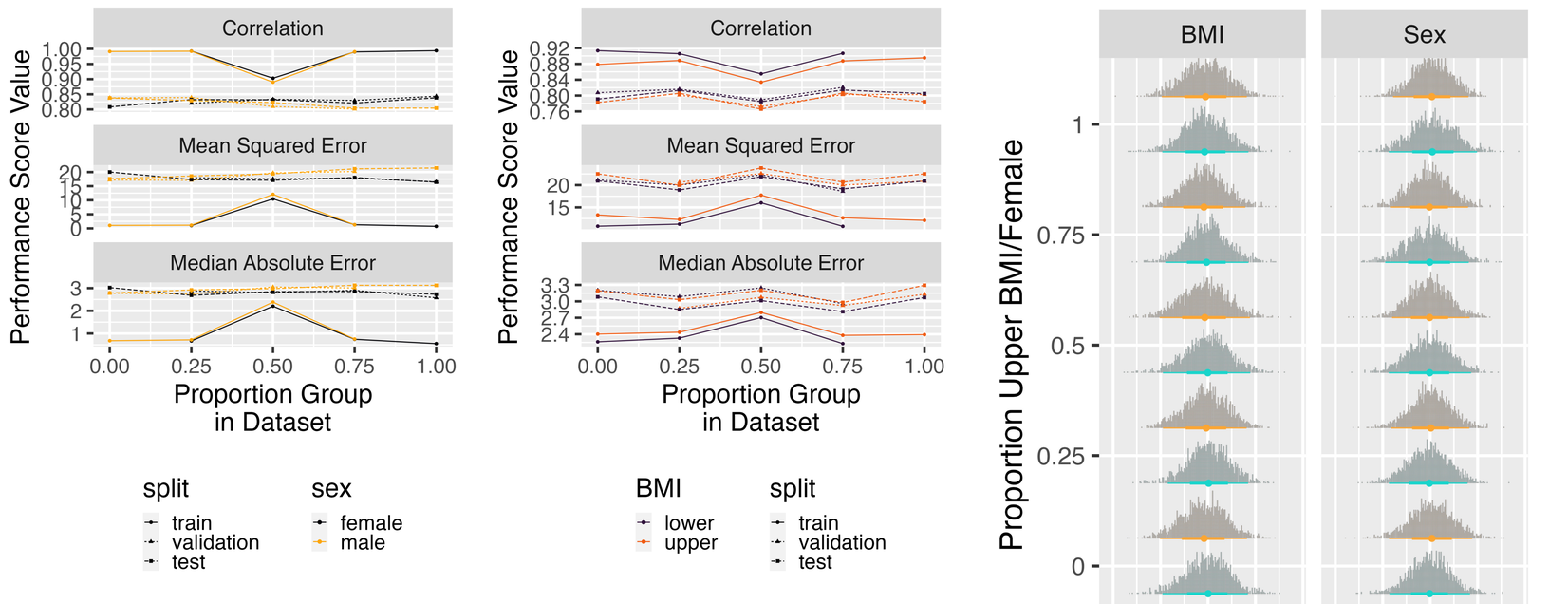

Figure 2 – Age Gap Generalization Across Sex and BMI

Upper Left) Model Performance Across Varied Proportion of Sex, **Upper Right)** Across Varied Proportion of BMI. **Right)** Absolute error of each participant in test split.

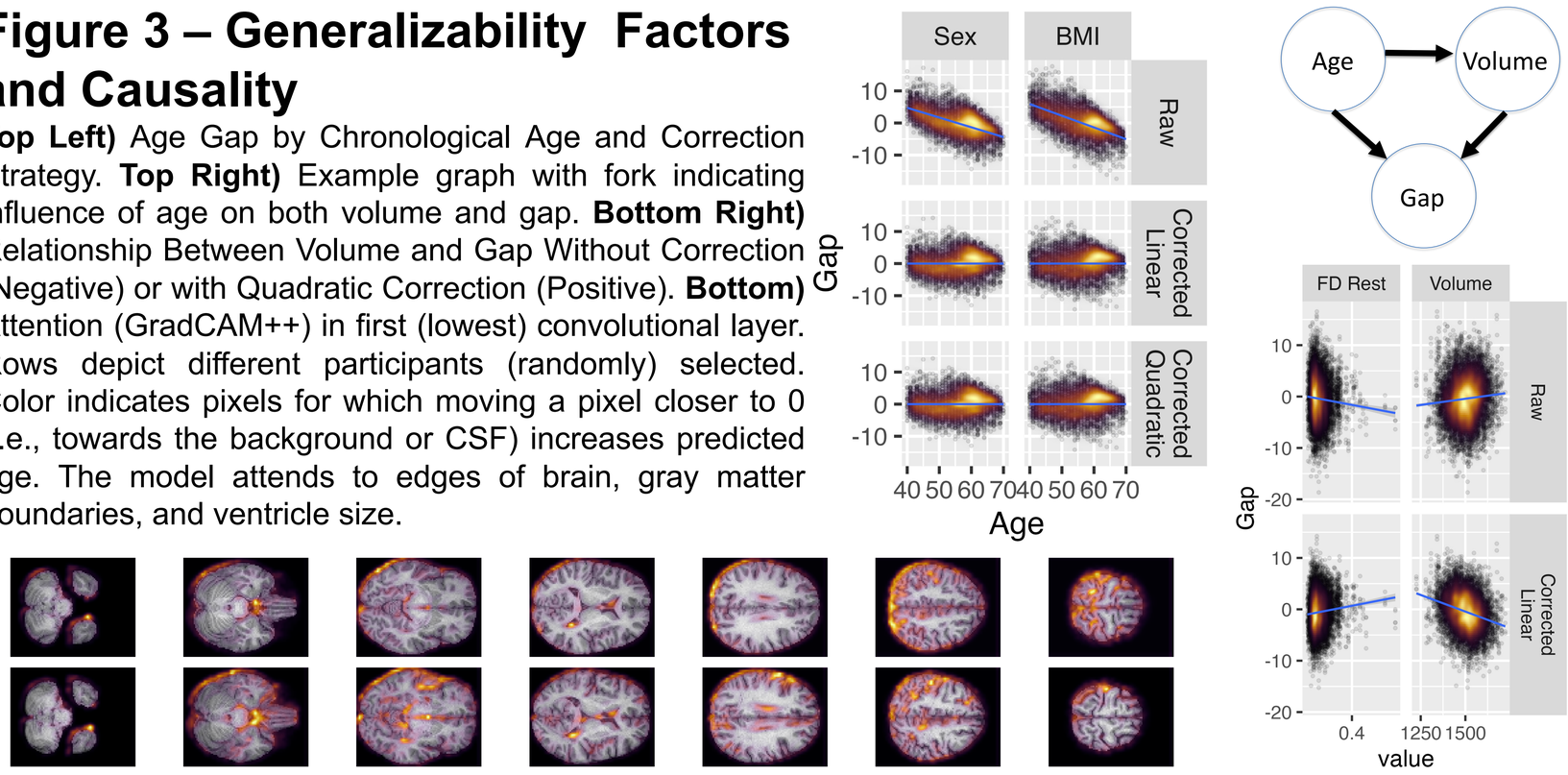

Figure 3 – Generalizability Factors and Causality

Top Left) Age Gap by Chronological Age and Correction Strategy. **Top Right)** Example graph with fork indicating influence of age on both volume and gap. **Bottom Right)** Relationship Between Volume and Gap Without Correction (Negative) or with Quadratic Correction (Positive). **Bottom)** Attention (GradCAM++) in first (lowest) convolutional layer. Rows depict different participants (randomly) selected. Color indicates pixels for which moving a pixel closer to 0 (i.e., towards the background or CSF) increases predicted age. The model attends to edges of brain, gray matter boundaries, and ventricle size.

Conclusion

To ensure biomarkers reliability and fairness, it will be important to understand when they fail. As a case study, we explored generalizability of age predictions across sex and BMI. A trained CNN could generalize across sex and BMI, though demographic shift influenced accuracy. The importance of the observed difference may depend on the clinical setting. Additionally, linking the model predictions to interpretable features may require causal tools.