Beyond Accuracy: Refining Brain-Age Models for Enhanced Disease Detection

Marc-Andre Schulz[1,2,★], Nys Tjade Siegel[1], Kerstin Ritter[1,2]

[1]Charité – Universitätsmedizin Berlin (corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health), Department of Psychiatry and Neurosciences, Berlin, Germany
[2]Bernstein Center for Computational Neuroscience, Berlin, Germany
★Corresponding author: Marc-Andre Schulz, marc-andre.schulz@charite.de

Abstract

This study critically reevaluates the utility of brain-age models within the context of detecting neurological and psychiatric disorders, challenging the conventional emphasis on maximizing chronological age prediction accuracy. Our analysis of T1 MRI data from 46,381 UK Biobank participants reveals a paradox: simpler machine learning models, and notably those with excessive regularization, demonstrate superior sensitivity to disease-relevant changes compared to their more complex counterparts. This counterintuitive discovery suggests that models traditionally deemed less accurate in predicting chronological age might, in fact, offer a more meaningful biomarker for brain health by capturing variations pertinent to disease states. Our findings challenge the traditional understanding of brain-age prediction as normative modeling, emphasizing the inadvertent identification of non-normative pathological markers over precise age prediction.

Introduction

The advancement of neuroimaging techniques has paved the way for innovative measures of brain aging, providing insights into biological aging processes that extend beyond mere chronological metrics (Cole & Franke, 2017). Among these, brain-age prediction models, underpinned by sophisticated machine learning algorithms, have emerged as promising tools for identifying deviations from normative aging trajectories (Baecker et al., 2021). Such deviations, manifesting as discrepancies between predicted and actual brain age—termed the brain-age gap—may signal the early onset of neurological or psychiatric disorders (Chen et al., 2022).
The brain age gap has been used to investigate various diseases of the central nervous system, including Alzheimer's (Gaser et al., 2013), schizophrenia (Koutsouleris et al., 2014), and major depressive disorder (Han et al., 2021). Evidence suggests that a wider brain-age gap correlates with more severe disease symptoms or faster progression, positioning brain age as a potentially sensitive biomarker for these conditions (Cole et al., 2019).

The pursuit of ever-more precise brain age models—from linear regression to convolutional neural networks to cutting-edge vision transformers—raises critical questions about the relationship between model accuracy and clinical utility. Initial observations indicated that models not fully optimized for precision might, paradoxically, reveal more pronounced brain-age gaps, hinting at a complex interplay between model complexity and diagnostic relevance. This has sparked a debate, challenging the assumption that increased accuracy inherently enhances the model's clinical value. The discussion extends to whether these models, in their quest for precision, might overlook subtle signs not directly related to aging yet crucial for identifying at-risk individuals (Bashyam et al., 2020; Hahn et al., 2021; Jirsaraie et al., 2023).

At the heart of this investigation lies a fundamental question: Do these models truly capture the essence of biological aging, or do they inadvertently highlight pathology-specific markers correlated with age? The ambiguity surrounding this question suggests a potential misalignment in the current

modeling approaches, which often prioritize fitting to normative data over capturing the full spectrum of brain health variations.

This study aims to dissect these intricacies, exploring the potential of less complex or intentionally over-regularized models in unearthing clinically relevant biomarkers. By reevaluating the efficacy of brain-age prediction models through this lens, we propose a recalibration of focus towards models that, while perhaps less accurate in conventional terms, may offer greater utility in early disease detection and understanding the multifaceted nature of brain aging.

Figure 1: Illustrative Overview of Brain Age Prediction Model Development and Analysis
a) Development of brain age prediction models using T1 brain images from a healthy control group. b) Comparison of model complexity and accuracy, showing Resnet-50 CNN and SWIN vision transformer models surpassing a regularized linear model in age estimation precision. c) Creation of matched patient and healthy control cohorts to examine brain age disparities across various neurological and psychiatric conditions. d) Inverse relationship observed between the machine learning models' prediction accuracy and the magnitude of brain age gap effect sizes in comparing patient and control groups. e-f) Exploration of the balance between model regularization strength, age prediction accuracy, and the ability to discriminate between patient and control groups, indicating that over-regularization, while reducing age prediction accuracy, enhances patient-control discrimination. g-h) Analysis of the impact of excluding outliers (undiagnosed patients) from the training dataset, which resulted in a decrease in observable brain age differences between patient and control groups, highlighting the influence of hard-to-predict outliers on model performance.

Results

Our investigation centers on discerning whether the brain-age gap mirrors normative aging or captures an alternative, yet significant, biological signal. This study analyzed T1 MRI images from 46,381 UK Biobank participants (Sudlow et al., 2015). We focused on a spectrum of psychiatric disorders—alcohol dependency, bipolar disorder, depression—and neurological disorders, such as Parkinson's disease, epilepsy, and sleep disorders, alongside cognitive and environmental factors including fluid intelligence, recent severe stress exposure, and levels of social support. To ensure rigorous comparison, we applied propensity score matching to generate matched healthy cohorts (Caliendo & Kopeinig, 2008), excluding individuals with any diagnosed neurological or psychiatric condition. Matching criteria included sex, age, education level, household income, the Townsend deprivation index, and genetic principal components, serving as an ethnicity proxy. Figure 1 illustrates our analysis workflow.

We trained three distinct models to explore these dynamics: a regularized linear regression model (Hastie et al., 2009), a 3D ResNet-50 CNN (Fisch et al., 2021; K. He et al., 2016; Kolbeinsson et al., 2020), and a 3D SWIN transformer (Hatamizadeh et al., 2022; Liu et al., 2021), each chosen for its unique approach to learning from imaging data. The linear model leveraged precomputed imaging-derived phenotypes (Alfaro-Almagro et al., 2018), while the deep learning models utilized raw, minimally processed T1 images, aiming to capture the most granular features possible.

Increased Model Accuracy May Diminish Biomarker Effectiveness

We examined how varying complexities of brain age models influence their ability to detect brain-age gaps associated with selected psychiatric and neurological conditions. We posited that simpler models, despite their potential limitations in capturing complex patterns, might exhibit heightened sensitivity to variations relevant to disease states. Conversely, more expressive models could potentially mask these critical signals due to overfitting to normative aging trajectories.

Our findings revealed a distinct pattern: the simplest model, based on ridge regression, demonstrated the most significant brain-age gap effect. This effect diminished with increasing model complexity, with the SWIN transformer (10.1M trainable parameters) showing a reduced effect and the CNN (46.2M trainable parameters) the least (Fig. 2). Intriguingly, the CNN, despite its superior performance in age prediction accuracy, exhibited the smallest disparity in brain-age gap. This counterintuitive result underscores a crucial insight: simpler models may possess a greater capacity to function as effective biomarkers, hinting that the brain-age gap does not solely encapsulate normative aging. The tendency of the more sophisticated CNN model to overfit aging patterns may obscure its ability to detect vital clinical deviations.

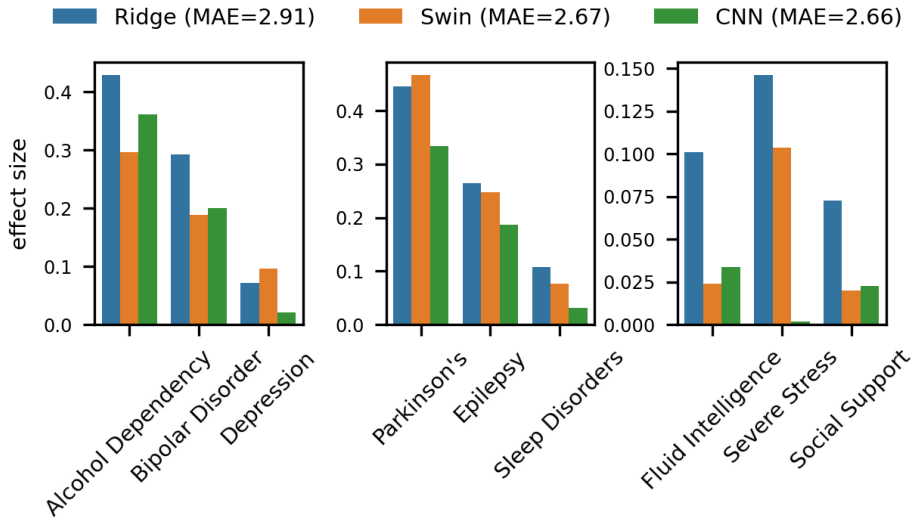

Figure 2: Visualization of Model Complexity vs. Sensitivity in Detecting Brain-Age Gaps
This figure illustrates the nuanced relationship between the complexity of machine learning models and their ability to detect brain-age gaps associated with psychiatric and neurological conditions. It highlights a key observation: models with lower complexity, such as ridge regression, exhibited a heightened sensitivity in identifying deviations that are clinically significant, compared to their more sophisticated counterparts like CNNs and SWIN Transformers. This unexpected finding challenges conventional expectations, suggesting that simplicity in model architecture may enhance biomarker efficacy by preserving sensitivity to critical, disease-relevant variations.

Over-Regularization of ML Models Increases Biomarker Sensitivity

Next, we investigated how increasing regularization in ridge regression models affects their ability to detect the brain-age gap. We hypothesized that excessive regularization might decrease age prediction accuracy but enhance the model's utility as a biomarker by improving its sensitivity to disease-relevant anomalies.

We systematically adjusted the L2 regularization strength in a sequence of ridge regression models tasked with brain age prediction, intentionally exceeding the regularization level that typically optimizes age prediction accuracy (Fig. 3). The findings confirmed that while over-regularization reduced the models' age estimation precision, it notably increased their sensitivity to the brain-age gap associated with selected psychiatric and neurological conditions.

The results suggest that increasing the level of regularization, effectively simplifying the model's representation of aging patterns, enhances its ability to identify variations pertinent to clinical conditions, distinct from conventional aging. This loss in precision for age prediction, counterintuitively, did not diminish the model's value; rather, it bolstered its performance as a biomarker. Such findings support the notion that the brain-age gap encompasses more than mere deviations from a normative aging process, affirming the potential of strategically simplified models to serve as more effective tools in disease detection and characterization.

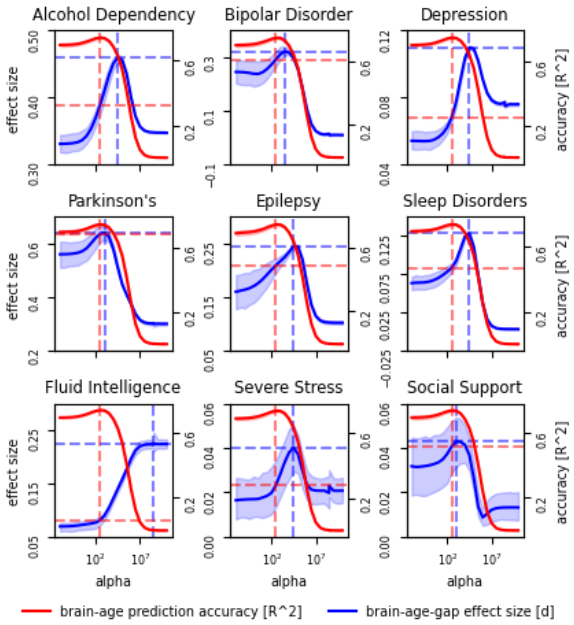

Figure 3: Impact of L2 Regularization on Brain-Age Gap Detection Sensitivity
This figure depicts how adjustments in L2 regularization levels within ridge regression models influence their ability to identify brain-age gaps. Notably, it demonstrates a counterintuitive trend: amplifying the regularization strength past the point optimal for age prediction accuracy significantly boosts the model's ability to detect clinically relevant deviations in cases of psychiatric and neurological conditions. Error bars indicate SEM.

Training on a Healthier Cohort Diminishes Brain-Age Gap Sensitivity

Finally, we aimed to ascertain whether optimizing the training dataset to more closely mirror a normative aging profile would enhance a model's disease-detecting sensitivity. We hypothesized that excluding outliers in favor of age- and sex-matched controls from the general population would produce a model that better represents typical aging patterns. A reduction in brain-age gaps following this data refinement would suggest that original estimates might have been skewed by extraneous variables, rather than indicating actual accelerated aging. Conversely, unchanged gaps would imply these disparities reflect true aging phenomena.

We refined the training dataset by iteratively excluding participants with the largest brain age prediction errors, substituting them with age- and sex-matched individuals from the broader cohort. After each refinement iteration, the models were retrained, and the observed brain-age gaps were evaluated (Fig. 4). Contrary to our initial hypothesis, the exclusion of outliers and the integration of matched controls consistently resulted in diminished brain-age gap effects for both psychiatric and neurological conditions under investigation.

This outcome challenges the prevailing belief that brain-age models predominantly track normative aging trajectories. Models trained on datasets that more accurately reflect a "healthier" aging process exhibited decreased sensitivity to the very deviations that are critical for disease detection, suggesting that prior models may have been inadvertently capturing pathological variations as indicators of aging.

Our collective findings underscore that brain-age gaps may not solely denote accelerated aging. Instead, it appears that the models might be unintentionally capitalizing on pathological markers correlated with age, diverging from a genuine normative aging framework.

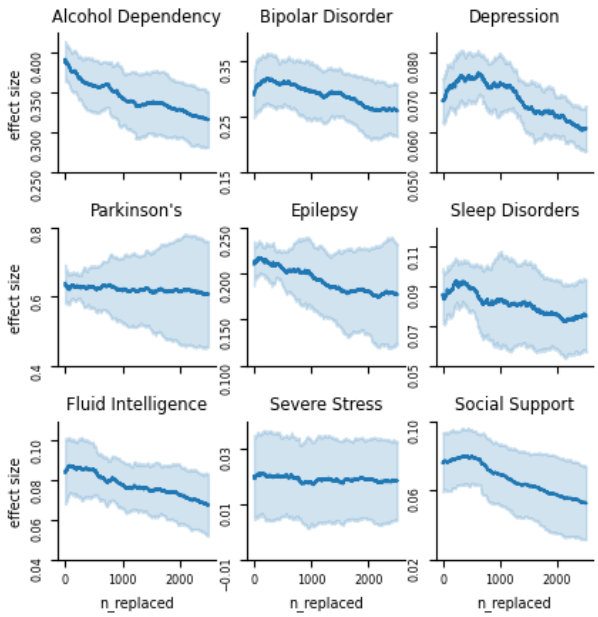

Figure 4: Influence of Dataset Refinement on Brain-Age Gap Detection
This figure illustrates the impact of adjusting the training dataset to better represent a typical aging population by excluding outliers and incorporating matched healthy controls. It reveals that such dataset refinement leads to a noticeable reduction in the effects of brain-age gaps. This observation raises questions about the underlying learning process of brain-age models, suggesting they might not solely delineate a normative aging pathway but could unintentionally emphasize pathological markers not typical of standard aging. Error bars indicate SEM.

Discussion

This study embarked on a journey to reassess the efficacy of brain-age models in detecting neurological and psychiatric disorders, revealing a counterintuitive yet significant paradox: models with lower complexity or those subjected to higher levels of regularization demonstrated enhanced sensitivity to disease-relevant variations compared to their more complex counterparts. Contrary to conventional wisdom, which equates model accuracy with clinical utility, our findings suggest that models traditionally considered less precise in chronological age prediction might, in fact, serve as more potent biomarkers for assessing brain health. This observation points towards a critical reassessment of the criteria used to evaluate the utility of brain-age models in clinical and research settings.

Our results challenge the prevailing narrative that the primary objective of brain-age models should be to mirror chronological aging with utmost accuracy. Instead, they underscore the need for a strategic realignment of model development priorities, emphasizing disease relevance over chronological precision. This shift has broad implications for the design and application of brain-age models, advocating for an approach that values the detection of clinically relevant deviations as a marker of brain health. By focusing on the utility of simpler models or those with intentional over-regularization, our study contributes to a growing body of evidence suggesting that the essence of effective brain-age modeling lies not in capturing the exact chronological age but in identifying variations indicative of underlying pathology.

The preference for simpler models or those with enhanced regularization as more effective biomarkers can be theoretically justified through several lenses. Firstly, complex models, by virtue of their design to capture intricate patterns, may inadvertently overfit to normative aging data. This overfitting can obscure subtle but clinically significant deviations, rendering these models less sensitive to variations that signal disease presence. Conversely, simpler models, with their constrained learning capacity, may avoid this pitfall, thereby remaining more attuned to pathological deviations from typical aging trajectories. Moreover, the principle of Occam's Razor, which advocates for simplicity in the face of complexity when both explain the data adequately, further supports the utility of less complex models in this context. These models, by focusing on broader patterns, might be better positioned to capture the stochastic nature of disease-related changes, which can be drowned out in the noise by more complex models striving for precision in age prediction. This theoretical framework provides a foundation for understanding the observed phenomena and points towards a reevaluation of model complexity in the context of brain-age gap detection.

Our findings challenge the foundational assumption that brain-age models predominantly reflect a normative aging process (Anatürk et al., 2021; Cole et al., 2019; Nguyen et al., 2024; Vidal-Pineiro et al., 2021; Wrigglesworth et al., 2022), highlighting instead that these models might capture a composite of normative and pathological aging signals. This suggests a reconceptualization of what is deemed 'normative,' indicating that the traditional binary distinction between normal and pathological aging may not fully explain the efficacy of brain age gap analyses. Our results underscore the need for a shift in interpreting brain-age gaps, advocating for an understanding that these gaps may signify a constellation of both typical and atypical aging patterns, rather than deviations from a universally normative trajectory.

In contextualizing our work within the existing literature, we build upon and extend several key discussions. Bashyam et al. (2020) initially proposed the idea that under certain conditions, particularly with specific deep neural network architectures, a model that is not fully optimized—termed as 'loosely fitted'—could serve as a more effective biomarker. This assertion, however, was met with skepticism and critiqued by Hahn (2021) who raised both methodological and theoretical concerns, leaving the debate unresolved. Our study addresses this gap by providing systematic evidence across a broader range of models and diseases, underscoring the potential of regularization as a deliberate strategy to enhance biomarker sensitivity.
Moreover, the fundamental inquiry into whether the concept of brain age strictly reflects normative aging or captures a different underlying phenomenon has been regularly mentioned in the literature (Anatürk et al., 2021; Cole et al., 2019; Nguyen et al., 2024; Vidal-Pineiro et al., 2021; Wrigglesworth et al., 2022), but lacks thorough experimental exploration. We venture into this territory, offering empirical evidence that supports a reevaluation of the brain age paradigm.
Additionally, our findings that over-regularized models can outperform their more complex counterparts resonate with a growing body of work questioning the overarching efficacy of deep learning over simpler linear models in brain imaging analyses. This line of inquiry, explored by Schulz (2020, 2024), Han (2022), and He (2020), suggests a critical rethinking of model choice in the context of neuroimaging data. Our contribution not only corroborates this skepticism towards the necessity of deep learning for all tasks within brain imaging but also provides a clearer empirical foundation for when and why simpler or deliberately constrained models might be more advantageous.

However, this study is not without its limitations. Our analysis was confined to a limited set of target phenotypes and utilized a limited range of machine learning models. Future research should aim to replicate these findings across a more diverse array of diseases and model architectures to validate the generalizability of our conclusions. Additionally, the mechanisms by which simpler or more regularized models achieve greater sensitivity to pathological signals warrant further investigation.

Disentangling the features and patterns these models prioritize could offer valuable insights into the biological underpinnings of the brain-age gap phenomenon.

In conclusion, our work contributes to a growing body of evidence that questions the exclusive focus on chronological age prediction accuracy in brain-age modeling. By highlighting the potential of simpler and more regularized models to serve as effective biomarkers for neurological and psychiatric disorders, we advocate for a shift towards models that prioritize clinical relevance and interpretability. This paradigm shift could pave the way for more effective diagnostic tools and intervention strategies, ultimately enhancing patient care in the realm of neurology and psychiatry.

Methods

Dataset and Preprocessing

The study utilized T1-weighted magnetic resonance imaging data from 39,000 participants enrolled in the UK Biobank project (Sudlow et al., 2015). This large-scale biomedical database and research resource contains in-depth genetic and health information from half a million UK participants, with a subset undergoing extensive brain imaging. Our analysis used T1 imaging-derived-phenotypes provided by the UK Biobank (Alfaro-Almagro et al., 2018), specifically 1425 descriptors defined in UK Biobank category 110. Psychiatric disorders (alcohol dependency, bipolar disorder, depression) and neurology-related disorders (Parkinson's disease, epilepsy, sleep disorders), as well as cognitive and environmental factors (fluid intelligence, exposure to severe stress in the last 2 years, level of social support) served as prediction targets. Disorders refer to ICD-10 codes F10, F31, F32, G20, G40, and G47, with time of first diagnosis preceding the date of imaging. Fluid intelligence refers to field 20016, binarized into top and bottom quartile. Exposure to severe stress refers to field 6145, binarized into having or not having an adverse life event in the last 2 years. Level of social support, specifically "able to confide", field 2110, was binarized into the majority category almost daily versus anything less frequent. All features were standard scaled, participants with missing data were excluded on an analysis-by-analysis basis.

Matching

To investigate the association between brain age gaps and various psychiatric and neurological conditions, we employed propensity score matching (Caliendo & Kopeinig, 2008), with a caliper of 0.25 standard deviations. This technique allowed us to create balanced cohorts of participants with and without specific diagnoses, thereby reducing confounding factors. Matching criteria included sex (field 31), age (field 21003), years of education (field 6138, translated into years of education via the 1997 International Standard Classification of Education), household income (field 738), Townsend deprivation index (field 22189), and the first three genetic principal components (field 22009) as proxies for ethnicity.

Experiment 1: Model Training and Architecture Specification

Three distinct machine learning models were trained for the task of brain age prediction: a regularized linear regression model (Hastie et al., 2001), a 3D ResNet-50 convolutional neural network (Fisch et al., 2021; K. He et al., 2016; Kolbeinsson et al., 2020), and a 3D SWIN transformer (Hatamizadeh et al., 2022; Liu et al., 2021). The linear model utilized precomputed IDPs, while the deep learning models were trained on minimally processed (skull stripped and linearly registered, mask and transformation matrix provided by UK Biobank) T1 images. Data was split into healthy participants (no F or G diagnoses) and patients (any F or G diagnoses). Patients (n=17671) and 1000 random healthy participants were kept as a held-out test set, patients with diagnoses that were not investigated in this work served as validation samples, the remaining healthy participants constituted the train set (n=27,538). Brain age prediction accuracy was computed on the healthy test set participants only. Architecture details and training hyperparameters of CNN and transformer are described in (Siegel et al. 2023). In brief, architectures and hyperparameters were kept close to literature or pytorch defaults, and both deep models were trained using AdamW, one-cycle learning rate scheduling, and over 150000 gradient update steps. The linear model relied on the ScikitLearn (Pedregosa et al., 2011) RidgeCV implementation (alpha 10^-5 to 10^5, 100 ticks, log spaced).

Experiment 2: Over-Regularization and Biomarker Sensitivity

We systematically varied the regularization strength in the ridge regression brain age model to study its impact on age prediction accuracy and the detection of brain-age gaps. This experiment aimed to understand how over-regularization, or intentionally reducing model complexity, might influence the model's utility as a biomarker for disease. The train set size was reduced to 5000 to ensure that both

over-regularisation and under-regularisaton effects could be visualized and easily distinguished. Supplementary analyses with larger train set sizes yielded comparable results.

Experiment 3: Impact of Training Cohort Composition
The final experiment sought to assess the effect of training data composition on model performance. By iteratively refining the training set to include healthier cohorts—specifically, by removing participants with large prediction errors and replacing them with age- and sex-matched controls—we evaluated how model sensitivity to disease markers changed. For maximal granularity, this outlier replacement and subsequent reanalysis of brain age gaps happened on the level of individual participants.

Statistical Analysis Details
Statistical analyses and the estimation of uncertainties were performed using bootstrapping techniques, which involved resampling the dataset with replacement to generate empirical confidence intervals. This method was chosen to robustly estimate the uncertainty associated with our predictions and the effect sizes of the brain-age gap, given the complex distribution of the data. All statistical computations were conducted using Python's SciPy and ScikitLearn libraries. Mean Absolute Error (MAE) served as our primary metric for assessing age prediction accuracy, while Cohen's d was used to quantify the effect sizes in a manner that is both standardized and interpretable.

Ethical Considerations
This study was conducted under the umbrella of the UK Biobank's ethics agreements. The UK Biobank has received ethical approval from the National Health Service National Research Ethics Service (16/NW/0274) to collect and distribute data to approved researchers. As our study utilized de-identified data provided by the UK Biobank, individual consent from participants was covered under the UK Biobank's broad consent model.

Data Availability
The dataset analyzed during the current study is sourced from the UK Biobank, a large-scale biomedical database and research resource containing in-depth genetic and health information from half a million UK participants. Due to the UK Biobank's data sharing policy, the data used in this study are available upon application to the UK Biobank by qualified researchers for approved purposes. We have provided comprehensive details on the data preprocessing steps and the features utilized to facilitate replication efforts using similarly structured datasets.

Code Availability
The code developed for this study, including data preprocessing, model training, evaluation, and the bootstrapping procedures used to estimate uncertainties, is available in a public GitHub repository at https://github.com/maschulz/brainage.

Acknowledgments

We thank Roshan Rane, Habakuk Hain, and Marcel Jühling, for feedback on the manuscript. We thank the UKBB participants for their voluntary commitment and the UKBB team for their work in collecting, processing, and disseminating these data for analysis. The manuscript was edited for style and grammar using GPT-4.
Research was conducted using the UKBB resource under project-ID 33073. Computation has been performed on the HPC for Research cluster of the Berlin Institute of Health. The project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project-ID

414984028 - CRC 1404. K.R. was supported by the DFG (389563835, 402170461 - TRR 265, 414984028 - CRC 1404, 459422098 - RU 5363, and 442075332 - RU 5187), the Brain & Behavior Research Foundation (NARSAD young investigator grant), the Manfred and Ursula-Müller Stiftung, and a DMSG research award.

Author Contributions

M.-A.S. designed the study. M.-A.S. and N.T.S. performed the experiments. M.-A.S., N.T.S, and K.R. analyzed the data. M.-A.S. wrote the manuscript draft. M.-A.S., N.T.S, and K.R. revised the manuscript.

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., … Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, *166*, 400–424.

Anatürk, M., Kaufmann, T., Cole, J. H., Suri, S., Griffanti, L., Zsoldos, E., Filippini, N., Singh-Manoux, A., Kivimäki, M., Westlye, L. T., Ebmeier, K. P., & de Lange, A.-M. G. (2021). Prediction of brain age and cognitive age: Quantifying brain and cognitive maintenance in aging. *Human Brain Mapping*, *42*(6), 1626–1640.

Baecker, L., Garcia-Dias, R., Vieira, S., Scarpazza, C., & Mechelli, A. (2021). Machine learning for brain age prediction: Introduction to methods and clinical applications. *EBioMedicine*, *72*, 103600.

Bashyam, V. M., Erus, G., Doshi, J., Habes, M., Nasralah, I., Truelove-Hill, M., Srinivasan, D., Mamourian, L., Pomponio, R., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., … Davatzikos, C. (2020). MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain: A Journal of Neurology*, *143*(7), 2312–2324.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Chen, C.-L., Hwang, T.-J., Tung, Y.-H., Yang, L.-Y., Hsu, Y.-C., Liu, C.-M., Lin, Y.-T., Hsieh, M.-H., Liu, C.-C., Chien, Y.-L., Hwu, H.-G., & Tseng, W.-Y. I. (2022). Detection of advanced brain aging in schizophrenia and its structural underpinning by using normative brain age metrics. *NeuroImage. Clinical*, *34*, 103003.

Cole, J. H., & Franke, K. (2017). Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. *Trends in Neurosciences*, *40*(12), 681–690.

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.

Fisch, L., Ernsting, J., Winter, N. R., Holstein, V., Leenings, R., Beisemann, M., Sarink, K., Emden, D., Opel, N., Redlich, R., Repple, J., Grotegerd, D., Meinert, S., Wulms, N., Minnerup, H., Hirsch, J. G., Niendorf, T., Endemann, B., Bamberg, F., … Hahn, T. (2021). Predicting brain-age from raw T 1 -weighted Magnetic Resonance Imaging data using 3D Convolutional Neural Networks. In *arXiv [eess.IV]*. arXiv. http://arxiv.org/abs/2103.11695

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., & Alzheimer’s Disease Neuroimaging Initiative. (2013). BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer’s Disease. *PloS One*, *8*(6), e67346.

Hahn, T., Fisch, L., Ernsting, J., Winter, N. R., Leenings, R., Sarink, K., Emden, D., Kircher, T., Berger, K., & Dannlowski, U. (2021). From “loose fitting” to high-performance, uncertainty-aware brain-age modelling [Review of *From “loose fitting” to high-performance, uncertainty-aware brain-age modelling*]. *Brain: A Journal of Neurology*, *144*(3), e31.

Han, J., Kim, S. Y., Lee, J., & Lee, W. H. (2022). Brain Age Prediction: A Comparison between Machine Learning Models Using Brain Morphometric Data. *Sensors* , *22*(20). https://doi.org/10.3390/s22208077

Han, L. K. M., Dinga, R., Hahn, T., Ching, C. R. K., Eyler, L. T., Aftanas, L., Aghajani, M., Aleman, A., Baune, B. T., Berger, K., Brak, I., Filho, G. B., Carballedo, A., Connolly, C. G., Couvy-Duchesne, B., Cullen, K. R., Dannlowski, U., Davey, C. G., Dima, D., … Schmaal, L. (2021). Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group. *Molecular Psychiatry*, *26*(9), 5124–5139.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY.

Hastie, T., Tibshirani, T., & Friedman, J. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2022). Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 272–284.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo, B. T. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, *206*, 116276.

Jirsaraie, R. J., Gorelik, A. J., Gatavins, M. M., Engemann, D. A., Bogdan, R., Barch, D. M., & Sotiras, A. (2023). A systematic review of multimodal brain age studies: Uncovering a divergence between model accuracy and utility. *Patterns (New York, N.Y.)*, *4*(4), 100712.

Kolbeinsson, A., Filippi, S., Panagakis, Y., Matthews, P. M., Elliott, P., Dehghan, A., & Tzoulaki, I. (2020). Accelerated MRI-predicted brain ageing and its associations with cardiometabolic and brain disorders. *Scientific Reports*, *10*(1), 19940.

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., Reiser, M., Pantelis, C., & Meisenzahl, E. (2014). Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. *Schizophrenia Bulletin*, *40*(5), 1140–1153.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2103.14030

Nguyen, H.-D., Clément, M., Mansencal, B., & Coupé, P. (2024). Brain structure ages-A new biomarker for multi-disease classification. *Human Brain Mapping*, *45*(1), e26558.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, *12*(Oct), 2825–2830.

Schulz, M.-A., Bzdok, D., Haufe, S., Haynes, J.-D., & Ritter, K. (2024). Performance reserves in brain-imaging-based phenotype prediction. *Cell Reports*, *43*(1), 113597.

Schulz, M.-A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., & Bzdok, D. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, *11*(1), 4238.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779.

Vidal-Pineiro, D., Wang, Y., Krogsrud, S. K., Amlien, I. K., Baaré, W. F. C., Bartres-Faz, D., Bertram, L., Brandmaier, A. M., Drevon, C. A., Düzel, S., Ebmeier, K., Henson, R. N., Junqué, C., Kievit, R. A., Kühn, S., Leonardsen, E., Lindenberger, U., Madsen, K. S., Magnussen, F., … Fjell, A. (2021). Individual variations in “brain age” relate to early-life factors more than to longitudinal brain change. *eLife*, *10*. https://doi.org/10.7554/eLife.69995

Wrigglesworth, J., Ryan, J., Ward, P. G. D., Woods, R. L., Storey, E., Egan, G. F., Murray, A., Espinoza, S. E., Shah, R. C., Trevaks, R. E., Ward, S. A., & Harding, I. H. (2022). Health-related heterogeneity in brain aging and associations with longitudinal change in cognitive function. *Frontiers in Aging Neuroscience*, *14*, 1063721.