PNAS RESEARCH ARTICLE | NEUROSCIENCE

Anatomically interpretable deep learning of brain age captures domain-specific cognitive impairment

Chenzhong Yin[a,1], Phoebe Imms[b,1], Mingxi Cheng[a,1], Anar Amgalan[b,1], Nahian F. Chowdhury[b,1], Roy J. Massett[b], Nikhil N. Chaudhari[b,c], Xinghe Chen[a], Paul M. Thompson[c,d,e,f,g,h,i,j], Paul Bogdan[a,1], Andrei Irimia[b,c,e,1,2], and the Alzheimer's Disease Neuroimaging Initiative[3]

Edited by Brian Wandell, Stanford University, Stanford, CA; received August 26, 2022; accepted November 10, 2022

The gap between chronological age (CA) and biological brain age, as estimated from magnetic resonance images (MRIs), reflects how individual patterns of neuroanatomic aging deviate from their typical trajectories. MRI-derived brain age (BA) estimates are often obtained using deep learning models that may perform relatively poorly on new data or that lack neuroanatomic interpretability. This study introduces a convolutional neural network (CNN) to estimate BA after training on the MRIs of 4,681 cognitively normal (CN) participants and testing on 1,170 CN participants from an independent sample. BA estimation errors are notably lower than those of previous studies. At both individual and cohort levels, the CNN provides detailed anatomic maps of brain aging patterns that reveal sex dimorphisms and neurocognitive trajectories in adults with mild cognitive impairment (MCI, $N = 351$) and Alzheimer's disease (AD, $N = 359$). In individuals with MCI (54% of whom were diagnosed with dementia within 10.9 y from MRI acquisition), BA is significantly better than CA in capturing dementia symptom severity, functional disability, and executive function. Profiles of sex dimorphism and lateralization in brain aging also map onto patterns of neuroanatomic change that reflect cognitive decline. Significant associations between BA and neurocognitive measures suggest that the proposed framework can map, systematically, the relationship between aging-related neuroanatomy changes in CN individuals and in participants with MCI or AD. Early identification of such neuroanatomy changes can help to screen individuals according to their AD risk.

brain age | cognitive impairment | Alzheimer's disease | deep learning

Significance

The phenotypic age of the human brain, as revealed via deep learning of anatomic magnetic resonance images, reflects patterns of structural change related to cognitive decline. Our interpretable deep learning estimates that the brain ages more accurately than any other approaches to date. Furthermore, compared to chronological age, our inferred brain ages are significantly more strongly associated with early signs of Alzheimer's disease. Maps conveying the importance of each brain region for estimating brain age reveal differences in patterns of neurological aging between males and females and between persons with and without cognitive impairment. These findings provide insight into early identification of persons at high risk of Alzheimer's disease.

Although chronological age (CA) reflects disease risk, the rate of aging varies across individuals, organs, tissues, and clinical conditions (1). Because CA does not capture this variation well, there is interest in estimating biological age to predict morbidity (2, 3). Among typically aging adults, in the absence of any clinical indications, biological age is expected to equal CA, on average (4). Neuroanatomic biological age inferred from MRI, henceforth referred to as brain age (BA), can quantify disease-related changes in aging and associated increases in mortality risk (5, 6). Thus, reliable BA estimators can help to stratify individuals according to disease risk (7, 8). The difference between BA and CA, known as age gap (AG), conveys whether aging is faster or slower than expected (9, 10). In clinical cohorts, improving BA estimates can translate into better estimates of participants' deviations from typical aging (11, 12). For example, BA has the potential to become an affordable and noninvasive preclinical indicator of mild cognitive impairment (MCI) and Alzheimer's disease (AD) (13) due to the strong association between BA and dementia risk (14, 15).

Deep learning (DL) methods can estimate BA by learning to estimate cognitively normal (CN) subjects' CAs from MRIs of their brain, while minimizing the mean absolute error (MAE) between BA and CA (16). Compared to other approaches, DL typically yields better BA estimates (17). However, its inherent black-box nature hinders the interpretability of its feature attribution (18), since the relative utility of regional brain features for BA estimation by DL methods is unknown. Furthermore, many DL estimators of BA are inaccurate and lack generalizability to cohorts not encountered during DL training. To address these shortcomings, we introduce an interpretable three-dimensional (3D) convolutional neural network (CNN) to estimate BA from T_1-weighted brain MRIs. To provide neuroanatomic interpretability, MRI feature attribution is achieved through saliency maps. These allow one to identify structural brain patterns of CN aging that reflect regional and sex-specific variations in neuroanatomic features reflecting BA. 3D-CNN generalizability to new cohorts is also illustrated. The translational potential of this study is reflected in the associations between estimated BAs and neurocognitive measures of CI.

Competing interest statement: The authors have research support to disclose, P.M.T. discloses research grant support from Biogen, Inc. for work unrelated to this study.

This article is a PNAS Direct Submission.

[1]C.Y., P.I., M.C., A.A., N.F.C., P.B., and A.I. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: irimia@usc.edu.

[3]A complete list of the Alzheimer's Disease Neuroimaging Initiative can be found in the *SI Appendix*.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2214634120/-/DCSupplemental.

Published January 3, 2023.

Results

Neuroanatomic Patterns of Aging. We use an interpretable 3D-CNN framework to estimate the BAs of 650 CN adults (age range: 18 to 88 y; 325 males) from the Cambridge Centre for Aging and Neuroscience (CamCAN, Fig. 1 *A* and *C*). BAs were also estimated in 359 participants with AD dementia (age range: 55 to 92 y; 198 males) and in 351 participants with MCI due to AD (age range 55 to 89; 230 males) from the Alzheimer's Disease Neuroimaging Initiative (ADNI, Fig. 1*A*). Among participants with MCI, 54% were diagnosed with dementia within 11 y from the acquisition of MRIs analyzed in this study. We generated 3D-CNN saliency maps of each participant's brain to determine how the 3D-CNN weighs each MRI voxel (Fig. 1 *B* and *D*). Saliency maps can help to identify brain locations whose MRI features are weighted more heavily during age estimation (Fig. 2 and *SI Appendix*, Figs. S1–S12). Using this strategy, we mapped CI-related aging patterns and studied their variation across sexes, brain regions, and subjects, as well as their association with neurocognitive outcome (Fig. 1*E*).

Our results in CN participants (Fig. 2 *A* and *B*) reveal typical neuroanatomic patterns of aging, including ventricular

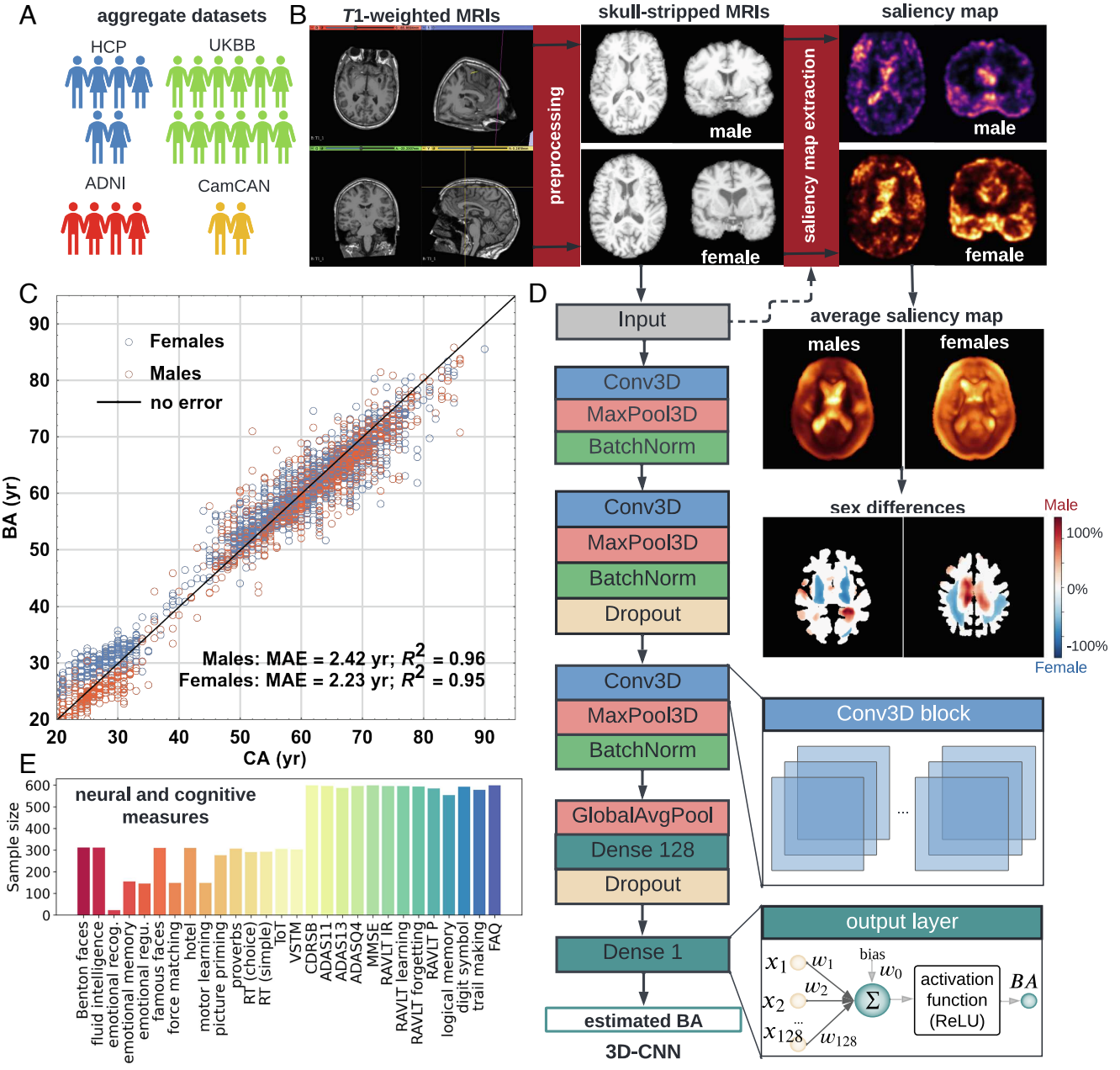

Fig. 1. Overview of BA estimation by an interpretable 3D-CNN. (*A*) Proportions of participants in the aggregate dataset (ADNI, UKBB, CamCAN, and HCP), where each human symbol represents ~300 participants. (*B*) T_1-weighted MRIs were skull-stripped and 3D saliency probability maps were generated from 3D-CNN output for each subject. (*C*) Prior to BA estimation using the 3D-CNN, participants were split by sex and assigned randomly into training and test sets. MAE was used to evaluate 3D-CNN performance from BA estimation results for test sets. The test set's CA histogram is displayed in an inset. (*D*) The 3D-CNN's input consists of T_1-weighted MRIs, and its output are BA estimates. Saliency maps are extracted from 3D-CNN output after training. A dropout rate of 0.3 is used in all dropout layers, and a ReLU activation function is used in all convolutional and dense layers. x_i is the feature map for input i and w_i is its weight. (*E*) Sample sizes for participants with neurocognitive measures.

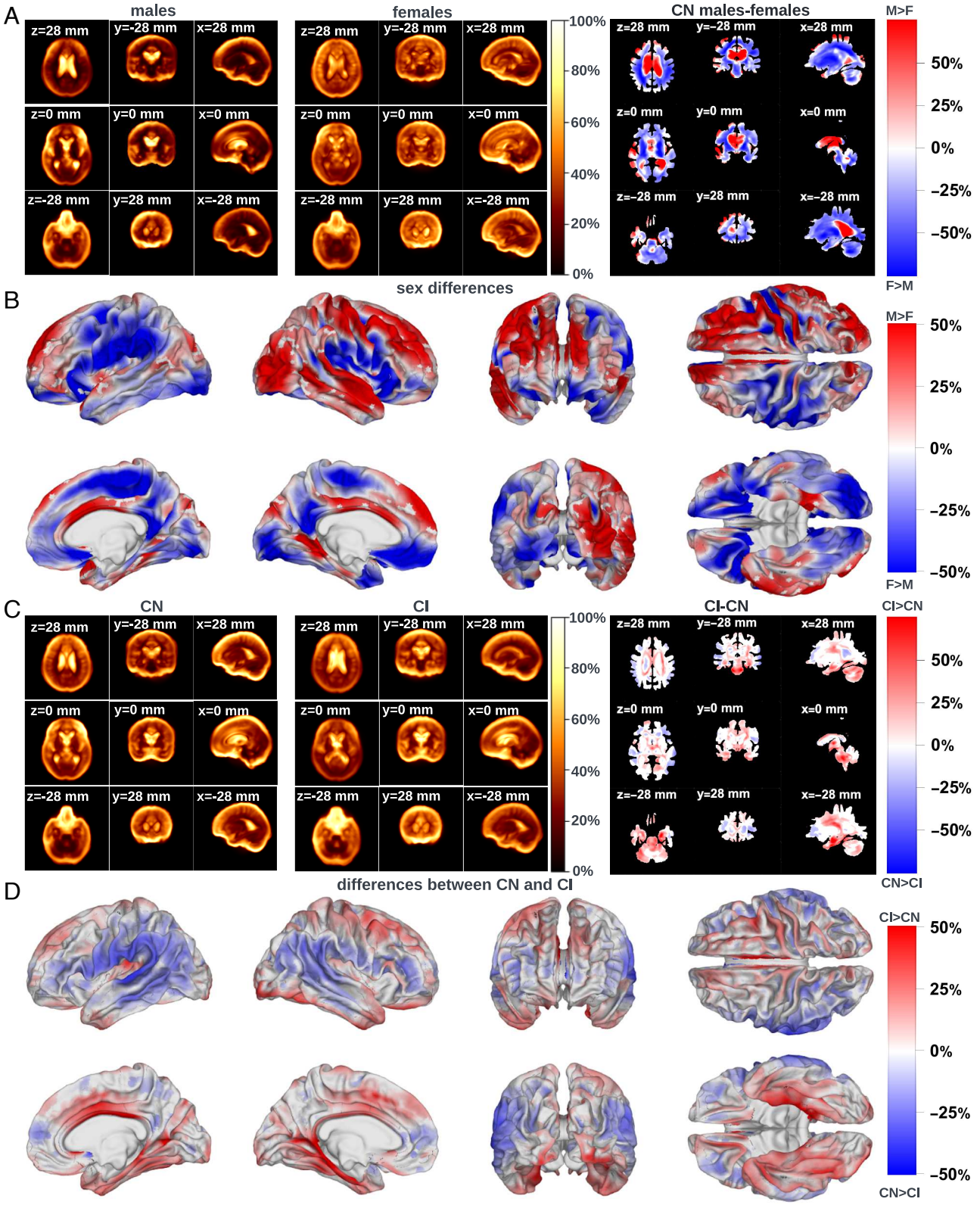

Fig. 2. Comparison of brain saliency maps across sexes and diagnoses. (*A*) Sex-specific mean saliency maps (P_M, P_F) and the sex dimorphism map $\Delta P = (P_M - P_F)/[(P_F + P_M)/2]$ of CN participants. In all cases, canonical cortical views (sagittal, axial, and coronal) are displayed in radiological convention. Higher saliencies (brighter regions) indicate neuroanatomic locations whose voxels contribute more to BA estimation. Regions drawn in red have higher saliencies in males ($P_M > P_F$); the reverse ($P_F > P_M$) is true for the regions drawn in blue. (*B*) Canonical views of the sex dimorphism map ΔP for CN participants. Sex-specific deviations of ΔP from its mean across sexes are expressed as percentages of the mean. Red indicates that $\Delta P_M > \Delta P_F$, i.e., males have higher saliency; blue indicates the reverse ($\Delta P_F > \Delta P_M$), i.e., females have higher saliency. (*C*) Like (*A*), for the comparison between CN participants and participants with CI, where $\Delta P = (P_{CI} - P_{CN})/P_{CN}$; red indicates $P_{CI} > P_{CN}$, blue indicates $P_{CN} > P_{CI}$. (*D*) Like (*B*), for the saliency difference ΔP between CN and CI participants. Images are displayed in radiological orientation convention (the right hand side of the reader is the left hand side of the participant and vice versa).

enlargement, atrophy of frontal, temporal, and hippocampal cortices, and cortical thinning. Cortical features are weighted differently across sexes (Fig. 2 *A* and *B*), which suggests that males' BA estimation is particularly reliant upon Sylvian fissure widening, ventricular enlargement, and cingulate cortex atrophy. Males' BA estimation is also weighted more heavily by features of the lateral temporal lobe and dorsolateral frontal lobe in the right hemisphere, a notable lateralization effect. By contrast, females' saliencies are higher in posterior and medial occipital regions (except the left calcarine sulcus), in the inferior and medial aspects of the parietal lobes, in the supramarginal gyrus and adjacent parietal structures, in the callosal sulcus, in the pars triangularis of the right inferior frontal gyrus, and in posterior insular regions (Fig. 2 *A* and *B*). In females, on average, white matter is weighted more heavily than gray matter when estimating BA.

Fig. 2 *C* and *D* compares subject-wise average saliency maps according to the cognitive status (CN vs. CI). This comparison reveals brain features upon which the 3D-CNN relies more when estimating age according to cognitive status. For this reason, such features may reflect how CI modifies regional brain aging. Many structures salient in CN aging are in the cortical gray matter and include the dorsolateral aspect of the right frontal lobe, the lateral aspect of the right temporal lobe, the posterolateral aspect of the right occipital lobe, as well as pericallosal regions in both hemispheres (Fig. 2 *C* and *D*). Cerebral white matter is more salient in aging with CI than in CN aging (Fig. 2 *C*), as is the brainstem, medial aspects of the temporal lobes (including parahippocampal and fusiform gyri), and the caudal portions of the anterior cingulate gyri (Fig. 2*D*). Appreciable lateralization of saliencies is noted when comparing CN participants to participants with CI, and the lateralization pattern is similar to that obviated by the sex comparison (Fig. 2*B*). Involved are lateral temporal areas, the angular and supramarginal gyri, middle cingulate cortex, parahippocampal areas, and both medial and dorsolateral prefrontal cortices.

Associations with Neurocognitive Endophenotypes. The ability of estimated BA to capture neurocognitive endophenotypes was contrasted to that of CA. This was achieved by comparing Spearman's correlations r_S between each age (BA and CA) and every neurocognitive measure of CN aging (Fig. 3 and *SI Appendix*, Tables S1–S5). For all neurocognitive measures, significant r_S values reflect typical aging effects on neurocognitive function (worse performance is correlated with older age). As expected, among CN participants, BA and CA reflect cognition to similar extents. For example, among CamCAN CN participants (Fig. 3*A* and *SI Appendix*, Table S1), older BA and CA are correlated with worse performance on word finding (picture priming), motor learning (force matching), motor response time [choice and simple response time (RT) tasks], face recognition (Benton's unfamiliar face recognition, famous faces test), Cattell's fluid intelligence, emotional memory, and visual short-term memory (VSTM) measures. In the ADNI CN cohort (Fig. 3*B* and *SI Appendix*, Table S2), no neurocognitive measure examined is significantly more correlated with BA than with CA.

Across participants with CI, BA is significantly more correlated than CA with neurocognitive measures. In participants with MCI (Fig. 3*C* and *SI Appendix*, Table S3), older BA (but not CA) is significantly correlated with worse scores on all measures of neurocognitive function examined, except 1) delayed verbal recall and learning on the Rey auditory verbal learning test (RAVLT), 2) delayed word recall measured by the AD assessment scale question 4 (ADAS Q4), and 3) logical memory. For the clinical dementia rating sum of boxes (CDR-SB) and the functional abilities questionnaire (FAQ), the difference in correlations between BA and CA is significant and BA outperforms CA in its ability to reflect neurocognitive function. In participants with AD, no significant correlations exist between BA and any neurocognitive measure apart from FAQ scores (*SI Appendix*, Table S4). Nevertheless, older CA is correlated with poorer delayed verbal memory (RAVLT forgetting). By contrast, among all participants with any type of CI (whether MCI or AD), BA (but not CA) is significantly correlated with all measures except delayed verbal recall (RAVLT, ADAS Q4) and logical memory (Fig. 3*D* and *SI Appendix*, Table S5). The difference in correlations between BA and CA is significant for the CDR-SB, mini-mental state exam (MMSE), RAVLT immediate recall (IR), and FAQ. When separating participants with CI by apolipoprotein E4 (APOE4) status, BA is not more correlated with any neurocognitive measure in carriers compared to noncarriers. The omnibus effect of a logistic regression accounting for all interactions between AG, CA, and sex is significant ($\chi^2_{343} = 29.500$, $P < 0.001$). AGs are significantly and positively associated with MCI participants' probability of conversion to AD ($\beta = 1.417$, $t_{343} = 2.240$, $P = 0.025$). The only significant interaction is between AG and sex ($\beta = -1.121$, $t_{343} = -2.129$, $P = 0.033$), i.e., MCI females with more negative AGs and MCI males with more positive AGs are significantly more likely to convert to AD. When including all interactions, the omnibus effect of the regression that predicts time to conversion is significant if AG is included as the predictor ($R^2 = 0.065$, $F_{8,181} = 2.880$, $P = 0.007$) but not if AG is excluded ($R^2 = 0.012$, $F_{4,185} = 1.790$, $P = 0.151$).

3D-CNN Benchmarking and Evaluation. We compare our 3D-CNN to an award-winning (19) state-of-the-art model, the simple fully convolutional network (SFCN) of Gong et al. (20, 21), by replicating its training, validation, and benchmarking. The SFCN was pretrained on 5,698 UK Biobank (UKBB) subjects, whereas our 3D-CNN was trained on 4,681 participants (2,513 females; age range: 22 to 95 y) aggregated across the UKBB, Human Connectome Project-Aging (HCP-A), Human Connectome Project-Young Adult (HCP-YA), and ADNI. In the testing set, our model's MAE between BA and CA is 2.41 y for males and 2.23 for females (*SI Appendix*, Fig. S13 *A* and *B*). The coefficient of determination R^2 is 0.96; the correlation coefficient r is 0.98. Across all external testing sets (UKBB, CamCAN, AD, and MCI), our model has a higher R^2 than the SFCN (*SI Appendix*, Table S6).

On identical UKBB data ($N = 518$), the 3D-CNN achieves MAEs of 2.27 y (males) and 2.31 y (females) (*SI Appendix*, Fig. S14 *A* and *E*), while the SFCN achieves an MAE of 2.14 y across both sexes (*SI Appendix*, Fig. S14 *Q* and *R*). In the independent CamCAN CN cohort, the SFCN's MAEs are 9.90 y (males) and 9.17 y (females) (*SI Appendix*, Fig. S14 *Q* and *R*). By contrast, our 3D-CNN achieves MAEs of 4.71 y (males) and 3.01 y (females) (*SI Appendix*, Fig. S14 *B* and *F*). During pretraining, the SFCN yields an MAE within 2% of the 3D-CNN's. However, in the independent test cohort of CN participants, our MAEs are 42% lower than the SFCN's in the same cohort. The SFCN yields MAEs of 7.72 y (males) and 7.50 y (females) for participants with MCI and 8.24 y (males) and 8.65 y (females) for participants with AD (*SI Appendix*, Fig. S14 *Q* and *R*). By contrast, our 3D-CNN model achieves an MAE of 5.26 y (males) and 4.33 y (females) for participants with MCI (*SI Appendix*, Fig. S14 *C* and *G*) and 6.48 y (males) and 5.98 y (females) for participants

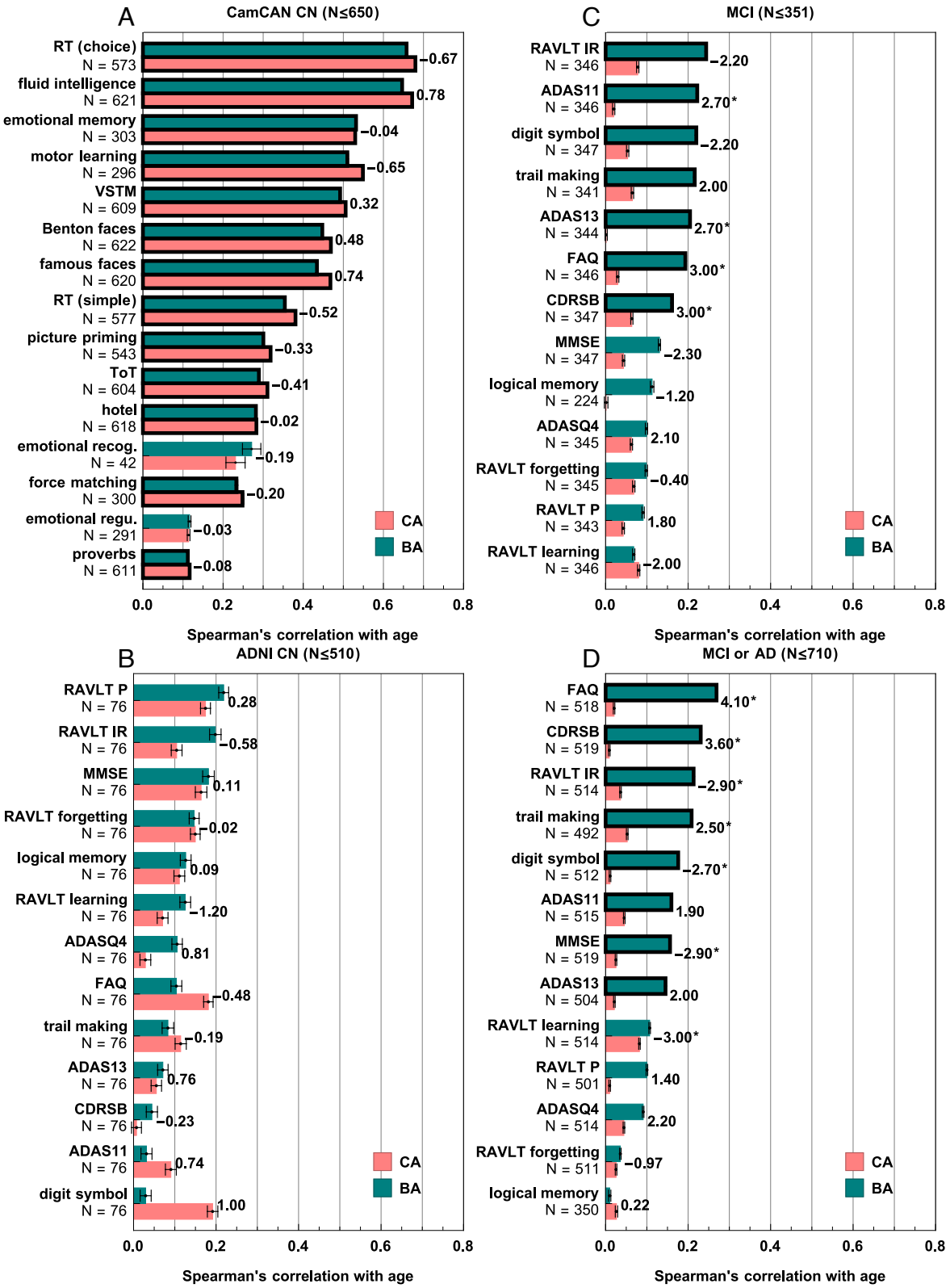

Fig. 3. Correlations between neurocognitive measures and both estimated BA and CA. Results are depicted for two independent test sets: CamCAN and ADNI. (*A*) displays CN participants from CamCAN, (*B*) displays CN participants from ADNI, (*C*) displays results only for participants with MCI, and (*D*) displays results for participants with either MCI or AD. For each independent test set, the sample size for each neurocognitive measure is listed below the measure name. Bar charts depict Spearman's correlations r_S (along x) between BA (green) or CA (red) and each neurocognitive measure (along y). Bars are contoured in black if r_S is significant. Error bar widths equate to one SE of the mean. For each neurocognitive measure, the corresponding bar pair is annotated with Fisher's z statistic. Asterisks indicate neurocognitive measures for which the difference in Spearman's correlations $r_S(BA) - r_S(CA)$ is significant.

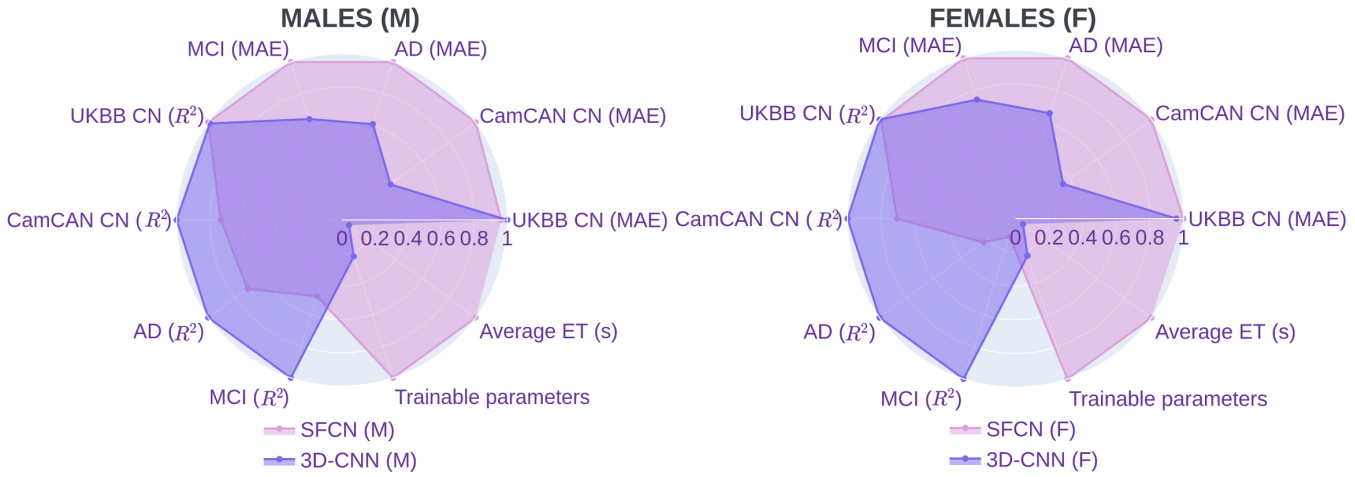

Fig. 4. Radar plots of sex-specific MAEs and performance parameters. Radar plots of MAE, R^2, and performance parameters (average ET and the number of trainable parameters) according to sex and diagnostic status (CN: UKBB, CamCAN; MCI or AD: ADNI). The SFCN of Gong et al. (20, 21) (purple) is compared to our 3D-CNN (blue). To facilitate simultaneous comparison, all values are normalized to range from 0 to 1, where the maximum value in each measurement was rescaled as 1 and 0 remained as 0.

with AD (*SI Appendix*, Fig. S14 *D* and *H*). Compared to the SFCN (*SI Appendix*, Fig. S14 *Q* and *R*), the 3D-CNN yields significantly larger mean AGs for A) females with MCI (t_{144} = 6.595, $P < 0.001$), B) males with AD (t_{195} = 4.710, $P < 0.001$) and C) females with AD (t_{162} = 6.200, $P < 0.001$). The 3D-CNN also yields significantly larger AG variances in participants with AD (males: $F_{197,197} = 1.857$, Pitman's $t_{196} = 4.440$, $P < 0.001$; females: $F_{162,162} = 2.493$, Pitman's $t_{161} = 6.006$, $P < 0.001$). Compared to the 3D-CNN, the SFCN yields significantly larger AG variances for the following CN groups: A) UKBB males ($F_{796,796} = 1.137$, Pitman's $t_{795} = 12.967$, $P < 0.001$); B) UKBB females ($F_{796,796} = 1.097$, Pitman's $t_{795} = 9.034$, $P < 0.001$); C) CamCAN females ($F_{309,309} = 7.576$, Pitman's $t_{308} = 21.082$, $P < 0.001$). As expected, the 3D-CNN's mean AG is ~75% larger in participants with CI than in CN participants (*SI Appendix*, Fig. S14 *I–P*), possibly reflecting faster brain aging in the former. The BA estimation parameters of the 3D-CNN and SFCN, evaluated without fine-tuning, are compared in Fig. 4 and *SI Appendix*, Table S6. The 3D-CNN has shorter execution times (ETs) and fewer trainable parameters, reflecting lower complexity (*SI Appendix*, Table S6 and Fig. 4). As *SI Appendix*, Table S6 and Fig. S14 suggest for participants with CI, our CNN yields higher R^2 and lower MAEs than the SFCN.

Discussion

Significance. While biological age can be computed for many phenotypic traits, BA summarizes MRI-derived neuroanatomic profiles using one number. This highlights both the appeal and caveats of this measure. Although straightforward to grasp, BA (as defined here) does not capture the nuances and complexity of brain aging. Nevertheless, with cautious interpretation, BA could assist the diagnosis and prognosis despite its limitations. Early screening for CI can help to monitor and improve the welfare of aging adults (22). Although positron emission tomography (PET) can aid the diagnosis of AD at the preclinical and prodromal stages (23), this technique is expensive, involves specialized tracers, and exposes participants to radiation (24). By contrast, MRI is noninvasive, more affordable, and safer. Thus, MRI-derived BAs that capture neurocognitive decline (25) could become affordable and noninvasive preclinical measures of CI risk (26).

The correlations of neurocognitive measures with our estimated BAs are, in many cases, significantly stronger than with CA, suggesting that our BAs better reflect neurocognitive functioning. These correlations are critical because one potential utility of BA estimation is to facilitate the early identification of persons at high risk of MCI and AD. AGs are predictive of AD conversion risk, as others reported (27–29). BA is not correlated with neurocognitive function in participants with AD, with the exception of informant-rated functional ability. One possible reason is that the 3D-CNN was trained on CN adults rather than participants with CI. Another reason could be that correlations are more difficult to detect due to lower statistical power (smaller sample size) in participants with AD compared to CI participants. Across persons with CI of any severity, BA (but not CA) is significantly correlated with measures used routinely (30) to screen for (or to diagnose) CI, including MCI (Fig. 3 and *SI Appendix*, Tables S4 and S5). Thus, our contributions can help to understand how CI-related neurocognitive changes within specific functional domains reflect neuroanatomic features that modify regional BAs.

Sex Differences in Anatomic Brain Aging. Of note, for patient-tailored profiling, our approach can generate subject-specific brain saliency maps reflecting individual neuroanatomic patterns of brain aging. Anatomic interpretability of BA is important because 1) brain regions age differently, 2) neuroanatomic alterations with age may reflect distinct disease processes paralleled by BA (31), and 3) individual neuroanatomic deviations may parallel neurocognitive endophenotypes. Sex differences in saliency confirm findings on the contributions of age to sex dimorphism in the pre- and post-central gyri (32, 33) and the pars triangularis of the left inferior frontal gyrus (34). Males, who are at higher risk of motor impairment due to Parkinson's disease (35), exhibit greater saliency in the primary motor cortex. Males' BA estimation relies more on the crowns of gyri on the lateral aspects of the frontal lobes, whereas females' BA estimation relies more on the troughs of sulci. These findings confirm prior reports on sex differences in older adults' cortical gyrification (36). Males' saliencies are higher along ventricular boundaries, indicating that BAs are disproportionately predicated upon ventricular enlargement in men, as reported elsewhere (37).

The right hemisphere's higher saliency in males is consistent with their lateralization of language function (38) and with lateralization trends in old age (39). Thus, in females, typical cortical aging may be relatively slower in the right hemisphere. By contrast, on average, most occipital and medial parietal areas exhibit age-related neuroanatomic patterns that are more salient in males. Males also have higher saliency in superior parietal and frontal regions, reflecting smaller gray matter volumes (40). By contrast, females have higher saliency at the occipital poles and in occipitoparietal regions, reflecting smaller gray matter volumes in these regions (40). Females' saliencies are higher across inferior parietal regions, where the cortex is thicker than in males (41). Thus, our approach to neuroanatomic saliency mapping can identify sex differences in the neuroanatomy of cortical aging.

Anatomy Changes According to Neurocognitive Status. Our interpretable 3D-CNN framework captures neuroanatomy changes related to both CN aging and aging with CI. In the case of CN aging, the estimated BAs of CN participants in our two independent samples (CamCAN and ADNI) are correlated with neurocognitive measures reflecting typical aging (e.g., motor learning, multitasking, and word finding). In ADNI CN participants, no significant associations were found between neurocognitive measures and either CA or BA. This was expected, as ADNI cognitive measures are sensitive to CI rather than to CN aging (42). In the case of CI, *SI Appendix*, Fig. S14 *I–P* confirms that participants with either MCI or AD have AGs considerably larger than those of sex- and age-matched CN adults (43), mostly due to older-than-expected brains (BA > CA). Atrophy of the parahippocampal gyrus is a strong structural correlate of MCI and AD (44); our 3D-CNN's greater reliance on this structure during BA estimation reflects this (Fig. 2*C*). Similarly, saliency differences between CN and CI participants are greater in parietal, occipital, and temporal cortices (Fig. 2*D*), whose atrophy is greater in participants with CI (45) and whose burdens of amyloid β plaques and τ neurofibrillary tangles are typically higher in AD (46). The brainstem, which is affected by amyloid deposition early during AD, is more salient in participants with CI than in CN adults (47). Comparison of the cortical patterns in Fig. 2 *B* and *D* indicates that saliency differences between sexes are largely paralleled by saliency differences across cognitive statuses (CN vs. CI). This may reflect females' higher risk for AD and supports the hypothesis according to which their higher risk is paralleled by faster cortical aging. Comparison of CN and CI cohorts suggests that the SFCN underestimates mean AG in the latter group and that the expected accuracy of BA estimation is lower for participants with CI. These findings highlight the importance of an accurate BA estimator when studying diseased populations. Some cortical structures that atrophy far more in CI than in CN aging are more salient in the latter (blue regions in Fig. 2*D*). This may reflect the fact that the 3D-CNN was trained on a CN adult cohort. During training on this cohort, our 3D-CNN likely relies on features whose variance is moderate in CN aging. When estimating the BAs of participants with CI, however, these features exhibit far greater variability. This may cause their relative saliency to decrease, such that the saliency difference ΔP between CN and CI aging is negative in such regions. Thus, although features with negative ΔP values can be useful for understanding how BA estimation relies on CI-related neuroanatomy features, the negative sign of ΔP must be interpreted cautiously.

Comparison to Other Methods. Our 3D-CNN alleviates major limitations of other approaches. The quantitative comparison below focuses on the SFCN because this open-source approach performed best in a competition (19) for which both training and testing data are available.

Accuracy. Our 3D-CNN estimates BA more accurately than the state of the art regardless of whether accuracy is quantified using MAE or R^2. In the test set, our model yields an MAE of ∼2.3 y; this is ∼1 y less than the SFCN, which is the second best. Other (published) BA estimators have MAEs that are even higher than that of the SFCN on their testing data. Presumably, since our MAE is ∼2.3 y, these estimators also perform more poorly than ours. However, we could not ascertain this because we did not have access to the testing sets on which others estimators were benchmarked. These estimators include a best linear unbiased predictor (MAE $\simeq$ 3.3 y) (48), a 3D residual neural network (3D-RNN, MAE $\simeq$ 3.3 y) (49), a graph CNN (MAE $\simeq$ 4.6 y) (50), Gaussian process regression (MAE $\simeq$ 4.1 y) (16), support vector regression synergized with a random forest classifier (MAE $\simeq$ 3.5 y) (51), and a 3D-DenseNet (MAE $\simeq$ 3.3 y) (52).

On the testing set, our model yields $R^2 \simeq 0.96$ and $r \simeq 0.98$. By contrast, the SFCN model yields $R^2 \simeq 0.92$ and $r \simeq 0.96$ (16, 20, 21). During testing, other (published) BA estimators achieve even lower R^2 than the SFCN. These include Gaussian process regression ($R^2 \simeq 0.91$) (16), a 3D-RNN ($R^2 \simeq 0.90$) (49), a graph CNN ($R^2 \simeq 0.87$) (50), and a 3D-DenseNet ($R^2 \simeq 0.85$) (52). Our R^2 is also higher than that of a BA estimator that used an optimized SFCN (53) with $R^2 = 0.94$. These comparisons suggest that even high R^2 can involve undesirably large MAE, such that it can be useful to consider both measures when evaluating accuracy.

In all females with CI and in AD males, our AGs are significantly larger than those estimated by the SFCN. As expected, the 3D-CNN's estimates of these subjects' CAs are consistently larger than their true CAs. Because CI involves more brain aging, this suggests that the 3D-CNN captures CI better than the SFCN. Females are at higher risk for AD and exhibit faster decline than males (54). Females already have a larger mean AG in the MCI stage, whereas this is not the case for males until the AD stage. Thus, our model captures known sex differences in AD risk.

Variances in AG between the 3D-CNN and SFCN are significantly different for the UKBB cohort even though $F = \sigma^2_{SFCN}/\sigma^2_{CNN} \simeq 1$, which usually implies lack of significant differences in variance. This finding can be explained by our use of Pitman's variance ratio test, which is justified here because the variances being compared pertain to correlated samples (CNN- and SFCN-computed AGs measured for the same cohort). Because the 3D-CNN and SFCN were both trained on UKBB CN participants, the abilities of these methods to estimate BA for new UKBB participants are likely better (and therefore more similar) than their ability to estimate BA for participants from altogether new cohorts. This similarity may explain the strong correlation r of UKBB BAs across the two methods (females: $r = 0.989$; males: $r = 0.990$). The dependence of Pitman's t on r (see *Methods*) satisfies $t \sim (1 - r^2)^{-1/2}$. A Maclaurin series expansion indicates that $t \to \infty$ as $r \to 1$. Thus, Pitman's t is large when $r \simeq 1$ even when $\sigma^2_{SFCN}/\sigma^2_{CNN} \simeq 1$. This explains our power to detect even moderate differences between σ^2_{SFCN} and σ^2_{CNN} in UKBB CN participants.

Complexity. Model complexity was quantified using mean ET and the number of trainable parameters in the model. By both measures, our 3D-CNN's execution complexity is lower than that of previous approaches. For example, the 3D-CNN features a ∼10 times shorter ET compared to the SFCN (20, 21) and ∼4 times fewer trainable parameters. The model of Leonardsen et al. (53), which is based on the SFCN (20, 21), has more trainable

parameters and is more challenging to fine-tune. The 3D-DenseNet (53) has ∼7 million trainable parameters (compared to our 682,881) and requires extensive fine-tuning on new validation datasets via grid searches for optimal hyperparameters.

Interpretability. Lee et al. (52) use a 3D-DenseNet to compute saliency by covering the brain with occlusion masks of size 11^3 $mm^3 = 1,331$ mm^3. According to these authors, their saliencies correlate with PET-mapped amyloid β and τ burdens. However, for participants with CI, the anatomic patterns of brain aging mapped by Lee et al. are broadly similar to ours (Fig. 2 *C* and *D*) across similar age ranges. This suggests the hypothesis that BA saliencies like ours can reflect dAD-related clinical PET findings. He et al. (55) used two-dimensional (2D) occlusions (box size: 32^2 $mm^2 = 1,024$ mm^2) to map saliency, whereas Wood et al. (56) monitored performance and saliency by occluding 3D masks (size: 5^3 $mm^3 = 125$ mm^3). Our study advances the state of the art by 1) providing voxelwise saliency maps to reveal detailed spatial variability at native MRI resolution (1 mm^3), 2) reporting comparisons by sex and cognitive status, and 3) conveying how cognitive status relates to neurocognitive function.

Generalizability. Most BA estimators are not typically tested across domain-specific neurocognitive measures, whereas our 3D-CNN features unique generalizability to independent cohorts in its ability to capture neurocognitive endophenotypes. Since the R^2 values achieved on independent and test data are similar, we surmise that overfitting was largely avoided. Compared to CA, the BA of participants with any type of CI is significantly more correlated with measures of neurocognitive function routinely used as clinical indicators of CI. Other published approaches have rarely been evaluated according to this critical performance benchmark. Because the 3D-CNN was trained on subjects aged 22 to 95 y, its utility extends across the age range of adulthood.

Limitations. Although we validated the 3D-CNN in cohorts independent from those used for its training, differences in acquisition sequences and scanners across MRIs can affect results (57). Like other dementia diagnosis criteria, ADNI's have limitations [e.g., a risk of false positive diagnoses (42)] that may affect the findings of studies like ours. Additionally, floor effects may affect cognitive measures in participants with AD by attenuating their correlation. Conceivably, our failure to find significant correlations between BA and neurocognitive measures in participants with AD could be due to our lower power to detect small effects in the AD sample, which is smaller ($N \leq 172$) compared to the MCI ($N \leq 347$) and combined CI (i.e., MCI or AD, $N \leq 519$) samples. These nonsignificant correlations, however, are not typically relevant for early CI screening because most participants with severe CI have been already diagnosed by the time brain MRIs are typically acquired. The nonuniform distribution of CAs in our aggregate sample translates into potential training data imbalance and inaccuracy in BA estimates. Nevertheless, our approach is more accurate than others currently available, as reflected by our test set's MAE and R^2, which are the lowest reported to date. Due to the lack of ground truth, there is no consensus on how the interpretability of approaches like ours ought to be evaluated (52, 55, 56, 58).

Methods

Participants and Neuroimaging. This study was undertaken in adherence with the US Code of Federal Regulations (45 C.F.R. 46) and the Declaration of Helsinki. MRIs analyzed in this study were acquired as part of other studies, with approval from the institutional review boards or similar ethical monitoring bodies at the respective institutions where data had been acquired for ADNI (30) and HCP (59). UKBB efforts were undertaken with ethical approval from the North West Multi-Centre Research Ethics Committee of the United Kingdom. Ethical approval for CamCAN was obtained by the Cambridgeshire 2 (now East of England–Cambridge Central) Research Ethics Committee. Informed written consent was obtained from all participants.

The aggregate dataset consists of 5,851 CN individuals (3,142 females) aged 22 to 95 yr sampled from ADNI ($N = 510$), HCP-A ($N = 508$), HCP-YA ($N = 1,112$), and UKBB ($N = 3,721$; Table 1). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. MRI acquisition protocols for HCP-A and HCP-YA are described elsewhere (60). For UKBB data, we used preprocessed images generated by a UKBB pipeline whose output included FreeSurfer reconstructions (61).

An independent test set of 650 CN participants aged 18 to 88 yr was obtained from CamCAN. Additionally, 408 participants with MCI and 359 participants with AD were obtained from ADNI (Table 1). CamCAN inclusion/exclusion criteria (62) and ADNI eligibility/diagnosis criteria are described elsewhere (30). $N = 75$ participants with MCI were excluded due to MCI diagnosis being unrelated to AD, leaving 351 participants with MCI (190 converted to AD, 161 did not). Of 524 CI (MCI or AD) participants whose correlations between BA and neurocognitive scores were computed, 307 participants were APOE4 carriers.

Neurocognitive Measures. We used neurocognitive measures available in CamCAN and ADNI to evaluate the utility of our estimated BAs to capture neurocognitive phenotypes (see *SI Appendix, Methods* for detailed task descriptions). Thirteen cognitive measures that assess emotional processing, executive function, memory, and motor function were obtained from the CamCAN repository (62). Emotional processing was measured via 1) Ekman's emotion expression recognition test, 2) the emotional memory test, and 3) the

Table 1. Participant demographics

			CA (y)				Ratio	FreeSurfer version			
Repository	Status	*N*	Min	Max	μ	σ	M:F	4.3.0	5.3.0	6.0.0	7.1.1
ADNI	CN	510	56	95	75.1	7.2	1:1.17	0	0	260	250
HCP-A	CN	508	36	80	55.8	12.0	1:1.38	0	0	309	199
HCP-YA	CN	1,112	22	37	28.8	3.7	1:1.17	0	1,112	0	0
UKBB	CN	3,721	45	83	62.7	10.1	1:1.13	0	0	3,721	0
CamCAN	CN	650	23	88	54.2	18.6	1:1.00	0	0	0	650
ADNI	AD	359	56	95	75.9	8.0	1:0.83	359	0	0	0
ADNI	MCI	351	55	89	75.2	7.3	1:0.53	351	0	0	0
All	all	7,211	23	95	58.4	10.3	1:1.09	710	1,112	4,290	1,099

Sample size, descriptive statistics (minimum, maximum, mean μ, and SD σ), the male-to-female (M:F) sex ratio, and breakdown by FreeSurfer version used for preprocessing. Demographics are listed for each repository and neurological/cognitive status.

emotional regulation test. Executive function was measured using 1) Cattell's fluid intelligence test, 2) the hotel test, and 3) a proverb comprehension task. Memory was measured using 1) Benton's face recognition test, 2) the famous faces test, 3) a picture priming task, 4) the tip of the tongue (ToT) test, and 5) a VSTM task. Motor function was assessed via 1) a force matching task, 2) a motor learning task, 3) a reaction time (RT) "choice" task, and 4) a RT "simple" task.

Nine cognitive measures that assess neural function, cognitive performance, and functional impairment were obtained from the ADNI repository (63). To eliminate systematic variability in FreeSurfer software versions, we limited correlation analysis for the CI cohort to subjects from $ADNI_1$ only (FreeSurfer v4.3). For neural function, four established dementia rating scales were obtained, including 1) the clinical dementia rating scale–sum of boxes (CDR-SB), 2) the diagnostic ADAS versions 11 and 13, and 3) the MMSE. Cognitive performance was measured via four neuropsychological measures: 1) the RAVLT, 2) delayed recall on the logical memory test, 3) the digit symbol substitution test, and 4) the trail-making test. Functional impairment was measured by the FAQ.

MRI Preprocessing. Freesurfer's recon-all function was used to reconstruct and segment T_1-weighted MRIs. This process includes skull-stripping, motion correction, normalization of nonuniform signal intensities, Talairach space transformation, removal of nonbrain tissues, and registration of all subjects' brains into a common coordinate space (64). FreeSurfer (FS) was used for three reasons: 1) UKBB makes FS reconstructions available; 2) the FS workflow is fully automated and thus convenient; 3) our study involves surface analyses and registrations across native and atlas spaces, which FS facilitates. During FS preprocessing using recon-all, all MRIs were affinely registered to the MNI_{305} atlas. Due to sourcing from several MRI repositories, enhancement of segmentation accuracy differed slightly between cohorts (Table 1). UKBB and HCP-YA reconstructions were enhanced using T_2-weighted MRIs, while ADNI, HCP-A, and CamCAN were enhanced using fluid-attenuated inversion recovery MRIs.

3D-CNN Architecture. We constructed a DL regression model using a 3D-CNN whose inputs are FS brain.mgz output files and whose outputs are estimated BAs. The DL architecture was implemented in Python 3.6 using TensorFlow 2.7.0 and executed on a computer with an Intel Core i7 processor (2.2 GHz clock speed) with 16 GB of RAM and a 12 GB NVIDIA Tesla K80 graphical processing unit. The 3D-CNN consists of three convolutional blocks followed by two dense layers. The input matrix size is $82 \times 86 \times 100$. Each convolutional (conv) block has a 3D conv layer, a batch normalization layer, a max-pooling layer, and an optional dropout layer. The filter sizes of the first three (conv) blocks are 64, 128, and 128, respectively. Conv block filter size determines the dimensionality of the output space. The rectified linear unit (ReLU) activation function is applied to all conv and dense layers. The ReLU activation function is defined as $g(x) = max(0, x)$ for input x. $g(x)$ can efficiently reduce the likelihood of a vanishing gradient and makes the output more sparse. After the conv blocks, the fourth block consists of one global average pooling layer (used for global average pooling of 3D data), one dense layer, and one dropout layer (dropout rate = 0.3). The resulting feature map, of size $18 \times 18 \times 18 \times 128$, is pooled to 128×1 and then projected onto the output dense layer, which has one output neuron to estimate BA using regression.

We choose MSE as loss function and use an Adam optimizer (learning rate = 0.001) (65). The advantage of outputting BAs as real numbers rather than assigning them to discrete age bins (20, 21) is that, in the former case, BA outputs are assigned within a continuous domain and range. Due to regression to the mean (66), estimated BAs exhibit a previously documented CA-dependent bias. To alleviate this effect, we use the zero correlation constraint method of Treder et al. (67) to regress out the bias from the BAs of testing set participants. This is done separately for each cohort. Bias-corrected BAs are used for all analyses. CN participants are aggregated from UKBB, HCP-A, HCP-YA, and ADNI. Participants were randomly assigned into training and test sets of sizes equal to 20% of the total sample size ($N = 5{,}851$).

3D-CNN Training and Testing. We optimized the CNN architecture and fine-tuned hyperparameters. 2D-CNNs use 2D kernels to estimate sliding windows across single slices, such that leveraging information from adjacent slices is challenging (68). We therefore chose a 3D-CNN that overcomes this deficit by using 3D kernels to estimate sliding windows for volumetric patches. The latter captures interslice image context and improves the model performance (68). We also included dropout and batch normalization layers because these help to alleviate overfitting (69). Grid and random searches determined suitable hyperparameter values (e.g., batch size, kernel size, weight decay). An n-dimensional grid was defined to map the n hyperparameters and to identify their ranges. We examined all possible 3D-CNN configurations to identify optimal values for each hyperparameter. Since we used MSE as a loss function, we selected a configuration with the lowest loss value (error).

We tested the 3D-CNN on independent cohorts to refine the 3D-CNN architecture, illustrate model generalizability, alleviate data overfitting, and compare the 3D-CNN to other approaches. The testing set was designed to include a random selection of participants from the same cohorts as the training set. To avoid overfitting the 3D-CNN to the training set, we monitored its performance on the testing set. To avoid overfitting on both training and testing sets, we tested our model on two independent cohorts (CamCAN and ADNI) that had not been used for the 3D-CNN design. The latter of these cohorts includes participants with a range of cognitive statuses (CN, MCI, or AD).

After computing AGs for identical samples using both our 3D-CNN and the SFCN, we performed Welch's t-tests for paired samples with unequal variances to compare the mean AGs obtained using the two methods. AG variances were compared using Pitman's variance ratio test for correlated samples, whereby $F = \sigma_1^2/\sigma_2^2$, Pitman's $t_{N-2} = [(F-1)\sqrt{N-2}]/[2\sqrt{F(1-r^2)}]$, and r is the correlation of AG_{SFCN} with AG_{CNN}. The AG variances are σ_1 and σ_2, whose subscripts {1, 2} denote the SFCN or CNN, as needed, to satisfy the inequality $\sigma_1 > \sigma_2$.

BA Associations with Sex and Neurocognition. Each neurocognitive measure m was obtained from CN participants (ADNI and CamCAN) and from participants with MCI or AD (ADNI). These measures were not normally distributed (Anderson–Darling goodness-of-fit test, *SI Appendix*, Tables S1–S5), so their Spearman rank correlations $r_S(m, BA)$ and $r_S(m, CA)$ were computed. These correlations were compared using Fisher's two-sided z-test after multiple comparison correction using the Benjamini–Hochberg procedure (70) (false discovery rate = 0.05). $|r_S|$ and $|z|$ were also calculated for measures whose lower scores indicate better performance. Test statistics and their degrees of freedom, confidence intervals, and effect sizes were tabulated (*SI Appendix*, Tables S1–S5). A logistic regression examined whether AG, CA, sex, and their interactions predicted the probability of *conversion* from MCI to AD. Another linear regression (independent variables: CA, sex, and their interactions) evaluated the ability of AG to predict the *interval* between MRI acquisition and AD conversion. AG and its interactions were added to this (reduced) model to examine how AG altered the significance of the regression.

Saliency Maps. A saliency map is a topographically organized depiction of the visual saliency in an MRI volume $\mathbf{V}_0$. Here, we extend a saliency approach for 2D-CNNs (71) to the 3D case. For an MRI brain volume $\mathbf{V}_0$ and a 3D-CNN model with score function $\mathbf{S}(\mathbf{V})$, we rank voxels in $\mathbf{V}_0$ based on their importance to $\mathbf{S}(\mathbf{V})$. We consider the linear score model $\mathbf{S}(\mathbf{V}) = \mathbf{w}^T\mathbf{V} + \mathbf{b}$, where the volume $\mathbf{V}$, weight $\mathbf{w}$, and bias $\mathbf{b}$ are in one-dimensional (vectorized) forms. Since the 3D-CNN and score function are highly nonlinear functions of $\mathbf{V}$, the linear score model cannot be applied directly. We approximate $\mathbf{S}(\mathbf{V})$ at $\mathbf{V}_0$ using the first-order Taylor series $\mathbf{S}(\mathbf{V}_0) \simeq \mathbf{w}_0^T\mathbf{V}_0 + \mathbf{b}_0$, where $\mathbf{w}_0 = \partial\mathbf{S}/\partial\mathbf{V} \mid_{\mathbf{V}_0}$ is the partial derivative of $\mathbf{S}(\mathbf{V})$ at $\mathbf{V}_0$ and $\mathbf{b}_0 = \mathbf{b} \mid_{\mathbf{V}_0}$ is the bias $\mathbf{b}$ at $\mathbf{V}_0$. The spatial and temporal distributions of saliencies contain unique patterns conveying information about BA.

Saliency Associations with Sex and Neurocognition. Two distinct workflows were used for volume- and surface-level transforms, respectively, to remove the confounding effects of subject differences in brain shape and size. For *volume*-level analysis, each saliency map was nonlinearly registered to the FS fsaverage

atlas. To this end, T_1-weighted brain volumes were first registered to the atlas in MATLAB using the imregister function, which applied the transformation from native space to the atlas, as provided by FS. MATLAB's imregdemons function was used to deform nonlinearly and to map T_1-weighted scans onto the atlas. The transformations above were applied to each subject's saliency map, resulting in its registration to the atlas. For *surface*-level analysis, saliencies were projected to the native cortical surface. To achieve this, each subject's saliency was projected onto the cortical mantle as a cortical overlay using a customized algorithm for volume-to-surface mapping (72, 73). Briefly, voxels assigned to the gray matter ribbon by FS were considered. At each vertex of the native mesh for the mid-thickness surface, ribbon voxels were selected within a cylinder that lay orthogonally with respect to the local surface. The cylinder was centered on the vertex; its height and radius were equal to the local cortical thickness. The saliency of ribbon voxels within the cylinder was averaged according to a Gaussian weighted function (full width at half maximum = 4 mm, $\sigma = 5/3$ mm) to compute a mean saliency value at the surface vertex in question. After cortical surface projection, each subject's saliency overlay was registered from native space onto the atlas. Subjects' saliency probability overlays were averaged into a cortical map of mean saliency.

For both volume- and surface-level analyses, each saliency map M was operationalized into a saliency probability map P by dividing saliency at each brain location by the sum of all brain saliencies. An average saliency probability map was computed for each sex and cognitive status, yielding P_M for males, P_F for females, P_{CN} for CN adults, and P_{CI} for participants with any form of CI. Both P_{CN} and P_{CI} were computed after averaging across sex effects. Relative sex differences in P were computed as $(P_M - P_F)/[(P_F + P_M)/2]$, i.e., as sex-specific deviations from the average across sexes. The relative deviation of participants with CI from CN participants was computed as $(P_{CI} - P_{CN})/P_{CN}$. Relative saliency differences between sexes or diagnostic statuses were mapped after thresholding to include only statistically significant values. For each salience value considered, significance was evaluated using a paired-sample t-test ($\alpha = 0.05$). Results were corrected for multiple comparisons using the Benjamini–Hochberg procedure (false discovery rate = 0.05).

Visualization. For *volume*-level visualization, CN participants' mean saliency maps were plotted for each sex along the coronal (x), sagittal (y), and axial (z) planes. For each coordinate, maps were generated along planes whose equations were specified by coordinate values of −28 mm, 0 mm, and 28 mm, respectively. In CN participants and participants with CI, the procedure was repeated after averaging across sexes. For *surface*-level visualization, gray matter saliencies were mapped onto the cortex to compare different cortical locations' relative importance to the 3D-CNN when estimating BA.

Data, Materials, and Software Availability. 3D-CNN software is available from https://github.com/irimia-laboratory/USC_BA_estimator MRI data are publicly available from ADNI (https://adni.loni.usc.edu/), UKBB (https://www.ukbiobank.ac.uk/), CamCAN (https://www.cam-can.org/), and HCP (https://www.humanconnectome.org/). There are no relevant accession codes required to access these data, and the authors had no special access privileges that others would not have to the data obtained from any of these databases. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

ACKNOWLEDGMENTS. A.I. is grateful to Kenneth H. Buetow, Caleb E. Finch, Margaret Gatz, and Mara Mather for useful discussions. The authors acknowledge Alexander S. Maher for assistance with data retrieval. P.E., A.A., N.F.C., R.J.M., N.N.C., and A.I. gratefully acknowledge the support from the (NIH) under grant R01 NS 100973, the US Department of Defense (DoD) under award W81-XWH-1810413, the James J. & Sue Femino Foundation, a Hanson-Thorell Research Scholarship, the Undergraduate Research Associate Program (URAP), and the Center for Undergraduate Research in Viterbi Engineering (CURVE) at the University of Southern California. C.Y., M.C., X.C., and P.B. gratefully acknowledge the support from the NSF under CAREER award CPS/CNS-1453860 and grants CCF-1837131, MCB-1936775, CNS-1932620, and CMMI-1936624, from the Okawa Foundation, from the Defense Advanced Research Projects Agency (DARPA) under a Young Faculty Award and a Director Award under grant N66001-17-1-4044, from a 2021 USC Stevens Center Technology Advancement Grant (TAG) award, from an Intel Faculty Award and from a Northrop Grumman grant. P.M.T. gratefully acknowledges the support from the NIH under grants R01 AG 058854, U01 AG 068057, R01 AG 060610, and P41 EB 015922. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) through NIH grant U01 AG 024904 and DoD ADNI through DoD award number W81-XWH-1220012. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development, LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the NIH (http://www.fnih.org). The ADNI grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data for this study were provided, in part, by the Human Connectome Project, WUMinn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1 U54 MH 091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research and by the McDonnell Center for Systems Neuroscience at Washington University. Research reported in this publication was also supported by the National Institute on Aging of the NIH under award U01 AG 052564. CamCAN funding was provided by the UK Biotechnology and Biological Sciences Research Council (grant number BB/H008217/1), together with support from the UK Medical Research Council and the University of Cambridge, UK. This research has been conducted using the UK Biobank Resource, under application numbers 11559 and 47656. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied by the NIH, DARPA, DoD, NSF, or any other entity acknowledged here. The funding sources had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Author affiliations: [a]Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089; [b]Ethel Percy Andrus Gerontology Center, Leonard Davis School of Gerontology, University of Southern California, Los Angeles, CA 90089; [c]Corwin D. Denney Research Center, Department of Biomedical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089; [d]Imaging Genetics Center, Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Marina del Rey, CA 90033; [e]Department of Quantitative & Computational Biology, Dana & David Dornsife College of Arts & Sciences, University of Southern California, Los Angeles, CA 90089; [f]Department of Ophthalmology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033; [g]Department of Neurology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033; [h]Department of Radiology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033; [i]Department of Psychiatry, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033; and [j]Department of Behavioral Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033

Author contributions: P.B. and A.I. designed research; C.Y. and A.A. performed research; C.Y., M.C., and A.A. contributed new reagents/analytic tools; C.Y., P.I., M.C., A.A., N.F.C., R.J.M., N.N.C., and X.C. analyzed data; A.N. data used in research was sourced from the ADNI repository; and C.Y., P.I., A.A., N.F.C., P.M.T., P.B., and A.I. wrote the paper.

1. A. Amgalan *et al*., Brain age estimation reveals older adults' accelerated senescence after traumatic brain injury. *GeroScience*, 1–17 (2022).
2. C. Lopez-Otin, M. A. Blasco, L. Partridge, M. Serrano, G. Kroemer, The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
3. A. Irimia, C. M. Torgerson, S.-Y. Matthew Goh, J. D. Van Horn, Statistical estimation of physiological brain age as a descriptor of senescence rate during adulthood. *Brain Imaging Behav.* **9**, 678–689 (2015).
4. I. Beheshti, S. Nugent, O. Potvin, S. Duchesne, Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme. *NeuroImage: Clin.* **24**, 102063 (2019).
5. J. H. Cole *et al*., Brain age predicts mortality. *Mol. Psychiatry* **23**, 1385–1392 (2018a).
6. R. J. Massett *et al*., Regional neuroanatomic effects on brain age inferred using magnetic resonance imaging and ridge regression. *J. Gerontol.: Ser. A* (2022).
7. K. Ning *et al*., Improving brain age estimates with deep learning leads to identification of novel genetic factors associated with brain aging. *Neurobiol. Aging* **105**, 199–204 (2021).
8. N. Opel *et al*., Brain structural abnormalities in obesity: Relation to age, genetic risk, and common psychiatric disorders. *Mol. Psychiatry* **26**, 4839–4852 (2021).
9. J. H. Cole *et al*., Increased brain-predicted aging in treated HIV disease. *Neurology* **88**, 1349–1357 (2017a).
10. J. H. Cole *et al*., No evidence for accelerated aging-related brain pathology in treated human immunodeficiency virus: Longitudinal neuroimaging results from the comorbidity in relation to AIDS (COBRA) project. *Clin. Infect. Dis.* **66**, 1899–1909 (2018b).
11. A. Irimia, Cross-sectional volumes and trajectories of the human brain, gray matter, white matter and cerebrospinal fluid in 9473 typically aging adults. *Neuroinformatics* **19**, 347–366 (2021).
12. A. Irimia, A. S. Maher, N. N. Chaudhari, N. F. Chowdhury, E. B. Jacobs, Acute cognitive deficits after traumatic brain injury predict Alzheimer's disease-like degradation of the human default mode network. *Geroscience* **42**, 1411–1429 (2020).
13. A. Salih *et al*., Brain age estimation at tract group level and its association with daily life measures, cardiac risk factors and genetic variants. *Sci. Rep.* **11**, 1–14 (2021).
14. W.-Y. I. Tseng, Y.-C. Hsu, T.-W. Kao, Brain age difference at baseline predicts clinical dementia rating change in approximately two years. *J. Alzheimer's Dis.* **86**, 613–627 (2022).
15. P. Imms, H. C. Chui, A. Irimia, Alzheimer's disease after mild traumatic brain injury. *Aging* **14**, 5292 (2022).
16. J. H. Cole *et al*., Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage* **163**, 115–124 (2017b).
17. H. Jiang *et al*., Predicting brain age of healthy adults based on structural MRI parcellation using convolutional neural networks. *Front. Neurol.* **10**, 1346 (2020).
18. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
19. L. Fisch *et al*., Predicting chronological age from structural neuroimaging: The predictive analytics competition 2019. *Front. Psychiatry* **12** (2021).
20. H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, S. M. Smith, Accurate brain age prediction with lightweight deep neural networks. *Med. Image Anal.* **68**, 101871 (2021).
21. W. Gong, C. F. Beckmann, A. Vedaldi, S. M. Smith, H. Peng, Optimising a simple fully convolutional network for accurate brain age prediction in the PAC 2019 challenge. *Front. Psychiatry* **12** (2021).
22. S. Borson *et al*., Improving dementia care: The role of screening and detection of cognitive impairment. *Alzheimers Dementia* **9**, 151–159 (2013).
23. C. Marcus, E. Mena, R. M. Subramaniam, Brain PET in the diagnosis of Alzheimer's disease. *Clin. Nucl. Med.* **39**, e413 (2014).
24. J. José Vaquero, P. Kinahan, Positron emission tomography: Current challenges and opportunities for technological advances in clinical and preclinical imaging systems. *Ann. Rev. Biomed. Eng.* **17**, 385–414 (2015).
25. J. Wrigglesworth *et al*., Brain-predicted age difference is associated with cognitive processing in later-life. *Neurobiol. Aging* **109**, 195–203 (2022).
26. L. E. Korthauer *et al*., Brain-behavior investigation of potential cognitive markers of Alzheimer's disease in middle age: A multi-modal imaging study. *Brain Imaging Behav.* **16**, 1–8 (2021).
27. J. Wang *et al*., Gray matter age prediction as a biomarker for risk of dementia. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21213–21218 (2019).
28. F. Biondo *et al*., Brain-age is associated with progression to dementia in memory clinic patients. *NeuroImage: Clin.* **36**, 103175 (2022).
29. C. Gaser *et al*., Brainage in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. *PloS One* **8**, e67346 (2013).
30. R. C. Petersen *et al*., Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization. *Neurology* **74**, 201–209 (2010).
31. J. H. Cole, R. E. Marioni, S. E. Harris, I. J. Deary, Brain age and other bodily: Implications for neuropsychiatry. *Mol. Psychiatry* **24**, 266–281 (2019).
32. D. Brennan, T. Wu, J. Fan, Morphometrical brain markers of sex difference. *Cereb. Cortex* **31**, 3641–3649 (2021).
33. L. Kogler *et al*., Sex differences in the functional connectivity of the amygdalae in association with cortisol. *NeuroImage* **134**, 410–423 (2016).
34. A. Veronica Witte, M. Savli, A. Holik, S. Kasper, R. Lanzenberger, Regional sex differences in grey matter volume are associated with sex hormones in the young adult human brain. *NeuroImage* **49**, 1205–1212 (2010).
35. G. F. Wooten, L. J. Currie, V. E. Bovbjerg, J. K. Lee, J. Patrie, Are men at greater risk for Parkinson's disease than women? *J. Neurol. Neuros. Psychiatry* **75**, 637–639 (2004).
36. T. Liu *et al*., The effects of age and sex on cortical sulci in the elderly. *NeuroImage* **51**, 19–27 (2010).
37. J. A. Kaye, C. DeCarli, J. S. Luxenberg, S. I. Rapoport, The significance of age-related enlargement of the cerebral ventricles in healthy men and women measured by quantitative computed X-ray tomography. *J. Am. Geriatr. Soc.* **40**, 225–231 (1992).
38. M. Hirnstein, K. Hugdahl, M. Hausmann, Cognitive sex differences and hemispheric asymmetry: A critical review of 40 years of research. *Laterality: Asymmetries Body Brain Cogn.* **24**, 204–252 (2019).
39. R. Cabeza, Hemispheric asymmetry reduction in older adults: The HAROLD model. *Psychol. Aging* **17**, 85 (2002).
40. S. Liu, J. Seidlitz, J. D. Blumenthal, L. S. Clasen, A. Raznahan, Integrative structural, functional, and transcriptomic analyses of sex-biased brain organization in humans. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 18788–18798 (2020).
41. E. R. Sowell *et al*., Sex differences in cortical thickness mapped in 176 healthy individuals between 7 and 87 years of age. *Cereb. Cortex* **17**, 1550–1560 (2007).
42. M. W. Bondi *et al*., Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *J. Alzheimer's Dis.* **42**, 275–289 (2014).
43. L. Christine Löwe, C. Gaser, K. Franke, Alzheimer's Disease Neuroimaging Initiative, The effect of the ApoE genotype on individual BrainAGE in normal aging, mild cognitive impairment, and Alzheimer's disease. *PloS One* **11**, e0157514 (2016).
44. H. I. L. Jacobs, M. P. J. Van Boxtel, J. Jolles, F. R. J. Verhey, H. B. M. Uylings, Parietal cortex matters in Alzheimer's disease: An overview of structural, functional and metabolic findings. *Neurosci. Biobehav. Rev.* **36**, 297–309 (2012).
45. A. M. Fjell *et al*., What is normal in normal aging? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus. *Progr. Neurobiol.* **117**, 20–40 (2014).
46. G. S. Day *et al*., Tau PET binding distinguishes patients with early-stage posterior cortical atrophy from amnestic Alzheimer's disease dementia. *Alzheimer Dis. Ass. Dis.* **31**, 87 (2017).
47. H. Braak, E. V. A. Braak, Staging of Alzheimer's disease-related neurofibrillary changes. *Neurobiol. Aging* **16**, 271–278 (1995).
48. B. Couvy-Duchesne *et al*., Ensemble learning of convolutional neural network, support vector machine, and best linear unbiased predictor for brain age prediction: Aramis contribution to the predictive analytics competition 2019 challenge. *Front. Psychiatry* **11**, 1451 (2020).
49. C.-Yuan Kuo *et al*., Improving individual brain age prediction using an ensemble deep learning framework. *Front. Psychiatry* **12** (2021).
50. P. Besson, T. Parrish, A. K. Katsaggelos, S. Kathleen Bandt, Geometric deep learning on brain shape predicts sex and age. *Comput. Med. Imaging Graph.* **91**, 101939 (2021).
51. F. Liem *et al*., Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage* **148**, 179–188 (2017).
52. J. Lee *et al*., Deep learning-based brain age prediction in normal aging and dementia. *Nature Aging* 10.1038/s43587-022-00219-7 (2022).
53. E. H. Leonardsen *et al*., Deep neural networks learn general and clinically relevant representations of the ageing brain. *NeuroImage* **256**, 119210 (2022).
54. R. Li, M. Singh, Sex differences in cognitive impairment and Alzheimer's disease. *Front. Neuroendocrinol.* **35**, 385–403 (2014).
55. S. He, P. Ellen Grant, Y. Ou, Global-local transformer for brain age estimation. *IEEE Trans. Med. Imaging* **41**, 213–224 (2021).
56. D. A. Wood *et al*., Accurate brain-age models for routine clinical MRI examinations. *NeuroImage* **249**, 118871 (2022).
57. H. Lee *et al*., Estimating and accounting for the effect of MRI scanner changes on longitudinal whole-brain volume change measurements. *NeuroImage* **184**, 555–565 (2019).
58. J. Stine Elam *et al*., The human connectome project: A retrospective. *NeuroImage* **244**, 118543 (2021).
59. D. C. Van Essen *et al*., The human connectome project: A data acquisition perspective. *NeuroImage* **62**, 2222–2231 (2012).
60. F. Alfaro-Almagro *et al*., Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* **166**, 400–424 (2018).
61. J. R. Taylor *et al*., The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage* **144**, 262–269 (2017).
62. L. Quitania Park *et al*., Confirmatory factor analysis of the ADNI neuropsychological battery. *Brain Imaging Behav.* **6**, 528–539 (2012).
63. B. Fischl, *Freesurfer. NeuroImage* **62**, 774–781 (2012).
64. Z. Zhang, "Improved Adam optimizer for deep neural networks" in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)* (IEEE, 2018), pp. 1–2.
65. T. T. Le *et al*., A nonlinear simulation framework supports adjusting for age when analyzing BrainAGE. *Front. Aging Neurosci.* **10**, 317 (2018).
66. M. S. Treder *et al*., Correlation constraints for regression models: Controlling bias in brain age prediction. *Front. Psychiatry* **12**, 25 (2021).
67. K. Kamnitsas *et al*., Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. image Anal.* **36**, 61–78 (2017).
68. P. Baldi, P. J. Sadowski, Understanding dropout. *Adv. Neural Inf. Process. Syst.* 26 (2013).
69. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995). http://www.jstor.org/stable/2346101.
70. K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv [preprint] (2013). http://arxiv.org/abs/1312.6034. Accessed 9 May 2022.
71. M. F. Glasser, D. C. Van Essen, Mapping human cortical areas in vivo based on myelin content as revealed by t1- and t2-weighted MRI. *J. Neurosci.* **31**, 11597–11616 (2011), 10.1523/JNEUROSCI.2180-11.2011.
72. S. O. Mahoney, N. F. Chowdhury, V. Ngo, A. Irimia, Mild traumatic brain injury results in significant and lasting cortical demyelination. *Front. Neurol.* **13**, 854396 (2022).
73. W. Zhang, L. Zhan, P. Thompson, Y. Wang, *Deep representation learning for multimodal brain networks in International Conference on Medical Image Computing and Computer-Assisted Intervention*. (Springer, pp. 613–624. 2020).