NEUROSCIENCE

Copyright © 2022 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

An uncertainty-aware, shareable, and transparent neural network architecture for brain-age modeling

Tim Hahn[1]*†, Jan Ernsting[1,2]†, Nils R. Winter[1], Vincent Holstein[1], Ramona Leenings[1,2], Marie Beisemann[3], Lukas Fisch[1], Kelvin Sarink[1], Daniel Emden[1], Nils Opel[1,4], Ronny Redlich[1,5], Jonathan Repple[1], Dominik Grotegerd[1], Susanne Meinert[1], Jochen G. Hirsch[6], Thoralf Niendorf[7], Beate Endemann[7], Fabian Bamberg[8], Thomas Kröncke[9], Robin Bülow[10], Henry Völzke[11], Oyunbileg von Stackelberg[12,13], Ramona Felizitas Sowade[12,13], Lale Umutlu[14], Börge Schmidt[14], Svenja Caspers[15,16], Harald Kugel[17], Tilo Kircher[18], Benjamin Risse[2], Christian Gaser[19], James H. Cole[20,21], Udo Dannlowski[1], Klaus Berger[22]

The deviation between chronological age and age predicted from neuroimaging data has been identified as a sensitive risk marker of cross-disorder brain changes, growing into a cornerstone of biological age research. However, machine learning models underlying the field do not consider uncertainty, thereby confounding results with training data density and variability. Also, existing models are commonly based on homogeneous training sets, often not independently validated, and cannot be shared because of data protection issues. Here, we introduce an uncertainty-aware, shareable, and transparent Monte Carlo dropout composite quantile regression (MCCQR) Neural Network trained on *N* = 10,691 datasets from the German National Cohort. The MCCQR model provides robust, distribution-free uncertainty quantification in high-dimensional neuroimaging data, achieving lower error rates compared with existing models. In two examples, we demonstrate that it prevents spurious associations and increases power to detect deviant brain aging. We make the pretrained model and code publicly available.

INTRODUCTION

Although aging is ubiquitous, the rate at which age-associated biological changes in the brain occur differs substantially between individuals. Building on this, the so-called "brain-age paradigm" (*1*) aims to estimate a brain's "biological age" (*2*) and posits that brain age may serve as a cumulative marker of disease risk, functional capacity, and residual life span (*3*). In a typical brain-age study, a machine learning model is trained on neuroimaging data—usually whole-brain structural T_1-weighted magnetic resonance imaging (MRI) data—to predict chronological age. This trained model is then used to evaluate neuroimaging data from previously unseen individuals and evaluated on the basis of the "brain-age gap" (BAG) as defined by the difference between predicted and chronological age.

A decade after its inception, this approach has developed into a major component of biological age research with a plethora of publications linking individual differences between chronological and brain age to genetic, environmental, and demographic characteristics in health and disease [for a comprehensive review, see (*4*)]. For example, a higher brain age compared to chronological age has been associated with markers of physiological aging (e.g., grip strength, lung function, walking speed), cognitive aging (*5*), life risk (*6*), and poor future health outcomes including progression from mild cognitive impairment to dementia (*7*, *8*), mortality (*5*), and a range of neurological diseases and psychiatric disorders [reviewed in (*9*)].

However, despite its scientific relevance and popularity, brain-age research faces numerous challenges, which hamper further progress and the translation of findings into clinical practice. First, the quantity and quality of the neuroimaging data upon which the underlying brain-age models are trained differ widely across studies, with only more recent studies training on more than 100 samples. Given the large number of voxels measured by modern structural MRI and the complexity of the multivariate machine learning models used in brain-age modeling, these small training sample sizes may lead to low-performance models with comparatively large errors. While studies drawing on larger training samples

[1]Institute for Translational Psychiatry, University of Münster, Münster, Germany. [2]Faculty of Mathematics and Computer Science, University of Münster, Münster, Germany. [3]Department of Statistics, TU Dortmund University, Dortmund, Germany. [4]Interdisciplinary Centre for Clinical Research (IZKF) of the Medical Faculty Münster, University of Münster, Münster, Germany. [5]Department of Psychology, University of Halle, Halle, Germany. [6]Fraunhofer MEVIS, Bremen, Germany. [7]Berlin Ultrahigh Field Facility (B.U.F.F.), NAKO imaging site Berlin, Max-Delbrueck Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. [8]Department of Radiology, Medical Center–University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. [9]Department of Diagnostic and Interventional Radiology, University Hospital Augsburg, Augsburg, Germany. [10]Institute of Diagnostic Radiology and Neuroradiology, University of Greifswald, Greifswald, Germany. [11]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany. [12]Department of Diagnostic and Interventional Radiology, University Hospital Heidelberg, Heidelberg, Germany. [13]Translational Lung Research Center, Member of the German Lung Research Center, Heidelberg, Germany. [14]Institute for Medical Informatics, Biometry and Epidemiology, University of Duisburg-Essen, Duisburg, Germany. [15]Institute for Anatomy I, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. [16]Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, 52425 Jülich, Germany. [17]Institute of Clinical Radiology, University of Münster, Münster, Germany. [18]Department of Psychiatry and Psychotherapy, Phillips University Marburg, Marburg, Germany. [19]Department of Psychiatry and Psychotherapy, and Department of Neurology, Jena University Hospital, Jena, Germany. [20]Department of Neuroimaging, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, UK. [21]Computational, Cognitive and Clinical Neuroimaging Laboratory, Department King's College, London, UK. [22]Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany.
*Corresponding author. Email: hahnt@wwu.de
†These authors contributed equally to this work.

exist, historical models are of limited value as they cover only a certain age range [cf., for example, the UK Biobank (*3*) dataset including subjects older than 45 years only]. In contrast, more recent studies train on several publicly available datasets that have reached training sample sizes of up to $N = 2001$ covering the full adult age range (*10*). Also, more recent studies [enhancing neuro imaging genetics through meta-analysis (ENIGMA) (*11*) or Kaufmann *et al.* (*12*)] have reached very good performance in large training and validation datasets despite focusing on a limited set of morphological features. These studies clearly provide major improvements with regard to previous studies.

Second, the assessment of the performance of the trained model is often severely hampered by small validation datasets. With so-called leave-one-out cross validation—a validation scheme that relies on averaging performance across N models using a validation dataset size of $N = 1$ in each iteration—the norm rather than the exception, performance estimates are often highly variable. If independent model validation on previously unseen datasets is conducted at all, validation samples are often small. Since validation sample sizes below $N = 100$ may yield a substantial percentage of spuriously inflated performance estimates, the stability of brain-age estimates underlying many brain-age studies might be questioned (*13*). Further, validation on large independent datasets including data from multiple recruitment centers and imaging sites has only improved recently with studies such as Bashyam *et al.* (*14*), who used a cross-disorder sample comprising $N = 11{,}729$ participants. Also, brain-age models are regularly not evaluated regarding potential biases such as gender, for which fitting separate models might be beneficial.

Third, although the initial brain-age framework suggested linear relevance vector regression (RVR), algorithms commonly used today also include (nonlinear) support vector machines (SVMs) and, more recently, Gaussian process regressors (*4*). Often, the choice of algorithm is not justified, and empirical comparisons of different approaches are rare. Especially problematic is the fact that Gaussian process regression (GPR) and RVR runtimes scale cubically with the number of samples, rendering training on large datasets difficult. Also, models trained using these algorithms cannot be shared among researchers because of data protection issues as they allow for a partial or even complete reconstruction of the training data, thereby hampering external validation. Moreover, the lack of publicly available brain-age models forces researchers to use a large portion of their data for brain-age model training and validation—either vastly decreasing statistical power for subsequent analyses or introducing additional variation because of cross-validation.

Fourth, an analysis regarding which characteristics of the brain drive brain-age predictions is either not conducted at all, or results are not comparable across studies because of different preprocessing and algorithm-specific importance score mapping. For example, mapping SVM weights as a proxy for feature importance may yield vastly different results, compared with deriving Gaussian process g-maps (*15*), rendering results incomparable. In addition, studies investigate the properties and performance of the machine learning model, not the interaction with single-subject data underlying predictions, thereby disregarding individual differences driving predictions. With ever-growing samples allowing for the application of ever more sophisticated machine learning algorithms, this underscores the need for a principled approach to transparency [cf. explainability (*16*)] that is comparable across algorithms.

The core metric of the field—i.e., the difference between chronological and predicted brain age (commonly referred to as BAG)—does not account for uncertainty in model predictions. Not adjusting the BAG for uncertainty, however, renders findings of altered aging confounded with data density and variability [cf. the concept of normative modeling (*17*) for an introduction]. Specifically, deviations between chronological age and BAG may arise not only from neural changes as intended but also erroneously from high uncertainty. Therefore, failing to properly model uncertainty may lead to spurious results, which depend on the characteristics of the training sample and properties of the model rather than on the underlying association of a variable with BAG. Uncertainty arising from noise inherent in the observations (i.e., aleatory uncertainty) and uncertainty arising from the model itself (i.e., epistemic uncertainty) must be considered. Figure 1 illustrates the concept of adjusting the BAG by individual uncertainty.

While not commonly used in brain-age research, algorithms often used in brain-age modeling such as RVR and the GPR are, in principle, capable of modeling aleatory and epistemic uncertainty. For high-dimensional inputs, however, uncertainty estimation becomes exceedingly difficult using these methods. This has sparked a plethora of research into alternative approaches, especially for neural networks (*18*–*20*). While interesting, most of these approaches

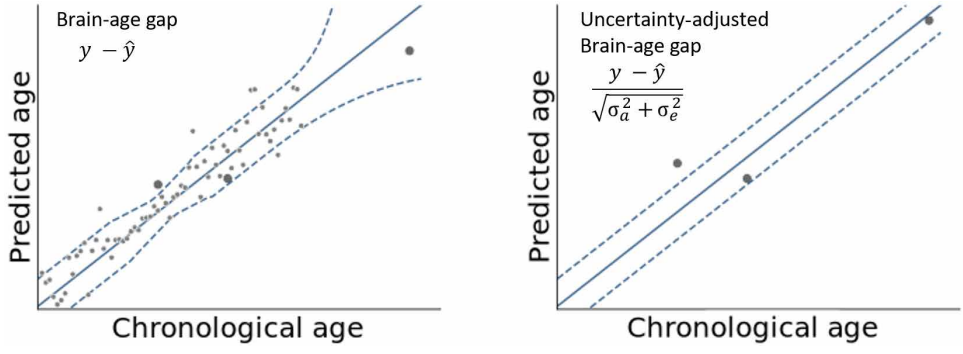

Fig. 1. Example data illustrating the effects of adjusting the BAG for individual uncertainty. Left: Regression model (solid line) with uncertainty estimate (e.g., 95% predictive interval; dotted lines) trained on toy data with varying density and variability (light grey) was applied to three test samples (dark gray). BAG is defined as a test sample's distance from the regression line. Right: Uncertainty adjustment increases BAG in areas of low uncertainty (left-most test sample) and decreases it in areas of high uncertainty (right-most test sample). σ_a, aleatory uncertainty; σ_e, epistemic uncertainty.

do not consider aleatory and epistemic uncertainty together, and none have been applied to brain-age research.

Here, we address these issues by (i) introducing a robust Monte Carlo dropout composite quantile regression (MCCQR) Neural Network architecture capable of estimating aleatory and epistemic uncertainty in high-dimensional data; (ii) training our model on anatomical MRI data of N = 10,691 individuals between 20 and 72 years of age from the German National Cohort (GNC); (iii) validating the resulting model using leave-site-out cross-validation across 10 recruitment centers and 3 additionally independent, external validation sets comprising a total of N = 4004 samples between 18 and 86 years of age; (iv) benchmarking the MCCQR model against the five most commonly used algorithms in brain-age modeling with regard to predictive performance and quantification of uncertainty; (v) systematically assessing model bias for gender, age, and ethnicity; and (vi) developing a unified explainability approach based on the combination of occlusion-sensitivity mapping and generalized linear multilevel modeling to identify brain regions driving brain-age predictions. Building on data from the GNC study, we apply the MCCQR model to predict uncertainty-adjusted brain-age gaps and investigate their association with body mass index (BMI) and major depressive disorder. As training data cannot be reconstructed from the MCCQR model, we make the pretrained model publicly available for validation and use in future research.

RESULTS

Model performance

We evaluated our MCCQR Neural Network model against five commonly used algorithms in brain-age modeling—namely, the RVR, linear SVM, SVM with a radial basis function kernel (SVM-rbf), GPR, and least absolute shrinkage and selection operator (LASSO) regression—regarding predictive performance. For comparison, we also evaluated a version of our neural network model without uncertainty quantification but with an otherwise identical network structure and hyperparameters [artificial neural network (ANN)].

We iteratively trained our model on data of patients recruited by 9 of the 10 recruitment centers contributing MRI data to the GNC (N = 10,691) and predicted brain age for all samples from the remaining center. This leave-site-out cross-validation showed a median absolute error (MAE) across all 10 recruitment centers of 2.94 years (SD = 0.22) for the MCCQR model. Performance of the other algorithms ranged from 3.05 (SD = 0.22) for both GPR and SVM to 4.25 (SD = 0.30) for LASSO regression. Results of 10-fold cross-validation corroborate this ranking of performance with the MCCQR, reaching an MAE of 2.95 years (SD = 0.16) with GPR and SVM MAE = 3.09 (SD = 0.11) and LASSO regression MAE = 4.19 (SD = 0.11). The ANN obtained MAE = 3.10 (SD = 0.14) for leave-site-out cross-validation and MAE = 3.02 (SD = 0.15) for 10-fold cross-validation.

While cross-validation performance—particularly across recruitment centers—usually provides good estimates of generalization performance, it does not consider additional sources of variability such as different data acquisition protocols, alterations in recruitment, or sample characteristics. Therefore, we validated all models in three independent samples (N = 4004), namely, the BiDirect study, the Marburg-Münster Affective Disorders Cohort Study (MACS), and the Information eXtraction from Images (IXI) dataset. To assess stability under real-world conditions of later use, these samples covered a larger age range than the training data (20 to 86 years versus 20 to 72 years in the GNC sample) as well as less restrictive exclusion criteria (for a detailed description, see Materials and Methods): For the BiDirect sample (N = 1460), the MCCQR model reached an MAE of 3.45 years. Performance of the other models ranged from 3.60 for the RVR to 4.79 for the SVM-rbf. The ANN reached an MAE of 3.76 years. In the MACS sample (N = 1986) (*21*), the MCCQR model reached an MAE of 3.92, while the other models obtained generalization performance between 4.15 (GPR and SVM) and 9.92 (SVM-rbf). The ANN reached MAE = 3.76 years. Last, we evaluated performance on the publicly available IXI dataset (www.brain-development.org, N = 561). The MCCQR model and the ANN reached an MAE of 4.57 and 4.48 years, respectively. The other models' performances ranged between MAE = 4.91 years (RVR) and MAE = 8.10 (SVM-rbf). Table 1 shows individual model performance for leave-site-out, 10-fold cross-validation, and the three independent validation samples. The MCCQR achieves lower MAE compared with all other algorithms commonly used in brain-age research. The ANN displays lower MAE than the MCCQR in two of five cases, indicating a slight advantage for the neural network architecture disregarding uncertainty in these cases.

Uncertainty quantification

Adjusting BAG for aleatory and epistemic uncertainty is crucial as findings of altered aging may otherwise be driven by, e.g., training data density and variability. While algorithms such as RVR and GPR are, in principle, capable of modeling aleatory and epistemic uncertainty, performance on high-dimensional data may be problematic. To evaluate the quality of the uncertainty quantification, we estimated uncertainty using the MCCQR, RVR, and GPR (note that the other algorithms do not readily provide uncertainty estimates). Then, we assessed prediction interval coverage probability (PICP), i.e., the probability that a sample's true value is contained within the predictive interval. Figure 2 depicts PICPs for given quantile values for cross-validation and the three independent validation datasets. Note that underestimation of uncertainty is highly problematic as samples may be erroneously characterized as deviants from the normal brain-aging trajectory. Overestimation of uncertainty decreases the ability to detect outliers, rendering the approach more conservative. While the GPR substantially overestimates uncertainty in all datasets, RVR and MCCQR provide high-quality

Table 1. MAE for all models, cross-validation schemes, and independent validation samples. CV, cross-validation. For cross-validation, SD across folds is given in parentheses.

Model	Leave-site-out CV	10-Fold CV	BiDirect	MACS	IXI
RVR	3.37 (0.16)	3.32 (0.13)	3.60	5.07	4.91
GPR	3.05 (0.22)	3.09 (0.11)	3.74	4.15	5.03
SVM	3.05 (0.22)	3.09 (0.11)	3.74	4.15	5.03
SVM-rbf	4.19 (0.27)	4.16 (0.16)	4.79	9.92	8.10
LASSO	4.25 (0.30)	4.19 (0.12)	4.44	8.35	6.94
ANN	3.10 (0.14)	3.02 (0.15)	3.56	3.76	4.48
MCCQR	2.94 (0.22)	2.95 (0.16)	3.45	3.91	4.57

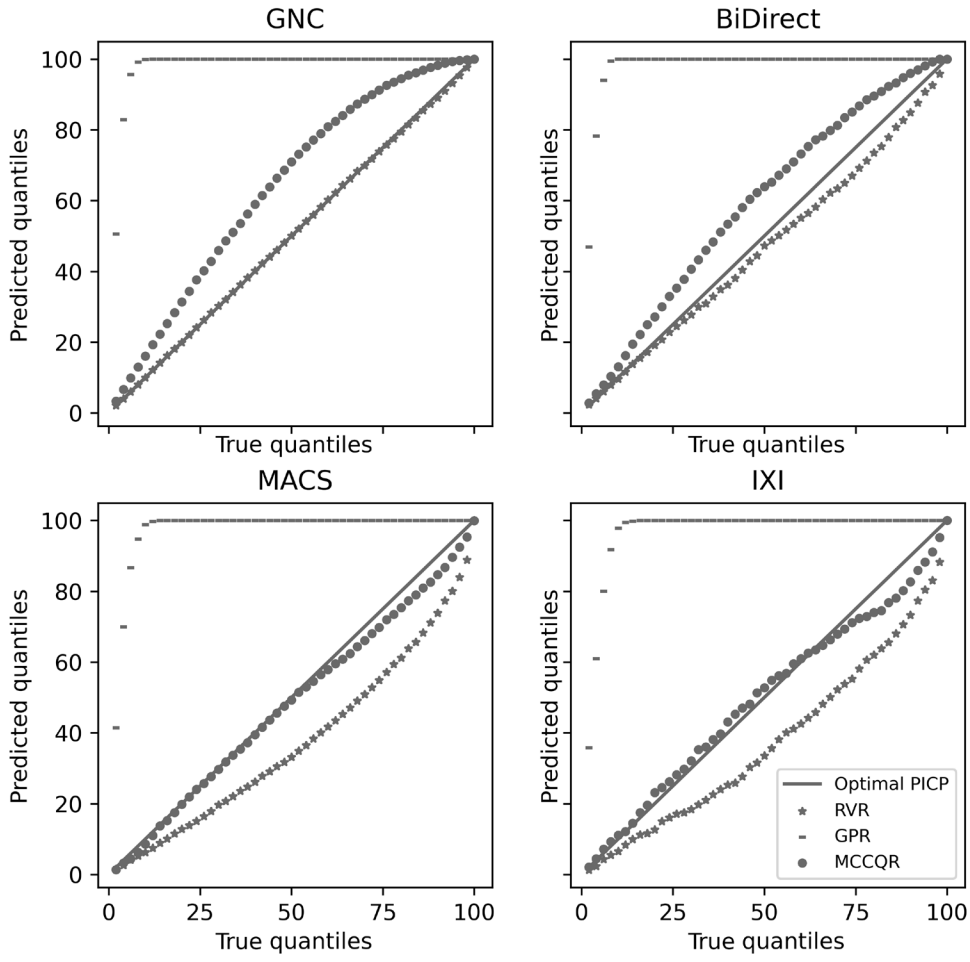

Fig. 2. PICP for leave-site-out GNC and independent validation samples (BiDirect, MACS, and IXI) for the RVR, the GPR, and our MCCQR Neural Network. Underestimation (overestimation) of uncertainty occurs if empirical PICPs are below (above) optimal PICP as indicated by the solid line.

uncertainty estimations for GNC and BiDirect datasets. MCCQR outperforms RVR in the MACS and IXI datasets. Note that incorporating epistemic or aleatory uncertainty alone—as has recently been suggested (*20*) for brain-age models—systematically underestimates uncertainty (see fig. S1).

Relevance of uncertainty adjustment

As outlined above, brain-age gaps may arise not only from brain changes as intended but also erroneously from high uncertainty, i.e., a person may have a large BAG not only due to actual changes in the brain but also due to properties of the underlying machine learning model that arise from characteristics of the training data such as data density and variability. Here, we empirically demonstrate two such cases.

First, we investigated the association of BAG and BMI in the GNC study data. On the basis of this, we find a significant association between BMI and BAG ($F_{1,10,344}$ = 7.06, P = 0.008). However, this effect is no longer observed if uncertainty is considered by scaling BAG with the SD of the individual predictive distribution ($F_{1,10,344}$ = 0.15, P = 0.697). Correspondingly, effect size (partial η^2) was reduced by 98%.

Second, we investigated the difference in BAG between a population sample (N = 1612) and patients suffering from major depressive disorder (N = 1541) from the MACS and the BiDirect sample. While the standard analysis based on BAG failed to reveal a significant difference ($F_{1,3148}$ = 1.61, P = 0.204), the same analysis based on uncertainty-corrected BAG detected a significant effect ($F_{1,3148}$ = 5.59, P = 0.018). Likewise, partial η^2 was increased by 247%.

Bias assessment

As machine learning models are not programmed but trained, they will mimic systematic biases inherent in their training data (*22*). While this potential algorithmic bias must be carefully investigated with regard to performance differences in specific subgroups to determine for which populations it yields robust estimates, it is often neglected in brain-age research [for an in-depth discussion, see (*23*)]. Here, we investigated model performance differences for gender, ethnicity, and age.

We found that model performance of the MCCQR did not substantially differ with regard to gender, reaching a standardized MAE of 0.459 for females and a standardized MAE of 0.456 for males in the BiDirect dataset (due to different age ranges between

males and females, we standardized MAE to make results comparable). In the MACS dataset, standardized MAE reached 0.298 in females and 0.264 in males. For the IXI dataset, the model reached a standardized MAE of 0.264 in females and 0.300 in males. Note that we defined gender as male and female as no further information on other genders was available. Regarding ethnicity, we tested our model on a publicly available dataset from Beijing Normal University (N = 179) between 18 and 28 years of age. Despite the low age range, the MCCQR failed to provide reasonable brain-age predictions with an MAE = 8.39. Last, we investigated model performance across different ages. As commonly reported in brain-age studies, we observed a correlation between BAG and age (r = −0.39) and between uncertainty-corrected BAG and age (r = −0.43) in the BiDirect sample. The same is true for the MACS dataset (BAG, r = −0.36; uncertainty-corrected BAG, r = −0.40) and the IXI dataset (BAG, r = −0.70; uncertainty-corrected BAG r = −0.75), implying that in all three validation datasets, performance is better in older participants (see fig. S2). Thus, we included age as a covariate in all statistical analyses involving the BAG in the "Relevance of uncertainty adjustment" section.

Explainability

Combining occlusion-sensitivity mapping and generalized linear multilevel modeling, we investigated which brain regions are relevant for accurate brain-age prediction in the MACS sample (N = 1986). We show that occlusion—i.e., the exclusion of data—of any of the 116 regions of the Automated Anatomical Labeling brain atlas significantly affected model performance (largest P < 0.001; see table S1). Occlusion of five regions leads to increased BAG, while occlusion of 107 regions decreased BAG. Strongest BAG-increasing effects were observed for left and right putamen and left Heschl's gyrus. Strongest BAG-decreasing effects were observed for right inferior temporal gyrus, left middle temporal gyrus, and left middle frontal gyrus (see fig. S3).

DISCUSSION

We trained an uncertainty-aware, shareable, and transparent MCCQR Neural Network on N = 10,691 samples from the GNC. This model achieves lower error rates compared with existing models across 10 recruitment centers and in three additional, independent validation samples (N = 4004). The model thus generalized well to independent datasets with larger age range than the training data. In contrast to currently used algorithms—which either do not provide uncertainty estimates or over- or underestimate uncertainty—the MCCQR model provides robust, distribution-free uncertainty quantification in high-dimensional neuroimaging data. Building on this, we show in N = 10,691 subjects from the GNC that a spuriously inflated brain-age deviation effect for BMI is found if uncertainty is ignored. Likewise, we demonstrate in the BiDirect sample that correcting for uncertainty also increases power to detect altered aging in major depressive disorder (N = 688) as compared with a population sample (N = 719). These findings underscore the need to adjust brain-age gaps for aleatory and epistemic uncertainty. Nonadjustment for these uncertainty components led to false-positive and false-negative findings in our studies.

While the importance of uncertainty quantification has been discussed at length in the context of normative modeling [cf. (*17*) for a discussion of the relevance of including aleatory and epistemic uncertainty] and two of the algorithms commonly used in brain-age research (mainly GPR) provide uncertainty quantification, it has largely been ignored in the brain-age literature to date. This could be due to several reasons: First, GPR and RVR do not scale well to large datasets as they require the inversion of the kernel matrix, and more scalable GPRs based on variational Bayes are still an active area of research. Second, reasonable uncertainty estimation for GPR models becomes exceedingly difficult in situations in which the number of features far exceeds the number of samples as is the case in virtually all MRI studies. This issue also arose in our study, leading to unreasonable uncertainty estimates (cf. Fig. 2). Third, GPR and RVR models require the full training sample to make predictions. While scalable GPR approaches only require a subset of the training data, the quality of predictions directly depends on the number and representativity of these so-called induction points. Note that in our study, the RVR algorithm's uncertainty quantification was of high quality in the GNC and BiDirect data but decreased in comparison to the MCCQR in the MACS and IXI datasets. This performance might render it a viable alternative with regard to uncertainty quantification in some cases. However, as literally the entire training dataset is required to make predictions using RVR, the data protection issues arising from this largely prohibit model sharing and thus independent validation.

A recent study also recognized the issue of uncertainty quantification for brain-age modeling and used quantile regression (QR) to estimate aleatory uncertainty in brain-age prediction (*20*). While this approach accounts for aleatory uncertainty induced by, e.g., measurement error, it does not consider epistemic uncertainty, i.e., uncertainty in the model weights. Empirically, we showed that accounting for aleatory uncertainty only substantially underestimated true uncertainty (cf. fig. S1). If data density differs over age groups, seemingly uncertainty-corrected brain-age gaps may still be confounded. Thereby, deviant brain ages might spuriously arise from differential training data density—an effect especially problematic given the relatively small training sample sizes used in most brain-age studies. In addition, it may be difficult to detect as only subsets of BAGs of a given sample are affected.

As mentioned above, our approach is intimately related to normative modeling (*17*). In contrast to this approach, however, we do not seek to quantify voxel-wise deviation, but deviation on the level of the individual. Hence, we do not predict single voxel data from chronological age, but chronological age from the multivariate pattern of whole-brain data. While this directly yields brain-age predictions on the level of the individual—hence circumventing the need to estimate individual-level predictions based on extreme value statistics or the combination of deviations across voxels as has been suggested for normative modeling—it cannot directly be used as a brain mapping method. To this end, we adopted an occlusion-sensitivity mapping approach, which quantifies regional importance as the reduction in BAG when features from a specific region are withheld (*24*). Compared with other approaches to explainability such as the visualization of the network weights using, e.g., layer-wise relevance propagation (*25*), occlusion-sensitivity mapping comes with the benefit of yielding relevant features for each individual. From a methodological point of view, the multilevel model used in this study holds crucial advantages over the commonly used mass-univariate approach. First, in the mass-univariate approach, estimates for the effect of the predictors on BAG are only informed by one region. In the multilevel framework, however, each estimate is

informed not only by its region but also by all other regions. Second, the multilevel model enables us to include theoretically necessary control variables, which vary over brain regions, but not over samples, such as the size of each region of interest (ROI). Third, the multilevel approach controls for dependency within the model, alleviating the need for multiple comparison correction as required in the mass-univariate case (*26*). This analysis revealed that occlusion of any region resulted in a significant change of brain-age predictions, underscoring the multivariate, distributed nature of aging in the brain. Also, our results mitigate concerns that high-performance brain-age models might focus on a small subset of features not affected by pathology (*14*).

Machine learning models are trained on data and may thus mimic systematic biases inherent in their training data. Investigating performance in specific subgroups revealed that the MCCQR performed comparably for females and males. As is the case for most brain-age models, we observed a correlation between BAG and age. This behavior is not unexpected, as errors in regression modeling will always tend toward the mean (age in this case) of the training set. Nonetheless, it requires modeling age as a covariate in all analyses aiming to associate BAG with variables of interest [for an introduction to age-bias correction approaches in brain-age modeling, see (*27*)]. Investigating ethnicity bias, we showed that our model fails to accurately predict age in a Chinese sample. On the one hand, this limitation was to be expected given the GNC aims to recruit a random population sample of Germany. One the other hand, this effect underscores the need for systematic bias assessment. The MCCQR also offers an opportunity to remedy this issue (see below).

The MCCQR—in contrast to most approaches used in brain-age research—does not allow for the reconstruction of individual samples from the training sample. Thus, we make it publicly available. This is beneficial for three reasons: First, it allows others to independently assess MCCQR model performance. Second, it may markedly increase power and robustness of smaller brain-age studies by circumventing the need to train a brain-age model and test associations with variables of interest in the same dataset. Third, it enables researchers to continue training the model with more data. For example, fine-tuning the MCCQR model with data from other ethnic groups—e.g., Asian—would help the model generalize better. In this regard, future studies using additional data from ethnic groups ought to clarify whether it is ethnicity per se or, e.g., different MRI scanning protocols that have led to lower performance. In the same vein, our model could be extended by adding, e.g., three-dimensional convolution layers [as done, e.g., in (*10*, *14*)] to the MCCQR to allow predictions directly from raw MRI data without the need for preprocessing. While this study constitutes a first step toward incorporating uncertainty in brain-age modeling, it is limited in several ways. First, we evaluated the model neither on a large sample of older participants (>72 years) nor on a sample of adolescents. Second, we did not explicitly model GNC imaging sites during training, and a benchmark against state-of-the-art deep learning approaches is missing. Future studies could therefore not only evaluate the model but also increase generalization and usability by training on more diverse datasets and developing model architecture. We facilitate this research by making the pretrained model publicly available with this publication.

In light of this, we provide the uncertainty-aware, shareable, and transparent MCCQR architecture and pretrained model with the intention to stimulate further research and increase power for small-sample analyses. The pretrained PHOTON-AI model and code can be downloaded from the PHOTON AI model repository (www.photon-ai.com/repo).

MATERIALS AND METHODS

Training and validation samples

Whole-brain MRI data from five sources were used. We trained the MCCQR on the GNC sample. Results for leave-site-out and 10-fold cross-validation are also based on the GNC sample. Independent validation was based on the BiDirect sample, the MACS data, and the IXI dataset. Ethnicity bias assessment was conducted using the Beijing Normal University dataset (see below). In the following, we describe each dataset in more detail. Also, table S2 provides further sample characteristics, including sample sizes, gender distribution, age minimum and maximum, and SD.

German National Cohort

This cohort is one of the population-based "megacohorts" and examines 205,000 Germans, aged 20 to 72 years, in 18 study centers across Germany between 2014 and 2019 using a comprehensive program. Specifically, this included a 3.0-Tesla whole-body MRI (T_1w-MPRAGE) in 30,000 participants, performed in five GNC imaging centers equipped with dedicated identical magnets (Skyra, Siemens Healthineers, Erlangen, Germany) examining participants from 11 of the centers. This analysis is based on the "data freeze 100K" milestone for the first 100,000 participants, which also included the first 10,691 participants with completed MRIs of sufficient quality [for a detailed protocol, see (*28*, *29*)]. We calculated BMI from directly measured height and weight (kg/m^2; mean Becks Depression Inventory (BDI) = 26.82; SD BDI = 4.76). To ensure that our models were not driven by data quality or total intracranial volume (TIV), we assessed the predictive power of the three data quality parameters provided by the Cat12 toolbox and TIV. We show that TIV, bias, noise, and weighted average interquartile range combined explain only 9.06% of variation in age using a linear SVM with 10-fold cross-validation compared to the 86% achieved by the MCCQR model.

BiDirect

The BiDirect study is an ongoing study that comprises three distinct cohorts: patients hospitalized for an acute episode of major depression, patients 2 to 4 months after an acute cardiac event, and healthy controls randomly drawn from the population register of the city of Münster, Germany. Baseline examination of all participants included a structural MRI of the brain, a computer-assisted face-to-face interview about sociodemographic characteristics, a medical history, an extensive psychiatric assessment, and collection of blood samples. Inclusion criteria for the present study were availability of completed baseline MRI data with sufficient MRI quality. All patients with major depressive disorder had an episode of major depression at the time of recruitment and were either currently hospitalized (>90%) or had been hospitalized for depression at least once during the 12 months before inclusion in the study (<10%). Further details on the rationale, design, and recruitment procedures of the BiDirect study have been described elsewhere (*30*).

Marburg-Münster Affective Disorder Cohort Study

Participants were recruited through psychiatric hospitals or newspaper advertisements. Inclusion criteria included mild, moderate, or partially remitted major depressive disorder episodes in addition to severe depression. Patients could be undergoing inpatient, outpatient, or no current treatment. The MACS was conducted at

two imaging sites: University of Münster, Germany, and University of Marburg, Germany. Further details about the structure of the MACS (*31*) and MRI quality assurance protocol (*21*) are provided elsewhere. Inclusion criteria for the present study were availability of completed baseline MRI data with sufficient MRI quality [see (*21*) for details].

Information eXtraction from Images

This dataset comprises images from normal, healthy participants, along with demographic characteristics, collected as part of the IXI project available for download (https://brain-development.org/ixi-dataset/). The data have been collected at three hospitals in London (Hammersmith Hospital using a Philips 3T system, Guy's Hospital using a Philips 1.5T system, and Institute of Psychiatry using a GE 1.5T system). Inclusion criteria for the present study were availability of completed baseline MRI data.

Beijing Normal University

This dataset includes 180 healthy controls from a community (student) sample at Beijing Normal University in China. Inclusion criteria for the present study were availability of completed baseline MRI data. Further details can be found online (http://fcon_1000.projects.nitrc.org/indi/retro/BeijingEnhanced.html).

MRI preprocessing

MRI data were preprocessed using the CAT12 toolbox (built 1450 with SPM12 version 7487 and Matlab 2019a; http://dbm.neuro.uni-jena.de/cat) with default parameters. Images were bias corrected, segmented using tissue classification, normalized to MNI-space using DARTEL normalization, smoothed with an isotropic Gaussian Kernel (8 mm FWHM), and resampled to 3-mm isomorphic voxels. Using the PHOTON AI software (see "Model training and validation" section below), a whole-brain mask comprising all gray matter voxels was applied, data were vectorized, features with zero variance in the GNC dataset were removed, and the scikit-learn Standard Scaler was applied.

MCCQR model

Commonly, regression models $f^W(x)$ are used to describe the relationship of features $X = (x,\ldots,x_n)$ and a target variable $y = (y_1,\ldots,y_n)$. We denote model predictions $\hat{y} = f^W(x)$ with W the model parameters. Commonly, these regression models provide predictions as point estimates rather predictive distributions. We thus consider two types of uncertainty in accordance with Kendall and Gal (*32*). The first—aleatory uncertainty—captures noise inherent in the observations. The second—epistemic uncertainty—accounts for uncertainty in the model. While the former is irreducible for a given model, the latter depends on data availability and training and is thus affected by data density and can be reduced by training (i) the model based on the loss function and (ii) with additional data. While numerous approaches to uncertainty quantification have been suggested, no commonly accepted approach exists. While Bayesian approaches such as GPR provide a mathematically elegant framework, such approaches often fail in high-dimensional settings or are not scalable to large datasets (*33*). In particular, capturing aleatory and epistemic uncertainty within the same neural network model remains challenging (*34*). Here, we suggest to combine composite QR and Monte Carlo dropout to model aleatory and epistemic uncertainty within a single framework, respectively.

Composite QR

QR provides an estimate not only of the conditional mean (as is done when optimizing for mean absolute error) but also of any quantile in the data. This comes with two advantages. First, we can estimate the median (instead of the mean), thereby obtaining a prediction more robust to potential outliers. Second, predicting quantiles can yield predictive intervals, thereby modeling aleatory uncertainty. For example, we cannot only predict a sample's target value but also the 95% confidence bounds of this prediction. This makes QR interesting whenever percentile curves are of interest, e.g., when screening for abnormal growth.

Commonly, conditional quantiles for predetermined quantile probabilities are estimated separately by different regression equations. These are then combined to build a piecewise estimate of the conditional response distribution. As this approach is prone to "quantile crossing," i.e., QR predictions do not increase with the specified quantile probability τ, composite QR was introduced (*35*). In composite QR, simultaneous estimates for multiple values of τ are obtained, and the regression coefficients are shared across the different QR models. In essence, we aim to approximate a single τ-independent function that best describes the function to be learned. Structurally, composite QR differs from QR only in that the QR error (tilted loss) function is summed over many, usually equally spaced values of τ. Specifically, QR is implemented using the QR error function (*36*) to optimize a neural network via

$$E_\tau = \frac{1}{N}\sum_{t=1}^{N}\rho_\tau(y(t)-\hat{y}(t))$$

with the tilted loss function ρ

$$\rho_\tau(\varepsilon) = \tau\cdot\varepsilon \text{ if } \varepsilon \geq 0, \text{else } (1-\tau)\,\varepsilon$$

Composite QR extends this idea to estimating multiple quantiles simultaneously. This is not only more stable but also computationally more efficient as fewer coefficients and fewer operations during weight updating are required. To achieve this, we modified the QR error function (above) to

$$E_{C\tau} = \frac{1}{\mathrm{KN}}\sum_{k=1}^{K}\sum_{t=1}^{N}\rho_{\tau_k}(y(t)-\hat{y}_{\tau_k}(t))$$

where τ_k are usually equally spaced, for example, $\tau_k = \frac{k}{K+1}$ for $k = 1,2,\ldots, K$ as suggested by Cannon (*18*). Specifically, we calculate 101 equally spaced quantile values for $0 < \tau < 1$. To allow for continuous sampling during prediction, we linearly interpolated these quantile values. Note that while this approach does not formally guarantee the absence of quantile crossover, composite QR reduced the likelihood so much that we never empirically observed quantile crossover for any of the more than 1.48 billion quantiles (arising from 101 quantiles, 14,695 samples, and 1000 draws from the predictive distribution per sample) estimated during independent and cross-validation in this study.

Monte Carlo dropout

While composite QR captures aleatory uncertainty, it does not account for epistemic uncertainty, i.e., uncertainty in the model parameters. For example, epistemic uncertainty should be higher in regions of the input space where little data are available, whereas it should be lower in regions with high data density. While we could, in principle, model each weight of a neural network as distribution from which to sample weight values during prediction, estimating these probabilistic models remains challenging for high-dimensional

data. Following Gal and Ghahramani (*19*), we therefore adopted a Monte Carlo dropout approach. While dropout is commonly used for regularization during neural network training, the Monte Carlo dropout approach enforces dropout during training and at test time. Thus, dropout can be used to obtain T predictions for the same input with different active neurons. This allows the estimation of $p(y|f^W(x))$, the mean probability of a prediction given a test input $X = (X_1,\ldots,X_n)$ for the neural network f with the according weights W. We therefore define our likelihood as a Gaussian with mean given by the model output according to Gal and Ghahramani (*19*)

$$p(y\,|\,f^W(x)) = N(f^W(x),\sigma^2)$$

We can then calculate the mean probability using Monte Carlo dropout performing T forward passes using randomly sampled weight values W_i from the neural network using dropout as

$$p(y|\,f^W(x)) = \frac{1}{T}\sum_{i=1}^{T} p(f^W(x) = k\,|\,x, W_i)$$

Monte Carlo dropout composite QR

To model epistemic and aleatory uncertainty simultaneously within the same framework, we combined Monte Carlo dropout and composite QR described above. Specifically, we implemented the resulting MCCQR Neural Network consisting of one hidden layer with 32 rectified linear units using Tensorflow 2.0 together with tensor flow probability for robust median MAE calculation over batches. Note that we used MAE instead of the more common mean absolute error as the median is more robust to outliers compared to the mean. We trained for 10 epochs with a learning rate of 0.01, a batch size of 64, and a dropout rate of 0.2 using the Adam Optimizer with default settings. Predictions were obtained by sampling 1000 times from each sample's predictive distribution with random τ values and dropout enabled. A sample's brain-age prediction was computed as the median of the resulting values. Likewise, uncertainty was computed as the SD of the resulting values.

Reconstruction of individual-level data

Machine learning models traditionally used in brain-age modeling allow for the reconstruction of individual-level data. For our model, however, only saving model parameters and network architecture is required. Reconstructing individual-level data from this information alone is not possible for quantitative and conceptual reasons: Considering the quantity of information, our final model contains 1,269,701 parameters, 64 of which are not trainable. The complete training dataset consists of 10,696 images containing 39,904 voxels each, resulting in 426,733,376 parameters in the dataset. This amounts to about 336 times the number of parameters in our model, rendering it incapable of "memorizing" the data. Considering the conceptual nature of the training process, our model applies the Monte Carlo dropout. This method—used to counter overfitting and to estimate epistemic uncertainty—randomly reduces the number of active units during a forward pass. As the network parameters are optimized for inference across the whole training set, the network is not capable of compressing the images. Thus, we cannot recover individual trainings samples from the network.

Benchmarking alternative machine learning models

We evaluated our MCCQR Neural Network model against five commonly used algorithms in brain-age modeling—namely, the RVR, linear SVM, SVM-rbf, GPR, and LASSO regression. We used the fast-rvm implementation from sklearn-bayes (https://github.com/AmazaspShumik/sklearn-bayes) and used SVM, GPR, and LASSO implementations from sci-kit learn (*37*) with default settings. For comparison, we also evaluated a version of our neural network model without uncertainty quantification but with an otherwise identical network structure and hyperparameters optimizing mean absolute error over predictions instead of the tilted loss function used for the MCCQR model.

Model training and validation

All models were trained and cross-validated using the PHOTON AI software (www.photon-ai.com) for leave-site-out and 10-fold cross-validation on the GNC sample. Independent validation was conducted using PHOTON AI's .photon format for pipeline-based prediction with the BiDirect, MACS, IXI, and Beijing Normal University samples.

Generalized linear multilevel modeling for occlusion-sensitivity mapping

Occlusion-sensitivity mapping—in analogy to occlusion-based approaches for two-dimensional images (*24*)—quantifies regional importance as the reduction in performance when features from a specific region are withheld. Occlusion was implemented by sequentially setting all voxels within each of the 116 ROIs of the AAL brain atlas (*38*) to zero. Occlusion sensitivity is widely used to gain insight into which regions of an image a machine learning model uses for prediction. A region is considered more important if model performance decreases more strongly when information from this region is withheld.

To quantify this notion of importance, we combined occlusion-sensitivity mapping with generalized linear multilevel modeling. Specifically, we used multiple linear regression using the R packages lme4 to model the difference between BAG based on whole-brain data and BAG if information from a specific atlas region is withheld.

To investigate regionally specific effects, we computed uncertainty-corrected BAG estimates for each individual and ROI (i.e., 116 regions from the AAL atlas distributed with statistical parametric mapping (SPM); www.fil.ion.ucl.ac.uk/spm/) independently. Then, we predicted uncertainty-corrected BAG from a group factor with AAL regions as factor levels, controlling for chronological age, site (where appropriate), gender, and ROI size. The ROI group factor was defined as treatment contrast with the whole-brain (no occlusion) level as reference factor. This way, ROI effect estimates can easily be interpreted as difference in BAG from the full model. P values for ROI factor level effects were computed using the Kenward-Roger approximation (*39*) as implemented in the afex R package. This approach allows us to test whether the occlusion of a given region results in an above-chance change of predictive performance. While beyond the scope of this paper, the combination of occlusion-sensitivity mapping and generalized linear multilevel modeling also allows for the investigation of regional effects between clinical groups.

More generally, our approach can be considered a form of model-independent explainable AI: In recent years, a large number of methods have been proposed that aim to shed light on the contribution of variables or groups of variables on a model prediction [for a review and best practice, see (*40*)]. Among those are algorithm-specific approaches such as layer-wise relevance propagation that work with deep neural networks or the weight mapping for SVMs.

In essence, these techniques provide (qualitative) heatmaps, i.e., help to understand the informational flow in the network. In contrast, more general approaches ask how the prediction for a single sample/participant would change if the data changed (e.g., if we did not have variable *x* or if its single-to-noise ratio changes). As we aimed to quantify the effect of omitting/occluding ROIs independent of a specific training sample, we opted for occlusion mapping here.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abg9471

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. J. H. Cole, K. Franke, Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends Neurosci.* **40**, 681–690 (2017).
2. F. C. Ludwig, M. E. Smoke, The measurement of biological age. *Exp. Aging Res.* **6**, 497–522 (1980).
3. J. H. Cole, Multimodality neuroimaging brain-age in UK biobank: Relationship to biomedical, lifestyle, and cognitive factors. *Neurobiol. Aging* **92**, 34–42 (2020).
4. K. Franke, C. Gaser, Ten years of brainage as a neuroimaging biomarker of brain aging: What insights have we gained? *Front. Neurol.* **10**, 789 (2019).
5. J. H. Cole, S. J. Ritchie, M. E. Bastin, M. C. V. Hernandez, S. M. Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, N. R. Wray, P. Redmond, R. E. Marioni, J. M. Starr, S. R. Cox, J. M. Wardlaw, D. J. Sharp, I. J. Deary, Brain age predicts mortality. *Mol. Psychiatry* **23**, 1385–1392 (2018).
6. N. Bittner, C. Jockwitz, K. Franke, C. Gaser, S. Moebus, U. J. Bayen, K. Amunts, S. Caspers, When your brain looks older than expected: Combined lifestyle risk and BrainAGE. *Brain Struct. Funct.* **226**, 621–645 (2021).
7. K. Franke, C. Gaser, Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. *GeroPsych* **25**, 235–245 (2012).
8. C. Gaser, K. Franke, S. Kloppel, N. Koutsouleris, H. Sauer; Alzheimer's Disease Neuroimaging Initiative, BrainAGE in Mild cognitive impaired patients: Predicting the conversion to Alzheimer's disease. *PLOS ONE* **8**, e67346 (2013).
9. J. H. Cole, R. E. Marioni, S. E. Harris, I. J. Deary, Brain age and other bodily 'ages': Implications for neuropsychiatry. *Mol. Psychiatry* **24**, 266–281 (2019).
10. J. H. Cole, R. P. K. Poudel, D. Tsagkrasoulis, M. W. A. Caan, C. Steves, T. D. Spector, G. Montana, Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage* **163**, 115–124 (2017).
11. L. K. M. Han, R. Dinga, T. Hahn, C. R. K. Ching, L. T. Eyler, L. Aftanas, M. Aghajani, A. Aleman, B. T. Baune, K. Berger, I. Brak, G. B. Filho, A. Carballedo, C. G. Connolly, B. Couvy-Duchesne, K. R. Cullen, U. Dannlowski, C. G. Davey, D. Dima, F. L. S. Duran, V. Enneking, E. Filimonova, S. Frenzel, T. Frodl, C. H. Y. Fu, B. R. Godlewska, I. H. Gotlib, H. J. Grabe, N. A. Groenewold, D. Grotegerd, O. Gruber, G. B. Hall, B. J. Harrison, S. N. Hatton, M. Hermesdorf, I. B. Hickie, T. C. Ho, N. Hosten, A. Jansen, C. Kahler, T. Kircher, B. Klimes-Dougan, B. Kramer, A. Krug, J. Lagopoulos, R. Leenings, F. P. MacMaster, G. MacQueen, A. McIntosh, Q. McLellan, K. L. McMahon, S. E. Medland, B. A. Mueller, B. Mwangi, E. Osipov, M. J. Portella, E. Pozzi, L. Reneman, J. Repple, P. G. P. Rosa, M. D. Sacchet, P. G. Samann, K. Schnell, A. Schrantee, E. Simulionyte, J. C. Soares, J. Sommer, D. J. Stein, O. Steinstrater, L. T. Strike, S. I. Thomopoulos, M. J. van Tol, I. M. Veer, R. R. J. M. Vermeiren, H. Walter, N. J. A. van der Wee, S. J. A. van der Werff, H. Whalley, N. R. Winter, K. Wittfeld, M. J. Wright, M. J. Wu, H. Volzke, T. T. Yang, V. Zannias, G. I. de Zubicaray, G. B. Zunta-Soares, C. Abe, M. Alda, O. A. Andreassen, E. Ben, C. M. Bonnin, E. J. Canales-Rodriguez, D. Cannon, X. Caseras, T. M. Chaim-Avancini, T. Elvsaashagen, P. Favre, S. F. Foley, J. M. Fullerton, J. M. Goikolea, B. C. M. Haarman, T. Hajek, C. Henry, J. Houenou, F. M. Howells, M. Ingvar, R. Kuplicki, B. Lafer, M. Landen, R. Machado-Vieira, U. F. Malt, C. McDonald, P. B. Mitchell, L. Nabulsi, M. C. G. Otaduy, B. J. Overs, M. Polosan, E. Pomarol-Clotet, J. Radua, M. M. Rive, G. Roberts, H. G. Ruhe, R. Salvador, S. Sarro, T. D. Satterthwaite, J. Savitz, A. H. Schene, P. R. Schofield, M. H. Serpa, K. Sim, M. G. Soeiro-de-Souza, A. N. Sutherland, H. S. Temmingh, G. M. Timmons, A. Uhlmann, E. Vieta, D. H. Wolf, M. V. Zanetti, N. Jahanshad, P. M. Thompson, D. J. Veltman, B. W. J. H. Penninx, A. F. Marquand, J. H. Cole, L. Schmaal, Brain aging in major depressive disorder: Results from the ENIGMA major depressive disorder working group. *Mol. Psychiatry* **26**, 5124–5139 (2020).
12. T. Kaufmann, D. van der Meer, N. T. Doan, E. Schwarz, M. J. Lund, I. Agartz, D. Alns, D. M. Barch, R. Baur-Streubel, A. Bertolino, F. Bettella, M. K. Beyer, E. Ben, S. Borgwardt, C. L. Brandt, J. Buitelaar, E. G. Celius, S. Cervenka, A. Conzelmann, A. Cordova-Palomera, A. M. Dale, D. J. F. de Quervain, P. D. Carlo, S. Djurovic, E. S. Drum, S. Eisenacher, T. Elvsaashagen, T. Espeseth, H. Fatouros-Bergman, L. Flyckt, B. Franke, O. Frei, B. Haatveit, A. K. Haaberg, H. F. Harbo, C. A. Hartman, D. Heslenfeld, P. J. Hoekstra, E. A. Hgestl, T. L. Jernigan, R. Jonassen, E. G. Jonsson, L. Farde, L. Flyckt, G. Engberg, S. Erhardt, H. Fatouros-Bergman, S. Cervenka, L. Schwieler, F. Piehl, I. Agartz, K. Collste, P. Victorsson, A. Malmqvist, M. Hedberg, F. Orhan, P. Kirsch, I. Koszewska, K. K. Kolskaar, N. I. Landr, S. L. Hellard, K. P. Lesch, S. Lovestone, A. Lundervold, A. J. Lundervold, L. A. Maglanoc, U. F. Malt, P. Mecocci, I. Melle, A. Meyer-Lindenberg, T. Moberget, L. B. Norbom, J. E. Nordvik, L. Nyberg, J. Oosterlaan, M. Papalino, A. Papassotiropoulos, P. Pauli, G. Pergola, K. Persson, G. Richard, J. Rokicki, A. M. Sanders, G. Selbk, A. A. Shadrin, O. B. Smeland, H. Soininen, P. Sowa, V. M. Steen, M. Tsolaki, K. M. Ulrichsen, B. Vellas, L. Wang, E. Westman, G. C. Ziegler, M. Zink, O. A. Andreassen, L. T. Westlye, Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat. Neurosci.* **22**, 1617–1623 (2019).
13. C. Flint, M. Cearns, N. Opel, R. Redlich, D. M. A. Mehler, D. Emden, N. R. Winter, R. Leenings, S. B. Eickhoff, T. Kircher, A. Krug, I. Nenadic, V. Arolt, S. Clark, B. T. Baune, X. Jiang, U. Dannlowski, T. Hahn, Systematic overestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology* **46**, 1510–1517 (2019).
14. V. M. Bashyam, G. Erus, J. Doshi, M. Habes, I. Nasralah, M. Truelove-Hill, D. Srinivasan, L. Mamourian, R. Pomponio, Y. Fan, L. J. Launer, C. L. Masters, P. Maruff, C. Zhuo, H. Volzke, S. C. Johnson, J. Fripp, N. Koutsouleris, T. D. Satterthwaite, D. Wolf, R. E. Gur, R. C. Gur, J. Morris, M. S. Albert, H. J. Grabe, S. Resnick, R. N. Bryan, D. A. Wolk, H. Shou, C. Davatzikos, MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain* **143**, 2312–2324 (2020).
15. A. Marquand, M. Howard, M. Brammer, C. Chu, S. Coen, J. Mourao-Miranda, Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. *NeuroImage* **49**, 2178–2189 (2010).
16. P. Voosen, How AI detectives are cracking open the black box of deep learning. *Science* (2017).
17. A. F. Marquand, I. Rezek, J. Buitelaar, C. F. Beckmann, Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies. *Biol. Psychiatry* **80**, 552–561 (2016).
18. A. J. Cannon, Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stoch. Env. Res. Risk A.* **32**, 3207–3225 (2018).
19. Y. Gal, Z. Ghahramani, in *Proceedings of The 33rd International Conference on Machine Learning*, K. Q. Weinberger, Maria Florina Balcan, Eds. (PMLR, 2016), vol. 48, pp. 1050–1059.
20. M. Palma, S. Tavakoli, J. Brettschneider, T. E. Nichols; Alzheimer's Disease Neuroimaging Initiative, Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression. *NeuroImage* **219**, 116938 (2020).
21. C. Vogelbacher, T. W. D. Mobius, J. Sommer, V. Schuster, U. Dannlowski, T. Kircher, A. Dempfle, A. Jansen, M. H. A. Bopp, The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data. *NeuroImage* **172**, 450–460 (2018).
22. M. Cearns, T. Hahn, B. T. Baune, Recommendations and future directions for supervised machine learning in psychiatry. *Transl. Psychiatry* **9**, 271 (2019).
23. M. L. FAT, Fairness, Accountability, and Transparency in Machine Learning. *Retrieved December*. **24**, 2018 (2018).
24. M. D. Zeiler, R. Fergus, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2014), vol. 8689 LNCS of *Lecture Notes in Computer Science*, pp. 818–833.
25. S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Muller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE* **10**, e0130140 (2015).
26. A. Gelman, J. Hill, M. Yajima, Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* **5**, 189–211 (2012).
27. A. M. G. de Lange, J. H. Cole, Commentary: Correction procedures in brain-age prediction. *Neuroimage Clin.* **26**, 24–26 (2020).
28. German National Cohort (GNC) Consortium, The German National Cohort: Aims, study des. *Eur. J. Epidemiol.* **29**, 371–382 (2014).
29. F. Bamberg, H.-U. Kauczor, S. Weckbach, C. L. Schlett, M. Forsting, S. C. Ladd, K. H. Greiser, M.-A. Weber, J. Schulz-Menger, T. Niendorf, T. Pischon, S. Caspers, K. Amunts, K. Berger, R. Bulow, N. Hosten, K. Hegenscheid, T. Kroncke, J. Linseisen, M. Gunther, J. G. Hirsch, A. Kohn, T. Hendel, H.-E. Wichmann, B. Schmidt, K.-H. Jockel, W. Hoffmann, R. Kaaks, M. F. Reiser, H. Volzke; German National Cohort MRI Study Investigators, Whole-body MR imaging in the german national cohort: Rationale, design, and technical background. *Radiology* **277**, 206–220 (2015).
30. H. Teismann, H. Wersching, M. Nagel, V. Arolt, W. Heindel, B. T. Baune, J. Wellmann, H. W. Hense, K. Berger, Establishing the bidirectional relationship between depression and subclinical arteriosclerosis - Rationale, design, and characteristics of the BiDirect Study. *BMC Psychiatry* **14**, 174 (2014).
31. T. Kircher, M. Wohr, I. Nenadic, R. Schwarting, G. Schratt, J. Alferink, C. Culmsee, H. Garn, T. Hahn, B. Muller-Myhsok, A. Dempfle, M. Hahmann, A. Jansen, P. Pfefferle, H. Renz,

M. Rietschel, S. H. Witt, M. Nothen, A. Krug, U. Dannlowski, Neurobiology of the major psychoses: A translational perspective on brain structure and function—the FOR2107 consortium. *Eur. Arch. Psychiatry Clin. Neurosci.* **269**, 949–962 (2019).
32. A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*. **2017-Decem**, 5575–5585 (2017).
33. H. Liu, Y.-S. Ong, X. Shen, J. Cai, When gaussian process meets big data: A review of scalable GPs. *Ieee T Neur Net Lear.* **31**, 4405–4423 (2020).
34. N. Tagasovska, D. Lopez-Paz, Single-Model Uncertainties for Deep Learning. arXiv:1811.00908 (2018).
35. Q. Xu, K. Deng, C. Jiang, F. Sun, X. Huang, Composite quantile regression neural network with applications. *Expert Syst. Appl.* **76**, 129–139 (2017).
36. R. Koenker, G. Bassett, Regression quantiles. *Econometrica* **46**, 33 (1978).
37. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python. *the. J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
38. N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**, 273–289 (2002).
39. U. Halekoh, S. Hjsgaard, A kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models-the R package pbkrtest. *J. Stat. Softw.* **59**, 1–32 (2014).
40. W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* **109**, 247–278 (2021).

Acknowledgments
Funding: This work was funded by the German Research Foundation (DFG grants HA7070/2-2, HA7070/3, and HA7070/4 to T.H.) and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grants Dan3/012/17 to U.D. and MzH 3/020/20 to T.H.). The analysis was conducted with data from the GNC (www.nako.de). The GNC is funded by the Federal Ministry of Education and Research (BMBF) (project funding reference numbers: 01ER1301A/B/C and 01ER1511D), the federal states, and the Helmholtz Association with additional financial support by the participating universities and the institutes of the Helmholtz Association and of the Leibniz Association. We thank all participants who took part in the GNC study and the staff in this research program. The BiDirect Study is supported by grants from the German Ministry of Research and Education (BMBF) to the University of Muenster (grant numbers 01ER0816 and 01ER1506). The MACS dataset used in this work is part of the German multicenter consortium "Neurobiology of Affective Disorders: A translational perspective on brain structure and function," funded by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG; Forschungsgruppe/Research Unit FOR2107). Principal investigators (PIs) with respective areas of responsibility in the FOR2107 consortium are as follows: Work Package WP1, FOR2107/MACS cohort and brain imaging: T. Kircher (speaker FOR2107; DFG grant numbers KI 588/14-1 and KI 588/14-2), U. Dannlowski (co-speaker FOR2107; DA 1151/5-1 and DA 1151/5-2), Axel Krug (KR 3822/5-1 and KR 3822/7-2), I. Nenadic (NE 2254/1-2), and C. Konrad (KO 4291/3-1). WP2, animal phenotyping: M. Wöhr (WO 1732/4-1 and WO 1732/4-2) and R. Schwarting (SCHW 559/14-1 and SCHW 559/14-2). WP3, miRNA: G. Schratt (SCHR 1136/3-1 and 1136/3-2). WP4, immunology, mitochondriae: J. Alferink (AL 1145/5-2), C. Culmsee (CU 43/9-1 and CU 43/9-2), and H. Garn (GA 545/5-1 and GA 545/7-2). WP5, genetics: M. Rietschel (RI 908/11-1 and RI 908/11-2), M. Nöthen (NO 246/10-1 and NO 246/10-2), and S. Witt (WI 3439/3-1 and WI 3439/3-2). WP6, multimethod data analytics: A. Jansen (JA 1890/7-1 and JA 1890/7-2), T. Hahn (HA 7070/2-2), B. Müller-Myhsok (MU1315/8-2), and A. Dempfle (DE 1614/3-1 and DE 1614/3-2). CP1, biobank: P. Pfefferle (PF 784/1-1 and PF 784/1-2) and H. Renz (RE 737/20-1 and 737/20-2). CP2, administration: T. Kircher (KI 588/15-1 and KI 588/17-1), U. Dannlowski (DA 1151/6-1), and C. Konrad (KO 4291/4-1). Data access and responsibility: All PIs take responsibility for the integrity of the respective study data and their components. All authors and coauthors had full access to all study data. The FOR2107 cohort project (WP1) was approved by the Ethics Committees of the Medical Faculties, University of Marburg (AZ: 07/14) and University of Münster (AZ: 2014-422-b-S). Financial support for the Beijing Normal University dataset used in this project was provided by a grant from the National Natural Science Foundation of China (30770594) and a grant from the National High Technology Program of China (863; 2008AA02Z405). We thank all study participants and staff at the GNC study centers, the data management center, and the GNC head office who enabled the conduction of the study and made the collection of all data possible. **Author contributions:** T.H. and J.E. conceived the idea, implemented the MCCQR method, analyzed the data, and wrote the paper. T.H., J.E., N.R.W., V.H., R.L., L.F., K.S., D.E., and B.R. developed and tested the MCCQR. N.R.W. and M.B. conceived and implemented the occlusion-based sensitivity mapping. T.H., J.E., N.R.W., C.G., and J.H.C. wrote the paper. K.S., V.H., and D.G. set up the preprocessing pipeline, preprocessed the data, and quality checked the results. N.O., R.R., J.R., D.G., S.M., T.K., H.K., and U.D. acquired, curated, quality checked, and provided the MACS data. U.D. and T.K. acquired funding, planned, and supervised the FOR2107 study as part of the MACS. J.G.H., T.N., B.E., F.B., T.K., R.B., H.V., O.v.S., R.F.S., L.U., B.S., S.C., and K.B. acquired, curated, quality checked, and provided the GNC data. K.B. acquired and provided the BiDirect data. T.H., J.E., U.D., K.B., C.G., and J.H.C. revised the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and the Supplementary Materials. In addition, IXI dataset is available from https://brain-development.org/ixi-dataset/. The GNC data can be provided by the GNC data transfer site based on a written data request, pending scientific review and a written cooperation agreement. Requests should be submitted to the GNC data transfer site (https://transfer.nako.de). The Beijing Normal dataset is available from http://fcon_1000.projects.nitrc.org/indi/retro/BeijingEnhanced.html. The MCCQR BrainAge model based on the GNC data as used in this study is available from https://photon-ai.com/model_repo/uncertainty-brain-age.

Submitted 8 February 2021
Accepted 11 November 2021
Published 5 January 2022
10.1126/sciadv.abg9471