Beyond brain age: Empirically-derived proxy measures of mental health

Kamalaker Dadi, Gael Varoquaux, Josselin Houenou, Danilo Bzdok, Bertrand Thirion, Denis Engemann

Beyond brain age: Empirically-derived proxy measures of mental health

Kamalaker Dadi[1], Gaël Varoquaux[1,4,6], Josselin Houenou[2,3], Danilo Bzdok[1,5,6], Bertrand Thirion[1], Denis Engemann[1,7*]

1 Inria, CEA, Neurospin, Parietal team, Univ. Paris Saclay, 91120 Palaiseau, France
2 CEA, NeuroSpin, Psychiatry Team, UNIACT Lab, Univ. Paris Saclay
3 APHP, Mondor University Hospitals, Psychiatry Dept, INSERM U955 Team 15 "Translational Psychiatry", Créteil, France
4 Montréal Neurological Institute, McGill University, Montreal, Canada
5 Department of Biomedical Engineering, Montreal Neurological Institute, Faculty of Medicine, McGill University, Montreal, Canada
6 Mila - Quebec Artificial Intelligence Institute, Canada
7 Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Germany

Correspondence:* denis-alexander.engemann@inria.fr

Abstract

Biological aging is revealed by physical measures, e.g., DNA probes or brain scans. Individual differences in personal functioning are instead explained by psychological constructs. Constructs such as intelligence or neuroticism are typically assessed by specialized workforce through tailored questionnaires and tests. Similar to how brain age captures biological aging, intelligence and neuroticism may provide empirical proxies for mental health. Could the combination of brain imaging and sociodemographic information yield measures for these constructs that do not rely on human judgment? Here, we built proxy measures by applying machine learning on multimodal MR images and rich sociodemographic information from the largest brain-imaging cohort to date: the UK Biobank. Objective comparisons revealed that all proxies captured the target constructs and related to health-contributing habits beyond the measures they were derived from. Our results demonstrate that proxies targeting classical psychological constructs reveal facets of mental health complementary to information conveyed by brain age.

Introduction

Individual assessments in psychology and psychiatry rely on observing behavior. Using biological insight to diagnose and treat mental disorders remains a hard problem despite substantial research efforts (Kapur et al., 2012). The field of psychiatry has struggled with purely descriptive and unstable diagnostic systems (Insel et al., 2010), small sample sizes (Szucs and Ioannidis, 2017), and reliance on dichotomized groups, *i.e.*, patients vs controls (Hozer and Houenou, 2016). Compared to somatic medicine, mental-health research faces the additional roadblock that mental pathologies cannot be measured the same way diabetes can be assessed through plasma levels of insulin or glucose. Psychological constructs, e.g., depressiveness or anxiety can only be probed indirectly through expert-built procedures such as specially-crafted questionnaires and structured interviews. Measuring reliably a given construct is difficult and questionnaires often remain the best option (Enkavi et al., 2019). While the field of psychometrics has thoroughly studied the validity of psychological constructs and their measures (Borsboom et al., 2004; Cronbach and Meehl, 1955; Eisenberg et al., 2019), the advent of new biophysical measurements on the brain brings new promises (Engemann et al., 2020; Kievit et al., 2018b; Nave et al., 2018). In particular, the growth of biobanks as well as the advances in statistical-learning techniques opens the door to large-scale validation of psychological constructs and measures for neuropsychiatric research (Collins, 2012).

In clinical neuroscience, machine learning is increasingly popular, driven by the hope to develop more generalizable models (Woo et al., 2017). Yet, to be reliable, machine learning needs large labeled datasets (Varoquaux, 2018). Its application to learn imaging biomarkers of neuropsychiatric disorders is limited by the availability of large cohorts with high-quality neuropsychiatric assessements (Bzdok and Meyer-Lindenberg, 2018). However, data on populations without diagnosed neuropsychiatric conditions is easier to collect. Such data has driven successes in developing brain-derived aging measures that capturing proxy information on mental health (Cole et al., 2015, 2018; Dosenbach et al., 2010; Engemann et al., 2020; He et al., 2020; Koutsouleris et al., 2014; Liem et al., 2017; Smith et al., 2020). Extrapolating from these successes, we propose to learn more of such *proxy measures* of health-related individual traits in large datasets. These could then enhance an analysis in a small dataset via links between the proxy measures and the actual clinical endpoint of interest, e.g., diagnosis or drug response. Emerging results validate the usefulness of age as one such proxy measure, leading to the so called *brain age delta*: the difference between predicted and actual age (Cole et al., 2015; Dosenbach et al., 2010; Smith et al., 2019a). The delta has been shown to reflect physical and cognitive impairment in adults and gives an index of neurodegenerative processes (Gonneaud et al., 2020; Liem et al., 2017). Can this strategy of biomarker-like proxy measures be extended beyond the construct of aging? Can measures derived from other targets than age serve as proxies for latent constructs?

Beyond aging, one high-stake target is intelligence, which is measured through socially administered tests and is one of the most extensively studied constructs in psychology. Fluid intelligence refers to the putatively culture-free, heritable and physiological component of intelligence (Cattell, 1963; Cattell and Scheier, 1961). Fluid intelligence is a latent construct designed to capture individual differences in cognitive capacity. It has been robustly associated with neuronal maturation and is typically reflected in cognitive-processing speed and working-memory capacity (Shelton et al., 2010). Compared to brain age, fluid intelligence may yield a proxy measure more specifically indexing cognitive function. It has been associated with psychiatric disorders such as psychosis, bipolar disorder and substance abuse (Keyes et al., 2017; Khandaker et al., 2018).

Neuroticism is a second promising target. As a key representative of the extensively studied Big Five personality inventory, neuroticism has a long-standing tradition in the psychology of individual differences (Costa and McCrae, 1992; Eysenck et al., 1985). Neuroticism is typically measured using self-assessment questionnaires and conceptualized as capturing dispositional negative emotionality including anxiety and depressiveness (Shackman et al., 2016). It has been inter-culturally validated (Cattell and Scheier, 1961; Lynn and Martin, 1997) and population-genetics studies have repeatedly linked variance in neuroticism to shared genes (Birley et al., 2006; Hettema et al., 2006; Pedersen et al., 1988). Neuroticism was shown useful in psychometric screening and supports predicting real-world behavior (Lahey, 2009; Tyrer et al., 2015). However, despite strong heritability at the population level (Power and Pluess, 2015; Vukasović and Bratko, 2015), the link with brain function at the level of large-scale network dynamics or the level of molecular mechanisms is being actively researched (Shackman et al., 2016; Yarkoni, 2015).

The advent of large MRI datasets has revealed the complexity of predicting personality traits from brain signals. Current attempts to predict fluid intelligence or neuroticism from thousands of MRI scans argue in favor of overwhelming heterogeneity and rather subtle effects that do not generalize strongly to unseen data (Dubois et al., 2018a,b). This stands in contrast to the remarkable performance obtained when predicting intelligence or neuroticism from other psychometric measures or semantic data qualitatively similar to psychometric questionnaires, e.g., Twitter and Facebook posts (Quercia et al., 2011; Youyou et al., 2015). As MRI acquisitions can be expensive and difficult in clinical settings or populations, the promises of social-media data is appealing. However, in clinical practice or research, such data can lead to measurement and selection biases difficult to control. On the other hand, background sociodemographic characteristics of individuals can be easily accessible and may help inform in similar ways on the heterogeneity of psychological traits, for instance capturing that fluid intelligence decreases with age (Horn et al., 1981). An important question is then whether this data can reveal non-redundant information on the constructs of interest.

Another challenge of quantifying psychological traits is the diversity of measurement scales, often categorical or on arbitrary non-physical units , e.g. education degree or monthly income. In fact, society treats individual differences as categorical or continuous, depending on the practical context. Personality has been proposed to span a continuum (Eysenck, 1958). Nevertheless, psychiatrists treat certain people as patients and not others (Perlis, 2011). The utility of any mental-health measure therefore depends on its practical context: When learning boundaries between qualitatively distinct groups, a measure that performs globally poorly as a continuous scale can nevertheless be sufficient to distinguish subgroups. In fact, a measure may be solely informative around the boundary region between certain classes, e.g., pilots who should fly and who should not. Importantly, the utility of any measure ultimately depends on its signal-to-noise ratio, which may be driven by measurement noise, heterogeneity, as well as the interesting variability of the particular construct measured, e.g., the type of test to assess intelligence.

Confronting the promises of population brain imaging with the challenges of measuring psychological traits raises the following questions. 1) How well can various health-related latent constructs be approximated from general-purpose inputs not designed to measure specific latent constructs? 2) What is the relative merit of brain imaging and sociodemographics for probing various latent constructs? 3) Can the success of brain age be extended to other proxy measures that capture complementary facets of health-contributing behavior? In this study, we tackled these questions by using machine learning to build *proxy measures*, crafted to approximate well-characterized *target measures* from brain-imaging and sociodemographic data. As target measures, we studied age, fluid intelligence, and neuroticism. Figure 1 summarizes our approach. We first assessed how well the proxy measures approximated the target measures, isolating the contributions of the different data types. Second, to assess the intrinsic value of the proxy measures, we studied their associations with health-related habits (alcohol consumption, cumulative tobacco consumption, sleep duration, physical activity). Results suggest that, as with brain age, proxy measures can bring value for the study of mental health that goes beyond approximating an available measure.

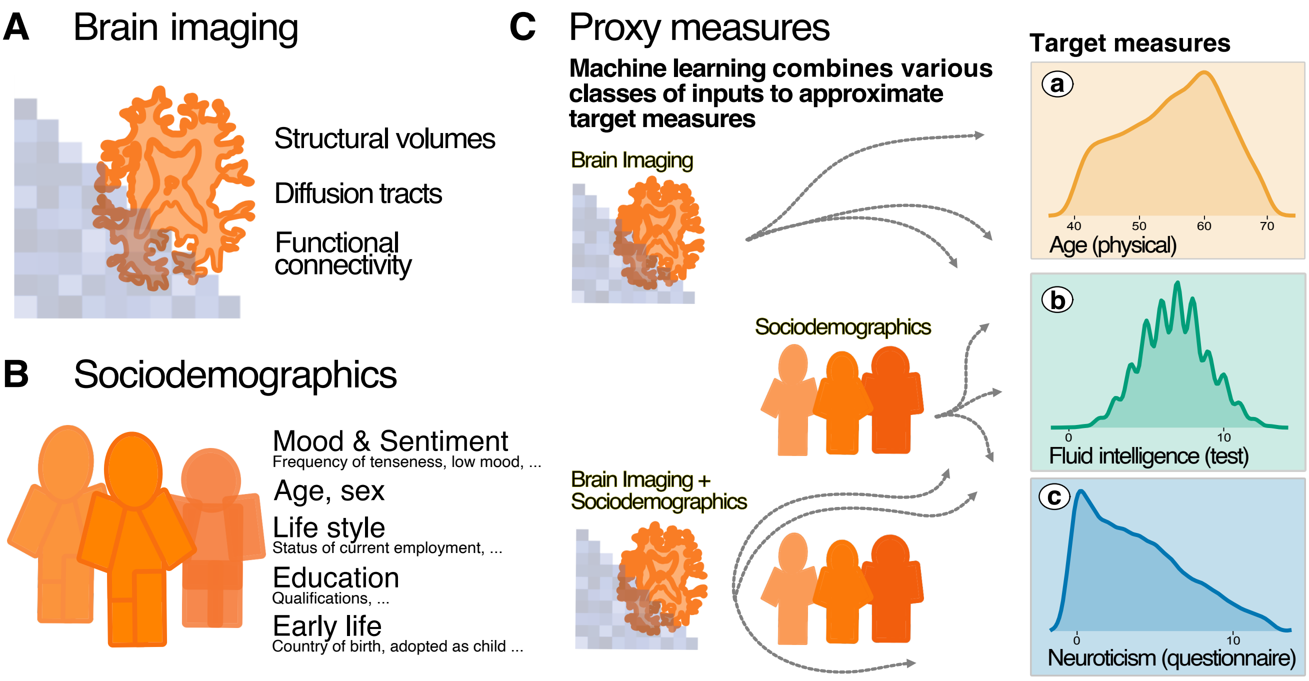
A Brain imaging
Structural volumes
Diffusion tracts
Functional connectivity
B Sociodemographics
Mood & Sentiment
Frequency of tenseness, low mood, ...
Age, sex
Life style
Status of current employment, ...
Education
Qualifications, ...
Early life
Country of birth, adopted as child ...
C Proxy measures
Machine learning combines various classes of inputs to approximate target measures
Brain Imaging
Sociodemographics
Brain Imaging + Sociodemographics
Target measures
a
40
50
60
70
Age (physical)
b
0
10
Fluid intelligence (test)
c
0
10
Neuroticism (questionnaire)

Figure 1. Methods overview: building and evaluating proxy measures We combined multiple brain-imaging modalities (**A**) with sociodemographic data (**B**) to approximate health-related biomedical and psychological constructs (**C**), *i.e.*, brain age (accessed through prediction of chronological age), cognitive capacity (accessed through a fluid-intelligence test) and the tendency to report negative emotions (accessed through a neuroticism questionnaire). We included the imaging data from the 10 000-subjects release of the UK biobank. Among imaging data (**A**) we considered features related to cortical and subcortical volumes, functional connectivity from rfMRI based on ICA networks, and white-matter molecular tracts from diffusive directions (see Table 3 for an overview about the multiple brain-imaging modalities). We then grouped the sociodemographic data (**B**) into five different blocks of variables related to self-reported mood & sentiment, primary demographics, lifestyle, education, and early-life events (Table 4 lists the number of variables in each block). Subsequently, we systematically compared the approximations of all three targets based on either brain images and sociodemographics in isolation or combined (**C**) to evaluate the relative contribution of these distinct inputs. Models were developed on 50% of the data (randomly drawn) based on random forest regression guided by Monte Carlo cross-validation with 100 splits (see section Model Development and Generalization Testing). We assessed generalization using the other 50% of the data as fully independent out-of-sample evaluations (see section Statistical Analysis).

Results

Traditional measures of mental health can be empirically approximated

We first performed model comparisons to evaluate the relative performance of proxy measures built from brain signals and distinct groups of sociodemographic variables. Figure 2 summarizes these model comparisons for approximating three targets: age, fluid intelligence and neuroticism. For the sociodemographic variables (Figure 2, dotted outlines), the analysis revealed that, for each target, there was one principal block of variables explaining most of the prediction performance. Combining all sociodemographic variables did not lead to obvious enhancements (Figure 2 – Figure supplement 2). For age prediction, variables related to current life-style showed by far the highest performance. For fluid intelligence, education performed by far best. Finally, for neuroticism, mood & sentiment clearly showed the strongest performance.

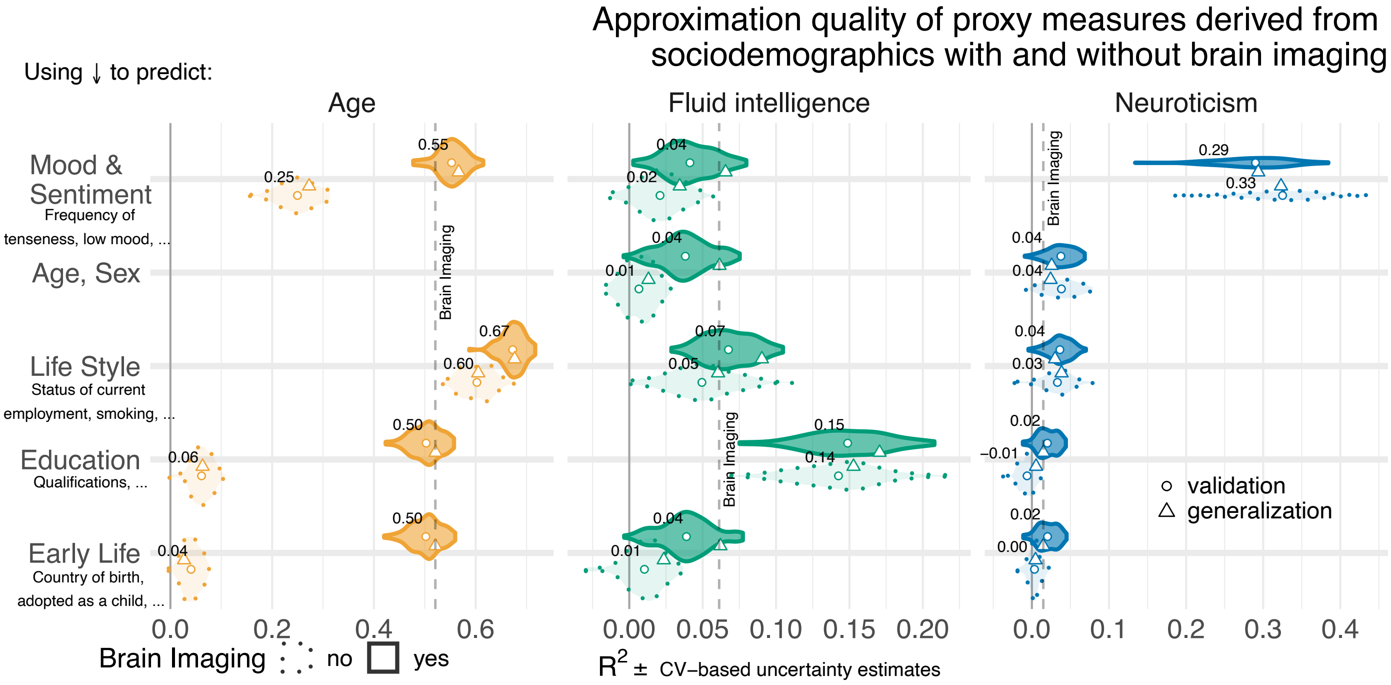

Figure 2. Approximation performance of proxy measures derived from sociodemographic data and MRI. To approximate age, fluid intelligence and neuroticism, we applied random-forest regression on sociodemographic data and brain images as inputs. The data was split into *validation data* for model construction (see section Model Development and Generalization Testing) and *generalization data* for statistical inference on out-of-sample predictions with independent data (see section Statistical Analysis). For each block of sociodemographic predictors models were fitted with and without additional predictors derived from brain images. We report the R^2 metric to facilitate comparisons across prediction targets. The cross-validation (CV) distribution (100 Monte Carlo splits) on the validation dataset is depicted by violins. Drawing style indicates whether brain imaging (solid outlines of violins) was included in addition or not (dotted outlines of violins). Dots depict the average performance on the validation data across CV-splits. Pyramids depict the performance of the average prediction (CV-bagging) on held-out generalization datasets. For convenience, the mean performance on the validation set is annotated for each plot. Vertical dotted lines indicate the average performance of the full MRI model. The validation and held-out datasets gave similar picture of approximation performance with no evidence for cross-validation bias Varoquaux et al. (2017a). One can readily see that approximation from sociodemographics (dotted violins) was often markedly better than purely brain-based models (dotted vertical lines) for all three targets. The most important blocks of sociodemographic predictors (annotated with red cross) were lifestyle for age, education for fluid intelligence, and mood & sentiment for neuroticism. The effect of combining sociodemographics with brain-data depended on the target measure. For age, overall performance improved beyond the purely sociodemographics-based or imaging-based analyses. The picture was less consistent for fluid intelligence and neuroticism showing weaker additive effects, if any. For the averaged out-of-sample predictions, the probability of the observed performance under the null-distribution and the uncertainty of effect sizes were formally probed using permutation tests and bootstrap-based confidence intervals (Table 1). Corresponding statistics for the baseline performance of models solely based on brain imaging (vertical dotted lines) are presented in Table S1. For additional findings please consider the supplement of Figure 2:

Figure 2 – Figure supplement 1. Prediction of individual differences in proxy measures from MRI.

Figure 2 – Figure supplement 2. Approximation performance using all sociodemographic data.

Table 1. Paired difference between purely sociodemographic and models including brain imaging on generalization data.

Target	sociodemographics	R^2_{diff}	p-value	CI_{low}	CI_{high}
Age	Early Life	0.494	0.0001	0.473	0.515
Age	Education	0.458	0.0001	0.437	0.479
Age	Life style	0.071	0.0001	0.058	0.085
Age	Mood & sentiment	0.294	0.0001	0.272	0.315
Fluid intelligence	Age, Sex	0.048	0.0001	0.040	0.057
Fluid intelligence	Early Life	0.039	0.0001	0.027	0.050
Fluid intelligence	Education	0.018	0.0001	0.010	0.025
Fluid intelligence	Life style	0.030	0.0001	0.020	0.040
Fluid intelligence	Mood & sentiment	0.031	0.0001	0.019	0.043
Neuroticism	Age, Sex	0.001	0.6789	−0.006	0.008
Neuroticism	Early Life	0.010	0.0697	−0.001	0.021
Neuroticism	Education	0.009	0.0817	−0.001	0.020
Neuroticism	Life style	−0.008	0.1750	−0.020	0.004
Neuroticism	Mood & sentiment	−0.030	0.0001	−0.041	−0.018

When combining MRI and sociodemographics (Figure 2, solid outlines), age prediction was enhanced in a systematic and visible way on all four blocks of variables (Table 1), suggesting that the observed differences should reproduce on future data and are unlikely to be due to chance. The benefit of including brain-imaging features, however, was less marked for prediction of fluid intelligence and neuroticism. With fluid intelligence, brain-imaging data improved the performance statistically significantly for all models, yet, with small effect sizes at the scale of a few percent or even lower (Table 1). Further, for neuroticism, no systematic advantage of including brain images alongside sociodemographics emerged. Instead, including brain images seemed to reduce generalization performance when predicting from mood & sentiment variables (Table 1, bottom row). Nevertheless, using only brain data was sufficient for statistically significant approximation of the target measures not only for age but also fluid intelligence and neuroticism (Table S1), suggesting that lifestyle and mood & sentiment explains at least some of the neurobiological variance. For neuroticism, variables on current mood & sentiment were strongly informative for prediction, reflecting that mood & sentiment is strongly related to neuroticism. Overall, predicting fluid intelligence or neuroticism was clearly more successful when sociodemographic was included (Table 1). For subsequent analyses we included all sociodemographic variables (Figure 2 – Figure supplement 2).

One important challenge with evaluating approximations of psychological measures is that such measures often come without physical scales and units (Stevens et al., 1946). In practice, clinicians and educators use them with specific thresholds for decision making. How useful proxy measures built with predictive models are to separate out discrete extreme groups? To address this question, we performed binary classification of extreme groups obtained from discretizing the targets using the 33_{rd} and 66_{th} percentiles. Moreover, we focused on the AUC as a performance metric which is only sensitive to ranking while ignoring the scale of the error. The results are comparable to the previous regression analysis. Classification performance for extreme groups visibly exceeded the chance level of an AUC of 0.5 for all models (Figure 3). Across proxy measures, models including sociodemographics performed best but the difference between purely sociodemographic and brain-based models was comparably weak, at the order of 0.01-0.02 AUC points (Table 2). Using only brain data resulted in proxy measures that perform less well, yet, still better than chance as revealed by permutation testing (Table S2). It is noteworthy that for both types of models the performance of discrimination reached levels above 0.8, which is considered clinically useful for biomarkers (Perlis, 2011). Overall, the results suggest that moving from the more difficult full-scale regression problem to extreme-group classification problem with purely ranking-based loss functions, the relative differences between brain-based and sociodemographics-based prediction gradually faded away.

Table 2. Difference statistics for classification on the held-out set for sociodemographic vs combined approximation.

Target	$AUC_{\text{diff observed}}$	p-value	CI_{low}	CI_{high}
Age	0.013	0.0008	0.006	0.021
Fluid intelligence	−0.031	0.0001	−0.044	−0.017
Neuroticism	−0.003	0.4818	−0.013	0.006

External validity: proxy measures capture ecological health-related factors

Results so far have shown that psychological constructs can be approximated from general-purpose inputs such as brain images and sociodemographic variables that are not tailored to measure these latent constructs. Beyond approximating target measures, which are themselves imperfect, can our empirically-derived proxy measures capture complementary facets of real-world behavior? To address this question we studied the link between the three proxy measures studied –built via brain age, fluid intelligence and neuroticism– and various health behavior (sleep, physical exercise, alcohol and tobacco consumption). These behaviors are more ecological probes of mental health than questionnaires or lab-based measures and are potentially linked in multiple ways to our proxy measures. We, hence, modeled them as weighted sums of predicted brain-age delta, fluid intelligence and neuroticism using multiple linear regression. To avoid any form of circularity, we used the out-of-sample predictions for all three proxy measures, applied on the generalization dataset that was not used for building the machine learning models. We derived the brain-age delta by subtracting the actual age from the predicted age. To mitigate brain age bias (Le et al., 2018), we deconfounded health-related habits for their association with actual age (Engemann et al., 2020; Smith

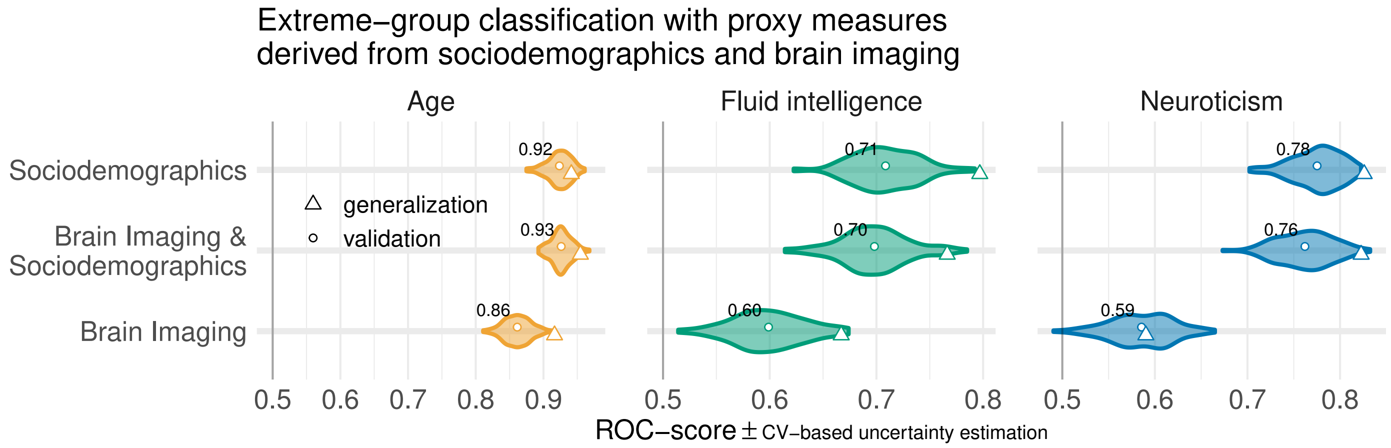

Figure 3. Classification analysis from imaging, sociodemographics and combination of both data. For classification of extreme groups instead of continuous regression, we split the data into low vs high groups based on 33_{rd} and 66_{th} percentiles. Visual conventions follow Figure 2. We report the accuracy in AUC. Models including sociodemographics performed visibly better than models purely based on brain imaging. Differences between brain-imaging and sociodemographics appeared less pronounced as compared to the fully-fledged regression analysis. For the average out-of-sample predictions, the probability of the observed performance under the null-distribution and the uncertainty of effect sizes were formally probed using permutation tests and bootstrap-based confidence intervals (Table 2). Corresponding statistics for the baseline performance of models solely based on brain imaging (vertical dotted lines) are presented in Table S2.

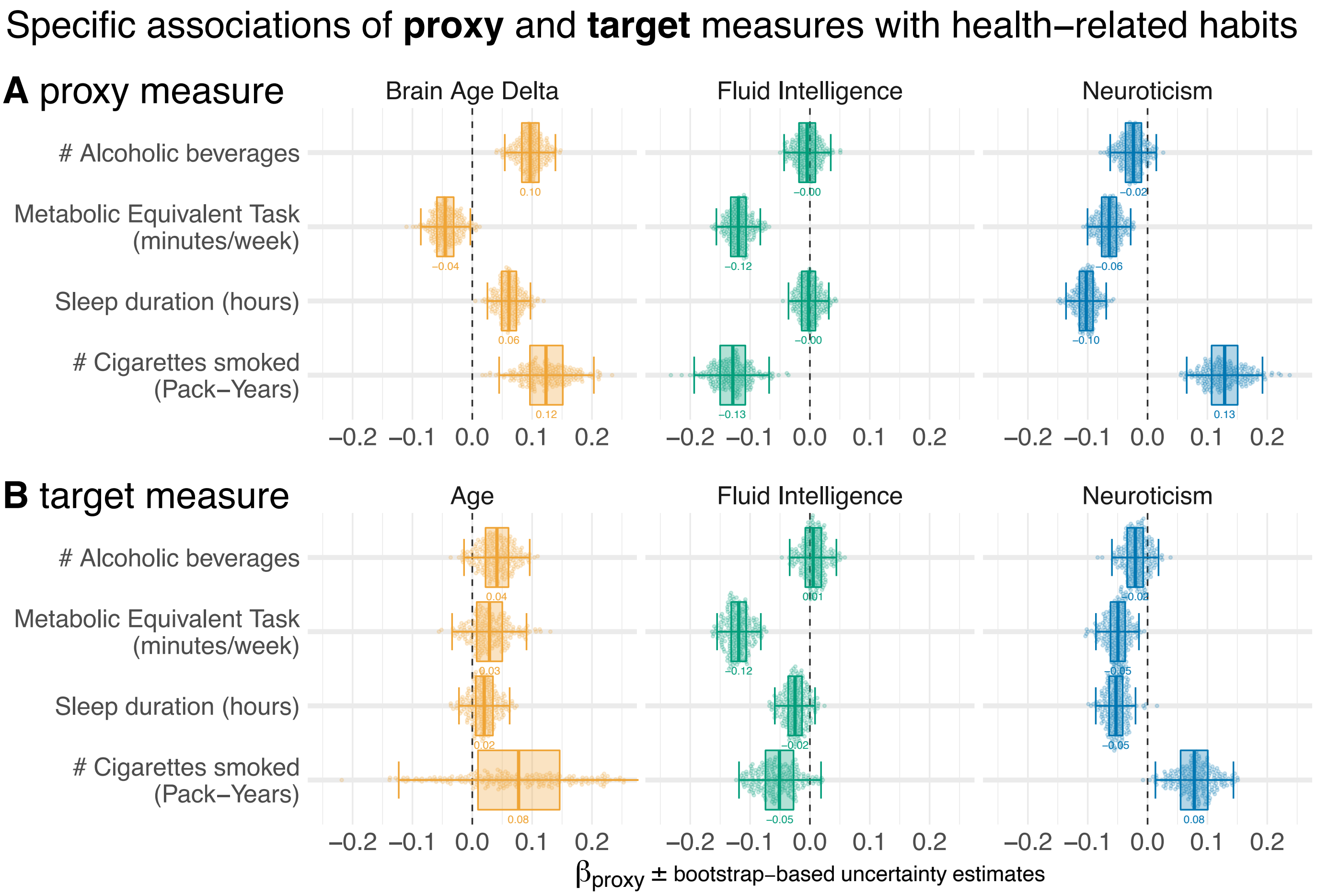

Figure 4. Proxy measures show systematic and complementary out-of-sample associations with health-related habits. To probe external validity of the proxy measures, we investigated their out-of-sample associations with ecological indicators of mental health (sleep duration, time spent with physical exercise, number of alcoholic beverages and cigarettes consumed). To tease apart complementary and redundant effects, we constructed multiple linear regression models on out-of-sample predictions combining the proxy measures **(A)** from Figure 2. For comparison, we repeated the analysis using the actual target measures **(B)** observed on the generalization data. Regression models are depicted rows-wise. Box plots summarize the uncertainty distribution of target-specific (color) regression coefficients with whiskers indicating two-sided 95% uncertainty intervals (parametric bootstrap). A random subset of 200 out of 10000 coefficient-draws is illustrated by dots. The average coefficient estimate is annotated for convenience. At least two distinct patterns emerged: either the health outcome was specifically associated with one proxy measures (brain age delta and number of alcoholic beverages) or multiple measures showed additive associations with the outcome (*e.g.* number of pack years smoked). Finally, target measures **(B)** show noisier associations than proxy measures **(A)**, though none of the significant associations changed direction. For additional findings, please consider the supplement of Figure 4:
Figure 4 – Figure supplement 1. Marginal associations between proxy measures and health-related habits.

et al., 2019a).

The estimated regression coefficients, capturing partial correlations, revealed specific as well as complementary associations between the proxy measures and health-related behavior (Figure 4). A marginal association analysis shows similar patterns, indicating that the relationships hold also when considering the proxy measures in isolation (Figure 4 – Figure supplement 1). Elevated brain-age delta was consistently associated with increased number of alcoholic beverages. These latter proxy measures showed no consistent association with alcohol consumption (Figure 4, first row). Level of physical exercise –measured through the number of minutes spent weekly with metabolic equivalent tasks– consistently associated with the scores in all three predicted targets, suggesting independent associations (Figure 4, second row). This may seem counter-intuitive but could simply point at the possibility that people with higher test scores, as a tendency, have a more sedentary life style. Sleep duration was independently associated with brain age delta and predicted neuroticism but in opposite directions (Figure 4, third row): increased sleep duration consistently went along with elevated brain age, but lower levels of predicted neuroticism. No consistent effect emerged for fluid intelligence. Increased cumulative numbers of cigarettes smoked was independently and consistently associated with all predicted targets (Figure 4, last row): Intensified smoking went along with elevated brain age delta and elevated neuroticism but lower fluid intelligence.

The question remains whether the proxy measures bring any additional value compared to the original target measures that they were derived from. Studying the association of these original target measures with the health-related habits shows similar trends: associations with the same signs as with the proxy measures (Figure 4, B). However, these associations were more noisy or less marked as those seen with the proxy measures.

These results demonstrates that the proxy measures capture well health-related habits, potentially better than the original target measures, and in a complementary way across the three measures.

Discussion

In this study, we have extended the brain-age approach for neuroimaging to the wider notion of empirical proxy measures. Guided by machine learning, we have derived empirical approximations of traditional, extensively validated target measures from psychology. Beyond biological age, we focused on cognitive capacity (accessed by the fluid-intelligence test) and negative emotionality (accessed by the neuroticism questionnaire). Our proxy measures were derived from data not explicitly designed to assess specific latent constructs: brain imaging data and heterogeneous sociodemographic descriptors. We observed that the combination of brain imaging and target-specific sociodemographic inputs often improved approximation performance. On the held-out data that was not used for model construction, we found important associations between all proxy measures and ecological health indicators. These associations were often complementary and useful beyond the information conveyed by the approximated targets.

Constructs of mental-health can be accessed from general-purpose data

Brain age has served as landmark in this study, both conceptually and empirically. It has been arguably the most discussed candidate for a surrogate biomarker in the neuroimaging literature so far (Cole et al., 2015; Dosenbach et al., 2010; Smith et al., 2019a). With mean absolute errors around 4 years, up to 67% variance explained, and AUC-scores up to 0.93 in the classification setting, our results compare favorably to the recent brain-age literature within the UK Biobank (Cole et al., 2017; Smith et al., 2020) and in other datasets (Engemann et al., 2020; Liem et al., 2017), though we relied on non-optimized standard inputs and algorithms and not deep learning (He et al., 2018). Applying the same approach to other behavioral outcomes that probe psychological constructs, namely fluid intelligence and neuroticism, we found that these were considerably harder to approximate from general brain imaging data or sociodemographic descriptors.

It is important to recapitulate that approximation quality on the three targets investigated has a different meaning, as these are measured differently. On the one hand, age is a physical variable measured with meaningful units (years) on a ratio scale (Stevens et al., 1946) (Selma is *twice as old* as Bob). On the other hand, psychometric scores such as fluid intelligence –measured via socially-administered performance tests– and neuroticism –measured by self-assessment via questionnaires– are unit-free scores resulting from operationalized counting, which provokes ambiguity regarding the level of measurement (Borsboom, 2005). Their implied scales may be considered as interval (the *difference between* Bob's and Selma's intelligence is -0.1 standard deviations) if not rather ordinal (Bob's intelligence was *ranked below* Selma's) (Stevens et al., 1946). In day-to-day psychological practice, these scores are often used via practically-defined thresholds, *e.g.* school admission or pilot candidate selection in aviation (Carretta, 2011; Carretta and Ree, 1994). Approximations of these measures via empirically-defined proxies should thus be subjected to different standards: Brain-age prediction should be gauged accordingly to its natural continuous scale; we observed more than 50% of the variance explained. Instead, approximation of psychometric scores might be more appropriately gauged via implicit thresholds, hence, discrimination tasks. With the corresponding metrics, the receiver-operator characteristics (ROC) and its AUC-score, all proxy measures approached or exceeded a performance of 0.80 deemed relevant in biomarker development (Perlis, 2011), though to be fair, they approximated established psychometric targets (proxy measures themselves) and not a medical condition.

Nevertheless, the out-of-sample associations of the approximated constructs –the proxy measures– with health-related habits (Figure 4) paint a more complete picture of their value. Sleep duration, minutes spent exercising, and the amount of alcoholic drinks or cigarettes consumed were specifically and complementarily associated with all proxy measures on more than 4000 held-out individuals. In other words, we found multiple statistically important associations with proxy measures fluid intelligence and neuroticism that were not accounted for by brain age. Compared to the traditional measures (Figure 4 **B**), the associations between these proxy measures and ecological behavioral traits were less noisy, hence more consistent, regardless of their approximation quality (Figure 4 **A**). This may seem surprising at first, but the target measures are themselves noisy and of imperfect ecological validity. Conversely, the proxy measures are assembled via a richer phenotyping than the target measures, drawing from both fine sociodemographics and brain signals, which can help refining them.

The benefits offered by brain data depend on the approximated construct

All brain-derived approximations were statistically meaningful. Yet, only for age prediction, imaging data by itself led to convincing performance levels. Combining brain-imaging data to sociodemographics led to systematically enhanced performance for predicting age and, less strongly, fluid intelligence (Table 1). On the other hand, for neuroticism, including brain imaging never substantially improved the approximation. Does this mean that brain imaging could be avoided in practice when approximating latent constructs? Such a view is probably misleading as the numerical quality of the approximation is not the only thing that matters in a proxy measure. The interest in building a proxy measure of age from brain imaging is justified by its interpretation as an index of precocious or accelerated biological aging (Cole et al., 2015, 2017; Smith et al., 2020). In contrast, it is not yet clear that an age delta built from sociodemographic inputs –along the lines of a "social age"– supports such interpretation. From this point of view one may even prefer purely brain-based assessment of individual aging, though sociodemographics probably provide important context to the brain images.

For fluid intelligence and neuroticism the situation seemed more complex. For both targets, the best performing sociodemographic model was based on inputs semantically close to the construct of interest, *i.e.*, education details for fluid intelligence and mood & sentiment for neuroticism. While those results reinforce the construct validity of the measure, they also come with a certain risk of circularity. In particular, the causal role of those predictors is not necessarily clear as better educational attainment is heritable itself (Krapohl et al., 2014) and may reinforce existing cognitive abilities rather than simply resulting from them. Similarly, prolonged emotional stress due to life events may exacerbate existing dispositions to experience negative emotions captured by neuroticism (Colodro-Conde et al., 2018), traits which in turn commonly help accumulate stressful life events (Lahey, 2009). Nevertheless, for fluid intelligence but not neuroticism, brain imaging added incremental value when combined with various sociodemographic predictors. This may suggest that the cues for neuroticism conveyed by brain imaging were already present in various sociodemographic predictors, potentially hinting at common causes.

It may be worthwhile to revisit the frequently reported difficulty to predict complex traits from brain imaging– especially fMRI (Dubois et al., 2018a,b; Liem et al., 2017; Maglanoc et al., 2020). This may not be entirely surprising at a theoretical level as it has even been argued that psychometric measures of complex traits may not map to biological mechanisms in simple ways (Yarkoni, 2015). Of course, this does not preclude the investigation of their brain correlates and mechanisms (Cole et al., 2015; Cox et al., 2019a; Kievit et al., 2018a; Shackman et al., 2016). It rather emphasizes the importance of searching for appropriate signals and representations supporting the given modeling goals (Bzdok and Ioannidis, 2019). As a speculation, some traits could be tightly linked to the current predominant behavior that may be poorly reflected by resting-state recordings. To consider an *extreme counter example*, disorders of consciousness— a stable trait induced by severe brain injuries— manifest themselves in systematically and intensely altered brain activity, hence, can be robustly detected from fMRI- and EEG-signals regardless of the present stimulation (Demertzi et al., 2019; Engemann et al., 2018). In this context, the recent turn towards naturalistic stimuli and movies (Hasson et al., 2010; Jääskeläinen et al., 2016; Nummenmaa et al., 2012; Sonkusare et al., 2019; Venkatesh et al., 2020) may be promising as trait-level differences in emotion and cognition may need to be systematically provoked by potent stimuli, *e.g.*, emotionally charging or cognitively demanding cinematic content.

Empirically-derived proxy measures: From validity to practical utility

The validity of constructs and their measures remains a challenging question (Borsboom, 2005; Borsboom et al., 2004; Cronbach and Meehl, 1955). Here, we have demonstrated reasonable out-of-sample generalization for our proxy measures. Yet, generalization performance in itself, arguably, only yields an upper bound for validity of the measure for a target construct, comparable internal-consistency checks and re-test reliability in classical psychometrics. Even a perfect approximation may be limited by the quality of the target measure as fluid intelligence and neuroticism are notoriously difficult to measure without noise. In our study, the construct validity of the corresponding proxy measures is supported by the substantial gain in prediction performance brought by related information, namely education history and mental-health variables respectively (Figure 2). Moreover, association with health-relevant habits brings external validity to the proxy (Figure 4). For example, the complementary patterns that emerged can be related to traditional construct semantics: High consumption of cigarettes is typically associated with neuroticism (Terracciano and Costa Jr, 2004) and excessive drinking may lead to brain atrophy and cognitive decline (Topiwala et al., 2017), both common correlates of brain age (Liem et al., 2017; Wang et al., 2019).

This raises the question of the practical utility of such empirically-derived proxy measures: Can these empirically-derived proxy measures substitute specific psychometric instruments? The present study does not claim to give an unequivocal answer to this question as the utility of proxy measures will depend on the practical context. A specialized mental-health professional may prefer an established routine for clinical assessment, relying on scores such as intelligence tests and personality-scales like neuroticism, and potentially applying implicit experience-based thresholds. Based on our findings, inclusion of brain imaging may even seem to yield diminishing returns when approximating high-level psychological traits. Yet, it mays simply be a matter of time until more effective acquisition protocols will be discovered alongside signal representations supporting predictive modeling. While the cost of including brain imaging may seem exorbitant, whenever available, its inclusion seems to be a “safe bet” as machine learning is capable at selecting relevant inputs (Engemann et al., 2020) and costs of MRI-acquisition can be amortized by baseline clinical usage. Moreover, our study shows that the associations of the proxy measures to health habits compare favorably to the original target measures. As such, the proxy measures may open new doors when tailored assessment of latent constructs is not applicable to due lack of specialized mental-health workforce or sheer cost. For instance, they may bring mental-health assessment in research endeavors on large populations, *e.g.*, for etiology, nosology, or typical epidemiology questions such as risk factors or treatment evaluation. In addition, results derived on large populations can be transferred to clinical data with finer mental-health assessment, *e.g.*, smaller cohorts, possibly leveraging dedicated methods (He et al., 2020; Pan and Yang, 2009). Relying on three proxy measures rather than the brain age alone promises a wider array of applications.

Limitations

This study has validated proxy measures of three target constructs. The selection of these targets was guided by literature review as well as the goal to find representative health-related measures with complementary semantics. Additional constructs and psychometric tools could have been visited. Intelligence can be characterized by multiple facets. The broader construct of intelligence as a general factor –g-factor– is often estimated using latent factor models on multiple correlated tests. While g-factor modeling can be interesting for its own sake, we are less interested in normative assessment of intelligence but rather in capturing inter-individual variance related to cognitive capacity as a situational fitness signal. Such variations have been repeatedly linked to mental-health conditions (Khandaker et al., 2018). Likewise, there is a wealth of questionnaires designed to measure negative emotionality and neuroticism specifically. Yet, we could study only that available in the UK-Biobank data, the EPQ neuroticism scale. A complementary approach, leading to different scientific questions, would be to estimate latent factors by pooling all non-imaging data semantically related to neuroticism (Maglanoc et al., 2020). Rather, we chose to consider established target measures “as is” instead of derivatives to avoid bringing in additional measure-validity considerations. Nevertheless, our framework encourages future studies targeting more sophisticated representations of latent constructs.

Second, while the study was clinically motivated, it falls short of directly testing the clinical relevance of estimated proxy measures. Indeed, even in a very large general-population cohort such as the UK Biobank, there are only a few hundred diagnosed cases of mental disorders (ICD-10 mental-health diagnoses from the F chapter) with brain-imaging data available. This challenge highlights the practical importance of studying mental as a continuous, in addition to diagnosed conditions. In this direction, our analysis of health-related habits does provide some clinical relevance.

Finally, our study falls short of presenting fine-grained spatial analysis of the imaging data. This work has focused on the approximation quality of proxy measures, relying on methods that are not designed for fine-grained inference on predictors (Bzdok et al., 2018), though future work could explore post-hoc explanations (Biecek, 2018). Our analysis comparing the quality of models helps isolating major explanatory factors, yet does not provide brain mapping (Cole, 2020; Cox et al., 2019a; Kievit et al., 2018a).

Conclusion

Empirically-derived proxy measures targeting age, fluid intelligence and neuroticism reveal complementary facets of real-world behavior that contribute to maintaining mental health. As the relative importance of brain imaging and sociodemographics varies with the approximated target, we recommend generously including all available data and approximating as many targets as possible while letting machine learning perform the labor of integration. We believe that further developing and using proxy measures for constructs that are difficult to assess is a promising agenda for mental-health research. Therefore, we have made all data analysis and visualization source code available on Github: https://github.com/KamalakerDadi/proxy_measures_2020.

Materials and Methods

Dataset

The United Kingdom Biobank (UKBB) database is to date the most extensive large-scale cohort aimed at studying the determinants of the health outcomes in the general adult population. The UKBB is openly accessible and has extensive data acquired on 500 000 individuals aged 40-70 years covering rich phenotypes, health-related information, brain-imaging and genetic data (Collins, 2012). Participants were invited for repeated assessments, some of which included MR imaging. For instance, cognitive tests that were administered during an initial assessment were also assessed during the follow-up visits. This has enabled finding for many subjects at least one visit containing all heterogeneous input data needed to develop the proposed proxy measures. The study was conducted using the UKBB Resource Applixaction 23827.

Participants

All participants gave informed consent. The UKBB study was examined and approved by the North West Multi-centre Research Ethics Committee. We considered participants who have responded to cognitive tests, questionnaires, and have access to their primary demographics and brain images (Sudlow et al., 2015). Out of the total size of UKBB populations, we found 11 175 participants who had repeated assessments overlapping with the first brain imaging release (Miller et al., 2016). The demographics are 51.6% female (5 572) and 48.3% male (5 403), and an age range between 40-70 years (with a mean of 55 years and standard deviation of 7.5 years). Out of the complete analysis set, 5 587 individuals were used in the study to train the model and remaining subjects were set aside as a held-out set for generalization testing (see section *Model development and generalization testing*).

To establish specific comparisons between models based on sociodemographics, brain data or their combinations we exclusively considered the cases for which MRI scans were available. The final sample sizes used for model construction and generalization testing then depended on the availability of MRI: For age and fluid intelligence, our random splitting procedure (*Model development and generalization testing*) yielded 4203 cases for model building and 4157 for generalization. For cases with valid neuroticism assessment, fewer brain images were available, which yielded 3550 cases for model building and 3509 for generalization.

Data acquisition

Sociodemographic data (non-imaging) was collected with self-report measures administered through touchscreen questionnaires, complemented by verbal interviews, physical measures, biological sampling and imaging data. MRI data were acquired with the Siemens Skyra 3T using a standard Siemens 32-channel RF receiver head coil (Alfaro-Almagro et al., 2018). We considered three MR imaging modalities as each of them potentially captures unique neurobiological details: structural MRI (sMRI/T1), resting-state functional MRI (rs-fMRI) and diffusion MRI (dMRI). For technical details about the MR acquisition parameters, please refer to Miller et al. (2016). We used image-derived phenotypes (IDPs) of those distinct brain-imaging modalities, as they provide actionable summaries of the brain measurements and encourage comparability across studies.

Target measures

As our target measures for brain age modelign, we use an individual's age at baseline recruitment (UKBB code "21022-0.0"). Fluid intelligence, was assessed using a cognitive battery designed to measure an individual's capacity to solve novel problems that require logic and abstract reasoning. In the UK Biobank, the fluid intelligence test (UKBB code "20016-2.0") comprises thirteen logic and reasoning questions that were administered via the touchscreen to record a response within two minutes for each question. Therefore, each correct answer is scored as one point with 13 points in total[1]. Neuroticism (UKBB code "20127-0.0") was measured using a shorter version of the revised Eysenck Personality Questionnaire (EPQ-N) comprised of 12-items (Eysenck et al., 1985). Neuroticism was assessed during Biobank's baseline visit. The summary of the individual's scores ranges from 0 to 12 that assess dispositional tendency to experience negative emotions [2].

Sociodemographic data

In this work, we refer to non-imaging variables broadly as sociodemographics excluding the candidate targets fluid intelligence and neuroticism. To approximate latent constructs from sociodemographics, we included 86 non-imaging inputs (Table S3) which are the collection of variables reflecting each participant's demographic and social factors *i.e.*, sex, age, date and month of birth, body mass index, ethnicity, exposures at early life –*e.g.* breast feeding, maternal smoking around birth, adopted as a child– education, lifestyle-related variables –*e.g.* occupation, household family income, household people living at the same place, smoking habits–, and mental-health variables. All these data were self-reported. We then assigned these 86 variables to five groups based on their relationships. Based on our conceptual understanding of the variables, we name assigned them to one out of five groups: **1)** mood & sentiment, **2)** primary demographics as age, sex, **3)** lifestyle, **4)** education, **5)** early life. We then investigated the intercorrelation between all 86 variables to ensure that the proposed grouping is compatible with their empirical correlation structure Figure S1.

The sociodemographic groups had varying amounts of missing data. For *e.g.* the source of missingness is concerned with the participants lifestyle habits such as smoking and mental health issues (Fry et al., 2017). To deal with this missingness in the data using imputation (Little and Rubin, 1986), we used column-wise replacement of missing information with the median value calculated from the known part of the variable. We subsequently included

[1] A complete overview of the 13 individual fluid intelligence items can be seen from this manual https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/Fluidintelligence.pdf

[2] For a complete list of Neuroticism questionnaires can be seen from this manual https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/MentalStatesDerivation.pdf

an indicator for the presence of imputed for down-stream analysis. Such imputation is well suited to predictive models (Josse et al., 2019).

Image processing to derive phenotypes for machine learning

MRI data preprocessing were carried out by UKBB imaging team. The full technical details are described elsewhere (Alfaro-Almagro et al., 2018; Miller et al., 2016). Below, we describe briefly the custom processing steps that we used on top of the already preprocessed inputs.

Structural MRI

This type of data analysis on T1-weighted brain images are concerned with morphometry of the gray matter areas *i.e.* the quantification of size, volume of brain structures and tissue types and their variations under neuropathologies or behavior (Lerch et al., 2017). For example, volume changes in gray matter areas over lifetime are associated with: brain aging (Ritchie et al., 2015), general intelligence (Cox et al., 2019b) and brain disease (Thompson et al., 2007). Such volumes are calculated within pre-defined ROIs composed of cortical and sub-cortical structures (Desikan et al., 2006) and cerebellar regions (Diedrichsen et al., 2009). We included 157 sMRI features consisting of volume of total brain and grey matter along with brain subcortical structures[3]. All these features are pre-extracted by UKBB brain imaging team (Miller et al., 2016) and are part of data download. We concatenated all inputs alongside custom-built fMRI features for predictive analysis (feature union).

Diffusion weighted MRI

Diffusion MRI enables to identify white matter tracts along principal diffusive direction of water molecules, as well as the connections between different gray matter areas (Behrens et al., 2003; Conturo et al., 1999). The study of these local anatomical connections through white matter are relevant to the understanding of neuropathologies and functional organization (Saygin et al., 2016). We included 432 dMRI skeleton features of FA (fractional anisotropy), MO (tensor mode) and MD (mean diffusivity), ICVF (intra-cellular volume fraction), ISOVF (isotropic volume fraction) and OD (orientation dispersion index) modeled on many brain white matter structures extracted from neuroanatomy[4]. For extensive technical details, please refer to de Groot et al. (2013). The skeleton features we included were from category134 shipped by the UKBB brain-imaging team and we used them without modification.

Functional MRI

Resting-state functional MR images capture low-frequency fluctuations in blood oxygenation that can reveal ongoing neuronal interactions in time forming distinct brain networks (Biswal et al., 1995). Functional connectivity within these brain network can be linked to clinical status (Greicius et al., 2004), to behavior (Miller et al., 2016), or to psychological traits (Dubois et al., 2018b). We also included resting-state connectivity features based on the time-series extracted from Independent Component Analysis (ICA) with 55 components representing various brain networks extracted on UKBB rfMRI data (Miller et al., 2016). These included the default mode network, extended default mode network and cingulo-opercular network, executive control and attention network, visual network, and sensorimotor network. We

[3] Regional grey matter volumes `http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=1101` Subcortical volumes `http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=1102`

[4] Diffusion-MRI skeleton measurements `http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=134`

Table 3. Imaging-based models.

Index	Name	# variables	# groups
1	brain volumes (sMRI)	157	1
2	white matter (dMRI)	432	1
3	functional connectivity (fMRI)	1485	1
4	sMRI, dMRI	589	2
5	sMRI, fMRI	1642	2
6	dMRI, fMRI	1917	2
7	sMRI, dMRI, fMRI (full MRI)	2074	3

Table 4. Non-imaging baseline models or sociodemographic models based on single group. Variables in each group are described at corresponding section: Sociodemographic data.

Index	Name	# variables
1	Mood & Sentiment (MS)	25
2	Age, Sex (AS)	5
3	Life style (LS)	45
4	Education (EDU)	2
5	Early Life (EL)	9

measured functional connectivity in terms of the between-network covariance. We estimated the covariance matrices using Ledoit-Wolf shrinkage (Ledoit and Wolf, 2004). To account for the fact that covariance matrices live on a particular manifold, *i.e.*, a curved non-Euclidean space, we used the tangent-space embedding to transform the matrices into a Euclidean space (Sabbagh et al., 2019; Varoquaux et al., 2010) following recent recommendations (Dadi et al., 2019; Pervaiz et al., 2020). For predictive modeling, we then vectorized the covariance matrices to 1 485 features by taking the lower triangular part. These steps were performed with `NiLearn` (Abraham et al., 2014).

Comparing predictive models to approximate target measures

Imaging-based models

First, we focused on purely imaging-based models based on exhaustive combinations of the three types of MRI modalities (see Table 3 for an overview). This allowed us to study potential overlap and complementarity between the MRI-modalities. Preliminary analyses revealed that combining all MRI data gave reasonable results with no evident disadvantage over particular combinations of MRI modalities (Figure 2 – Figure supplement 1), hence, for simplicity, we only focused on the full MRI model in subsequent analyses.

Sociodemographic models

We composed predictive models based on non-exhaustive combinations of different types of sociodemographic variables. To investigate the relative importance of each class of sociodemographic inputs, we performed systematic model comparisons. We were particularly interested in studying the relative contributions of early-life factors as compared to factors related to more recent life events such as education as well as factors related to current circumstances such as mood & sentiment and life-style. The resulting models based on distinct groups of predictors are listed in Table 4 (for additional details see Table S3 and Figure S1).

Table 5. Random forest hyperparameters and tuning with grid search (5 fold cross-validation).

Hyperparameter	Values
Impurity criterion	Mean squared error
Maximum tree depth	5, 10, 20, 40, full depth
Fraction of features for split	1, 5, “log2”, “sqrt”, “complete”
Number of trees	250

Table 6. Number of samples for classification analysis (N).

# groups	Age	Fluid intelligence	Neuroticism
1	1335	1108	1054
2	1200	898	1020

Combined imaging and sociodemographic models

In the next step, we were interested in how brain-related information would interact within each of these sociodemographic models. For example, information such as the age of an individual, or the level of education, may add important contextual information to brain images. We therefore considered an alternative variant for each of the models in Table 4 that included all MRI-related features (2 074 additional features) as described at section Image processing to derive phenotypes for machine learning.

Predictive model

Linear models are recommended as default choice in neuroimaging research (Dadi et al., 2019; Poldrack et al., 2020) especially when datasets include fewer than 1000 data points. In this study approximated targets generated by distinct underlying mechanisms based on multiple classes of heterogenous input data with several thousands of data points. We hence chose the non-parametric random forest algorithm that can be readily applied on data of different units for non-linear regression and classification (Breiman, 2001) with mean squared error as impurity criterion. To improve computation time we fixed tree-depth to 250 trees, a hyper-parameter that is not usually not tuned but set to a generous number as performance plateaus beyond a certain number of trees (Hastie et al., 2005, ch. 15). Preliminary analyses suggested that additional trees would not have led to substantial improvements in performance. We used nested cross-validation (5-fold grid search) to tune the depth of the trees as well as the number of variables considered for splitting (see Table 5 for a full list of hyper-parameters considered).

Classification analysis. We also performed classification analysis on the continuous targets. For this purpose, we discretized the targets into extreme groups based on the 33_{rd} and 66_{th} percentiles (see Table 6 for the number of classification samples per group). We were particularly interested in understanding whether model performance would increase when moving toward classifying extreme groups. For this analysis, we considered all three types of models (full MRI 2074 features from imaging-based models see section Imaging-based models, all sociodemographics variables, total 86 variables see section Sociodemographic models), combination of full MRI and all sociodemographics, a total 2160 variables see section Combined imaging and sociodemographic models. When predicting age, we excluded the age & sex sociodemographic block from all sociodemographic variables which then yielded a total of 81 variables. To assess the performance for classification analysis, we used the area under the curve (AUC) of the receiver operator characteristic (ROC) as an evaluation metric (Poldrack et al., 2020).

Model development and generalization testing

Before any empirical work, we generated two random partitions of the data, one validation dataset for model construction and one held-out generalization dataset for studying out-of-sample associations using classical statistical analyses.

For cross-validation, we then subdivided the validation set into 100 training- and testing splits following the Monte Carlo resampling scheme (also referred to as shuffle-split) with 10% of the data used for testing. To compare model performances based on paired tests, we used the same splits across all models. Split-wise testing performance was extracted and carried forward for informal inference using violin plots (Figure 2,Figure 3). For generalization testing, predictions on the held-out data were generated from all 100 models from each cross-validation split.

On the held-out set, unique subject-wise predictions were obtained by averaging across folds and occasional duplicate predictions due to Monte Carlo sampling which could produce multiple predictions per subject[5]. Such strategy is known as CV-bagging (Varoquaux et al., 2017b) and can improve both performance and stability of results[6]. The resulting averages were reported as point estimates in Figures 2,3, and 2 – Figure supplement 1 and used as proxy measures in the analysis of health-related behaviors Figure 4.

Statistical analysis

Resampling statistics for model comparisons on the held-out data

To assess the statistical significance of the observed model performance and the differences in performance between the models, we computed resampling statistics of the performance metrics on the held-out generalization data not used for model construction (Gemein et al., 2020). Once unique subject-wise predictions were obtained on the held-out generalization data by averaging the predictions emanating from each fold of the validation set (cv-bagging), we computed null- and bootstrap-distributions of the observed test statistic on the held-out data, i.e., R^2 score for regression and *AUC* score for classification.

Baseline comparisons. To obtain a p-value for baseline comparisons (*could the prediction performance of a given model be explained chance?*) on the held-out data, we permuted targets 10 000 times and then recomputed the test statistic in each iteration. P-values were then defined as the probability of the test statistic under null distribution being larger than the observed test statistic. To compute uncertainty intervals, we used bootstrap, recomputing the test statistic after resampling 10 000 times with replacement and reporting the 2.5 and 97.5 percentiles of the resulting distribution.

Pairwise comparisons between models. For model comparisons, we considered the out-of-sample difference in R^2 or *AUC* between any two models. To obtain a p-value for model comparisons (*could the difference in prediction performance between two given models be explained chance?*) on the held-out data, we permuted the scores predicted by model A and model B for every single prediction 10 000 times and then recomputed the test statistic in each iteration. We omitted all cases for which only predictions from one of the models under comparison was present. P-values were then defined as the probability of the absolute of the test statistic under null distribution being larger than the absolute observed test statistic. The absolute was considered to account for differences in both directions. Uncertainty intervals were obtained from computing the 2.5 and 97.5 percentiles of the bootstrap distribution

[5] We ensured prior to computation that with 100 CV-splits, predictions were available for all subjects.

[6] The use of CV-bagging can explain why on figures 2,3, and 2 – Figure supplement 1 the performance was sometimes slightly better on the held-out set compared to the cross-validation on the validation test.

Table 7. Extra health variables used for correlation analysis with subject-specific predicted scores.

Family	eid	Variables
Alcohol*	1568-0.0	Average weekly red wine intake
	1578-0.0	Average weekly champagne plus white wine intake
	1588-0.0	Average weekly beer plus cider intake
	1598-0.0	Average weekly spirits intake
	1608-0.0	Average weekly fortified wine intake
	5364-0.0	Average weekly intake of other alcoholic drinks
Physical activity	22040-0.0	Summed MET minutes per week for all activity
Smoking	20161-0.0	Pack years of smoking
Sleep	1160-0.0	Sleep duration

*We computed a compound drinking score by summing up all variables from the alcohol family

based on 10 000 iterations. Here, predictions from model A and model B were resampled using identical resampling indices to ensure a meaningful paired difference.

Out-of-sample association between proxy measures and health-related habits

Computation of brain age delta and de-confounding For association with health-contributing habits (Table 7), we computed the brain age delta as the difference between predicted age and actual age:

$$BrainAge\Delta = Age_{predicted} - Age \tag{1}$$

As age prediction is rarely perfect, the residuals will still contain age-related variance which commonly leads to brain age bias when relating the brain age to an outcome of interest, e.g., sleep duration (Le et al., 2018). To mitigate leakage of age-related information into the statistical models, we employed a de-confounding procedure in line with Smith et al. (2019b) and (Engemann et al., 2020, eqs. 6-8) consisting in residualizing a measure of interest (e.g. sleep duration) with regard to age through multiple regression with quadratic terms for age. To minimize computation on the held-out data, we first trained a model relating the score of interest to age on the validation set to then derive a de-confounding predictor for the held-out generalization data. The resulting de-confounding procedure for variables in the held-out data amounts to computing an age-residualized predictor $measure_{resid}$ from the measure of interest (*e.g.* sleep duration) by applying the following quadratic fit on the validation data:

$$measure_{validation} = age_{validation} \times \beta_{val1} + age^2_{validation} \times \beta_{val2} + \epsilon \tag{2}$$

The de-confounding predictor was then obtained by evaluating the weights β_{val1} and β_{val2} obtained from Equation 2 on the generalization data:

$$measure_{deconfounding} = age_{generalization} \times \beta_{val1} + age^2_{generalization} \times \beta_{val2} \tag{3}$$

We performed this procedure for all target measures, to study associations not driven by the effect of age.

Health-related habits regression We then investigated the joint association between proxy measures of interest and health-related habits (Table 7) using multiple linear regression. For simplicity, we combined all brain imaging and all sociodemographics variables (Figure 2, Figure 2 – Figure supplement 1, Figure 2 – Figure supplement 2). The ensuing model can be denoted as

$$measure = measure_{deconfounding} \times \beta_1 + BrainAge \times \Delta\beta_2 + PredFluidInt \times \beta_3 + PredNeurot \times \beta_4 + \epsilon, \tag{4}$$

where *outcome*$_{resid}$ is given by Equation 2. Prior to model fitting, rows with missing inputs were omitted. For comparability, we then applied standard scaling on all outcomes and all predictors.

The parametric bootstrap was a natural choice for uncertainty estimation, as we used standard multiple linear regression which provides a well defined procedure for mathematically quantifying its implied probabilistic model. Computation was carried out using `sim` function from the `arm` package as described in Gelman and Hill (2006, Ch.7,pp.142-143). This procedure can be intuitively regarded as yielding draws from the posterior distribution of the multiple linear regression model under the assumption of a uniform prior. For consistency with previous analyses, we computed 10000 draws.

Software

Preprocessing and model building were carried out using Python 3.7. The `NiLearn` library was used for processing MRI inputs (Abraham et al., 2014). We used the *scikit-learn* library for machine learning (Pedregosa et al., 2011). For statistical modeling and visualization we used the R-language (R Core Team, 2019) (version 3.5.3) and its ecosystem: `data.table` for high-performance manipulation of tabular data, `ggplot` (Clarke and Sherrill-Mix, 2017; Wickham, 2016) for visualization and the `arm` package for parametric bootstrapping (Gelman and Su, 2020). All data analysis code is shared on GitHub: `https://github.com/KamalakerDadi/proxy_measures_2020`.

Acknowledgments

We would like to thank Dr. Stefania de Vito and Dr. Benjamin de Haas for the critical review and helpful discussion of previous versions of the manuscript. We would like to thank Dr. Julien Dubois and Prof. Ralph Adolphs for helpful discussions in the course of this research project.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8.

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

Behrens, T., Woolrich, M., Jenkinson, M., Johansen-Berg, H., Nunes, R., Clare, S., Matthews, P., Brady, J., and Smith, S. (2003). Characterization and propagation of uncertainty in diffusion-weighted mr imaging. *Magnetic Resonance in Medicine*, 50:1077–1088.

Biecek, P. (2018). Dalex: explainers for complex predictive models in r. *The Journal of Machine Learning Research*, 19(1):3245–3249.

Birley, A. J., Gillespie, N. A., Heath, A. C., Sullivan, P. F., Boomsma, D. I., and Martin, N. G. (2006). Heritability and nineteen-year stability of long and short epq-r neuroticism scales. *Personality and individual differences*, 40(4):737–747.

Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4):537–541.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4):1061–1071.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Bzdok, D., Engemann, D., Grisel, O., Varoquaux, G., and Thirion, B. (2018). Prediction and inference diverge in biomedicine: Simulations and real-world data.

Bzdok, D. and Ioannidis, J. P. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends in neurosciences*, 42(4):251–262.

Bzdok, D. and Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3):223–230.

Carretta, T. R. (2011). Pilot candidate selection method. *Aviation Psychology and Applied Human Factors*.

Carretta, T. R. and Ree, M. J. (1994). Pilot-candidate selection method: Sources of validity. *The International Journal of Aviation Psychology*, 4(2):103–117.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1.

Cattell, R. B. and Scheier, I. H. (1961). The meaning and measurement of neuroticism and anxiety.

Clarke, E. and Sherrill-Mix, S. (2017). *ggbeeswarm: Categorical Scatter (Violin Point) Plots*. R package version 0.6.0.

Cole, J. H. (2020). Multi-modality neuroimaging brain-age in uk biobank: relationship to biomedical, lifestyle and cognitive factors. *Neurobiology of Aging*.

Cole, J. H., Leech, R., Sharp, D. J., and Initiative, A. D. N. (2015). Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of neurology*, 77(4):571–581.

Cole, J. H., Poudel, R. P., Tsagkrasoulis, D., Caan, M. W., Steves, C., Spector, T. D., and Montana, G. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124.

Cole, J. H., Ritchie, S. J., Bastin, M. E., Hernández, M. V., Maniega, S. M., Royle, N., Corley, J., Pattie, A., Harris, S. E., Zhang, Q., et al. (2018). Brain age predicts mortality. *Molecular psychiatry*, 23(5):1385.

Collins, R. (2012). What makes UK Biobank special? *The Lancet*, 379(9822):1173–1174.

Colodro-Conde, L., Couvy-Duchesne, B., Zhu, G., Coventry, W. L., Byrne, E. M., Gordon, S., Wright, M. J., Montgomery, G. W., Madden, P. A., Ripke, S., et al. (2018). A direct test of the diathesis–stress model for depression. *Molecular psychiatry*, 23(7):1590–1596.

Conturo, T. E., Lori, N. F., Cull, T. S., Akbudak, E., Snyder, A. Z., Shimony, J. S., McKinstry, R. C., Burton, H., and Raichle, M. E. (1999). Tracking neuronal fiber pathways in the living human brain. *Proceedings of the National Academy of Sciences*, 96:10422–10427.

Costa, P. T. and McCrae, R. R. (1992). *Neo Pi-R*. Psychological Assessment Resources Odessa, FL.

Cox, S., Ritchie, S., Fawns-Ritchie, C., Tucker-Drob, E., and Deary, I. (2019a). Structural brain imaging correlates of general intelligence in uk biobank. *Intelligence*, 76:101376.

Cox, S., Ritchie, S., Fawns-Ritchie, C., Tucker-Drob, E., and Deary, I. (2019b). Structural brain imaging correlates of general intelligence in UK Biobank. *Intelligence*, 76:101376.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., and Varoquaux, G. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, 192:115–134.

de Groot, M., Vernooij, M. W., Klein, S., Ikram, M. A., Vos, F. M., Smith, S. M., Niessen, W. J., and Andersson, J. L. R. (2013). Improving alignment in Tract-based spatial statistics: Evaluation and optimization of image registration. *NeuroImage*, 76:400–411.

Demertzi, A., Tagliazucchi, E., Dehaene, S., Deco, G., Barttfeld, P., Raimondo, F., Martial, C., Fernández-Espejo, D., Rohaut, B., Voss, H., et al. (2019). Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Science advances*, 5(2):eaat7603.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.

Diedrichsen, J., Balsters, J. H., Flavell, J., Cussans, E., and Ramnani, N. (2009). A probabilistic mr atlas of the human cerebellum. *NeuroImage*, 46(1):39 – 46.

Dosenbach, N. U., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., et al. (2010). Prediction of individual brain maturity using fmri. *Science*, 329(5997):1358–1361.

Dubois, J., Galdi, P., Han, Y., Paul, L. K., and Adolphs, R. (2018a). Resting-State Functional Brain Connectivity Best Predicts the Personality Dimension of Openness to Experience. *Personality Neuroscience*, 1.

Dubois, J., Galdi, P., Paul, L. K., and Adolphs, R. (2018b). A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1756):20170284.

Eisenberg, I. W., Bissett, P. G., Enkavi, A. Z., Li, J., MacKinnon, D. P., Marsch, L. A., and Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, 10(1):1–13.

Engemann, D. A., Kozynets, O., Sabbagh, D., Lemaître, G., Varoquaux, G., Liem, F., and Gramfort, A. (2020). Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *eLife*, 9:e54055.

Engemann, D. A., Raimondo, F., King, J.-R., Rohaut, B., Louppe, G., Faugeras, F., Annen, J., Cassol, H., Gosseries, O., Fernandez-Slezak, D., Laureys, S., Naccache, L., Dehaene, S., and Sitt, J. D. (2018). Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain*, 141(11):3179–3192.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., and Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, 116(12):5472–5477.

Eysenck, H. J. (1958). The continuity of abnormal and normal behavior. *Psychological Bulletin*, 55(6):429–432.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, 6:21–29.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N. E. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology*, 186(9):1026–1034.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A. and Su, Y.-S. (2020). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.11-1.

Gemein, L. A., Schirrmeister, R. T., Chrabąszcz, P., Wilson, D., Boedecker, J., Schulze-Bonhage, A., Hutter, F., and Ball, T. (2020). Machine-learning-based diagnostics of eeg pathology. *NeuroImage*, 220:117021.

Gonneaud, J., Baria, A. T., Binette, A. P., Gordon, B. A., Chhatwal, J. P., Cruchaga, C., Jucker, M., Levin, J., Salloway, S., Farlow, M., et al. (2020). Functional brain age prediction suggests accelerated aging in preclinical familial alzheimer's disease, irrespective of fibrillar amyloid-beta pathology. *bioRxiv*.

Greicius, M., Srivastava, G., Reiss, A., and Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proceedings of the National Academy of Sciences*, 101:4637.

Hasson, U., Malach, R., and Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in cognitive sciences*, 14(1):40–48.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

He, T., An, L., Feng, J., Bzdok, D., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. T. (2020). Meta-matching: a simple framework to translate phenotypic predictive models from big to small data. *bioRxiv*.

He, T., Kong, R., Holmes, A. J., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., and Yeo, B. T. (2018). Is deep learning better than kernel regression for functional connectivity prediction of fluid intelligence? In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4. IEEE.

Hettema, J. M., Neale, M. C., Myers, J. M., Prescott, C. A., and Kendler, K. S. (2006). A population-based twin study of the relationship between neuroticism and internalizing disorders. *American journal of Psychiatry*, 163(5):857–864.

Horn, J. L., Donaldson, G., and Engstrom, R. (1981). Apprehension, memory, and fluid intelligence decline in adulthood. *Research on Aging*, 3(1):33–84.

Hozer, F. and Houenou, J. (2016). Can neuroimaging disentangle bipolar disorder? *Journal of affective disorders*, 195:199–214.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, 167(7):748–751.

Jääskeläinen, I. P., Pajula, J., Tohka, J., Lee, H.-J., Kuo, W.-J., and Lin, F.-H. (2016). Brain hemodynamic activity during viewing and re-viewing of comedy movies explained by experienced humor. *Scientific reports*, 6:27741.

Josse, J., Prost, N., Scornet, E., and Varoquaux, G. (2019). On the consistency of supervised learning with missing values. working paper or preprint.

Kapur, S., Phillips, A. G., and Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, 17(12):1174–1179.

Keyes, K. M., Platt, J., Kaufman, A. S., and McLaughlin, K. A. (2017). Association of Fluid Intelligence and Psychiatric Disorders in a Population-Representative Sample of US Adolescents. *JAMA psychiatry*, 74(2):179–188.

Khandaker, G. M., Dalman, C., Kappelmann, N., Stochl, J., Dal, H., Kosidou, K., Jones, P. B., and Karlsson, H. (2018). Association of Childhood Infection With IQ and Adult Nonaffective Psychosis in Swedish Men: A Population-Based Longitudinal Cohort and Co-relative Study. *JAMA Psychiatry*, 75(4):356–362.

Kievit, R. A., Fuhrmann, D., Borgeest, G. S., Simpson-Kent, I. L., and Henson, R. N. (2018a). The neural determinants of age-related changes in fluid intelligence: a pre-registered, longitudinal analysis in uk biobank. *Wellcome open research*, 3.

Kievit, R. A., Fuhrmann, D., Borgeest, G. S., Simpson-Kent, I. L., and Henson, R. N. A. (2018b). The neural determinants of age-related changes in fluid intelligence: a pre-registered, longitudinal analysis in UK Biobank. *Wellcome Open Research*, 3.

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

Krapohl, E., Rimfeld, K., Shakeshaft, N. G., Trzaskowski, M., McMillan, A., Pingault, J.-B., Asbury, K., Harlaar, N., Kovas, Y., Dale, P. S., et al. (2014). The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proceedings of the national academy of sciences*, 111(42):15273–15278.

Lahey, B. B. (2009). Public health significance of neuroticism. *American Psychologist*, 64(4):241.

Le, T. T., Kuplicki, R. T., McKinney, B. A., Yeh, H.-W., Thompson, W. K., Paulus, M. P., , T. . I., Aupperle, R. L., Bodurka, J., Cha, Y.-H., Feinstein, J. S., Khalsa, S. S., Savitz, J., Simmons, W. K., and Victor, T. A. (2018). A nonlinear simulation framework supports adjusting for age when analyzing brainage. *Frontiers in Aging Neuroscience*, 10:317.

Ledoit, O. and Wolf, M. (2004). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

Lerch, J., van der Kouwe, A., Raznahan, A., Paus, T., Johansen-Berg, H., Miller, K., Smith, S., Fischl, B., and Sotiropoulos, S. (2017). Studying neuroanatomy using mri. *Nature neuroscience*, 20:314 – 326.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S. K., Huntenburg, J. M., Lampe, L., Rahim, M., Abraham, A., Craddock, R. C., et al. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage*, 148:179–188.

Little, R. J. A. and Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA.

Lynn, R. and Martin, T. (1997). Gender differences in extraversion, neuroticism, and psychoticism in 37 nations. *The Journal of social psychology*, 137(3):369–373.

Maglanoc, L. A., Kaufmann, T., Meer, D. v. d., Marquand, A. F., Wolfers, T., Jonassen, R., Hilland, E., Andreassen, O. A., Landrø, N. I., and Westlye, L. T. (2020). Brain Connectome Mapping of Complex Human Traits and Their Polygenic Architecture Using Machine Learning. *Biological Psychiatry*, 87(8):717–726.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P. M., and Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536.

Nave, G., Jung, W. H., Linnér, R. K., Kable, J. W., and Koellinger, P. D. (2018). Are Bigger Brains Smarter? Evidence From a Large-Scale Preregistered Study:. *Psychological Science*.

Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., Hari, R., and Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences*, 109(24):9599–9604.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pedersen, N. L., Plomin, R., McClearn, G. E., and Friberg, L. (1988). Neuroticism, extraversion, and related traits in adult twins reared apart and reared together. *Journal of personality and social psychology*, 55(6):950.

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

Perlis, R. H. (2011). Translating biomarkers to clinical practice. *Molecular Psychiatry*, 16(11):1076–1087.

Pervaiz, U., Vidaurre, D., Woolrich, M. W., and Smith, S. M. (2020). Optimising network modelling methods for fmri. *NeuroImage*, 211:116604.

Poldrack, R. A., Huckins, G., and Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*, 77(5):534–540.

Power, R. A. and Pluess, M. (2015). Heritability estimates of the Big Five personality traits based on common genetic variants. *Translational psychiatry*, 5(7):e604.

Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. pages 180–185.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ritchie, S. J., Dickie, D. A., Cox, S. R., Valdes Hernandez, M. d. C., Corley, J., Royle, N. A., Pattie, A., Aribisala, B. S., Redmond, P., Muñoz Maniega, S., Taylor, A. M., Sibbett, R., Gow, A. J., Starr, J. M., Bastin, M. E., Wardlaw, J. M., and Deary, I. J. (2015). Brain volumetric changes and cognitive ageing during the eighth decade of life. *Human Brain Mapping*, 36(12):4910–4925.

Sabbagh, D., Ablin, P., Varoquaux, G., Gramfort, A., and Engeman, D. A. (2019). Manifold-regression to predict from meg/eeg brain signals without source modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Saygin, Z., Osher, D., Norton, E., Youssoufian, D., Beach, S., Feather, J., Gaab, N., Gabrieli, J., and Kanwisher, N. (2016). Connectivity precedes function in the development of the visual word form area. *Nature neuroscience*, 19.

Shackman, A. J., Tromp, D. P., Stockbridge, M. D., Kaplan, C. M., Tillman, R. M., and Fox, A. S. (2016). Dispositional negativity: An integrative psychological and neurobiological perspective. *Psychological bulletin*, 142(12):1275.

Shelton, J. T., Elliott, E. M., Matthews, R. A., Hill, B., Gouvier, W., and others (2010). The relationships of working memory, secondary memory, and general fluid intelligence: working memory is special. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(3):813.

Smith, S. M., Elliott, L. T., Alfaro-Almagro, F., McCarthy, P., Nichols, T. E., Douaud, G., and Miller, K. L. (2020). Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. *Elife*, 9:e52677.

Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., and Miller, K. L. (2019a). Estimation of brain age delta from brain imaging. *NeuroImage*.

Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., and Miller, K. L. (2019b). Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528 – 539.

Sonkusare, S., Breakspear, M., and Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in cognitive sciences*, 23(8):699–714.

Stevens, S. S. et al. (1946). On the theory of scales of measurement.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10.

Szucs, D. and Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology*, 15(3):e2000797.

Terracciano, A. and Costa Jr, P. T. (2004). Smoking and the five-factor model of personality. *Addiction*, 99(4):472–481.

Thompson, P., Hayashi, K., Dutton, R., Chiang, M.-C., Leow, A., Sowell, E., De Zubicaray, G., Becker, J., Lopez, O., Aizenstein, H., and Toga, A. (2007). Tracking alzheimer's disease. *Annals of the New York Academy of Sciences*, 1097:183–214.

Topiwala, A., Allan, C. L., Valkanova, V., Zsoldos, E., Filippini, N., Sexton, C., Mahmood, A., Fooks, P., Singh-Manoux, A., Mackay, C. E., et al. (2017). Moderate alcohol consumption as risk factor for adverse brain outcomes and cognitive decline: longitudinal cohort study. *bmj*, 357:j2353.

Tyrer, P., Reed, G. M., and Crawford, M. J. (2015). Classification, assessment, prevalence, and effect of personality disorder. *The Lancet*, 385(9969):717–726.

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77.

Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., and Thirion, B. (2010). Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 13(Pt 1):200–208.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017a). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166 – 179. Individual Subject Prediction.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017b). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145(August 2015):166–179.

Venkatesh, M., JaJa, J., and Pessoa, L. (2020). Capturing brain dynamics: latent spatiotemporal patterns predict stimuli and individual differences. *bioRxiv*.

Vukasović, T. and Bratko, D. (2015). Heritability of personality: a meta-analysis of behavior genetic studies. *Psychological bulletin*, 141(4):769.

Wang, J., Knol, M. J., Tiulpin, A., Dubost, F., de Bruijne, M., Vernooij, M. W., Adams, H. H., Ikram, M. A., Niessen, W. J., and Roshchupkin, G. V. (2019). Gray matter age prediction as a biomarker for risk of dementia. *Proceedings of the National Academy of Sciences*, 116(42):21213–21218.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Woo, C.-W., Chang, L. J., Lindquist, M. A., and Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, 20(3):365–377.

Yarkoni, T. (2015). Neurobiological substrates of personality: A critical overview. *APA handbook of personality and social psychology*, 4:61–83.

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.

Supporting Information 939

Appendix 1: Additional results 940

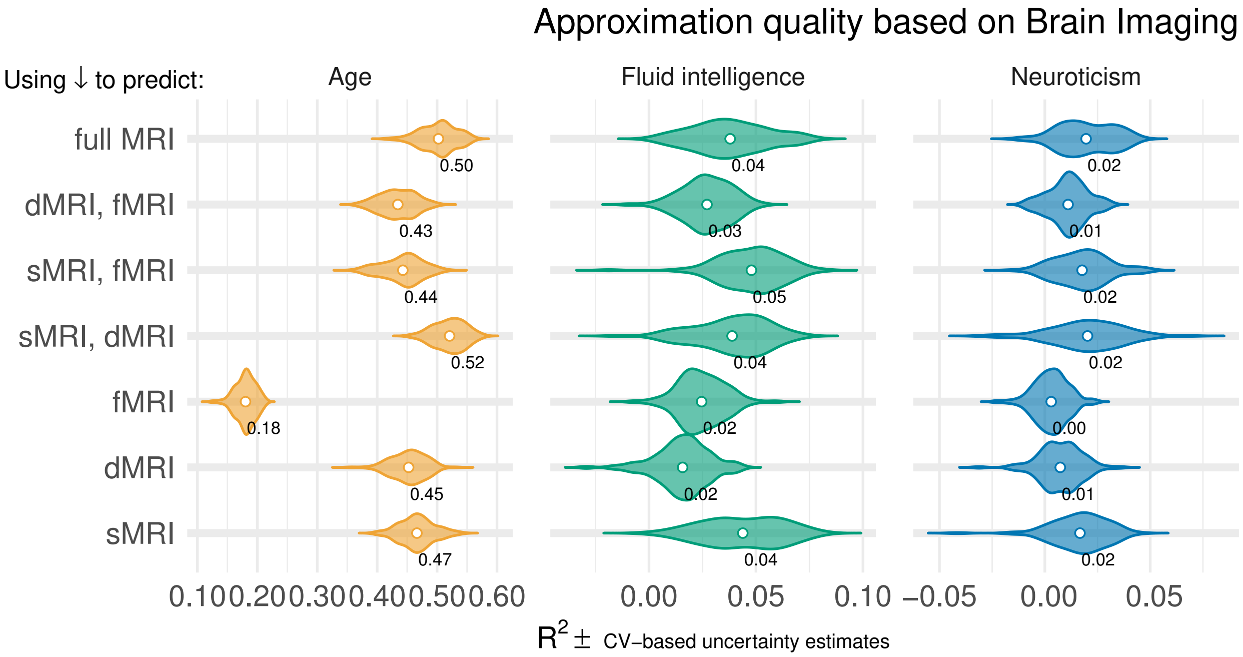

Figure 2 – Figure supplement 1. Prediction of individual differences in proxy measures from MRI. Approximation performance using multiple MR modalities on the validation dataset: sMRI, dMRI, rfMRI and their combinations (see Table 3). Visual conventions as in Figure 2. One can see that prediction of age was markedly stronger than prediction of fluid intelligence or prediction of neuroticism. As a general trend, models based on multiple MRI modalities tended to yield better prediction. For simplicity, we based subsequent analyses on the full model based on all MRI data.

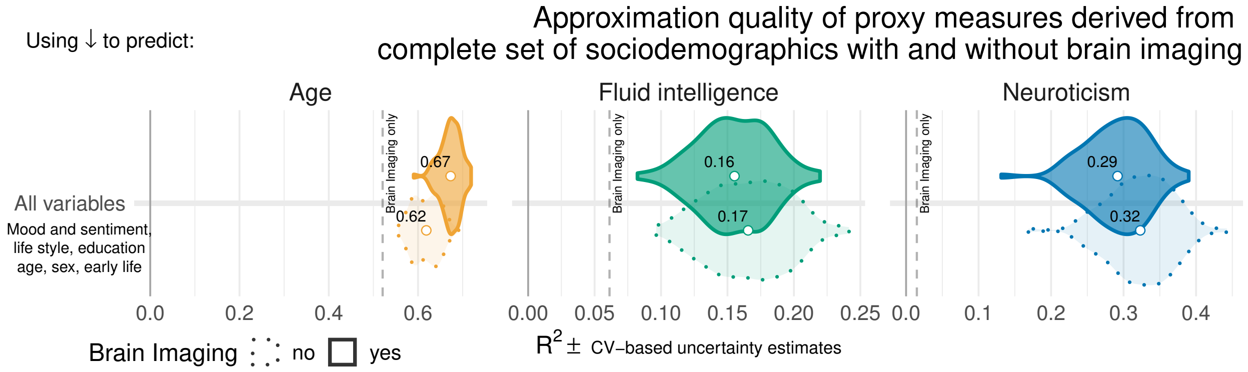

Figure 2 – Figure supplement 2. Approximation performance using all sociodemographic data. Approximation performance using all sociodemographic variables with or without brain imaging included on the validation dataset. Visual conventions as in Figure 2. The performance was highly to the best performing models within each target Figure 2, *i.e.*, life style for age, education for fluid intelligence and mood & sentiment for neuroticism. This suggests that for each target those specific blocks of predictors were sufficiently explaining the performance. For simplicity, we based subsequent analyses in Figure 3 and Figure 4 on all sociodemographic variables.

Table S1. Regression statistics on the held-out set for purely MRI-based approximation.

Target	R^2_{observed}	p-value	$\mathrm{CI}_{\mathrm{low}}$	$\mathrm{CI}_{\mathrm{high}}$
Age	0.521	1×10^{-4}	0.502	0.538
Fluid intelligence	0.061	1×10^{-4}	0.052	0.070
Neuroticism	0.015	1×10^{-4}	0.005	0.024

Table S2. Classification difference statistics on the held-out set for MRI-based approximation.

Target	$\mathrm{AUC}_{\mathrm{observed}}$	p-value	$\mathrm{CI}_{\mathrm{low}}$	$\mathrm{CI}_{\mathrm{high}}$
Neuroticism	0.590	1×10^{-4}	0.566	0.614
Age	0.916	1×10^{-4}	0.905	0.927
Fluid intelligence	0.667	1×10^{-4}	0.643	0.690

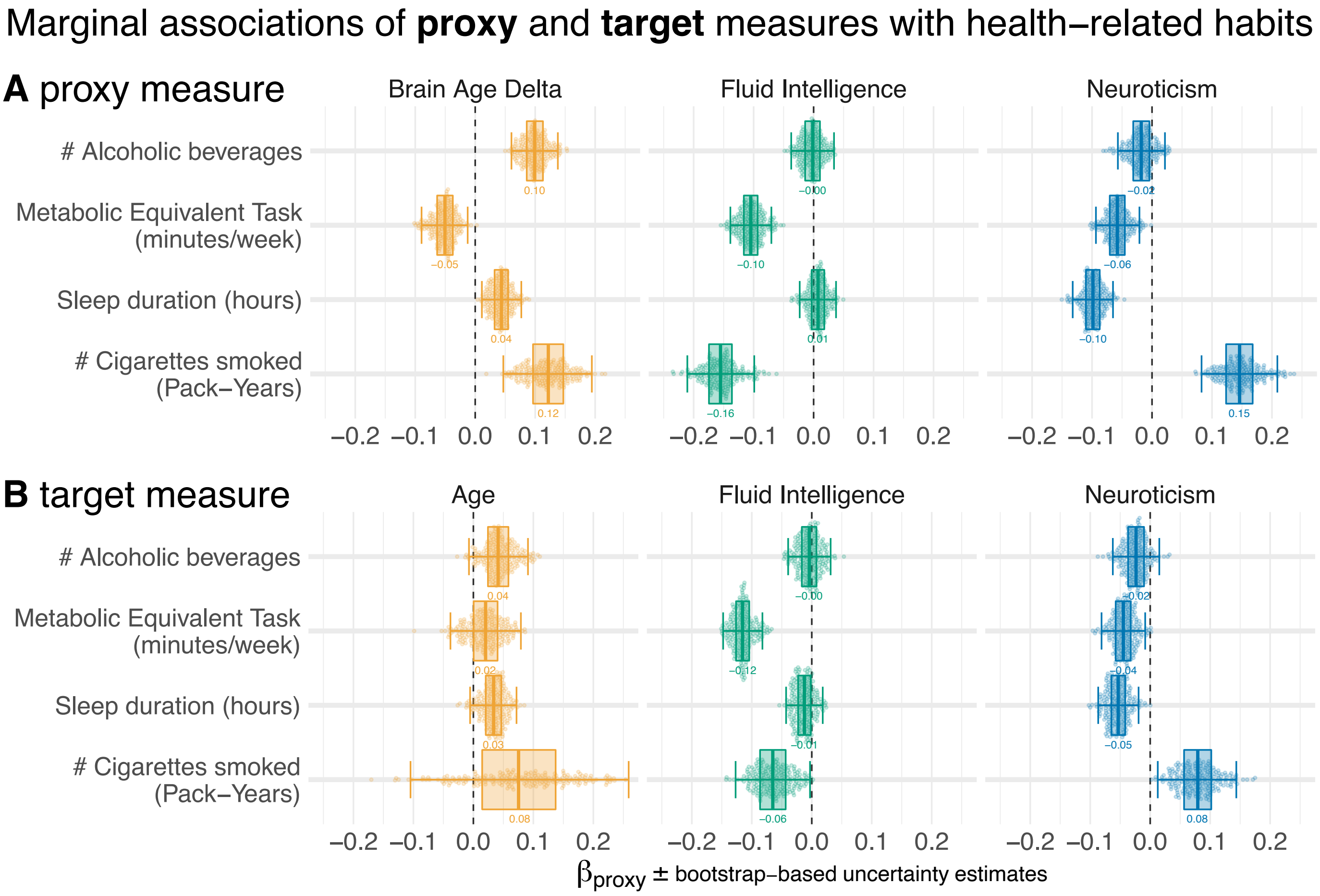

Figure 4 – Figure supplement 1. Marginal associations between proxy measures and health-related habits. Marginal (instead of conditional) estimates using univariate regression. Same visual conventions as in Figure 4.

Appendix 2: Sociodemographic variables

Table S3. List of variables contained in each block of sociodemographic models: mood & sentiment (MS), Age, Sex (AS), Education (EDU), Early life (EL).

Group	UKBB code	Variables
Mood & Sentiment	2040-2.0	Risk taking
	4526-2.0	Happiness
	4537-2.0	Work/job satisfaction
	4548-2.0	Health satisfaction
	4559-2.0	Family relationship satisfaction
	4570-2.0	Friendships satisfaction
	4581-2.0	Financial situation satisfaction
	4598-2.0	Ever depressed for a whole week
	4609-2.0	Longest period of depression
	4620-2.0	Number of depression episodes
	4631-2.0	Ever unenthusiastic/disinterested for a whole week
	4642-2.0	Ever manic/hyper for 2 days
	4653-2.0	Ever highly irritable/argumentative for 2 days
	2050-2.0	Frequency of depressed mood in last 2 weeks
	2060-2.0	Frequency of unenthusiasm / disinterest in last 2 weeks
	2070-2.0	Frequency of tenseness / restlessness in last 2 weeks
	2080-2.0	Frequency of tiredness / lethargy in last 2 weeks
	2090-2.0	Seen doctor (GP) for nerves, anxiety, tension or depression
	2100-1.0	Seen a psychiatrist for nerves, anxiety, tension or depression
	5375-2.0	Longest period of unenthusiasm / disinterest
	5386-2.0	Number of unenthusiastic/disinterested episodes
	5663-2.0	Length of longest manic/irritable episode
	5674-2.0	Severity of manic/irritable episode
	6145-2.0	Illness, injury, bereavement, stress in last 2 years
	6156-2.0	Manic/hyper symptoms
Age, Sex	31-0.0	Sex
	34-0.0	Year of birth
	52-0.0	Month of birth
	21022-0.0	Age at recruitment
	21003-2.0	Age when attended assessment centre
Education	6138-2.0	Qualifications
	845-2.0	Age completed full time education
Early life	1647-2.0	Country of birth (UK/elsewhere)
	1677-2.0	Breastfed as a baby
	1687-2.0	Comparative body size at age 10
	1697-2.0	Comparative height size at age 10
	1707-2.0	Handedness (chirality/laterality)
	1767-2.0	Adopted as a child
	1777-2.0	Part of a multiple birth

Table S3 continued

	1787-2.0	Maternal smoking around birth
Lifestyle	670-2.0	Type of accommodation lived in
	680-2.0	Own or rent accommodation lived in
	6139-2.0	Gas or solid-fuel cooking/heating
	699-2.0	Length of time at current address
	709-2.0	Number in household
	6141-2.0	How are people in household related to participant
	728-2.0	Number of vehicles in household
	738-2.0	Income before tax
	796-2.0	Distance between home and job workplace
	757-2.0	Time employed in main current job
	767-2.0	Length of working week for main job
	777-2.0	Freq. of travelling from home to job workplace
	6143-2.0	Transport type for commuting to job workplace
	6142-2.0	Current employment status
	806-2.0	Job involves mainly walking or standing
	816-2.0	Job involves heavy manual or physical work
	826-2.0	Job involves shift work
	3426-2.0	Job involves night shift work
	1031-2.0	Freq. of friend/ family visits
	6160-2.0	Leisure/social activities
	2110-2.0	Able to confide
	1239-2.0	Current tobacco smoking
	1249-2.0	Past tobacco smoking
	1259-2.0	Smoking/smokers in household
	1269-2.0	Exposure to tobacco smoke at home
	1279-2.0	Exposure to tobacco smoke outside home
	2644-2.0	Light smokers, at least 100 smokes in lifetime
	2867-2.0	Age started smoking in former smokers
	2877-2.0	Type of tobacco previously smoked
	2887-2.0	Number of cigarettes previously smoked daily
	2897-2.0	Age stopped smoking
	2907-2.0	Ever stopped smoking for 6+ months
	2926-2.0	Number of unsuccessful stop-smoking attempts
	2936-2.0	Likelihood of resuming smoking
	3436-2.0	Age started smoking in current smokers
	3446-2.0	Type of tobacco currently smoked
	3456-2.0	Number of cigarettes currently smoked daily (current cigarette smokers)
	3466-2.0	Time from waking to first cigarette
	3476-2.0	Difficulty not smoking for 1 day
	3486-2.0	Ever tried to stop smoking

Table S3 continued

3496-2.0	Wants to stop smoking
3506-2.0	Smoking compared to 10 years previous
5959-2.0	Previously smoked cigarettes on most/all days
6157-2.0	Why stopped smoking
6158-2.0	Why reduced smoking

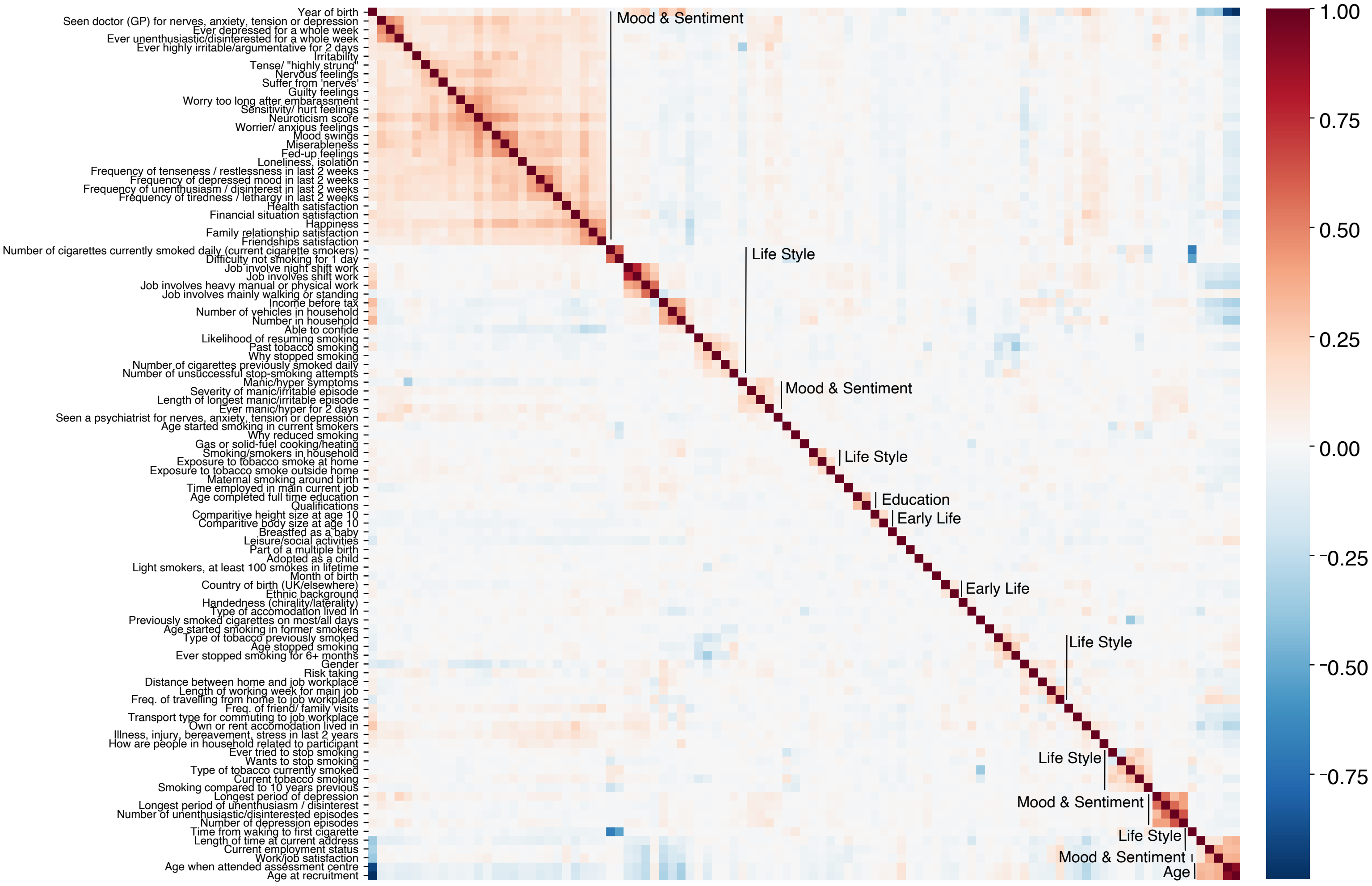

Figure S1. Intercorrelations between sociodemographic inputs. To check the plausibility of the proposed grouping of variables into blocks, we investigated the inter-correlations among the sociodemographic inputs (Table S3). We first applied Yeo-Johnson power transform to the variables yield approximately symmetrical distributions. Then we computed Pearson correlations. One can see that a large majority of variables shows low if any inter-correlations. Strongly inter-correlated blocks emerged, in particular for Mood & Sentiment and Life Style. Note that within the Life Style category many smaller blocks with strong inter-correlation occurred, some of which were obviously related to the circumstance of living such as household or employment status.