6-1-2023

Population Modeling with Machine Learning can Enhance Measures of Mental Health - Open-Data Replication

Ty Easley
Washington University School of Medicine in St. Louis

Ruiqi Chen
Washington University in St. Louis

Kayla Hannon
Washington University School of Medicine in St. Louis

Rosie Dutt
Washington University School of Medicine in St. Louis, rdutt@smith.edu

Janine Bijsterbosch
Washington University School of Medicine in St. Louis

Neuroimage: Reports 3 (2023) 100163

Contents lists available at ScienceDirect

Neuroimage: Reports

journal homepage: www.sciencedirect.com/journal/neuroimage-reports

Population modeling with machine learning can enhance measures of mental health - Open-data replication

Ty Easley [a], Ruiqi Chen [b], Kayla Hannon [a], Rosie Dutt [a], Janine Bijsterbosch [a,*]

[a] Department of Radiology, Washington University School of Medicine, Saint Louis, Missouri, 63110, USA
[b] Division of Biology and Biomedical Sciences, Washington University in St. Louis, Saint Louis, Missouri, 63110, USA

ARTICLE INFO

Original content: UK BioBank (Reference data)

Keywords:
Replication
Prediction
Neuroticism
Intelligence
Data pollution
Resting state fMRI

ABSTRACT

Efforts to predict trait phenotypes based on functional MRI data from large cohorts have been hampered by low prediction accuracy and/or small effect sizes. Although these findings are highly replicable, the small effect sizes are somewhat surprising given the presumed brain basis of phenotypic traits such as neuroticism and fluid intelligence. We aim to replicate previous work and additionally test multiple data manipulations that may improve prediction accuracy by addressing data pollution challenges. Specifically, we added additional fMRI features, averaged the target phenotype across multiple measurements to obtain more accurate estimates of the underlying trait, balanced the target phenotype's distribution through undersampling of majority scores, and identified data-driven subtypes to investigate the impact of between-participant heterogeneity. Our results replicated prior results from Dadi et al. (2021) in a larger sample. Each data manipulation further led to small but consistent improvements in prediction accuracy, which were largely additive when combining multiple data manipulations. Combining data manipulations (i.e., extended fMRI features, averaged target phenotype, balanced target phenotype distribution) led to a three-fold increase in prediction accuracy for fluid intelligence compared to prior work. These findings highlight the benefit of several relatively easy and low-cost data manipulations, which may positively impact future work.

1. Introduction

In recent years, studies with sufficiently large participant samples ($N > 2000$) have reported very low effect sizes for associations between neuroimaging measures and trait-level phenotypes (Marek et al., 2020; Dutt et al., 2021; Dadi et al., 2021). For example, Dadi et al. used Random Forest Regression in approximately 9000 participants from the UK Biobank (UKB) data to predict trait-level phenotypes of neuroticism and fluid intelligence, and reported a maximum R^2 of 0.04 when using neuroimaging features (Dadi et al., 2021). Despite explaining a small amount of variance, results from these large participant samples are robust against sampling variability (Marek et al., 2022) and have good out-of-sample generalizability (Dadi et al., 2021; Dutt et al., 2021). Nevertheless, low brain-phenotype effect sizes are somewhat surprising given the presumed brain basis of phenotypic traits such as neuroticism and fluid intelligence. The goal of this paper is to replicate the existing findings (Dadi et al., 2021) in a larger UKB sample and test multiple hypotheses regarding potential factors that may influence low brain-phenotype effect sizes. Specifically, we test whether addressing potential data pollution challenges (De Nadai et al., 2022) such as noisy phenotypes, noisy neuroimaging measures, skewed (imbalanced) phenotypic distributions, and population heterogeneity lead to improvements in R^2. We focus on resting state functional MRI (rfMRI) neuroimaging measures to control the scope of this work, but relative improvements in R^2 are expected to generalize to other modalities.

The first potential explanation for low brain-phenotype effect sizes is that the neuroimaging features used for prediction are noisy measures of brain function. Previous work reported UKB test-retest reliability results for neuroimaging measures ranging from 0.3 for rfMRI measures to 0.9 for structural measures (Dutt et al., 2021). The relatively lower test-retest reliability of rfMRI measures likely results from both dynamic state fluctuations and measurement error due to the relatively small number of timepoints ($t = 490$ per subject). An interesting question is therefore whether residuals from the task fMRI scan (after removing task activation effects) can be combined with rfMRI data to reduce measurement error and improve prediction accuracy. This possibility is supported by previous work showing shared trait-level connectivity information between task residual and rfMRI data, although state-related

* Corresponding author.
E-mail address: janine.bijsterbosch@wustl.edu (J. Bijsterbosch).

https://doi.org/10.1016/j.ynirp.2023.100163
Received 2 May 2022; Received in revised form 21 September 2022; Accepted 2 March 2023
Available online 13 March 2023

connectivity differences were also observed (Fair et al., 2007). In addition, recent work has shown that combining task and resting-state data improves the reliability and predictive power of intrinsic connectivity measures (Elliott et al., 2019; Gao et al., 2019). Furthermore, it may be possible to improve prediction accuracy by leveraging other types of features that can be extracted from rfMRI and task residual data (J. D. Bijsterbosch et al., 2021; J. Bijsterbosch et al., 2020). For example, network amplitudes have higher test-retest reliability than connectivity information (Dutt et al., 2021), and have been shown to capture individual differences in behavioral traits (J. Bijsterbosch et al., 2017; Miller et al., 2016). We test whether adding task residual connectivity features and/or including additional amplitude features will result in a higher R^2.

The second potential explanation for low brain-phenotype effect sizes is that the target phenotype for prediction may be a noisy measure of the underlying trait. This potential explanation is supported by the substantially larger observed maximum R^2 of 0.52 for age when compared to $R^2 < 0.04$ for neuroticism and fluid intelligence (Dadi et al., 2021). Importantly, the phenotype of age is known without error (apart from potential database entry mistakes), whereas neuroticism and fluid intelligence scores are obtained from self-report questionnaires and test questions, respectively, which are prone to intra-individual response instability, and potential social desirability bias in the case of neuroticism (McKelvie, 2004). In the UKB these factors are reflected in previous estimates of test-retest reliability of neuroticism (0.85) (Dutt et al., 2021) and fluid intelligence (0.65) (Lyall et al., 2016). Response instability can be reduced by averaging over multiple available repeats of the same measure, thereby reducing noise and obtaining a more accurate estimate of the underlying trait. We test whether using phenotypes that have been averaged across all available UKB instances as the prediction target will result in a higher R^2.

The third potential explanation for low brain-phenotype effect sizes is that the distribution of the target phenotype may be skewed, leading to an imbalanced regression problem (Yang et al. 2021 18–24 Jul 2021). For example, neuroticism is strongly positively skewed such that the distribution peaks at zero (low neuroticism) with a long tail to the maximum score of 12 (high neuroticism; Fig. 1). Such underrepresentation of high neuroticism scores in population data will be propagated into the training sample, which likely results in inaccurate predictions of the underrepresented scores in the test sample, leading to a lower R^2. A simple solution to this challenge is to flatten the distribution of neuroticism by undersampling the majority scores to ensure that all possible scores have equal representation (Dal Pozzolo, Caelen, and Bontempi, 2015). We test whether undersampling the UKB dataset to flatten the distribution of the target phenotype will result in a higher R^2, despite the reduced overall sample size.

The fourth potential explanation for low brain-phenotype effect sizes is potential between-participant biological heterogeneity. For example, two individuals with the same high score for neuroticism may not share the same brain-basis for their high neuroticism. Hence, different subgroups of participants may have divergent brain-phenotype associations, suggesting the presence of biological subtypes (also known as 'biotypes'). If multiple biotypes are combined in the same analysis, the overall population effect size estimate may be diluted or canceled out (Ferrante et al., 2019). We adapt a previously developed data-driven pipeline to identify biotypes (Drysdale et al., 2017) based on canonical correlation analysis (Hotelling, 1936; Winkler et al., 2020) combined with rigorous cross-validation to test for biotype stability (Dinga et al., 2019). We test whether subsequently performing phenotype predictions separately within each biotype will result in a higher R^2, despite the reduced sample size.

Taken together, the goal of this study is to better understand potential data pollution drivers of low effect sizes for brain-phenotype associations and to identify avenues for improvements to inform future research. Strengthening brain-phenotype associations is important for personalized medicine efforts and maximizing the impact of research into the brain basis of clinical, cognitive, and behavioral traits of interest (Paulus and Thompson, 2019, 2021). Our results reveal small but significant increases in R^2 from each of the tested avenues for improvements, which when combined, resulted in a boost in R^2 from 0.03 to 0.06 for fluid intelligence and from 0.00 to 0.03 for neuroticism.

2. Materials and methods

2.1. Sample

Data from the UKB 20,000 neuroimaging release were used for this study. N = 18,550 participants had complete resting state and task neuroimaging data, which is approximately double the participants used previously (Dadi et al., 2021). The demographic characteristics were 52.9% female (9,822) and 47.1% male (8,728), with an age range at the time of scanning of 44–80 years (mean ± standard deviation: 62.5 ± 7.5 years). The sample was split in half to generate separate datasets for model training and model generalization. This research was performed under UK Biobank application number 47267.

2.2. Target phenotypes

Neuroticism was calculated for each instance using a summary score based on the sum across 12 binary (yes/no) questions (Table 1) developed by (Smith et al., 2013). This is identical to the field ID 20127 used

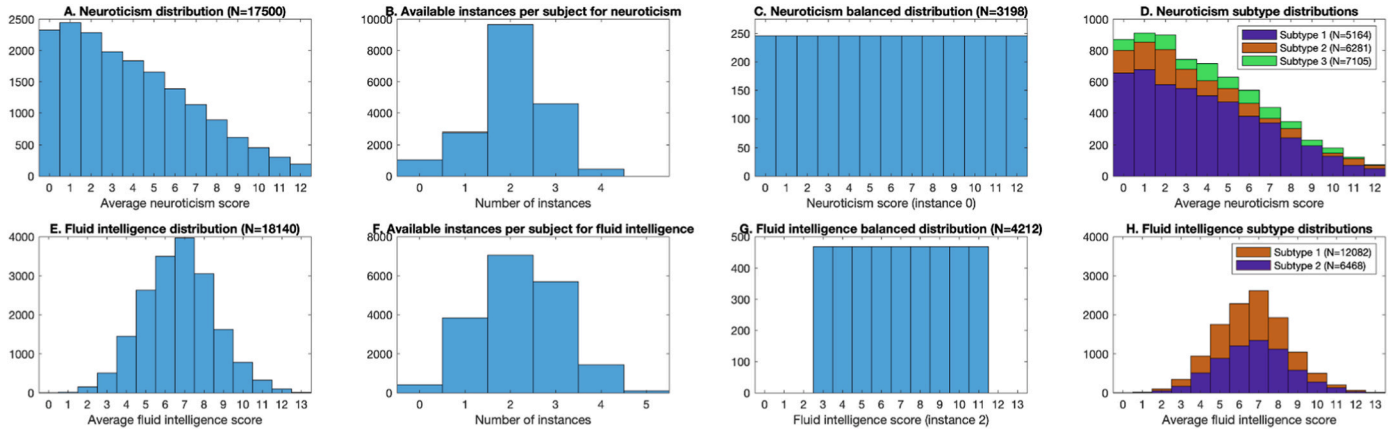

Fig. 1. Target phenotype distributions. Top row (A–D) shows neuroticism and bottom row (E–H) shows fluid intelligence. The first column (A, E) displays the distribution of averaged phenotypes, the second column (B, F) displays the number of available instances per subject, the third column (C, G) displays the balanced phenotype distributions, and the fourth column (D, H) displays the distributions of data-driven subtypes.

Table 1
UKB variables used as target phenotypes. Field IDs 1920–2030 were used to calculate neuroticisms. Each question is answered yes (1) or no (0) and the neuroticism score represents the sum across all responses. Field IDs 20016 and 20191 are identical measures of fluid intelligence.

Phenotype	Field ID	Question/description
Neuroticism	1920	Does your mood often go up and down?
	1930	Do you ever feel 'just miserable' for no reason?
	1940	Are you an irritable person?
	1950	Are your feelings easily hurt?
	1960	Do you often feel 'fed-up'?
	1970	Would you call yourself a nervous person?
	1980	Are you a worrier?
	1990	Would you call yourself tense or 'highly strung'?
	2000	Do you worry too long after an embarrassing experience?
	2010	Do you suffer from 'nerves'?
	2020	Do you often feel lonely?
	2030	Are you often troubled by feelings of guilt?
Fluid Intelligence	20016	Unweighted sum of the number of correct answers given to the 13 fluid intelligence questions completed on touch screen during assessment center visit
	20191	Unweighted sum of the number of correct answers given to the 13 fluid intelligence questions completed in web-based online follow-up

in (Dadi et al., 2021), but enables calculation of Neuroticism scores for each of the available instances.

Within the UKB, the same fluid intelligence test was repeated at each assessment center visit and as part of a cognitive function online follow-up questionnaire (Table 1). Both these variables provide an unweighted sum of the number of correct answers given to the same set of 13 fluid intelligence questions. Participants were given 2 min to answer the questions, and those who did not finish scored zero for each of the unattempted questions.

Participants with missing data for individual questions (resulting from answers such as 'do not know' or 'prefer not to answer') were removed for neuroticism and fluid intelligence. Specifically, neuroticism was only calculated for participants and instances with complete data on all 12 questions. For replications of original work (Dadi et al., 2021), we used the target phenotypes of 20016 at instance 2 for fluid intelligence, and the sum neuroticism score (corresponding to 20127) at instance 0.

2.2.1. Phenotype averaging

To test whether averaging the target phenotypes improved prediction accuracy, neuroticism and fluid intelligence scores were averaged across all instances available for each participant respectively. The range of available instances per participant was 0–4 for neuroticism and 0–5 for fluid intelligence (Fig. 1B and F).

2.2.2. Phenotype flattened distribution

To test whether flattening the phenotype distributions improved prediction accuracy, we calculated the number of participants with the least common value (neuroticism = 12 for 246 participants, Fig. 1A; fluid intelligence = 11 for 468 participants, Fig. 1E) and randomly selected the matching number of participants for each score. To achieve a sufficiently large sample size, fluid intelligence was truncated to range from 3 to 11 (inclusive) for the flat distribution tests (i.e., participants with fluid intelligence score 0, 1, 2, 12, 13, 14 were excluded). The total size of the flat distribution samples was N = 3198 for neuroticism (Fig. 1C) and N = 4212 for fluid intelligence (Fig. 1G).

We also compared prediction on phenotype-flattened data to prediction on subsampled data with a matched sample size and the same feature distribution as the full dataset. For the phenotype-flattened distributions of both fluid intelligence (N = 4212) and neuroticism (N = 3198), a random subset of matched size was uniformly subsampled from the initial data, leaving the full data's original distribution unmodified; see the supplement for the results of this analysis.

2.3. Resting state features

Preprocessed resting state data was downloaded from the UKB, as described previously (Miller et al., 2016; Alfaro-Almagro et al., 2018). The task residual data was not released with the UKB task data to limit the download size, so we repeated the fit of the task general linear model for each participant and saved the residuals. Previous work performed a group ICA on processed rfMRI data of N = 5000 UKB participants at a dimensionality of 100, out of which 55 components represented signal resting state networks (Miller et al., 2016). Dual regression was performed separately for rfMRI and task residual data after normalization to MNI space using all 100 canonical group components to obtain participant-specific resting state time series (Nickerson et al., 2017). From the 55 signal component dual regression time series, we estimated covariance matrices using Ledoit-Wolf shrinkage (Ledoit and Wolf, 2004) and used tangent-space embedding to transform the matrices into a Euclidean space (Dadi et al., 2019; Sabbagh et al., 2019; Pervaiz et al., 2020; Pennec et al., 2006; Varoquaux et al., 2010; Ng et al., 2014). We then vectorized the matrices' lower triangles, obtaining 1485 connectivity features.

2.3.1. Combining rfMRI and task residual data

We separately calculated the feature vectors from rfMRI and task residual data and performed joint tangent-space embedding. To test whether combining rfMRI and task residual connectivity information improved prediction accuracy, we compared the effect of averaging the resulting feature vectors and of concatenating the vectors, doubling the feature space to 2970 features.

2.3.2. Adding amplitude features

To estimate the network amplitudes, we calculated the standard deviation of each signal network dual regression time series (J. Bijsterbosch et al., 2017), resulting in 55 additional features per scan per participant.

2.4. Predictive model

We used code from the original paper (Dadi et al., 2021), released on GitHub, to implement random forest regression: https://github.com/KamalakerDadi/empirical_proxy_measures, which is based on scikit-learn. Nested 5-fold cross-validation was used to tune the depth of the trees and the number of variables considered for splitting, and the number of trees was fixed to 250. We follow the hyperparameter tuning strategy used in the original paper, as laid out in Table 2. The full sample was randomly split in half to generate a validation dataset for model construction and a held-out generalization dataset. Following the original paper, the validation set was randomly split 100 times into training (90%) and test (10%) subdivisions in a standard shuffle-split (or Monte Carlo) cross-validation. In the generalization dataset, predictions were generated for each of the 100 models from each cross-validation split. Split validation performance was used to generate the violin plots.

Table 2
Random forest hyperparameters and tuning with grid search (5-fold cross-validation). Reproduced from the original paper (Dadi et al., 2021).

Hyperparameter	Values
Impurity criterion	Mean squared error
Maximum tree depth	5, 10, 20, 40, full depth
Fraction of features for split	1, 5, "log 2," "sqrt," "complete"
No. of trees	250

2.5. Biotype definition

We adopted a method that enables brain-phenotype interactions to drive the definition of subtypes. To this end, Canonical Correlation Analysis (CCA (Hotelling, 1936),) was used to separately relate the individual questions that comprise the phenotypes of neuroticism and fluid intelligence to the rfMRI features (Drysdale et al., 2017), and the canonical scores were subsequently used to drive K-means clustering to identify biotype clusters. The pipeline included feature selection, CCA fitting, permutation testing, and estimation of the optimal number of clusters using bootstrapping as explained in further detail below. The full pipeline was repeated across 100 independent cross-validations (90/10 split), and the Adjusted Rand Index (ARI) was calculated using shared participants for each pair of the 100 cross-validations to assess cluster stability (Dinga et al., 2019; Varol et al., 2017). After finding optimal cluster mode numbers and feature selections across 100 cross-validations, participants were each assigned to a biotype by pushing the full dataset through the pipeline described above. The predictive model was then applied separately in each biotype.

2.5.1. Feature selection

Feature selection was performed prior to CCA, keeping the top 25% of features to ensure a ratio of approximately 50 participants per feature, which is important to ensure robustness of the CCA (Helmer et al., 2021). To this end, separate unpaired t-tests were performed for each individual question making up the phenotype to compare participants who scored 0 (corresponding to 'no' in neuroticism and 'incorrect' in fluid intelligence) to participants who scored 1 (corresponding to 'yes' in neuroticism and 'correct' in fluid intelligence). The absolute t-statistics were summed across all questions within the phenotype and the 25% of features with the highest combined t-statistic were entered into the CCA (i.e., 371 features out of 1485). Feature selection was repeated within each cross-validation fold. The 25% features that were most commonly selected across the 100 cross-validations were used in the final run on the full dataset.

2.5.2. CCA and permutation testing

Within each cross-validation fold, CCA was performed between the selected rfMRI features and all separate questions within the phenotype using permCCA (Winkler et al., 2020). For each cross-validation split, permutation testing was performed (2000 permutations) to identify significant canonical covariates; family-wise error (FWE) correction was used to control for multiple comparisons across canonical covariates. Each canonical score has two values per participant related to rfMRI features and phenotype questions, respectively (which were correlated as they represent the canonical correlation). All canonical scores with correlations yielding a FWE-corrected p-value below 0.05 were used for subsequent clustering. CCA fitting and permutation testing was repeated within each cross-validation fold. The distribution of canonical correlations over all CV folds was used to determine significance.

2.5.3. Estimation of optimal number of clusters

Bootstrapping with replacement was performed 1000 times to test cluster solutions with $k = 2^{-10}$. For each bootstrap, the k with the highest silhouette score was recorded. The mode across all 1000 bootstrap was subsequently implemented as the optimal number of clusters. Estimation of the optimal number of clusters was repeated within each cross-validation fold. The optimal number of clusters was that which occurred most commonly across the 100 cross-validations and was set in the final run on the full dataset. K-means clustering with the optimal number of clusters was performed to assign each participant to a cluster. K-means clustering was performed separately in each cross-validation fold and in the full dataset. To assess the stability of the clusters, the Adjusted Rand Index (ARI; implementation from McComb, https://github.com/cmccomb/rand_index (Vinh et al., 2010);) was computed from the cluster assignment for each of the 100 cross-validation folds.

2.6. Statistical comparison of prediction results

A one-sided Welch's *t*-test was used to calculate *p*-values for the comparison between the R^2 distribution from the replication analysis against distributions from additional (manipulation) analyses designed to test the impact of data pollution. We note that *p*-values from differences in generalization R^2 distributions were less than or equal to those corresponding to significant differences in validation R^2 distributions. Unless explicitly stated otherwise, reported *p*-values are computed from validation R^2 distributions.

3. Results

3.1. Replication of previous results

Our replication of both neuroticism and fluid intelligence using a larger UKB sample resulted in significant increases in R^2 (0.01 vs 0.00 for neuroticism, $p = 3e\text{-}10$; 0.03 vs 0.02 for fluid intelligence, $p = 6e\text{-}9$; Fig. 2) for both phenotypes compared with previous work (Dadi et al., 2021). These results are consistent with the gradual monotonic increase with sample size shown in Supplementary Fig. 1 in (Dadi et al., 2021).

3.2. Effect of including additional fMRI features

In addition to rest connectivity features, we considered the addition of signal amplitude and task connectivity residual as prediction features. Furthermore, when considering task connectivity, we compared averaging task connectivity and rest connectivity data against their concatenation.

The inclusion of additional fMRI features yielded modest increases in R^2 for fluid intelligence and small increases in R^2 for neuroticism. Concatenating tangent-space projected residual task features provided the largest increase in R^2 of any single-feature manipulation for fluid intelligence (0.05 vs. 0.03, $p = 5e\text{-}24$; Fig. 2). Adding amplitude data produced a significant increase in fluid intelligence R^2 ($p = 7e\text{-}5$), as did averaging rest data and task residuals ($p < 1e\text{-}10$); combining these manipulations prompted an increase of slightly greater significance ($p = 6e\text{-}17$), though none of these boosted fluid intelligence R^2 above 0.04 (Fig. 2). No significant change in average prediction accuracy for neuroticism was observed when adding amplitude data ($p > 0.5$; Fig. 2). After including task residual data, all manipulations produced significant (if negligible) increases, though averaging ($0.002 < p < 0.01$) was less successful than concatenation ($2e\text{-}16 < p < 2e\text{-}5$). The most significant increase in prediction R^2 occurred when all fMRI features were combined ($p = 2e\text{-}16$; Fig. 2).

Using only signal amplitudes as features yielded prediction accuracies similar to those on full connectivity data in smaller sample sizes (0.02 vs. 0.02 in fluid intelligence; 0.01 vs. 0.00 in neuroticism) in prior work (Dadi et al., 2021). In fact, in the case of neuroticism, amplitude-only feature prediction significantly improved prediction relative to full connectivity in the smaller sample ($p = 6e\text{-}6$). However, rest-amplitude and amplitude-only predictions showed a decrease in generalization vs. validation performance relative to other manipulations (Fig. 2).

3.3. Effect of phenotype averaging

The use of average phenotype values as prediction targets yielded small but significant increases in prediction accuracy. Targeting average fluid intelligence prompted small but significant improvement in prediction accuracy (from 0.03 to 0.04, $p = 3e\text{-}13$; Fig. 2). In fluid intelligence, the combination of all manipulations (including phenotype averaging) gave the largest and most significant improvement, from 0.03 to 0.06 ($p = 7e\text{-}46$; Fig. 2). Targeting average neuroticism yielded a significant improvement in R^2 (0.01–0.02, $p = 2e\text{-}11$; Fig. 2) only when combined with the inclusion of additional resting state features;

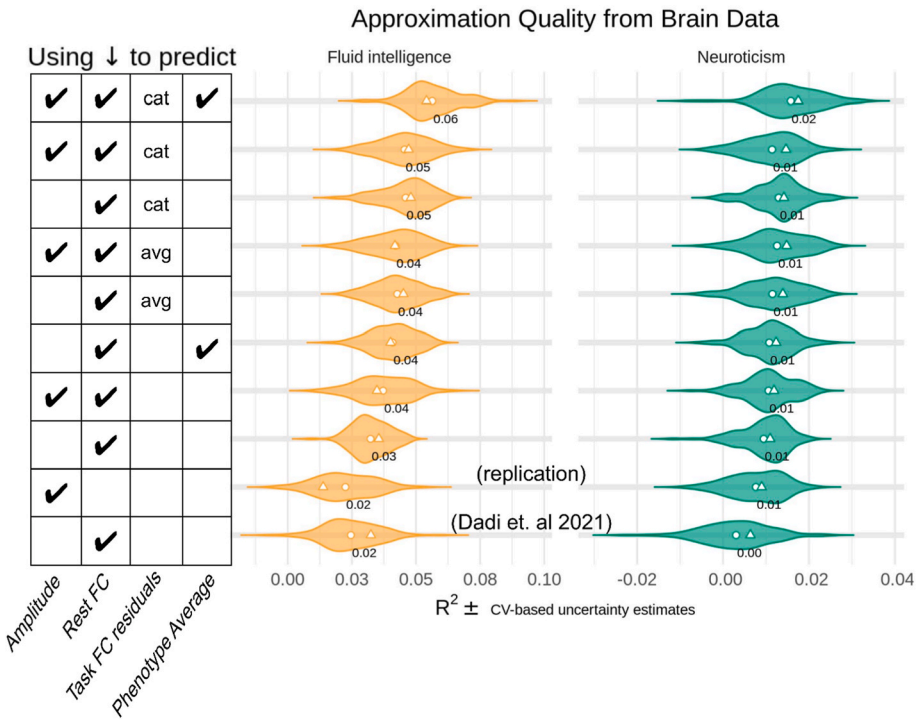

Fig. 2. Including additional resting state features led to improvements in prediction accuracy. We report the R^2 metric to facilitate comparisons across prediction targets. The cross-validation (CV) distribution (100 splits) on the validation dataset is depicted by violin plots. Circles depict the average performance on the validation data across CV-splits, and triangles depict the performance of the average prediction (CV-bagging) on held-out generalization datasets. For reference, the mean performance on the validation set is written on each violin plot. Circles denote mean validation R^2 and triangles represent mean generalization R^2. In the table on the left 'cat' indicates that resting state and task connectivity features were concatenated and 'avg' indicates that resting state and task connectivity features were averaged.

phenotype averaging alone did not produce a significant increase in neuroticism prediction ($p = 0.051$; Fig. 2).

3.4. Effect of flattened (balanced) phenotype distribution

Selecting subjects to generate a flat distribution of phenotype prediction targets yielded small but significant increases in validation R^2 for fluid intelligence (0.04 vs. 0.03, $p = 0.002$), but not for neuroticism (0.00 vs 0.01, $p \sim 1$) when compared with predictions on the same features (Fig. 3). In fluid intelligence, models trained on balanced data showed a much more significant increase in generalization R^2 than in validation R^2, which holds both when using all features (0.06 vs. 0.05, $p = 1e\text{-}55$) and considering only rest FC data (0.04 vs. 0.03, $p = 1e\text{-}40$).

In comparison to unmodified phenotype distributions of matched sample size, phenotype flattening yielded significant increases in generalization performance across all interventions in both neuroticism and fluid intelligence (see supplement and Fig. S1 for details).

3.5. Effect of separate predictions in biotypes

3.5.1. Evaluation of biotype stability

Fig. 4 reports distributions of canonical correlations and their corresponding p-values. Over each of 100 independent cross-validation folds, p-values were computed and family-wise error (FWE) corrected from 2000 permutations of the data within each fold. Despite the fact that canonical fluid intelligence correlations were generally lower than neuroticism (Fig. 4c-d), a larger number of them were significant (Fig. 4a-b). A canonical coefficient is "significant" if it yields a p-value

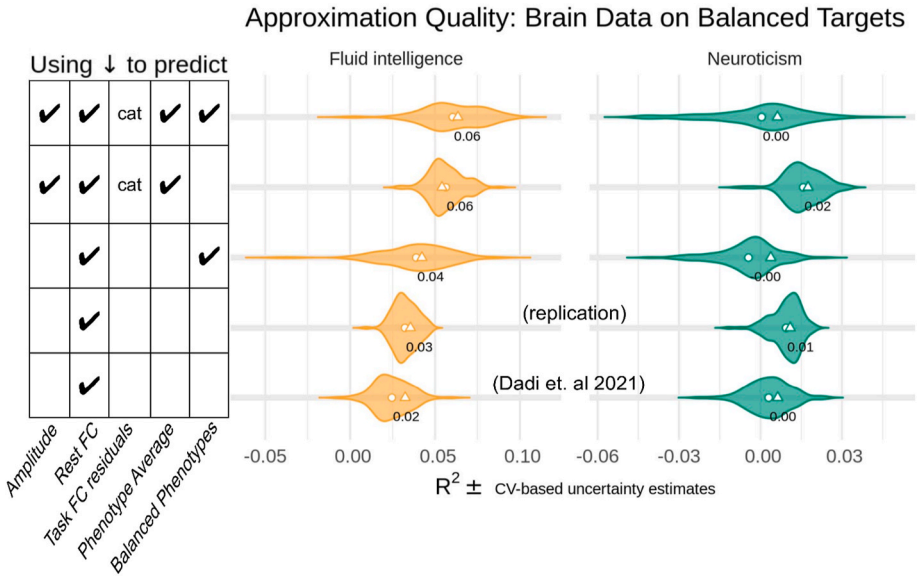

Fig. 3. Averaging the phenotypes and balancing the sample led to additional improvements in prediction accuracy for fluid intelligence, over and above improvements achieved by including additional resting state measures. We report the R^2 metric to facilitate comparisons across prediction targets. The cross-validation (CV) distribution (100 splits) on the validation dataset is depicted by violin plots. Circles depict the average performance on the validation data across CV-splits, and triangles depict the performance of the average prediction (CV-bagging) on held-out generalization datasets. For reference, the mean performance on the validation set is written on each violin plot. Circles denote mean validation R^2 and triangles represent mean generalization R^2. In the table on the left, 'cat' indicates that resting state and task connectivity features were concatenated.

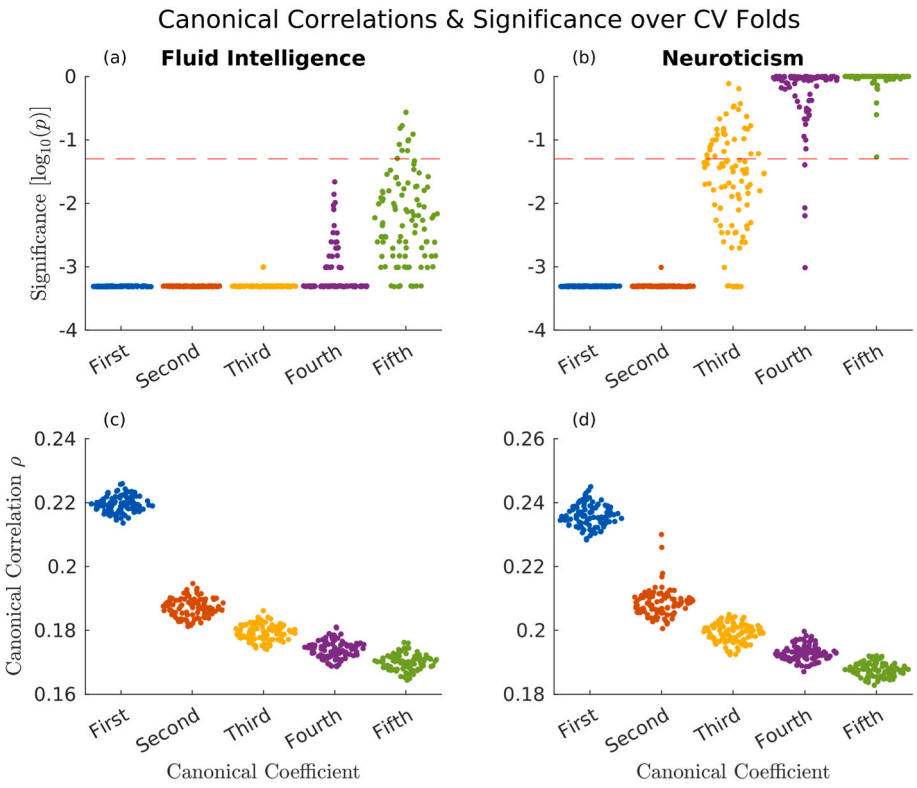

Fig. 4. A swarm plot of distributions (over 100 CV folds) of canonical coefficients (c for fluid intelligence & d for neuroticism) and their corresponding p-values after family-wise error correction (a for fluid intelligence & b for neuroticism). The significance threshold is denoted by the dotted red line in (a) and (b); a coefficient was significant if it yielded p-values below the significance threshold in all CV folds. Four components of the fluid intelligence CCA were significant, and two components of the neuroticism CCA were significant. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

under 0.05 (after family-wise error correction) in every CV fold; this occurs in four canonical components in fluid intelligence and two in neuroticism (Fig. 4a-b). In addition, note that the fifth canonical coefficient in fluid intelligence gives *p*-values under 0.05 in 90% of CV folds; thus, fluid intelligence clustering was performed on 5-dimensional data in 90% of CV folds and 4-dimensional data in 10% of folds. Similarly, slightly more than half of neuroticism clustering was performed on 3-dimensional data.

Across the 1000 bootstrap iterations for *k*-means clustering, the mode cluster number with the highest silhouette score was two for fluid intelligence and three for neuroticism. The Adjusted Rand Index (ARI) across cross-validation folds indicated moderate stability for both fluid intelligence clusters (ARI = 0.76) and all three neuroticism clusters (ARI = 0.72). The small difference in cluster stability may be partly driven by the greater consistency of clustering dimensions in fluid intelligence (5 dimensions for ~90% of CV folds) than in neuroticism (3 for ~35%) across CV folds (Fig. 4a-b). Sample sizes for the resulting subtypes are reported in Fig. 1D and H.

3.5.2. *Effect of biotype on prediction*

Repeating the random forest regression separately within each subtype did not result in improved prediction accuracy in either fluid intelligence or neuroticism (relative to full-data prediction). Prediction accuracy suffered in comparison to prediction from the full dataset (Fig. 5), but was comparable to accuracies observed in prior work on a smaller sample (Dadi et al., 2021). Notably, the generalization (held--out) prediction in Neuroticism subtypes 1 and 3 was significantly higher than in prior work on a smaller sample (p = 0.01 and p = 0.006,

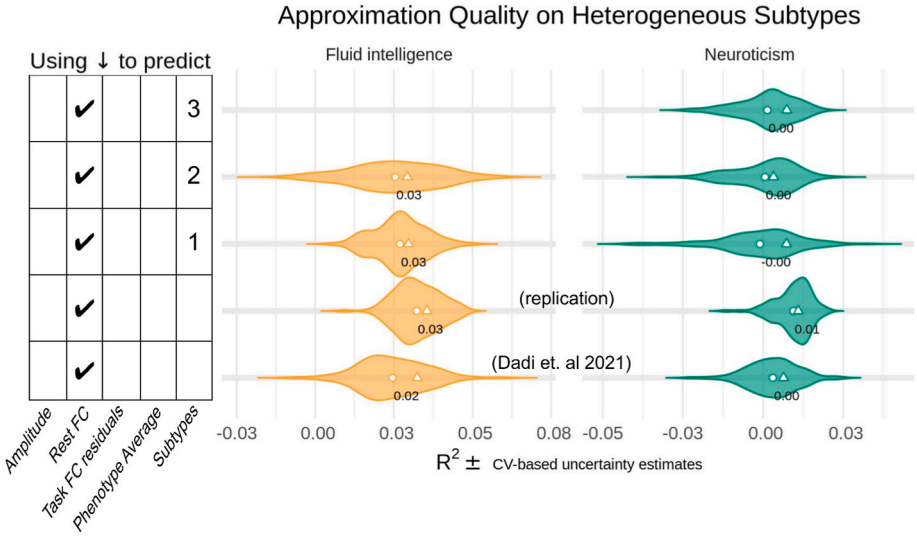

Fig. 5. Identifying homogeneous subgroups did not result in major improvements in prediction accuracy. We report the R^2 metric to facilitate comparisons across prediction targets. The cross-validation (CV) distribution (100 splits) on the validation dataset is depicted by violin plots. Circles depict the average performance on the validation data across CV-splits, and triangles depict the performance of the average prediction (CV-bagging) on held-out generalization datasets. Circles denote mean validation R^2 and triangles represent mean generalization R^2. For reference, the mean performance on the validation set is written on each violin plot.

respectively; Fig. 5). By contrast, prediction on the validation set did not significantly improve.

4. Discussion

In this paper we replicated and extended previous efforts (Dadi et al., 2021) to predict trait phenotypes of neuroticism and fluid intelligence based on rfMRI neuroimaging features. Our goal was to test different manipulations of the data to address data pollution challenges (De Nadai et al., 2022). Our replication findings using a larger sample achieved higher R^2 without any further manipulations, pointing to the benefit of larger sample sizes than N = 10,000. Beyond this boost from sample size, our results revealed that most manipulations led to small but significant increases in R^2, which were largely additive when multiple manipulations were combined (Figs. 2–4). The manipulations to address data pollution largely fall into three categories: input fMRI feature additions, target prediction phenotype averaging, and input participant changes. In the remainder of the discussion, we summarize the results and provide recommendations for each of these categories.

Firstly, we tested different manipulations of the input fMRI feature data by combining resting state and task residual datasets, and incorporating additional amplitude rfMRI features. Our findings showed small but significant increases in R^2 (Fig. 2). Averaging connectivity features between resting state and task residuals (i.e., keeping the total number of features the same) led to increases in R^2, suggesting the presence of shared trait-relevant information. Notably, the amount of combined functional MRI data per person in the UKB (5 min rfMRI + 5 min tfMRI (Miller et al., 2016);) is lower than other datasets such as the Human Connectome Project (1 h rfMRI + 1 h tfMRI (Glasser et al., 2016);), and substantially lower than densely sampled datasets such as the Midnight Scan Club (5 h rfMRI + 6 h tfMRI (Gordon et al., 2017);), so it is possible that further gains may be achievable with more data. However, since the UKB test-retest reliability estimates for rfMRI measures (Dutt et al., 2021) are in line with other datasets (Noble et al., 2019), further gains may be limited. Although the addition of task fMRI features led to improvements, it required substantially more time and computational resources than the other manipulations discussed below. Future work will need to carefully consider the trade-off between maximizing prediction accuracy and investment of time and other resources.

Secondly, we tested whether averaging the target phenotype for prediction across multiple measurements to obtain a less noisy estimate of the underlying trait improved R^2. The results showed improvements, which were additive when combined with other manipulations (Fig. 2). Indeed, the largest improvement for neuroticism was obtained when combining extended fMRI features and averaging the target phenotype, which increased R^2 from 0.00 in Dadi et al. to 0.02 for Neuroticism. If multiple measurements are readily available (as in the UKB), we recommend leveraging the data and using averaged phenotypes as an easy and low-cost manipulation that leads to consistent improvements in R^2.

Thirdly, we tested two options to alter the participant sample: balancing the distribution of the target phenotype through undersampling the majority scores and splitting the sample into separate biological subtypes. Notably, each of these approaches substantially reduced the sample size from 18,000 to samples ranging from 3000 to 12,000 participants. Despite reducing sample sizes, the balanced sample manipulation yielded highly significant improvements in fluid intelligence generalization R^2 (Fig. 3), highlighting the importance of a balanced sample over and above pure sample size. The largest improvement for fluid intelligence was obtained when combining extended fMRI features, averaging the target phenotype, and using a balanced sample, which tripled R^2 from 0.02 (in Dadi et al.) to 0.06 for fluid intelligence. Splitting the sample up into homogeneous subtypes did not result in major improvements in R^2 (Fig. 4), which may be driven by smaller sample sizes. Importantly, our findings showed improvements in the held-out generalization sample (but not the CV validation fold) in two out of three neuroticism subtypes, suggesting that subtypes may be valuable to improve homogeneity. Further research is therefore warranted to further develop and evaluate subtyping manipulations.

5. Conclusion

We tested various data manipulations to address data pollution in an attempt to improve the prediction accuracy of neuroticism and fluid intelligence traits based on large-scale neuroimaging data from the UK Biobank. Most data manipulations led to a small but significant increase in R^2. Combining all data manipulations achieved a three-fold increase in R^2 for fluid intelligence and a nonzero predictive accuracy for neuroticism compared to prior work (Dadi et al., 2021).

Accessibility and data repository

All analysis code for this article is publicly available at: https://github.com/PersonomicsLab/popmodel_ML_MH. UK Biobank data (Miller et al., 2016; Sudlow et al., 2015) are available following an access application process, for more information please see: https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access.

Ethics statement

The authors report no conflicting interests.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the code and data in the Attach Files step.
UK BioBank (Reference data) (GitHub)

Acknowledgements

We are grateful to UK Biobank and the UK Biobank participants for making the resource data possible, and to the data processing team at Oxford University for producing the shared processed data. This research was performed under UK Biobank application number 47267. This research was supported by the NIH (1 R34 NS118618-01), the McDonnell Center for Systems Neuroscience, and the Division of Biology and Biomedical Sciences.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ynirp.2023.100163.

References

Alfaro-Almagro, Fidel, Jenkinson, Mark, Bangerter, Neal K., Andersson, Jesper L.R., Griffanti, Ludovica, Douaud, Gwenaëlle, Sotiropoulos, Stamatios N., et al., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166 (February), 400–424.

Bijsterbosch, Janine D., Valk, Sofie L., Wang, Danhong, Glasser, Matthew F., 2021. Recent developments in representations of the connectome. Neuroimage 243 (November), 118533.

Bijsterbosch, Janine, Harrison, Samuel, Duff, Eugene, Alfaro-Almagro, Fidel, Woolrich, Mark, Smith, Stephen, 2017. Investigations into within- and between-subject resting-state amplitude variations. Neuroimage 159 (October), 57–69.

Bijsterbosch, Janine, Harrison, Samuel J., Jbabdi, Saad, Woolrich, Mark, Beckmann, Christian, Smith, Stephen, Duff, Eugene P., 2020. Challenges and future

directions for representations of functional brain organization. Nat. Neurosci. October, 1–12.
Dadi, Kamalaker, Rahim, Mehdi, Abraham, Alexandre, Chyzhyk, Darya, Milham, Michael, Bertrand, Thirion, Varoquaux, Gaël, Alzheimer's Disease Neuroimaging Initiative, 2019. Benchmarking functional connectome-based predictive models for resting-state fMRI. Neuroimage 192 (May), 115–134.
Dadi, Kamalaker, Varoquaux, Gaël, Houenou, Josselin, Bzdok, Danilo, Bertrand, Thirion, Engemann, Denis, 2021. Population modeling with machine learning can enhance measures of mental health. GigaScience 10 (10). https://doi.org/10.1093/gigascience/giab071.
Dal Pozzolo, Andrea, Olivier Caelen, Bontempi, Gianluca, 2015. When is undersampling effective in unbalanced classification tasks?. In: Machine Learning and Knowledge Discovery in Databases, 200–215. Springer International Publishing.
De Nadai, A.S., Hu, Y., Thompson, W.K., 2022. Data Pollution in Neuropsychiatry-An Under-Recognized but Critical Barrier to Research Progress. JAMA Psychiatr 79 (2), 97–98. https://doi.org/10.1001/jamapsychiatry.2021.2812.
Dinga, Richard, Schmaal, Lianne, Penninx, Brenda W.J. H., Jose van Tol, Marie, Veltman, Dick J., van Velzen, Laura, Mennes, Maarten, van der Wee, Nic J.A., Marquand, Andre F., 2019. Evaluating the Evidence for Biotypes of Depression: methodological Replication and Extension of Drysdale et Al, 2017 Neuroimage: Clinic 22 (January), 101796.
Drysdale, Andrew T., Grosenick, Logan, Downar, Jonathan, Dunlop, Katharine, Mansouri, Farrokh, Yue, Meng, Fetcho, Robert N., et al., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. 23 (1), 28–38.
Dutt, Rosie K., Kayla Hannon, Ty O. Easley, Griffis, Joseph C., Zhang, Wei, Bijsterbosch, Janine D., 2021. Mental health in the UK Biobank: a roadmap to self-report measures and neuroimaging correlates. Hum. Brain Mapp. https://doi.org/10.1002/hbm.25690. October.
Elliott, Maxwell L., Knodt, Annchen R., Cooke, Megan, Kim, M. Justin, Melzer, Tracy R., Ross, Keenan, Ireland, David, et al., 2019. General functional connectivity: shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. Neuroimage 189 (April), 516–532.
Fair, Damien A., Schlaggar, Bradley L., Cohen, Alexander L., Miezin, Francis M., Dosenbach, Nico U.F., Wenger, Kristin K., Fox, Michael D., Snyder, Abraham Z., Raichle, Marcus E., Petersen, Steven E., 2007. A method for using blocked and event-related fMRI data to study 'resting state' functional connectivity. Neuroimage 35 (1), 396–405.
Ferrante, Michele, David Redish, A., Oquendo, Maria A., Averbeck, Bruno B., Kinnane, Megan E., Gordon, Joshua A., 2019. Computational psychiatry: a report from the 2017 nimh workshop on opportunities and challenges. Mol. Psychiatr. 24 (4), 479–483.
Gao, Siyuan, Greene, Abigail S., Todd Constable, R., Scheinost, Dustin, 2019. Combining multiple connectomes improves predictive modeling of phenotypic measures. Neuroimage 201 (November), 116038.
Glasser, Matthew F., Smith, Stephen M., Marcus, Daniel S., Andersson, Jesper L.R., Auerbach, Edward J., Behrens, Timothy E.J., Coalson, Timothy S., et al., 2016. The human connectome project's neuroimaging approach. Nat. Neurosci. 19 (9), 1175–1187.
Gordon, Evan M., Laumann, Timothy O., Gilmore, Adrian W., Newbold, Dillan J., Greene, Deanna J., Berg, Jeffrey J., Mario, Ortega, et al., 2017. Precision functional mapping of individual human brains. Neuron 95 (4), 791–807.e7.
Helmer, Markus, Warrington, Shaun, Mohammadi-Nejad, Ali-Reza, Ji, Jie Lisa, Howell, Amber, Rosand, Benjamin, Anticevic, Alan, Sotiropoulos, Stamatios N., Murray, John D., 2021. On stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. bioRxiv. https://doi.org/10.1101/2020.08.25.265546.
Hotelling, Harold, 1936. Relations between two sets of variates. Biometrika 28 (3–4), 321–377.
Ledoit, Olivier, Wolf, Michael, 2004. Honey, I shrunk the sample covariance matrix. J. Portfolio Manag. 30 (4), 110–119.
Lyall, Donald M., Cullen, Breda, Allerhand, Mike, Smith, Daniel J., Mackay, Daniel, Evans, Jonathan, Anderson, Jana, et al., 2016. Cognitive test scores in UK Biobank: data reduction in 480,416 participants and longitudinal stability in 20,346 participants. PLoS One 11 (4), e0154222.
Marek, Scott, Tervo-Clemmens, Brenden, Calabro, Finnegan J., Montez, David F., Kay, Benjamin P., Hatoum, Alexander S., Rose Donohue, Meghan, et al., 2020. Towards Reproducible Brain-wide Association Studies. https://doi.org/10.1101/2020.08.21.257758.
Marek, Scott, Tervo-Clemmens, Brenden, Calabro, Finnegan J., Montez, David F., Kay, Benjamin P., Hatoum, Alexander S., Rose Donohue, Meghan, et al., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603 (7902), 654–660.
McKelvie, S.J., 2004. Is the Neuroticism Scale of the Eysenck Personality Inventory contaminated by response bias? Pers. Indiv. Differ. 36 (4), 743–755. https://doi.org/10.1016/S0191-8869(02)00348-3.
Miller, Karla L., Alfaro-Almagro, Fidel, Bangerter, Neal K., Thomas, David L., Yacoub, Essa, Xu, Junqian, Bartsch, Andreas J., et al., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19 (11), 1523–1536.
Ng, Bernard, Dressler, Martin, Varoquaux, Gaël, Poline, Jean Baptiste, Greicius, Michael, Bertrand, Thirion, 2014. Transport on riemannian manifold for functional connectivity-based classification. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, 405–12. Springer International Publishing.
Nickerson, Lisa D., Smith, Stephen M., Öngür, Döst, Beckmann, Christian F., 2017. Using dual regression to investigate network shape and amplitude in functional connectivity analyses. Front. Neurosci. 11 (March), 115.
Noble, Stephanie, Scheinost, Dustin, Todd Constable, R., 2019. A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. Neuroimage 203 (September), 116157.
Paulus, M.P., Thompson, W.K., 2019. The Challenges and Opportunities of Small Effects: The New Normal in Academic Psychiatry [Review of *The Challenges and Opportunities of Small Effects: The New Normal in. In: Psychiatry]*. JAMA Psychiatry 76, 353–354. Academic. https://doi.org/10.1001/jamapsychiatry.2018.4540.
Paulus, Martin P., Thompson, Wesley K., 2021. Computational approaches and machine learning for individual-level treatment predictions. Psychopharmacology 238 (5), 1231–1239.
Pennec, Xavier, Fillard, Pierre, Ayache, Nicholas, 2006. A riemannian framework for tensor computing. Int. J. Comput. Vis. 66 (1), 41–66.
Pervaiz, Usama, Vidaurre, Diego, Woolrich, Mark W., Smith, Stephen M., 2020. Optimising network modelling methods for fMRI. Neuroimage 211 (May), 116604.
Sabbagh, David, Ablin, Pierre, Varoquaux, Gael, Gramfort, Alexandre, Engemann, Denis A., 2019. Manifold-regression to predict from MEG/EEG brain signals without source modeling. In: Advances in Neural Information Processing Systems, 32. In: https://proceedings.neurips.cc/paper/2019/file/d464b5ac99e74462f321c06ccacc4bff-Paper.pdf.
Smith, Daniel J., Nicholl, Barbara I., Cullen, Breda, Martin, Daniel, Zia Ul-Haq, Evans, Jonathan, Gill, Jason M.R., et al., 2013. Prevalence and characteristics of probable major depression and bipolar disorder within UK Biobank: cross-sectional study of 172,751 participants. PLoS One 8 (11), e75362.
Sudlow, Cathie, Gallacher, John, Allen, Naomi, Beral, Valerie, Burton, Paul, Danesh, John, Paul, Downey, et al., 2015. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12 (3), e1001779.
Varol, Erdem, Sotiras, Aristeidis, Davatzikos, Christos, Alzheimer's Disease Neuroimaging Initiative, 2017. HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. Neuroimage 145 (Pt B), 346–364.
Varoquaux, Gaël, Baronnet, Flore, Kleinschmidt, Andreas, Fillard, Pierre, Bertrand, Thirion, 2010. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010, 200–208. Springer Berlin Heidelberg.
Vinh, Nguyen Xuan, Epps, Julien, Bailey, James, 2010. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res.: JMLR 11 (95), 2837–2854.
Winkler, Anderson M., Renaud, Olivier, Smith, Stephen M., Nichols, Thomas E., 2020. Permutation inference for canonical correlation analysis. Neuroimage 220 (October), 117065.
18–24 Jul Yang, Yuzhe, Zha, Kaiwen, Chen, Yingcong, Wang, Hao, Katabi, Dina, 2021. Delving into deep imbalanced regression. In: Meila, Marina, Zhang, Tong (Eds.), Proceedings of the 38th International Conference on Machine Learning, 139. Proceedings of Machine Learning Research. PMLR, 11842–51.