NeuroImage 256 (2022) 119275

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Reporting details of neuroimaging studies on individual traits prediction: A literature survey

Andy Wai Kan Yeung[a,*], Shammi More[b,c], Jianxiao Wu[b,c], Simon B. Eickhoff[b,c,*]

[a] *Oral and Maxillofacial Radiology, Applied Oral Sciences and Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong, China*
[b] *Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany*
[c] *Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany*

ARTICLE INFO

Keywords:
Individual trait
Prediction
Neuroimaging
Predictive modeling
Survey

ABSTRACT

Using machine-learning tools to predict individual phenotypes from neuroimaging data is one of the most promising and hence dynamic fields in systems neuroscience. Here, we perform a literature survey of the rapidly work on phenotype prediction in healthy subjects or general population to sketch out the current state and ongoing developments in terms of data, analysis methods and reporting. Excluding papers on age-prediction and clinical applications, which form a distinct literature, we identified a total 108 papers published since 2007. In these, memory, fluid intelligence and attention were most common phenotypes to be predicted, which resonates with the observation that roughly a quarter of the papers used data from the Human Connectome Project, even though another half recruited their own cohort. Sample size (in terms of training and external test sets) and prediction accuracy (from internal and external validation respectively) did not show significant temporal trends. Prediction accuracy was negatively correlated with sample size of the training set, but not the external test set. While known to be optimistic, leave-one-out cross-validation (LOO CV) was the prevalent strategy for model validation ($n = 48$). Meanwhile, 27 studies used external validation with external test set. Both numbers showed no significant temporal trends. The most popular learning algorithm was connectome-based predictive modeling introduced by the Yale team. Other common learning algorithms were linear regression, relevance vector regression (RVR), support vector regression (SVR), least absolute shrinkage and selection operator (LASSO), and elastic net. Meanwhile, the amount of data from self-recruiting studies (but not studies using open, shared dataset) was positively correlated with internal validation prediction accuracy. At the same time, self-recruiting studies also reported a significantly higher internal validation prediction accuracy than those using open, shared datasets. Data type and participant age did not significantly influence prediction accuracy. Confound control also did not influence prediction accuracy after adjusted for other factors. To conclude, most of the current literature is probably quite optimistic with internal validation using LOO CV. More efforts should be made to encourage the use of external validation with external test sets to further improve generalizability of the models.

Introduction

Individual traits prediction (e.g. cognition abilities, personality traits, emotional feeling, and motor performance) using neuroimaging data is an upcoming hotspot in cognitive neuroscience (Shen et al., 2017; Sui et al., 2020). The term prediction refers to the ability to predict outcomes successfully in data sets other than the original one used to construct the model (Poldrack et al., 2020). It is better for translational or prediction purposes than the traditional univariate brain mapping analysis, as the latter focused on within-sample fit of correlational relationships that tends to be overfitting and not generalizable (Poldrack et al., 2020). The overall scheme usually begins with collecting structural or functional (resting-state or task-induced) neuroimaging data and personal trait measures from a large sample. Then the neuroimaging data should be preprocessed and entered into a machine-learning model. The model will be trained to find out the link between the neuroimaging data (brain features) and the personal traits. Finally, the trained model can be generalized to predict the traits in a new sample. Its accuracy can be computed by comparing with the ground truth (reality) (Eickhoff and Langner, 2019). In short, there are four stages: model building, internal validation, external validation, and generalizability and transposability (Bzdok and Ioannidis, 2019). There are many ap-

[*] Corresponding authors at: Oral and Maxillofacial Radiology, Applied Oral Sciences and Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong, China (Andy Wai Kan Yeung); Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany (Simon B. Eickhoff)

E-mail addresses: ndyeung@hku.hk (A.W.K. Yeung), Simon.Eickhoff@uni-duesseldorf.de (S.B. Eickhoff).

https://doi.org/10.1016/j.neuroimage.2022.119275.
Received 28 February 2022; Received in revised form 27 April 2022; Accepted 29 April 2022
Available online 2 May 2022.

proaches to individual traits prediction. One famous approach is called the connectome-based predictive modeling (CPM) approach, developed by (Finn et al., 2015), the term CPM established by (Rosenberg et al., 2016), its protocol published by (Shen et al., 2017), codes deposited at https://github.com/YaleMRRC/CPM. Moreover, (Finn et al., 2015) introduced the now widely used Shen 268 atlas, which was produced based on the parcellation method and the 100-, 200-, and 300-node atlases introduced by (Shen et al., 2013). In that study by Rosenberg et al., it was reported that functional connectivity derived from task-induced functional magnetic resonance imaging (fMRI) could be used to train a model that predicted a previously unseen individual's performance in sustained attention and even symptoms of attention deficit hyperactivity disorder based on his/her resting-state fMRI signals (Rosenberg et al., 2016). The predictive modeling field in neuroscience has seen a rapid growth and accumulated many papers since then.

Of course, there are a number of factors that affect the validity or generalizability of predictive models in neuroimaging, spanning from sample size, processing, features, learning, to validation. To begin with, it was recommended that a dataset of over 100 individuals should be used for predictive modeling (Scheinost et al., 2019). Data from (He et al., 2020) even suggested that 500–1000 subjects should be the minimum. Small sample sizes would lead to underestimated errors and vibration effects, meaning that methodological choices could have a drastic impact on the analysis outcome based on few samples (Varoquaux, 2018). Subject recruitment and financial constraints could be potential issues, and might be circumvented by the use of large, open, shared datasets as training or test set. During data processing, potential confounding factors should be accounted for, such as physiological and head motion artifacts (Murphy et al., 2013). At the stage of features input, one needs to consider what data to be entered. For instance, for a model that predicts behavior based on brain connectivity data, connectomes from multiple sources could improve the prediction accuracy compared to a single connectome (Gao et al., 2019). Finally, external validation is the best practice, meaning testing the model with an independently collected (external) data set (Scheinost et al., 2019). Out-of-sample generalization and later cross-validation (CV) is less ideal, as the portion of the sample taken out from the same dataset will inevitably share similar subject and imaging features with the training set and create bias. Since generally it is relatively difficult to obtain a separate test set, doing a CV has been a popular approach, meaning that the whole dataset is divided into subsets that train and test the model respectively. CV is generally fine, but it should be noted that CV in small samples may render the models too optimistic (Whelan and Garavan, 2014).

In this work, we performed an updated general literature survey on the study design and analytic pipeline of the individual traits prediction among healthy individuals or general population (not purely clinical), and aimed to evaluate the published studies on individual traits prediction based on regression, to reveal if their generalizability could be undermined by the caveats mentioned above.

Methods

Literature search strategy

PubMed and Web of Science Core Collection online databases were queried on 16 December 2021 with the following search string: ((("machine learning") OR ("predict* model*") OR ("support vector machine*") OR ("LASSO*") OR ("elastic net*") OR ("random forest*") OR ("cross validat*") OR ("artificial intelligen*")) AND ((brain behavior*) OR (brain behavior*) OR (neuromarker*) OR (brain biomarker*) OR ("individual difference*"))). The search covered "all fields" for PubMed and "title/abstract/keywords" for Web of Science. We also performed reference tracing from the yielded publications and previous review articles. A total of 7153 publications were identified after removing duplicates. The full text of these them were inspected and publications were excluded due to the following reasons: irrelevant (e.g. within-sample correlation instead of predictive modeling; $n = 6018$), classification instead of regression (e.g. sex classification; $n = 692$), involved patients only ($n = 154$), review/opinion paper ($n = 121$), method papers ($n = 17$), age prediction instead of individual traits ($n = 43$), conference abstract without full text ($n = 0$), and unspecific phenotype ($n = 0$). In the end, 108 articles entered the survey (Supplementary Table 1). For completeness, a list of excluded papers could be found in Supplementary Table 2.

To assess the reporting details and identify patterns/trends among these papers, we examined the content of them carefully. The surveyed contents involved sampling, processing strategy, feature selection, learning algorithm, and validation. Sample size is a critical aspect of the papers, as smaller samples may be underpowered and overfit the models, and hence producing false positives (Varoquaux et al., 2017). Meanwhile, papers dealing with large open-source neuroimaging datasets should report the dataset details well enough, as each dataset has its unique demographic factor, imaging and behavioral measures (Horien et al., 2021). For processing, accounting for confounds such as head motion is an important step in modeling, as their presence may make the model less meaningful (Rao et al., 2017). Other details of processing such as dimensionality reduction, feature selection, learning algorithm, hyperparameter tuning, and validation strategy were also evaluated and recorded as these are important for fellow researchers to replicate their results. Finally, prediction accuracy was noted to evaluate the model performances. Because of these rationales, the parameters recorded for each study were listed in the following paragraph.

Parameters recorded

The following parameters were recorded for each study: sample size (training set and test set), data source, type of subject (minor vs adult), amount of data for each subject (number of volumes), input data (e.g. what kind of connectome and matrix size), data type (task, rest, naturalistic, vs structural), target phenotype (e.g. intelligence), processing strategy, reference to the Yale approach (connectome-based predictive modeling, c.f. (Finn et al., 2015; Rosenberg et al., 2016; Shen et al., 2017)), brain atlas referred to, confounding variables accounted for (e.g. head motion), dimensionality reduction if relevant, feature selection, learning algorithm, hyperparameter tuning, validation strategy (e.g. external validation or CV), and prediction accuracy (from internal and external validation, respectively). The temporal trends of the statistics were tested across studies if they were continuous variables (e.g. prediction accuracy), and across years if they were categorical (e.g. ratio of studies using external test set). Additional analyses were performed for fMRI and structural MRI (sMRI) studies separately.

Results

General bibliographic information

The annual publication count showed a sharp increase in year 2018 (Fig. 1A). Prior to 2018, there were fewer than 6 papers published per year.

Prediction accuracy

In brief, 81 studies reported Pearson r as the prediction accuracy value from internal validation, and 16 studies reported so from external validation. The accuracy from internal validation ranged from 0.098 to 0.978, whereas the accuracy from external validation ranged from 0.220 to 0.736. Though the prediction accuracy from either validation method seemed to show a slight decreasing trend by year (Fig. 1B), no significant linear correlation was observed (Pearson correlation test, internal validation: $n = 81$, $r = -0.201$, $p = 0.071$; external validation: $n = 16$, $r = -0.482$, $p = 0.059$). For the studies that did not report any Pearson r as the prediction accuracy, Spearman rho was the most popular metric

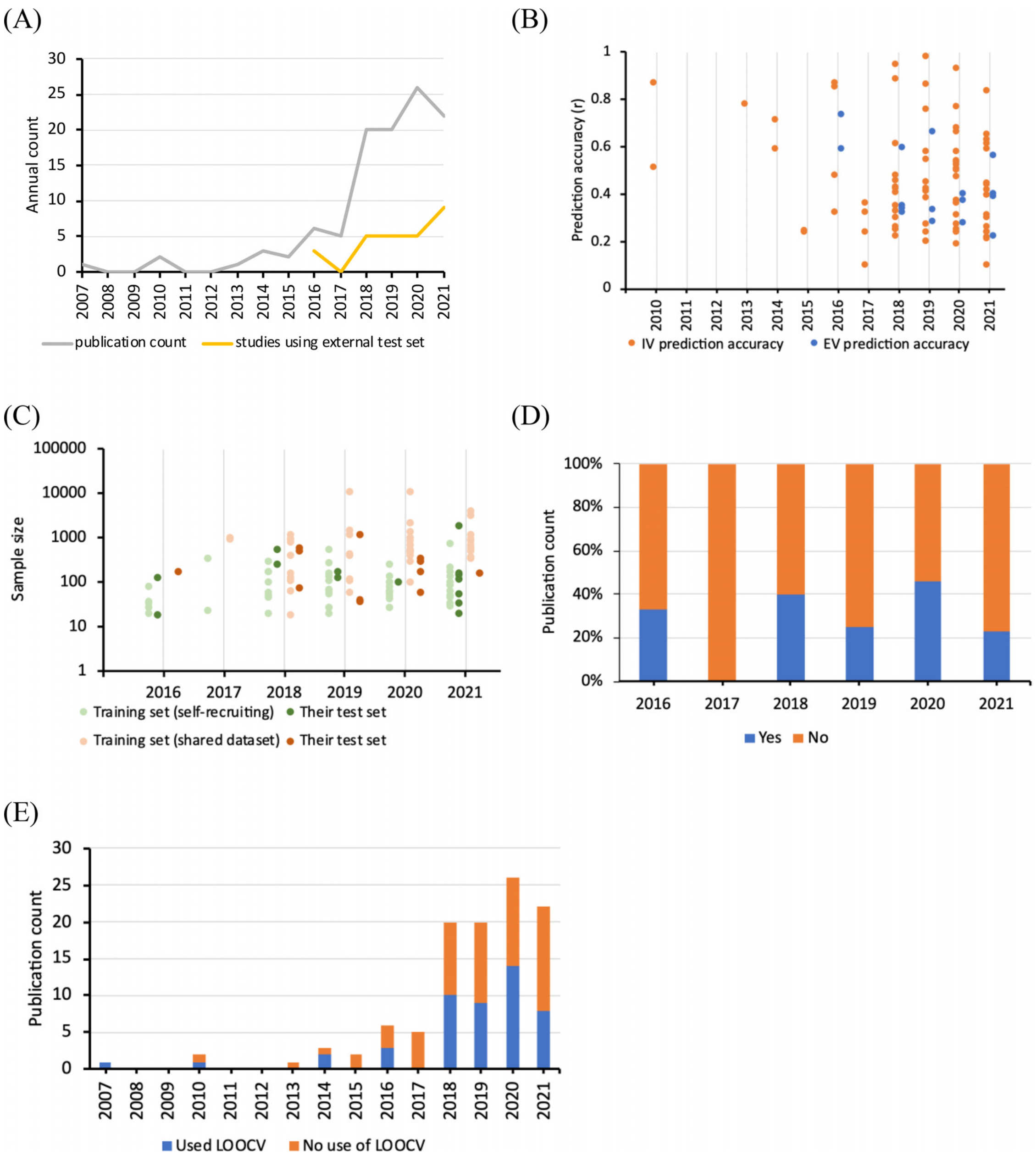

Fig. 1. Graphical summary of the surveyed articles. (A) Annual publication count and studies using external test set. (B) Prediction accuracy of models using internal validation (IV) and external validation (EV) respectively. (C) Sample sizes of training set and external test set. Studies published in 2015 or before were not plotted as they did not recruit external test set. (D) Proportion of studies mentioning that they followed the Yale approach ("connectome-based predictive modeling" [CPM], e.g. from (Finn et al., 2015; Rosenberg et al., 2016; Shen et al., 2017)). (E) The use of leave-one-out cross-validation (LOO CV).

(n = 11). Other reported metrics included standardized mean squared error, mean absolute error (MAE), root mean square error (RMSE), prediction R^2, and adjusted R^2.

Sample size

Before year 2016, the few surveyed studies only recruited subjects for their training set, without external test set or involvement of open, shared dataset. The mean sample size of their training set was 64, 25.5, 40, 49, and 185.5 for year 2007, 2010, 2013, 2014, and 2015, respectively. Fig. 1C illustrated the sample size since year 2016. For training set, self-recruiting studies had a mean sample size of 108 during the period of 2016–2021, whereas studies using open, shared datasets had a much larger mean sample size of 1140. For test set, however, the former group and the latter group had a similar mean sample size (251 vs 278). The sample size of self-recruiting studies did not show significant linear correlation with year (Pearson correlation test, training set: $n = 52$, $r = 0.078$, $p = 0.581$; test set: $n = 14$, $r = 0.158$, $p = 0.590$). The same held true for studies using open, shared datasets (training set: $n = 45$, $r = 0.106$, $p = 0.489$; test set: $n = 12$, $r = -0.093$, $p = 0.773$).

Selection of data source and data type

In terms of data source of the training set, 61 studies recruited their own subjects, whereas the Human Connectome Project (HCP) was used by 21 studies (Table 1). HCP may refer to various versions of the HCP dataset, such as "unrelated 100", S500, S900, and S1200. Readers should be aware that the more recent datasets (e.g. S1200) not only had a larger sample size, but also contained updated data on family structures of the subjects (e.g., relationships as twins or non-twin siblings, but excluding

Table 1
Data sources of the training set.

Data source	Number of studies
Recruited subjects	61
HCP S1200	6
HCP S900	6
HCP S500	4
ABCD	4
PNC	4
HCP (unclear version)	3
SLIM	3
Previously trained model	2
HCP "unrelated 100"	2
UESTC	2
UKB	2

ABCD, Adolescent Brain Cognitive Development Study. HCP, Human Connectome Project. PNC, Philadelphia Neurodevelopmental Cohort. SLIM, SLIM dataset from Liu et al. (2017). UESTC, University of Electronic Science and Technology of China. UKB, UK Biobank. The data sources below were each used once and hence not listed in the table: ABIDE-II (Autism Brain Imaging Data Exchange), ADNI-2 (Alzheimer's Disease Neuroimaging Initiative), ADNI-GO, AHAB-2 (Adult Health and Behavior project—Phase 2), ATR dataset from Yamashita et al. (2015), BBP (Behavioral Brain Research Project of Chinese Personality), BCAS (Brain and Cognition Aging Study), CamCAN (Cambridge centre for Ageing and Neuroscience), CBDC (Cognition and Brain Development in Children), DIAMOND (Dimensions of Affect, Mood, and Neural Circuitry Underlying Distress Study), Duke Neurogenetics Study, GUSTO (Growing Up in Singapore Towards healthy Outcomes), IMAGEN dataset from Schumann et al. (2010), NKI-RS (Nathan Kline Institute Rockland Sample), OASIS-3 (Open Access Series of Imaging Studies), PING (Pediatric Imaging, Neurocognition, and Genetics), PIP (Pittsburgh Imaging Project), TTC (Tokyo TEEN Cohort Study), UNC Early Brain Development Study.

Table 2
Frequency of common types of input data.

Input data	Frequency (n)
Resting state functional connectivity (RSFC) connectome	32
Both task-induced FC and RSFC connectome	15
Task-induced FC connectome	10

For other less common types of input data, please refer to Supplementary Table 1 for the details of each study.

birth order). Therefore, the umbrella term HCP did not necessarily imply an identical sample used across the studies. Among the 21 studies using HCP data, S1200 was the most popular dataset (Table 1). Meanwhile, different target phenotypes were investigated, with some of the recurring ones being fluid intelligence/intelligence quotient (n = 15), attention (n = 12), and memory (n = 11).

Regarding input data, resting state functional connectivity (RSFC) connectome was much more common than task-induced FC connectome or the combination of both (Table 2). The brain atlas used for connectome data were also diverse, with the canonical Shen 268 atlas (Finn et al., 2015; Shen et al., 2013) being most prevalent (Table 3).

Dealing with confounding factors during data processing

Over half of the studies (61 out of 108) did not control for potential confounding factors such as age, sex, head motion. Studies controlled for them mainly entered them as regressors in the regression models.

Varied feature selection and learning

Thirty-two papers (32.3%) published since 2016 followed the Yale approach pioneered by Finn et al. (2015) mentioned in the Introduction, which achieved a brain-behavior prediction by means of an approach called connectome-based predictive modeling (CPM) (Fig. 1D). In year 2020, almost half of the studies followed this approach. It basically involved a linear regression and the feature selection method was typically choosing FCs with significant correlation (e.g. $p < 0.01$) with the predicted measure [and then the sums of selected positive or negative edges (the summary measure), is used as input features for linear regression]. Other papers had very diverse feature selection methods, with two recurring methods including feature selection from regions-of-interest (ROIs, n = 4) and principal component analysis (PCA, n = 2). Some common learning algorithm used by these non-CPM papers were multiple linear regression, relevance vector regression (RVR), support vector regression (SVR), partial least squares regression (PLSR), least absolute shrinkage and selection operator (LASSO), and elastic net. Most studies did not require hyperparameters tuning, and nested k-fold CV [in descending order of frequency: (n = 8) 10-fold, (n = 6) 3-fold, (n = 3) 5-fold, and (n = 2) 20-fold] was the predominant choice. See Supplementary Table 1 for details.

Diversity of validation

For validation strategy, 48 studies (44.4%) involved leave-one-out cross-validation (LOO CV). Twenty-four of these LOO CV papers mentioned they used the Yale approach, suggesting a dependency of this CV strategy on the CPM modeling approach. The annual ratio of studies using LOO CV fluctuated around 50% and showed no obvious trend against year (Pearson correlation test on period 2013–2021, n = 9, r = 0.341, p = 0.369; Fig. 1E). Ten-fold CV and 4-fold CV were involved in 19 and 10 studies respectively. Meanwhile, 27 studies involved external test sets (26 were cross-dataset whereas one was cross-site) and they were published in 2016–2021 (Fig. 1A). The proportion of studies using an external test set has remained largely constant across years with no discernible trend (Pearson correlation test, n = 6, r = 0.037, p = 0.945). Readers should be aware of the few data points used in these two tests.

Potential influencing factors on prediction accuracy

Table 4 shows that sample size could influence the prediction accuracy. Precisely, the smaller the sample size of the training set, the higher the internal validation prediction accuracy was found (n = 81, r = −0.265, p = 0.017). On the contrary, the sample size of the external test set did not show significant correlation with external validation prediction accuracy. Meanwhile, the amount of data from self-recruiting studies (but not studies using open, shared dataset) was positively correlated with internal validation prediction accuracy (n = 30, r = 0.651, p < 0.001). At the same time, self-recruiting studies also reported a significantly higher internal validation prediction accuracy than those using open, shared datasets (Mean ± SD, self-recruiting: n = 46, 0.509 ± 0.229, shared dataset: n = 35, 0.386 ± 0.190, p = 0.012). Besides, internal validation reported higher accuracy than external validation within studies that used both types of validation. Meanwhile, studies using models that were uncontrolled for confounds reported a significantly higher internal validation prediction accuracy than those using models that were controlled for confounds (Mean ± SD, controlled studies: n = 33, 0.395 ± 0.197; uncontrolled studies: n = 48, 0.498 ± 0.228, p = 0.038). Data type and participant age did not significantly influence prediction accuracy. Table 4 shows the detailed results of the statistical tests. It should be noted that some studies may provide more than one data point whereas some studies may have missing data for the statistical tests, and hence the n reported may not correspond to the number of studies involved. Readers should refer to Supplementary Table 3 for the data used.

When the significant factors were considered together by partial correlation tests, it was found that training set sample size remained significant after adjusted for confound control, but became insignificant after considering participant source or amount of data. Meanwhile, participant source remained significant after adjusted for confound control, but became insignificant after considering training set sample size or amount of data. In turn, amount of data for studies recruiting subjects

Table 3
Brain atlases referred by studies using connectome data.

Brain atlas	No. of nodes	Coverage (whole brain, cortex only, cerebellum only, etc.)	Functionally defined vs anatomically defined	Number of studies
Shen 268 atlas, see (Finn et al., 2015; Shen et al., 2013)	268	Whole brain	Functionally defined	29
(Power et al., 2011)	264	Whole brain	Functionally defined	8
(Fan et al., 2016)	246	Whole brain	Anatomically defined	5
Independent component analysis (ICA) components	Variable	Variable	Variable	5
(Dosenbach et al., 2010)	160	Whole brain	Functionally defined	4
(Tzourio-Mazoyer et al., 2002)	116	Cortex only	Anatomically defined	3
(Glasser et al., 2016)	360	Cortex only	Functionally and anatomically defined	3
(Gordon et al., 2016)	333	Cortex only	Functionally defined	3
(Schaefer et al., 2018)	100–1000 (400 version used in 2 studies)	Cortex only	Functionally defined	2
(Desikan et al., 2006)	68	Cortex only	Anatomically defined	2
(Gilmore et al., 2012)	78	Cortex only	Anatomically defined	1
(Diedrichsen, 2006)	28	Cerebellum	Anatomically defined	1
(Fischl et al., 2002)	37 (14 used in 1 study)	Whole brain	Anatomically defined	1
(Buckner et al., 2011)	7 or 17	Cerebellum	Functionally defined	1
(Feng et al., 2019)	52	Whole brain (neonatal)	Anatomically defined	1
(Yeo et al., 2011)	114 (39 used in 1 study)	Cortex only	Functionally defined	1
(Destrieux et al., 2010)	148	Cortex only	Anatomically defined	1

Some studies referred to multiple atlas and they were counted within the table.

Table 4
Influencing factors of prediction accuracy.

Factor	Test	Stat	P value
Sample size	Pearson correlation		
a. of external test set (prediction accuracy from external validation)		$n = 17$, $r = -0.302$ (i.e. larger test set, lower prediction accuracy)	0.239
b. of training set (prediction accuracy from internal validation)		$n = 81$, $r = -0.265$ (i.e. larger test set, lower prediction accuracy)	0.017
Amount of data (total number of volumes per individual)	Pearson correlation		
a. for studies recruiting subjects		$n = 30$, $r = 0.651$ (i.e. more data, higher prediction accuracy)	< 0.001
b. for studies using open, shared dataset		$n = 28$, $r = -0.095$ (i.e. less data, higher prediction accuracy)	0.629
Data type (task, rest, naturalistic, structural vs mixed)	One-way ANOVA	Mean ± SD Task ($n = 20$): 0.510 ± 0.218, rest ($n = 36$): 0.421 ± 0.166, structural ($n = 18$): 0.410 ± 0.270, mixed ($n = 10$): 0.567 ± 0.243 (No study used naturalistic data)	0.129
Participant source (self-recruiting vs open, shared dataset)	T-test	Mean ± SD Self-recruiting ($n = 46$): 0.509 ± 0.229, shared dataset ($n = 35$): 0.386 ± 0.190	0.012
Participant age (involved minor vs adult only)	T-test	Mean ± SD Involved minor ($n = 16$): 0.476 ± 0.233, adult only ($n = 65$): 0.451 ± 0.219	0.688
Control for confounds (yes vs no)	T-test	Mean ± SD Yes ($n = 33$): 0.395 ± 0.197, no ($n = 48$): 0.498 ± 0.228	0.038
Validation type (internal vs external)			
a. for studies that involved both types	Paired *t*-test	Mean ± SD ($n = 13$) Internal: 0.536 ± 0.242, external: 0.427 ± 0.158	0.014
b. across all studies	T-test	Mean ± SD Internal ($n = 81$): 0.456 ± 0.220, external ($n = 16$): 0.426 ± 0.153	0.606

Unless otherwise specified, prediction accuracy referred to Pearson's correlation r value resulted from internal validation. Studies without reporting r value were omitted. It should be noted that some studies may provide more than one data point whereas some studies may have missing data for the statistical tests, and hence the n reported may not correspond to the number of studies involved. Readers should refer to Supplementary Table 3 for the data used.

remained significant after adjusted for training set sample size and confound control. On the contrary, confound control and validation type (for studies that involved both internal and external validation) were not significant after adjusted for other factors. Readers should refer to Supplementary Table 4 for the full results of the partial correlation tests.

Additional analyses for only fMRI studies have shown that, the amount of data from self-recruiting studies (but not studies using open, shared dataset) was positively correlated with internal validation prediction accuracy ($n = 28$, $r = 0.654$, $p < 0.001$). Self-recruiting studies also reported a significantly higher internal validation prediction accuracy than those using open, shared datasets (Mean ± SD, self-recruiting: $n = 31$, 0.534 ± 0.211, shared dataset: $n = 27$, 0.375 ± 0.150, $p = 0.002$). Internal validation reported higher accuracy than external validation within studies that used both types of validation ($n = 10$, Mean ± SD, internal: 0.524 ± 0.255, external: 0.408 ± 0.146, $p = 0.045$). Sample size did not correlate with prediction accuracy (Supplementary Table 5). No partial correlation test was conducted for this subset, as there were very little or no overlap between studies involving these significant factors.

Meanwhile, additional analyses for only sMRI studies have shown that none of the factors correlated with prediction accuracy (Supplementary Table 6).

Discussion

Based on 108 neuroimaging studies on individual traits prediction published mainly in the late 2010s, it was found that sample size of the training set was negatively correlated with prediction accuracy from studies using internal validation. Meanwhile, amount of data of recruited subjects was positively correlated with internal validation prediction accuracy. Recurring target phenotypes were memory, attention, and intelligence. Half of the studies recruited their own subjects whereas HCP was the dominant open, shared dataset to be used. The most typical method for working with connectome data was “FCs with significant correlation (e.g. $p < 0.01$) with the predicted measure were selected as features”. The most popular learning algorithms were CPM, multiple linear regression, RVR, SVR, PLSR, LASSO, and elastic net. Most studies did not require hyperparameters tuning, and nested k-fold CV was the predominant choice for those required. LOO CV was the commonest validation strategy. Only a quarter of studies used external validation.

Our results showed a negative correlation between internal validation prediction accuracy and sample size. This negative correlation was similar to what was reported for mental disorders and health (Sui et al., 2020). Small sample size could lead to overfitting and hence the higher prediction accuracy particularly for CV cases, so that the trained model might explain little of the variance from an external test set (Varoquaux et al., 2017). This is particularly problematic for studying patients with uncommon diseases or medical conditions, or evaluating clinical outcomes of certain treatments (Gabrieli et al., 2015). However, the use of external validation could overcome this problem, as such negative correlation vanished when the external validation prediction accuracy and the sample size of the external test set were evaluated.

Meanwhile, the small sample size issue could be partially addressed by using large open neuroimaging datasets. Currently there are multiple open datasets available to researchers, covering structural MRI, diffusion MRI, resting-state MRI, task-based fMRI, behavioral data, genomics data, and occasionally physiological and angiographic data from a single subject up to 100,000 subjects (Horien et al., 2021; Madan, 2021). Examples included HCP, UK Biobank, and Adolescent Brain Cognitive Development (ABCD) study. However, the use of these large datasets tended to encourage some researchers to use CV or hold-out test sets (a priori split of the dataset) that could be optimistic. It will be more challenging, but the results will be more robust if researchers share data and evaluate model performance on new sites or unseen datasets. Also, researchers should know how the data have been pre-processed and manipulated, so that it could better match the characteristics of the neuroimaging data from their own recruited subjects. Otherwise, the trained model might not give good predictions on an external test set. The users of HCP data should report precisely which dataset was used, as different datasets went through different processing pipelines and contained different subjects.

Here, we reported a 25% of surveyed studies using external test sets. Consistent to a previous review reporting that only 25% of predictive modeling studies on treatment response to addictions (alcohol and substance use) included external validation (Yip et al., 2020). The small number of studies reporting external validation / unseen test sets could be due to generalization failure (of the models) or lack of additional independent data (Sui et al., 2020). It was not possible for us to know if generalization failure did occur for the models, but such failure, if existed, could be accounted by model selection-related issues and homogeneous sample (Boeke et al., 2020). As very few datasets actually collected behavioral measures by implementing the same psychometric tests, it remains to be investigated whether similar behavioral measure (e.g. fluid intelligence/intelligence quotient) from different datasets can be predicted with the same predictive model. External validation is recommended, as it will avoid confusion from reporting in-sample model fit indices as predictive accuracy and avoid inappropriate CV procedure such as post hoc CV (Poldrack et al., 2020). Therefore, open data sharing initiatives should be encouraged to make external validation more feasible beyond a single laboratory or study site, before the models would be ultimately tested in a large-scale, diverse population-level (Woo et al., 2017).

The connectome data and brain atlases used by the surveyed studies were heterogeneous. This created a variation in the methodology used by different studies, rendering it a potential confounding factor in comparing results across different predictive models. Together with the studies using recruited subjects instead of open datasets, the variability of the analysis pipeline would influence the results as it would for single dataset or group analysis (Botvinik-Nezer et al., 2020; Carp, 2012).

Conclusion

Based on this work, it was found that the literature currently largely fails to adhere to the recommended best practices, for instance, as outlined by (Scheinost et al., 2019; Woo et al., 2017). Few studies employed external validation for their trained predictive model. Without external validation, internal validation requires very careful planning and considerations with regard to sample size and CV method, which might be subjects of debate to avoid predictive models being too optimistic. Therefore, the authors recommended that future predictive modeling studies should always consider incorporating external validation. When using training and test sets from the same datasets, it is crucial to make them completely independent.

Funding

SBE acknowledges funding by the European Union’s Horizon 2020 Research and Innovation Program (grant agreements 945539 (HBP SGA3) and 826421 (VBC)), the Deutsche Forschungsgemeinschaft (DFG, SFB 1451 & IRTG 2150) and the National Institute of Health (R01 MH074457).

Data and code availability statement

Data used in this study is provided in the Supplementary Tables 1–3.

The code for the partial correlations used in this study can be found at: https://github.com/jadecci/partialcorr_factors.

Conflict of interest

None to declare.

Credit authorship contribution statement

Andy Wai Kan Yeung: Conceptualization, Methodology, Writing – original draft. **Shammi More:** Data curation, Writing – review & editing. **Jianxiao Wu:** Data curation, Writing – review & editing. **Simon B. Eickhoff:** Conceptualization, Methodology, Writing – review & editing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119275.

References

Boeke, E.A., Holmes, A.J., Phelps, E.A, 2020. Toward robust anxiety biomarkers: a machine learning approach in a large-scale sample. Biol. Psychiatry: Cognit. Neurosci. Neuroimaging 5, 799–807.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., Adcock, R.A, 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582, 84–88.

Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., Yeo, B.T, 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 2322–2345.

Bzdok, D., Ioannidis, J.P, 2019. Exploration, inference, and prediction in neuroscience and biomedicine. Trends Neurosci. 42, 251–262.

Carp, J., 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. Front. Neurosci. 6, 149.

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, 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980.

Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1–15.

Diedrichsen, J., 2006. A spatially unbiased atlas template of the human cerebellum. Neuroimage 33, 127–138.

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., 2010. Prediction of individual brain maturity using fMRI. Science 329, 1358–1361.

Eickhoff, S.B., Langner, R., 2019. Neuroimaging-based prediction of mental traits: road to utopia or Orwell? PLoS Biol. 17, e3000497.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A.R., 2016. The human brainnetome atlas: a new brain atlas based on connectional architecture. Cereb. Cortex 26, 3508–3526.

Feng, L., Li, H., Oishi, K., Mishra, V., Song, L., Peng, Q., Ouyang, M., Wang, J., Slinger, M., Jeon, T., 2019. Age-specific gray and white matter DTI atlas for human brain at 33, 36 and 39 postmenstrual weeks. Neuroimage 185, 685–698.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T, 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18, 1664–1671.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Gabrieli, J.D., Ghosh, S.S., Whitfield-Gabrieli, S., 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron 85, 11–26.

Gao, S., Greene, A.S., Constable, R.T., Scheinost, D., 2019. Combining multiple connectomes improves predictive modeling of phenotypic measures. Neuroimage 201, 116038.

Gilmore, J.H., Shi, F., Woolson, S.L., Knickmeyer, R.C., Short, S.J., Lin, W., Zhu, H., Hamer, R.M., Styner, M., Shen, D., 2012. Longitudinal development of cortical and subcortical gray matter from birth to 2 years. Cereb. Cortex 22, 2478–2485.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178.

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., Petersen, S.E, 2016. Generation and evaluation of a cortical area parcellation from resting-state correlations. Cereb. Cortex 26, 288–303.

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

Horien, C., Noble, S., Greene, A.S., Lee, K., Barron, D.S., Gao, S., O'Connor, D., Salehi, M., Dadashkarimi, J., Shen, X., 2021. A hitchhiker's guide to working with large, open–source neuroimaging datasets. Nat. Hum. Behav. 5, 185–193.

Liu, W., Wei, D., Chen, Q., Yang, W., Meng, J., Wu, G., Bi, T., Zhang, Q., Zuo, X.-.N., Qiu, J., 2017. Longitudinal test-retest neuroimaging data from healthy young adults in southwest China. Sci. Data 4, 170017.

Madan, C.R., 2021. Scan once, analyse many: using large open-access neuroimaging datasets to understand the brain. Neuroinformatics doi:10.1007/s12021-12021-09519-12026, [Epub ahead of print].

Murphy, K., Birn, R.M., Bandettini, P.A., 2013. Resting-state fMRI confounds and cleanup. Neuroimage 80, 349–359.

Poldrack, R.A., Huckins, G., Varoquaux, G., 2020. Establishment of best practices for evidence for prediction: a review. JAMA Psychiatry 77, 534–540.

Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L, 2011. Functional network organization of the human brain. Neuron 72, 665–678.

Rao, A., Monteiro, J.M., Mourao-Miranda, J., Initiative, A.s.D, 2017. Predictive modelling using neuroimaging data in the presence of confounds. Neuroimage 150, 23–49.

Rosenberg, M.D., Finn, E.S., Scheinost, D., Papademetris, X., Shen, X., Constable, R.T., Chun, M.M, 2016. A neuromarker of sustained attention from whole-brain functional connectivity. Nat. Neurosci. 19, 165–171.

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-.N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T, 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb. Cortex 28, 3095–3114.

Scheinost, D., Noble, S., Horien, C., Greene, A.S., Lake, E.M., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D.S, 2019. Ten simple rules for predictive modeling of individual differences in neuroimaging. Neuroimage 193, 35–45.

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., Conrod, P., Dalley, J., Flor, H., Gallinat, J., 2010. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol. Psychiatry 15, 1128–1139.

Shen, X., Finn, E.S., Scheinost, D., Rosenberg, M.D., Chun, M.M., Papademetris, X., Constable, R.T, 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat. Protoc. 12, 506–518.

Shen, X., Tokoglu, F., Papademetris, X., Constable, R.T, 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. Neuroimage 82, 403–415.

Sui, J., Jiang, R., Bustillo, J., Calhoun, V., 2020. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. Biol. Psychiatry 88, 818–828.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289.

Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180, 68–77.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage 145, 166–179.

Whelan, R., Garavan, H., 2014. When optimism hurts: inflated predictions in psychiatric neuroimaging. Biol. Psychiatry 75, 746–748.

Woo, C.-.W., Chang, L.J., Lindquist, M.A., Wager, T.D, 2017. Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci. 20, 365–377.

Yamashita, M., Kawato, M., Imamizu, H., 2015. Predicting learning plateau of working memory from whole-brain intrinsic network connectivity patterns. Sci. Rep. 5, 7622.

Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R, 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 1125–1165.

Yip, S.W., Kiluk, B., Scheinost, D., 2020. Toward addiction prediction: an overview of cross-validated predictive modeling findings and considerations for future neuroimaging research. Biol. Psychiatry: Cognit. Neurosci. Neuroimaging 5, 748–758.