Contents lists available at ScienceDirect

Intelligence

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

On the prediction of human intelligence from neuroimaging: A systematic review of methods and reporting

Bruno Hebling Vieira[a,b,c,*], Gustavo Santo Pedro Pamplona[d,e], Karim Fachinello[f,g], Alice Kamensek Silva[g], Maria Paula Foss[g], Carlos Ernesto Garrido Salmon[a]

[a] InBrain Lab, Departamento de Física, FFCLRP, Universidade de Sao Paulo, Ribeirão Preto, Brazil
[b] Methods of Plasticity Research, Department of Psychology, University of Zurich, Zurich, Switzerland
[c] Neuroscience Center Zurich (ZNZ), University of Zurich & ETH Zurich, Zurich, Switzerland
[d] Sensory-Motor Lab (SeMoLa), Department of Ophthalmology-University of Lausanne, Jules Gonin Eye Hospital-Fondation Asile des Aveugles, Lausanne, Switzerland
[e] Rehabilitation Engineering Laboratory (RELab), Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland
[f] PsiCog Lab, Psicobiologia, FFCLRP, Universidade de São Paulo, Ribeirão Preto, Brazil
[g] Neuropsicologia, Setor de Distúrbios do Movimento e Neurologia Comportamental, Departamento de Neurociencias e Ciências do Comportamento, FMRP, Universidade de São Paulo, Ribeirão Preto, Brazil

ARTICLE INFO

Keywords:
Behavior
fMRI
Resting-state
Deep learning
Intelligence
Prediction
Systematic review

ABSTRACT

Reviews and meta-analyses have proved to be fundamental to establish neuroscientific theories on intelligence. The prediction of intelligence using invivo neuroimaging data and machine learning has become a widely accepted and replicated result. We present a systematic review of this growing area of research, based on studies that employ structural, functional, and/or diffusion MRI to predict intelligence in cognitively normal subjects using machine learning. We systematically assessed methodological and reporting quality using the PROBAST and TRIPOD in 37 studies. We observed that fMRI is the most employed modality, resting-state functional connectivity is the most studied predictor. A meta-analysis revealed a significant difference between the performance obtained in the prediction of general and fluid intelligence from fMRI data, confirming that the quality of measurement moderates this association. Studies predicting general intelligence from Human Connectome Project fMRI averaged $r = 0.42$ ($\mathrm{CI}_{95\%} = [0.35, 0.50]$) while studies predicting fluid intelligence averaged $r = 0.15$ ($\mathrm{CI}_{95\%} = [0.13, 0.17]$). We identified virtues and pitfalls in the methods for the assessment of intelligence and machine learning. The lack of treatment of confounder variables and small sample sizes were two common occurrences in the literature which increased risk of bias. Reporting quality was fair across studies, although reporting of results and discussion could be vastly improved. We conclude that the current literature on the prediction of intelligence from neuroimaging data is reaching maturity. Performance has been reliably demonstrated, although extending findings to new populations is imperative. Current results could be used by future works to foment new theories on the biological basis of intelligence differences.

1. Introduction

Intelligence is a broad construct comprising multiple components, which can be estimated with a range of well-established tests (Urbina, 2011). Regardless of the instrument, scores in intelligence tests are positively correlated. G was postulated to be the “general factor” explaining this phenomenon by Spearman (1904), whose evidence “[...] can be said to be overwhelming" (Carroll, 1997). Albeit originally terming it “general intelligence”, Spearman later in his life adopted a critical view of the term and ceased to associate it with G (Spearman, 1927). Henceforth, to avoid ambiguities in this review we will employ the widely used term “intelligence”. Even though G successfully captures the overall positive correlation (Spearman, 1904), there is controversy regarding its validity as a single, all-encompassing, measure of intelligence. An alternative view posits that intelligence comprises multiple factors (Thurstone, 1938). Posteriorly, an integrated model for intelligence called G_F-G_C was proposed by Cattell (Cattell, 1941; Cattell, 1971). G_F stands for fluid intelligence and is associated with inductive and deductive reasoning, covering non-verbal components; therefore, it does not depend on previously acquired knowledge and the influence of

[*] Corresponding author at: Methods of Plasticity Research, Department of Psychology, University of Zurich, Zurich, Switzerland.
E-mail address: bruno.heblingvieira@uzh.ch (B.H. Vieira).

https://doi.org/10.1016/j.intell.2022.101654
Received 22 November 2021; Received in revised form 19 April 2022; Accepted 28 April 2022
Available online 12 May 2022

culture. Concept formation and recognition, identification of complex relationships, understanding of implications, and making inferences are examples of tasks related to G_F. On the other hand, G_C, crystallized intelligence, comprises the knowledge acquired through life experience and education related to cultural experiences. Hence, crystallized capacities are demonstrated, for example, in tasks regarding the recognition of the meaning of words (Schelini, 2006). While the scientific construction of G is based on correlations between test scores, intelligence quotient (IQ) is based on the sum of standardized scores of commonly used cognitive batteries, such as Wechsler scales with full scale IQ (FSIQ), verbal IQ (VIQ), and performance IQ (PIQ). FSIQ scores are excellent measures of G (Gignac & Bates, 2017) representing the general level of cognitive functioning. VIQ relates to verbal comprehension, acquired knowledge, language processing, verbal reasoning, attention, verbal learning, and memory. In sharp contrast, PIQ is connected to perceptual organization, processing visual, planning ability, non-learning-verbal and thinking skills, and manipulating visual stimuli with speed.

Studies show associations between brain and behavior measurements. The first finding was the positive correlation between brain volume or intracranial volume and intelligence (Luders, Narr, Thompson, & Toga, 2009; McDaniel, 2005). Other structural MRI (sMRI) correlates of intelligence include fine-grained morphometry, such as callosal thickness (Luders et al., 2007), striatal volume (Grazioplene et al., 2015) and regional gray and white matter volumetry (Haier, Jung, Yeo, Head, & Alkire, 2005). Functional connectivity (FC), as measured by functional MRI (fMRI), has reliably been shown to correlate with G and IQ. This includes correlations between resting-state FC (RSFC) network organization and FSIQ (Pamplona, Santos Neto, Rosset, Rogers, & Salmon, 2015; Song et al., 2008) and regional global connectivity and G_F (Cole, Yarkoni, Repovš, Anticevic, & Braver, 2012). The topography of task fMRI (T-fMRI) statistical maps have been found to correlate with intelligence as well (Choi et al., 2008; Graham et al., 2010). Correlates of intelligence extend beyond fMRI RSFC and task activations as well, to include measures such as amplitude of low frequency fluctuations (ALFF) and dynamic functional connectivity (dynFC). Using multimodal magnetic resonance imaging (MRI) Ritchie et al. (2015) demonstrates a plethora of correlates of G, including diffusion MRI (dMRI). For extensive literature reviews, see Dizaji et al. (2021), Basten and Fiebach (2021).

Previous reviews (Barbey, 2018; Jung & Haier, 2007) and meta-analyses (Basten, Hilger, & Fiebach, 2015; McDaniel, 2005; Pietschnig, Penke, Wicherts, Zeiler, & Voracek, 2015) were fundamental in the development of theories of biological intelligence (for an overview on theories, see Euler & McKinney, 2021). At the time studies performing predictive analyses were scarcer than today. This type of analysis enjoys growing popularity in neuroimaging (Bzdok, 2017; Bzdok, Altman, & Krzywinski, 2018). Machine learning (ML)-based predictive analyses allow one to test a much more complex hypothesis space than univariate, group-based testing. The multivariate nature of ML allows interactions and commonalities between predictors to be taken into account. It also “tests” such hypotheses on the basis of individualized predictions, taking into account heterogeneity that is diluted in group-based analyses (Sui, Jiang, Bustillo, & Calhoun, 2020). Data-driven studies based on ML are fundamental to understand the degree that variability in brain phenotypes explain variability in intelligence. ML-based studies also address the question of generalizability patterns at the forefront. For these reasons, this type of study is widely used in the investigation of behavior, with cognition and, specifically, human intelligence as the most studied domains (Sui et al., 2020).

While the literature of brain correlates on intelligence covers various techniques, such as sMRI, fMRI, dMRI, positron emission tomography (PET), electroencephalography (EEG), magnetoencephalography (MEG), predictive studies are limited in this regard. Availability is one of the main factors behind that choice, because ML benefits from large amounts of data (Cui & Gong, 2018). Small data samples have been identified as a source of optimistic bias in error-bars (Varoquaux, 2018), and leads to non-reproducible results. Increasing the amount of data available potentially leads to multifold increases in performance (Schulz, Bzdok, Haufe, Haynes, & Ritter, 2022). Large-scale open-data imaging cohorts are often centered on fMRI, with sMRI and dMRI providing complimentary information. For this reason, we opted to focus on fMRI, sMRI and dMRI, anticipating a small incidence of studies using other imaging modalities.

A large number of studies on the prediction of intelligence was published in recent years. To the best of the authors’ knowledge, no systematic review on this application of ML to predict human intelligence from brain imaging has been previously published. The purpose of this review is to identify existing literature, critically appraise reporting and methodology. We hope that our work will promote the establishment of best practices and prospects for future research in this field of research.

2. Methods

This review was developed following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for transparent reporting of systematic reviews (Moher et al., 2009). See Table B.4 for the PRISMA checklist. Choice of methods and search strategy are based on a protocol we developed and registered at Open Science Framework (Vieira et al., 2021). Post-hoc adaptations are mentioned below, when applicable.

2.1. *Eligibility criteria*

Eligibility criteria were peer-reviewed original articles written in English that performed individualized prediction of intelligence using at least one of fMRI, sMRI and dMRI in neurotypical human subjects using ML and include evaluation of generalizability, i.e. cross-validation, bootstrapping, or external validation. We adapted the preregistered protocol to include only studies published before 1st January 2021.

2.2. *Information sources*

We performed a systematized search in Scopus (scopus.com), dating to 29th March 2022. Additional documents were retrieved from a recent literature review (Dizaji et al., 2021), co-authored by B.H.V. and C.E.G. S., and another study (Fan, Jianpo, Qin, Hu, & Shen, 2020, Table 1) that provide a comparison between similar studies.

2.3. *Search strategy*

We retrieved all documents in Scopus that contained at least one of the following terms in their title, abstract, or keywords: “morphometry”, “cortical thickness”, “functional connectivity”, “MRI”, “fMRI”, “structural connectivity”, or “effective connectivity”. Simultaneously, the document should contain at least one of the following terms: “predict*”, “multivariate pattern analysis”, “bases”, “CPM”, “variability”, “mvpa”, or “machine learning”. The documents should also contain in their title one of the following terms: “intellig*”, “behavior*”, “cognitive ability” or “IQ”. This search was modified post-hoc from the preregistered one due to the erroneous omission of some terms. See Appendix A for the actual search string used.

After removal of duplicates, all records had title and abstract screened. Records were discarded if we could identify disagreement with inclusion criteria, and kept otherwise. Remaining records were retained for full-text inspection. If in accordance with the inclusion criteria, these were retained as eligible for qualitative synthesis. Otherwise discarded with reasons, e.g., non-human subjects, no validation or other generalizability evaluation, did not predict intelligence, did not use neurotypical subjects.

Table 1
General characteristics of documents retrieved using based on our data extraction form.

Studies	Number of subjects	Input	ML models	Validation strategy	Target
Choi et al. (2008)[b]	408 for FA, 225 for prediction from NRI/KAIST (training data: 116 sMRI and 61 fMRI; test data: 48);	fMRI & sMRI Cortical thickness, T-fMRI activation in a fluid reasoning task, gray matter volume, sex	Linear modeling (derived from separate structural and functional samples)	Independent test sample	G
Yang et al. (2013)[a,b]	78 from NRI	sMRI Cortical thickness, surface area, sulcal depth, mean curvature	PLSR	LOOCV	FSIQ
Finn et al. (2015)	118 from HCP (Q2 release)	fMRI RSFC under various preprocessing pipelines	CPM	LOOCV	G_F
Wang, Wee, Suk, Tang, and Shen (2015)[a,b]	164 from ABIDE	sMRI Regional gray and white matter volume	Multi-kernel KSVR following multiple feature selection	Repeated (10×) 10-fold CV, with inner CV (unspecified) for parameter tuning	IQ
Park, Hong, Lee, and Park (2016)	56 (non-ADHD-affected) from ADHD-200	fMRI Degree values from 1 (out of 10 RSN) obtained from group ICA (33 ICs kept)	Linear regression	LOOCV	FSIQ, VIQ and PIQ
Ferguson, Anderson, and Spreng (2017)[b]	830 from HCP (S900 release) (600 for training, 230 for testing)	fMRI Scaled eigenvalues from spectral decomposition of concatenated RS-fMRI, products of eigenvalues	LASSO	Independent test sample	G_F
Powell, Garcia, Yeh, Vettel, and Verstynen (2017)[a]	841 from HCP	dMRI Local Connectome Fingerprints and intracranial volume	LASSO PCR	5-fold CV	G_F
Noble, Spann, Tokoglu, and Shen (2017)[a]	606 from HCP (S900 release)	fMRI RSFC	CPM	LOOCV	G_F
Greene, Gao, Scheinost, and Constable (2018)[a]	515 from HCP; 571 from PNC	fMRI RSFC and T-fMRI activation (7 tasks in HCP, 2 in PNC)	CPM	LOOCV (within samples) and between samples/ between conditions validation	G_F
Dubois, Galdi, Han, Paul, and Adolphs (2018)[a]	884 from HCP (S1200 release)	fMRI RSFC under various preprocessing pipelines	CPM & elastic net following univariate filtering	LOFOCV (410 families)	G_F
Dubois, Galdi, Paul, and Adolphs (2018)[a,b]	884 for CV, 1181 for FA from HCP	fMRI RSFC	Elastic Net after univariate filtering	LOFOCV	G
Li, Yang, Li, and Li (2018)[a,b]	100 from HCP (Unrelated subjects)	fMRI ALFF following voxelwise univariate filtering, seed-based FC	L2SVR	LOOCV	G_F
Alnæs et al. (2018)[b]	6487 for FA, 748 for prediction from PNC	dMRI Linked ICA (LICA), from 8 maps of dMRI properties	Shrinkage (Schafer and Strimmer) linear regression	Repeated (1000×) 10-fold CV	G_F
Cox, Ritchie, Fawns-Ritchie, Tucker-Drob, and Deary (2019)[b]	27,100 for FA and 4768 for training, 2510 for testing with fractional anisotropy; 4707 for training, 2494 for testing with mean diffusion; cortical: 5246 for training, 2589 for testing with cortical volume; 5253 for training, 2595 for testing with subcortical volume; from the UK Biobank	sMRI & dMRI ROI white matter mean diffusivity and fractional anisotropy and gray matter cortical and subcortical volumes	MIMIC	Independent test sample (Manchester = training data, Newcastle = test data)	G
Yang et al. (2019)[b]	68 from HCP-Q1	fMRI RS-fMRI temporal variances of temporal autocorrelations (sulci, gyri, undefined cortices) from four ROIs	Linear regression	LOOCV	G_F
Zhang, Allen, Zhu, and Dunson (2019)	1065 from HCP	dMRI & fMRI Structural connectivity tensor (weighted according to 12 factors based on diffusion, endpoints and geometry); RSFC; local structural connectivity	Linear regression (after tensor network PCA with k = 60)	5-fold CV	G_F
Gao, Greene, Constable, and Scheinost (2019)[a]	515 from HCP; 571 from PNC	fMRI RSFC and T-fMRI FC	rCPM, GFC-ridge, cCPM, CPM, GFC-CPM	Repeated (100×) 10-fold CV; External Validation	G_F
Dadi et al. (2019)[a]	443 from HCP (213 High IQ, 230 Low IQ, based on terciles)	fMRI RSFC	K-Nearest Neighbors ($K = 1$, Euclidean distance metric),	Repeated (100×) Stratified Holdout (75%)	G_F

(continued on next page)

Table 1 (*continued*)

Studies	Number of subjects	Input	ML models	Validation strategy	Target
			Gaussian Naïve Bayes, Random Forests, L1-SVC L1-LogReg, Ridge classification, L2-SVC, L2-LogReg, 10%-univariate ANOVA SVC		
Elliott et al. (2019)	298 from HCP; 591 from DMHDS	fMRI RSFC, GFC	CPM	LOOCV (within samples) and between samples validation	Cognitive Ability
Yoo, Rosenberg, Noble, and Scheinost (2019)	316 unrelated subjects (out of 563) from HCP (S1200 release)	fMRI Bivariate and multivariate (distance correlation) RSFC	CPM	Repeated (5000×) 10-fold CV	G_F
Li et al. (2019)	862 from BGSP, 953 from HCP	fMRI RSFC	KRR (correlation kernel)	20-fold nested family-aware CV, inner 20-fold CV for selection of parameters	G_F
Kashyap et al. (2019)[a]	803 from HCP	fMRI RSFC with and without Common Orthogonal Basis Extraction (COBE)	Elastic Net after univariate filtering	20-fold nested CV, with inner CV for tuning	G_F
Kong et al. (2019)[a]	577 from HCP	fMRI Dice overlap kernel of different parcellation algorithms (ICA back-projection algorithm, individual-specific parcellation algorithm of Gordon, parcellation algorithm of Wang, multi-session hierarchical Bayesian model (MS-HBM))	KRR (dice overlap kernel)	Repeated (100×) 20-fold family-aware CV nested with inner tuning 20-fold CV	G_F
Xiao, Stephen, Wilson, Calhoun, and Wang (2019)[a,b]	224 from PNC (134 High IQ, 90 Low IQ, based on *Z*-scores)	fMRI RSFC and T-fMRI (emotion and fractal N-back) FC	SVC, vectorized or with DM (diffusion map) or with ADM (alternating DM), under different kernels (log-Euclidean, Euclidean or Cholesky distance)	Repeated (20×) 5-fold CV with nested inner tuning 5-fold CV	IQ
Dryburgh, McKenna, and Rekik (2020)[a,b]	226 from ABIDE-I	fMRI RSFC	CPM	LOOCV	FSIQ and VIQ
Jiang et al. (2020)[a,b]	326 from UESTC	fMRI & sMRI RSFC, cortical thickness (vertexwise)	CPM	LOOCV	FSIQ
Hilger et al. (2020)[a,b]	308 from NKI (Enhanced)	sMRI Gray matter volume (voxel and regionwise)	PCA-SVR & Atlas-SVR	10-fold CV (with nested inner 3-fold CV for parameter tuning) stratified for intelligence	FSIQ
Fan et al. (2020)[a,b]	1050 from HCP	fMRI dynFC	Deep neural network (CNN-LSTM), SVR	10-fold CV (no splitting runs from the same subject)	G_F and Crystalized Intelligence
Sripada, Angstadt, Rutherford, Taxali, and Shedden (2020)[a,b]	967 for T-fMRI, 903 for RS-fMRI from HCP (S1200 release)	fMRI Brain Basis Set (BBS) modeling decomposition of task contrasts, RSFC	Linear regression (75 components/coefficients)	10-fold family-aware CV	General Cognitive Ability (computed for each fold)
Wei, Jing, and Li (2020)[a]	1003 (812 "recon2" used as discovery set; 191 "recon1" as validation set) from HCP (S1200 release)	fMRI RSFC	CPM, SVR, LASSO, and Ridge regression, after Bootstrapping Feature selection	10-fold stratified CV & independent validation set	G_F
He et al. (2020)[a]	953 from HCP (S1200 release); 8868 from UK Biobank	fMRI RSFC	KRR, FNN, BrainNetCNN, GNN	HCP: 20-fold family-aware CV nested with inner tuning; UK Biobank: Holdout (6868 training, 1000 validation and 1000 test)	G_F
Jiang et al. (2020)[a,b]	360 from UESTC; 200 from HCP (Q3 release); 120 from COBRE (60 HCs)	fMRI RSFC	LASSO	LOOCV (with nested 10-fold CV for tuning)	FSIQ and G_F
Wu, Li, and Jiang (2020)	922 from HCP (S1200 release) (830 for training and 92 for test)	fMRI T-fMRI activation in seven HCP tasks (emotion, gambling, language, motor, relational, social, and working memory)	PLS	Independent test sample	G_F and Fluid, Crystalized and Total Scores
Lin, Baete, Wang, and Boada (2020)[a]	143 (1 subject with missing data) from HCP (S900 release)	dMRI & fMRI RSFC and structural connectivity (quantitative anisotropy, mean streamline length, and normalized number of streamlines)	CPM	LOOCV	G_F
Li et al. (2020)[a,b]		fMRI White matter RSFC	CPM	LOOCV and Repeated (100×) 20-fold CV for	G_F and PIQ

(*continued on next page*)

Table 1 (*continued*)

Studies	Number of subjects	Input	ML models	Validation strategy	Target
	326 for internal validation from SLIM, 53 for external validation from SXMU			internal validation, LOOCV for external validation	
Xiao, Stephen, Wilson, Calhoun, and Wang (2020)[a,b]	355 from PNC	fMRI T-fMRI (emotion and fractal N-back) FC	Single modality LASSO, MTL (Multi-Task Learning), M2TL (Manifold Regularized MTL), NM2TL (new M2TL). All follow univariate filtering	Repeated (10×) 5-fold CV with nested inner tuning 5-fold CV	IQ
Jiang et al. (2020)[a,b]	463 from HCP (S500 release)	fMRI RSFC and T-fMRI FC	PLS	Repeated (100×) 10-fold CV	G_F

[a] Primarily about predictive modeling.

[b] Primarily about intelligence.

2.4. Data collection process

We originally planned to use the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) checklist (Moons et al., 2014), but ultimately it became clear that we needed a form tailored for our research question. We constructed our own data extraction form, borrowing from CHARMS.

An online document was created and shared between authors B.H·V., K.F. and A.K.S. All three authors performed data extraction, including: (1) identification (title, year, source title, digital object identi (2) study population (dataset, number of subjects for psychometric assessment, number of subjects for prediction, age and sex characteristics of the sample), (3) methods (imaging modality, input features, number of features, ML models, validation strategy, performance metrics, a priori feature selection, construct, instrument, components of intelligence, scores, quality of cognitive assessment), (4) results (performance of individual methods/data combinations, best performance).

Regarding the quality of intelligence measurement, retrieved items included, when applicable: number of subtests, number of dimensions, time duration of test application, test completeness. These are important to assess whether the test properly measures intelligence and is applicable to the construct.

We additionally retrieved citations among identified documents, to build a citation network. Due to differences in formats and unreliability of automatic searching, we opted to perform a manual search over all documents. For each studie identified, we searched for the names of first authors of every other document. For consistency, we opted to consider citations of pre-print versions (same authors and title) of identified documents.

We used the Prediction model risk of bias assessment tool (PROBAST) to assess risk of bias (RoB) and concerns regarding applicability in individual studies. This was a choice made post-hoc to the registration of the study. We originally planned to create an RoB assessment checklist for the reviewed studies, but after registration we became aware of PROBAST, which fulfilled this role, requiring minimal adaptations. This assessment was performed at the result-level.

It is critical to ensure that reporting is transparent in order to ensure that findings can be replicated. We also used the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) (Collins, Reitsma, Altman, & Moons, 2015; Moons et al., 2015) checklist assessment tool (Heus et al., 2019) to evaluate reporting quality. We used a modified version tailored for ML predictions (Wang et al., 2020), including three modified items, shown in Table D.5. Several items in TRIPOD were not adequate for our research question, and were removed from the questionnaire for our evaluation. A few items and subitems were deemed not applicable or not important to our review question, and their assessments do not appear in this review. Namely, 1.i, 1.iii, 2.iii, 2.iv, 2.xi, 3b, 4a, 4b, 5c, 6b, 7a.iv, 7b, 10a, 10b.iv, 10b.v, 10c, 10d.ii, 10e, 11, 13a, 13b.iii, 13b.iv, 13c.ii, 15a.ii, 15b, 16.iii, 17, and 20.i, 22.ii. Items 1 and 16.i were edited to allow NA entries, due to studies that had broader scopes than the one pertaining to this review's question. Item 13b, pertaining to demographics, requires description of the actual data being used, and not from the original sample before exclusions. Items 13b and 14a, that should be assessed based on "Results" sections, were extended to "Methods" sections as well. We performed the TRIPOD assessment at the study-level and performed across-studies summarization of reporting quality ratings.

Authors G.S.P.P. and B.H.V. completed PROBAST and TRIPOD independently. To ensure both reviewers' interpretations were aligned, calibration was performed twice, using one study from each checklist on each occasion. Interrater agreement was then computed based on the Kappa statistic, at the score-level, for the remaining documents, excluding the two used for calibration.

The quality of measurement of intelligence is linked to validity and can interfere with results of each study. For example, Gignac and Bates (2017) demonstrated that the quality of measurement moderates the association between intelligence and brain volume. The guide for categorization of measurement quality by Gignac and Bates (2017) proposes four quality criteria: the number of tests, the number of group-level dimensions, testing time, and correlation with G. Authors K.F. and A.K.S. performed the assessment of measurement quality based on these criteria. "Number of tests" is categorized into 1, 1–2, 2–8, and 9+ which signal "poor", "reasonable", "good", and "excellent" measures of G, respectively, in the absence of any other information. Therefore, a minimum of nine tests is needed to represent an excellent G. The "number of group-level dimensions" criterion is divided into 1, 1–2, 2–3, and 3+ test dimensions, leading to the respective classifications "poor", "reasonable", "good", and "excellent" measures of G, in the absence of any other information. So, an excellent measure of G is expected to present at least three group-level dimensions, e.g., G_F, G_C, and processing speed. "Testing times" of 3–9 min, 10–19 min, 20–39 min, and 40+ minutes are respectively classified as possibly "poor", "reasonable", "good", and "excellent" measures of G. The last criterion, "correlation with G", is the best indicator of measurement quality and takes precedence over the others. However, this correlation is scarcely reported. Gignac and Bates (2017) recommends substituting the correlation with G with the three other criteria.

The primary measure of prediction performance evaluation was chosen to be the Pearson correlation coefficient, R-squared and mean squared error (MSE). See Appendix C for a mathematical description of different performance measures. The Pearson correlation coefficient is the most used measure in the literature. It is scale- and location-invariant, which means that high values can be obtained with arbitrarily large errors. R-squared, when properly evaluated, is a less biased measure of explained variance than the correlation coefficient squared. However, it also suffers from its own biases that will be discussed below, requiring proper care regarding the variance of the sample. Ideally, MSE or mean absolute error (MAE) should be used when comparing different models applied to the same data (Poldrack, Huckins, & Varoquaux, 2020). Regardless of the choice of the performance measure, comparisons between modeling approaches using different data can be ambiguous, since intrinsic variation can differ between datasets.

2.5. Synthesis of results

To determine the level of performance expected for each modality, we estimated a mixed-effects meta-analytic model using the package “metafor” in R 4.0.5 (Viechtbauer, 2010) using results that were rated with both low RoB and low concerns regarding applicability in PROBAST. The number of samples was taken to be the total number of subjects used in the estimation of performance with pooled or unpooled means. We employed the Hunter-Schmidt estimator to deal with the sampling variance, which entails a homogeneity assumption. Different datasets and measurements of intelligence were treated as fixed effects. The same procedures were used for the R-squared, except that the Hunter-Schmidt estimator was not applied, since it pertains exclusively to correlation coefficients. Residual heterogeneity, i.e. the variability unaccounted for by the model and covariates, was measured by the I^2 statistic.

Standard errors are seldom reported in the literature. Moreover, due to the nature of cross-validation (CV), where resulting models across folds are not independent, standard errors are underestimated (Varoquaux, 2018).

Assessment of within-study selective reporting is unfeasible in our setting, due to the lack of pre-registrations. Due to computational resources available today, the risk of selective reporting is real, leading to overfitting of the validation set. For an in-depth exposition, see Hosseini et al. (2020).

The funnel-plot was used to qualitatively assess the risk of publication bias.

Since one of the biggest bottlenecks for ML is sample size, we compared the number of training samples used with measured performances across studies. Training set size is often not homogeneous within studies. For CV-based studies, including leave-one-family-out CV (LOFOCV), we chose to approximate it as $N \times (K-1)/K$, where N is the total amount of data available for training and K is the number of groupings, i.e. folds or families. The formula holds true for leave-one-out CV (LOOCV) as well. For Holdout-based studies, the actual number of training data is given by the studies.

3. Results

3.1. Search results and study characteristics

Our search strategy identified 689 records in Scopus. Additionally, 17 records were identified from Dizaji et al. (2021) and 7 in Fan et al. (2020). 74 records remained after removal of duplicates and screening. These were submitted to full-text eligibility analysis. 37 records were considered eligible for qualitative synthesis. See Fig. 1. The number of studies per year is shown in Fig. 2. General characteristics from each document obtained with our data extraction form are reported in Table 1.

A co-citation network is shown in Fig. 3. Arrows point from the cited to the citing document. In total, 87 citations were identified. This network systematically demarks highly influential works in the sample. Finn et al. (2015) is cited by 23 studies, out of 26 studies that were published posteriorly to it.

Regarding data sources, 25 (68%) studies used different releases of the Human Connectome Project (HCP). Among these, 19 (51%) studies use solely HCP data. 6 (16%) studies used the HCP together with other datasets, such as the Philadelphia Neurodevelopmental Cohort (PNC) data, the Dunedin Multidisciplinary Health and Development Study (DMHDS), Center for Biomedical Research Excellence (COBRE) and University of Electronic Science and Technology of China (UESTC), the UK Biobank, and Brain Genomics Superstruct Project (BGSP). Other sources of data included the Neuroscience Research Institute (NRI) (Choi et al., 2008; Yang et al., 2013), Korea Advanced Institute of Science and Technology (KAIST) (Choi et al., 2008), Autism Brain Imaging Data Exchange (ABIDE) (Dryburgh et al., 2020; Wang et al., 2015), UK Biobank (Cox et al., 2019), Nathan Kline Institute - Rockland Sample (NKI) (Hilger et al., 2020), UESTC (Jiang, Calhoun, Cui, et al., 2020), ADHD-200 (Park et al., 2016), and Southwest University Longitudinal Imaging Multimodal Brain Data Repository (SLIM) and Shanxi Medical University (SXMU) (Li et al., 2020). All sources of data provide images acquired with 3 T MRI scanners, with the exception of NRI, that only includes data 286 acquired with 1.5 T. See Fig. 4a.

Regarding imaging modality, 27 (73%) studies only used fMRI data. sMRI was the only imaging modality in 3 (8%) studies. 2 (5%) study concerned only dMRI. Multimodality was also explored, with fMRI and sMRI in 2 (5%) studies, dMRI and fMRI in 2 (5%) studies, and sMRI and

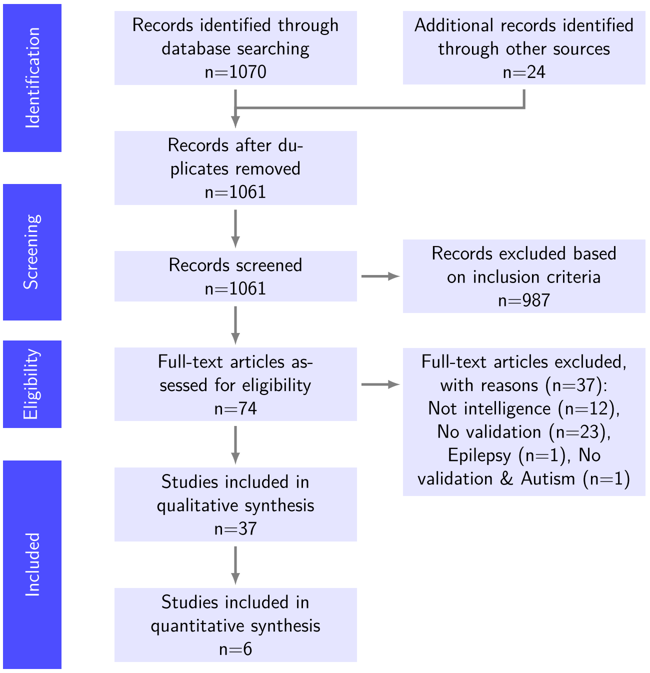

Identification
Screening
Eligibility
Included
Records identified through database searching n=1070
Additional records identified through other sources n=24
Records after duplicates removed n=1061
Records screened n=1061
Records excluded based on inclusion criteria n=987
Full-text articles assessed for eligibility n=74
Full-text articles excluded, with reasons (n=37): Not intelligence (n=12), No validation (n=23), Epilepsy (n=1), No validation & Autism (n=1)
Studies included in qualitative synthesis n=37
Studies included in quantitative synthesis n=6

Fig. 1. Systematic review flow diagram. See PRISMA statement (Moher et al., 2009).

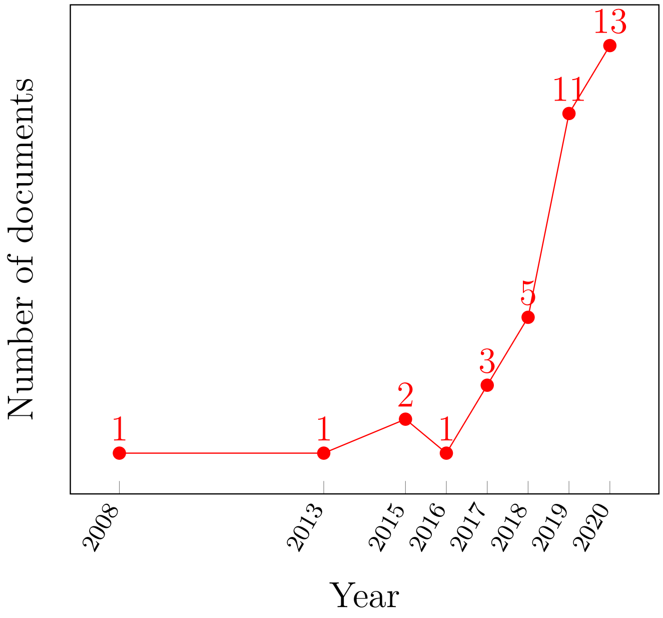


Fig. 2. Year of publication of the 37 studies identified. An upward tendency is demonstrated, with 24 studies being published in 2019 and 2020 alone.

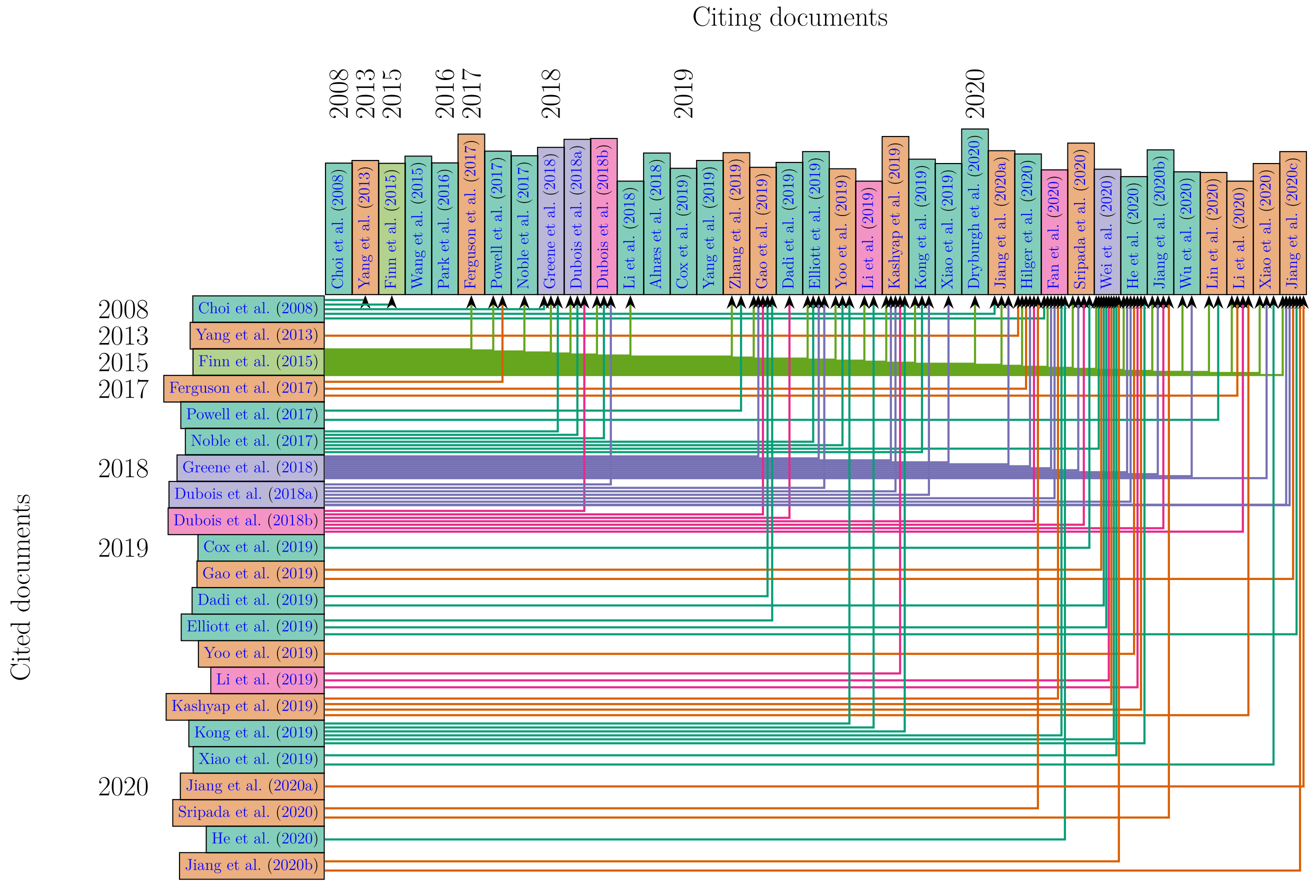
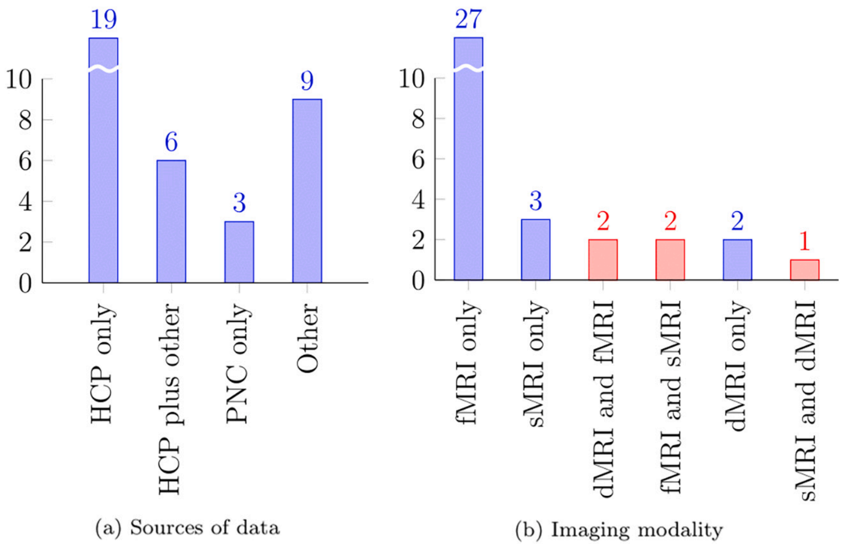


Fig. 3. A citation network with 87 citations relating all 35 studies identified in Fig. 1. Colors are used to better differentiate studies and carry no meaning. Arrows are colored according to parent nodes and point from the cited work to the one citing it.



(a) Sources of data

(b) Imaging modality

Fig. 4. General characteristics of eligible studies. (a) shows the main sources of data identified in the sample. 25 (68%) studies employed different releases of the HCP, with 19 (51%) based solely on HCP data. (b) shows the use of different imaging modalities. Shown in blue, 32 studies were based on unimodal data: 27 (73%) used fMRI, 3 (8%) used sMRI and 2 (5%) used dMRI exclusively. Shown in red, the remaining five studies employed multimodal data: 2 (5%) used fMRI and sMRI, 1 (3%) used sMRI and dMRI, 2 (5%) used dMRI and fMRI. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

dMRI in 2 (5%) study. No study performed multimodal prediction based on fMRI, sMRI and dMRI simultaneously. Also, all studies used solely MRI data, i.e. no additional imaging such as PET, EEG or MEG is used. See Fig. 4b.

We identified four constructs reported as outcomes. G_F is an outcome in 24 (65%) studies, IQ in 10 (27%), PIQ and VIQ in 2 (5%) each, general intelligence, general cognitive ability or G appears in 4 (11%) studies, crystallized ability appears in 2 (5%) studies, and cognitive ability appears in 1 (3%) study. 2 (5%) studies reported results on G_F and other NIH Toolbox for Assessment of Neurological and Behavioral Function

(NIHTB) cognition scores (Fan et al., 2020; Wu et al., 2020), i.e. total, fluid and/or crystallized cognition scores. 1 (3%) study includes measures of both IQ and G_F as outcomes (Jiang, Calhoun, Fan, et al., 2020).

The most common reported instrument is the 24-item Raven's Progressive Matrices (RPM), appearing in 25 (68%) studies. In all these studies, the RPM employed is the Penn Matrix Test (PMAT), from the University of Pennsylvania Computerized Neurocognitive Battery (PennCNB), which also appears in 18-item format in 3 (8%) studies (Alnæs et al., 2018; Gao et al., 2019; Greene et al., 2018). The 36-item Raven's Advanced Progressive Matrices Set II appears in 1 (3%) study (Choi et al., 2008), as one test in the estimation of G. The Combined Raven's Test (Chinese Revision), which combines aspects of Raven's Colored Progressive Matrices and Raven's Standard Progressive Matrices (RSPM), appears in 1 (3%) study (Li et al., 2020). All 25 (68%) studies that studied G_F reported the usage of RPM. 4 (11%) of these studies also studied additional scores, either due to availability in specific datasets or as parallel measures. These are the Wechsler Adult Intelligence Scale (WAIS) matrix reasoning test score, as a substitute for G_F in the BGSP (Li et al., 2019), NIHTB fluid cognition scores (Fan et al., 2020; Wu et al., 2020) and WAIS (Chinese Revision) PIQ (Li et al., 2020). RPM-like tests also appear in studies that derive analytical decompositions of test scores, such as G (Choi et al., 2008; Dubois, Galdi, Paul, & Adolphs, 2018; Sripada et al., 2020) and G_F (Alnæs et al., 2018). See Table 2.

We used qualitative cues in titles and abstracts to determine the overall scope of studies. 13 (35%) studies had prediction of intelligence as their primary objective. Other 11 (30%) studies were concerned primarily with predictive modeling, although not focused on intelligence. 6 (16%) studies focused primarily on intelligence, but not primarily on predictive modeling. The remaining 7 (19%) studies did not focus primarily on intelligence and primarily on predictive modeling, albeit including results on both.

Out of the 31 (84%) studies employing fMRI, 25 (68%) explored FC. All but one of these include RSFC-based analyses, while 14 (38%) studied RSFC exclusively. In 22 (59%), the only fMRI data was resting-state fMRI (RS-fMRI). 9 (24%) studies used T-fMRI, with task FC and/or spatial topographies as inputs (Choi et al., 2008; Elliott et al., 2019; Gao et al., 2019; Greene et al., 2018; Jiang, Zuo, Ford, et al., 2020; Sripada et al., 2020; Wu et al., 2020; Xiao et al., 2019; Xiao et al., 2020). Choi et al. (2008) employed a fluid reasoning task. Greene et al. (2018), Wu et al. (2020), Sripada et al. (2020), Elliott et al. (2019), Gao et al. (2019), Jiang, Zuo, Ford, et al. (2020) employed seven tasks from the HCP Additionally, Greene et al. (2018), Gao et al. (2019), Xiao et al. (2019) used the working-memory and emotion identification tasks from the PNC, Xiao et al. (2020) used the working-memory task only from the PNC, and Elliott et al. (2019) employed the emotion processing, color Stroop, monetary incentive delay and episodic memory tasks from the DMHDS.

Not counting intracranial volume, which is used both as a predictor and as a confounder in several studies, all 6 (16%) studies reporting usage of sMRI employ morphometric measurements as predictors. The small sample of dMRI-including studies included as predictors mean diffusivity and fractional anisotropy, structural connectivity, local connectome fingerprints, structural connectivity tensors and local structural connectivity, and linked Independent Component Analysis (ICA) components obtained across diffusion descriptor dimensions.

Regression based on linear models was reported in 33 (89%) studies. Among these, 14 (38%) reported use of some form of penalized linear modeling. 12 (32%) reported using Connectome Predictive Modeling (CPM). 4 (11%) reported using Support Vector Regression. 6 (16%) reported using linear regression, either on inputs or on extracted components, e.g., Principal Components Regression. 3 (8%) reported using Partial Least Squares Regression. Regression based on nonlinear models was reported in 5 (14%) studies. These include polynomial Kernel SVR (Wang et al., 2015), correlation kernel ridge regression (KRR) (He et al., 2020; Li et al., 2019), dice overlap KRR (Kong et al., 2019) and deep

Table 2
On the quality of the measurement of intelligence. This categorization follows a set of rules established in Gignac and Bates (2017).

Studies	Measurement	Number of tests	Dimensions	Testing time (min)	Rating
Choi et al. (2008)	G (principal component of 36-item RPM and K-WAIS-R subtests)	9+	3+	40+	4
Dubois, Galdi, Han, et al. (2018), Sripada et al. (2020)	G (FA of 10 tests in the NIHTB and PennCNB)	9+	3+	40+	4
Cox et al. (2019)	G (FA of 4 tests in the UK Biobank)	2–8	3+	20–39	3
Yang et al. (2013)	FSIQ (K-WAIS-R)	9+	3+	40+	4
Jiang, Calhoun, Cui, et al. (2020), Jiang, Calhoun, Fan, et al., 2020	FSIQ (WAIS Chinese revision)	9+	3+	40+	4
Wang et al. (2015)	IQ (WISC-IV in ABIDE)	9+	3+	40+	4
Xiao et al. (2019, 2020)	IQ (WRAT in PennCNB)	9+	3+	40+	4
Park et al. (2016)	FSIQ (WASI)	9+	3+	40+	4
Hilger et al. (2020)	FSIQ (WASI in NKI)	2–8	3+	40+	3
Wang et al. (2015)	IQ (WASI in ABIDE)	2–8	3+	40+	3
Dryburgh et al. (2020)	IQ (Unclear)	?	?	?	?
Alnæs et al. (2018)	G_F (Principal component of 12 tests)	9+	3+	40+	4
Ferguson et al. (2017), Fan et al. (2020), Gao et al. (2019), Yang et al. (2019), Jiang, Calhoun, Fan, et al. (2020), Greene et al. (2018), Li et al. (2019, 2018), Kashyap et al. (2019), Yoo et al. (2019), He et al. (2020), Lin et al. (2020), Wei et al. (2020), Finn et al.	G_F (24-item RPM number of correct responses in HCP)	1–2	1–2	3–19	2

(continued on next page)

Table 2 (*continued*)

Studies	Measurement	Number of tests	Dimensions	Testing time (min)	Rating
(2015), Dubois, Galdi, Han, et al. (2018), Zhang et al. (2019), Dadi et al. (2019), Kong et al. (2019), Wu et al. (2020), Powell et al. (2017), Noble et al. (2017), Jiang, Zuo, Ford, et al. (2020)					
He et al. (2020)	G_F (13-item test number of correct responses in the UK Biobank)	1–2	1–2	2	?
Li et al. (2019)	G_F (WAIS - Matrix Reasoning test)	1–2	1–2	?	2
Greene et al. (2018), Gao et al. (2019)	G_F (18-item RPM in PNC)	1–2	1–2	3–19	2
Greene et al. (2018), Gao et al. (2019)	G_F (24-item RPM in PNC)	1–2	1–2	3–19	2
Powell et al. (2017)	G_F (24-item RPM total skipped items in HCP)	1–2	1–2	3–19	2
Powell et al. (2017)	G_F (24-item RPM median reaction time for correct responses in HCP)	1–2	1–2	3–19	2
Li et al. (2020)	G_F (CRT Chinese revision)	1–2	1–2	40+	2
Li et al. (2020)	G_F (WAIS Chinese revision PIQ)	2–8	1–2	20–39	3
Park et al. (2016)	VIQ (WASI)	2–8	1–2	3–19	2
Park et al. (2016)	PIQ (WASI)	2–8	1–2	3–19	2
Dryburgh et al. (2020)	VIQ (Unclear)	?	?	?	?
Wu et al. (2020)	Total cognition score (composite score from the NIHTB)	2–8	3+	40+	3
Wu et al. (2020)	Fluid cognition score (composite score from the NIHTB)	2–8	3+	40+	3
Wu et al. (2020), Fan et al. (2020)	Crystallized cognition score (composite score from the NIHTB)	2–8	3+	40+	3
Elliott et al. (2019)	Cognitive ability (WAIS-IV in the DMHDS)	9+	3+	40+	4
Elliott et al. (2019)	Cognitive ability (24-item RPM in HCP)	1–2	1–2	3–19	2

1 = poor, 2 = fair, 3 = good, 4 = excellent, ? = unclear. FA = factor analysis; K-WAIS-R = Korean WAIS-R; WASI = Wechsler Abbreviated Scale of Intelligence; WISC-IV = Wechsler Intelligence Scale for Children - 4th edition; WRAT = Wide Range Achievement Test.

learning, based on convolutional neural networks (CNNs), graph neural networks and fully connected deep networks (He et al., 2020) or recurrent neural networks (RNNs) (Fan et al., 2020).

In 35 (95%) studies, prediction of intelligence was implemented as regression, i.e. prediction of a continuous variable. 2 (5%) studies performed classification, subdividing subjects into two groups, one with high and the other with low IQ. Dadi et al. (2019) reports using Support Vector Classification (SVC) and Penalized Logistic Regression, as linear models, and 1-Nearest Neighbor, Naïve Bayes and Random Forest, as non-linear models. Xiao et al. (2019) reports using SVC, with and without diffusion map and alternating diffusion map, using kernels based on log-Euclidean, Euclidean, and Cholesky distances.

Regarding the level of spatial abstraction of input data, 31 (84%) studies presented inputs at the regional level, either intra-regional features 7 (19%), e.g., regional cortical thickness estimates, or inter-regional features in 25 (68%), e.g., RSFC. Inter-voxel predictors appear in 2 (5%) studies (Powell et al., 2017; Zhang et al., 2019), e.g., local dMRI structural connectivity, which is measured between adjacent voxels. Intra-voxel predictors appear 5 (14%) studies (Hilger et al., 2020; Jiang, Calhoun, Cui, et al., 2020; Kong et al., 2019; Li et al., 2018; Wu et al., 2020), e.g., seed-based FC or voxelwise morphometry, ALFF, or T-fMRI statistical maps, which measure properties pertaining to individual voxels. Global features appear in 2 (5%) studies, i.e. linked ICA components (Alnæs et al., 2018) and graph-theoretical degree from a resting-state network (Park et al., 2016). No study used raw or minimally preprocessed imaging data directly as input to ML models.

In total, discounting censored and unclear results, e.g., results presented only graphically, 269 sets of results are presented across 32 studies, encompassing 12 performance metrics. These are Pearson correlation coefficient, Spearman rank correlation coefficient, R-squared, square root of R-squared, MAE, MSE, root MSE (RMSE), normalized RMSE (NRMSE), normalized root mean squared deviations (nRMSD), percentage error, area under the ROC curve (AUC), and classification accuracy. See Appendix C for the mathematical definition of each.

3.2. *Risk of bias within-studies*

We could identify the tests used by all but one study by consulting the text, supplementary materials and citations offered. See Table 2. There was, however, little information about the measurement validity for the populations under study. 3 studies cited references deemed adequate (Ferguson et al., 2017; Wei et al., 2020; Yang et al., 2019), whereas partial references were cited in 2 studies (Hilger et al., 2020; Lin et al., 2020).

Regarding measurement quality, 9 measurements were rated as excellent, distributed across 12 (32%) studies. 7 measurements were rated as good, distributed across 6 (16%) studies. 10 measurements were rated as fair, distributed across 25 (68%) studies. 1 study has a measurement of IQ which we could not identify, based on pre-processed ABIDE, which includes multiple instruments. 1 study has a measurement of G_F based on the UK Biobank, whose rating is not clear. See Table 2 for detailed ratings.

The PROBAST assessments of RoB and applicability are shown in Table 3. Fair interrater agreement was obtained in this analysis, estimated with a Cohen's Kappa equal to 0.353. In total, 8 (22%) studies were rated with low overall RoB and low concern regarding applicability. This includes five development-only studies (Dubois, Galdi, Han, et al., 2018; Dubois, Galdi, Paul, & Adolphs, 2018; He et al., 2020; Li et al., 2019; Sripada et al., 2020), one development-validation study (Cox et al., 2019), and the validation portions of two development-validation studies (Elliott et al., 2019; Greene et al., 2018). These are eligible for quantitative synthesis, i.e. meta-analysis. Li et al. (2019) does not present prediction results in text format however, and thus was not used. Results pertaining to sMRI and dMRI encompass only the 4 results in Cox et al. (2019), and thus these modalities were ineligible for quantitative synthesis, per our protocol. 87 results sets identified among the remaining 6 studies were suitable for quantitative synthesis: 3 in Dubois, Galdi, Paul, and Adolphs (2018), 8 in He et al. (2020), 16 in Sripada et al. (2020), 39 in Dubois, Galdi, Han, et al. (2018), 6 in Greene et al. (2018), and 15 in Elliott et al. (2019). All of these employed fMRI solely and reported either the Pearson Correlation Coefficient or R-squared, with the exception of Greene et al. (2018), which reported squared Spearman Rank Correlation. We opted to group this result with R-squared.

3.3. *Synthesis of results*

Forest plots with individual results are shown in Fig. 5. For the Correlation coefficient obtained from fMRI, both G and G_F have expected correlations significantly different from zero, based on 66 results from 5 studies (Dubois, Galdi, Han, et al., 2018; Dubois, Galdi, Paul, & Adolphs, 2018; Elliott et al., 2019; He et al., 2020; Sripada et al., 2020). For G, the expected correlation was 0.42 ($CI_{95\%} = [0.35, 0.50]$, $p < 0.001$) in the HCP. For G_F, the expected correlation was 0.15 ($CI_{95\%} = [0.13, 0.17]$, $p < 0.001$) in the HCP. Both are significantly different ($p < 0.001$). A significant difference between the HCP and UK Biobank was found: studies score, on average, 0.086 ($CI_{95\%} = [0.011, 0.16]$, $p = 0.026$) higher when using the latter. Residual heterogeneity was estimated at $I^2 = 77{:}8\%$ for this analysis.

For R-squared, only G has expected R-squared significantly different from zero, based on 34 results from 6 studies (Dubois, Galdi, Han, et al., 2018; Dubois, Galdi, Paul, & Adolphs, 2018; Elliott et al., 2019; Greene et al., 2018; He et al., 2020; Sripada et al., 2020). No significant differences between HCP and PNC or between HCP and UK Biobank were

Table 3
PROBAST = Prediction model Risk Of Bias ASsessment Tool; ROB = risk of bias; D = Development; V = Validation. expresses low ROB/low concern regarding applicability; – expresses high ROB/high concern regarding applicability; and? expresses unclear ROB/unclear concern regarding applicability.

Studies	D/V	ROB				Applicability			Overall	
		Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Choi et al. (2008)	D	–	–	?	+	–	–	–	+	–
Yang et al. (2013)	D	–	–	–	+	–	–	–	+	–
Finn et al. (2015)	D	–	–	–	+	–	–	–	+	–
Wang et al. (2015)	D	?	–	–	+	?	–	?	+	?
Park et al. (2016)	D	–	–	–	+	–	–	–	+	–
Ferguson et al. (2017)	D	–	–	–	?	–	–	–	?	–
Powell et al. (2017)	D	?	–	–	?	–	–	–	?	–
Noble et al. (2017)	D	–	–	–	?	–	–	–	?	–
Greene et al. (2018)	D	–	–	–	?	–	–	–	?	–
Greene et al. (2018)	V	–	–	–	–	–	–	–	–	–
Dubois, Galdi, Han, et al. (2018)	D	–	–	–	–	–	–	–	–	–
Dubois, Galdi, Paul, and Adolphs (2018)	D	–	–	–	–	–	–	–	–	–
Li et al. (2018)	D	?	–	–	+	–	–	–	+	–
Alnæs et al. (2018)	D	–	–	–	?	?	–	–	?	?
Cox et al. (2019)	D	–	–	–	–	–	–	–	–	–
Cox et al. (2019)	V	–	–	–	–	–	–	–	–	–
Yang et al. (2019)	D	?	–	–	+	–	–	–	+	–
Zhang et al. (2019)	D	?	–	–	+	–	–	–	+	–
Gao et al. (2019)	D	–	–	–	+	–	–	–	+	–
Gao et al. (2019)	V	–	–	–	?	–	–	–	?	–
Dadi et al. (2019)	D	?	–	–	+	–	–	–	+	–
Elliott et al. (2019)	D	–	–	–	?	–	–	–	?	–
Elliott et al. (2019)	V	–	–	–	–	–	–	–	–	–
Yoo et al. (2019)	D	–	–	–	+	–	–	–	+	–
Li et al. (2019)	D	–	–	–	–	–	–	–	–	–
Kashyap et al. (2019)	D	–	–	–	+	–	–	–	+	–
Kong et al. (2019)	D	?	–	–	?	–	–	–	?	–
Xiao et al. (2019)	D	–	–	–	?	–	–	–	?	–
Dryburgh et al. (2020)	D	–	–	?	+	+	–	–	+	+
Jiang, Calhoun, Cui, et al. (2020)	D	–	–	–	+	?	–	–	+	?
Hilger et al. (2020)	D	–	–	–	+	–	–	–	+	–
Fan et al. (2020)	D	–	–	–	?	–	–	–	?	–
Sripada et al. (2020)	D	–	–	–	–	–	–	–	–	–
Wei et al. (2020)	D	–	–	–	?	–	–	–	?	–
He et al. (2020)	D	–	–	–	–	–	–	–	–	–
Jiang, Calhoun, Fan, et al. (2020)	D	–	–	–	?	–	–	–	?	–
Jiang, Zuo, Ford, et al. (2020)	V	–	–	–	?	–	–	–	?	–
Wu et al. (2020)	D	–	–	–	+	–	–	–	+	–
Lin et al. (2020)	D	–	–	–	+	–	–	–	+	–
Li et al. (2020)	D	?	–	–	?	–	–	–	?	–
Li et al. (2020)	V	?	–	–	?	–	–	–	?	–
Xiao et al. (2020)	D	–	–	–	?	–	–	–	?	–
Jiang, Calhoun, Cui, et al. (2020)	D	–	–	–	?	–	–	–	?	–

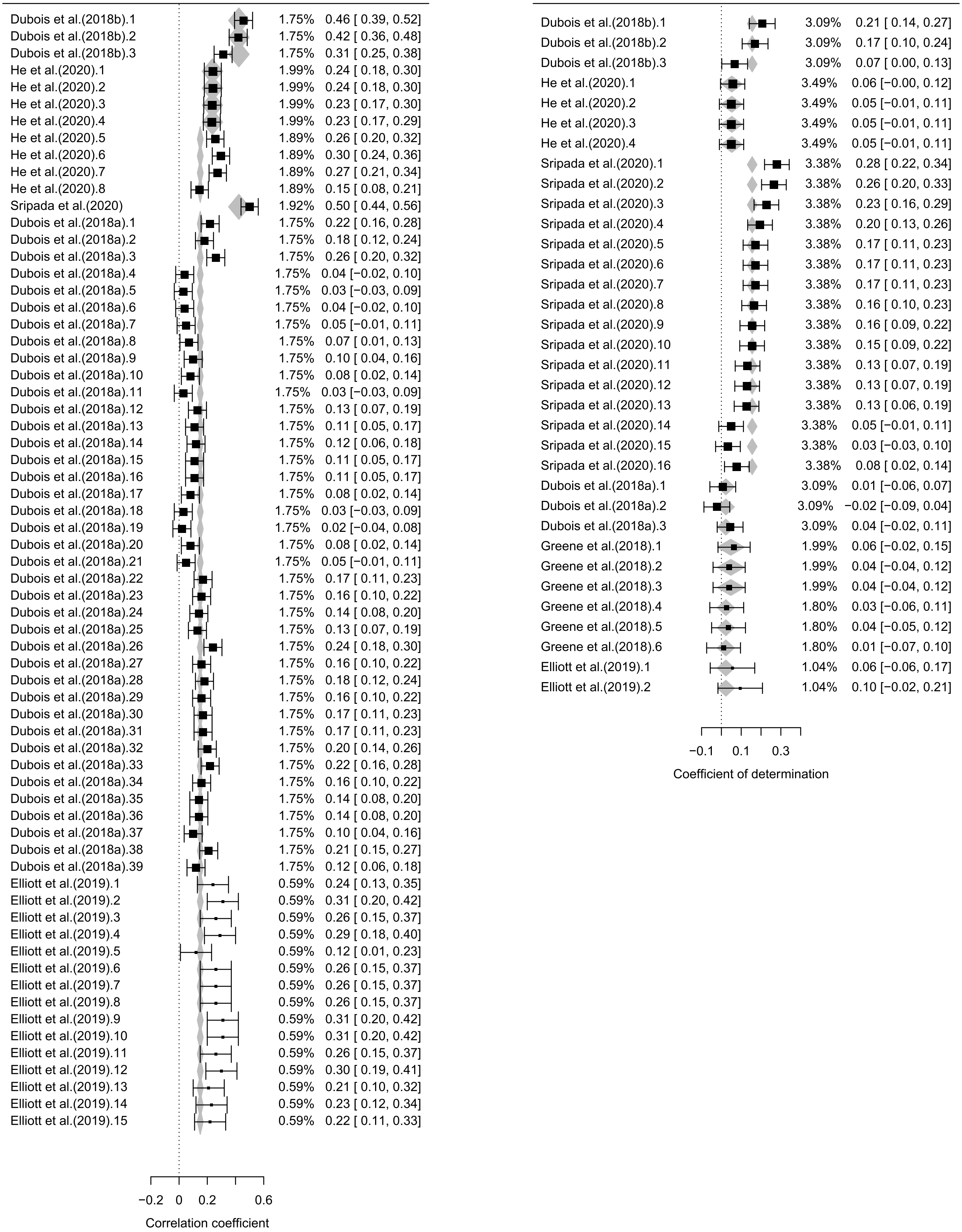


Fig. 5. Forest plots for (a) the correlation coefficient and (b) the R-squared meta-analyses. The outcome, either G or G_F, and the dataset, either HCP, PNC or the UK Biobank, were included as moderators.

found (ANOVA $p = 0.717$). For G, the estimated mean of R-squared was 0.16 ($CI_{95\%} = [0.13,0.18]$, $p < 0.001$), in a model without the dataset moderator. For G_F, the estimated marginal mean of R-squared was 0.038 ($CI_{95\%} = [0.0074, 0.0685]$, $p = 0.0165$). Both are significantly different (mean difference 0.1175, $CI_{95\%} = [0.078,0.16]$, $p < 0.001$). Residual heterogeneity was estimated at $I^2 = 59{:}3\%$ for this analysis. See Appendix E for additional meta-regressions without low RoB results, where the impact of inclusion of results without low RoB lead to substantial differences.

TRIPOD has items that apply only to either validation or development of models. Here, all studies included development of models, while a few also included external validation. Good interrater agreement was obtained in this analysis, estimated with a Cohen's Kappa equal to 0.6. We chose to represent results together in Fig. 6, with the caveat that a few items (10e, 12, 13c, 17, 19a) only apply to studies that include validation of models.

The histogram of TRIPOD ratings is shown in Fig. 7.

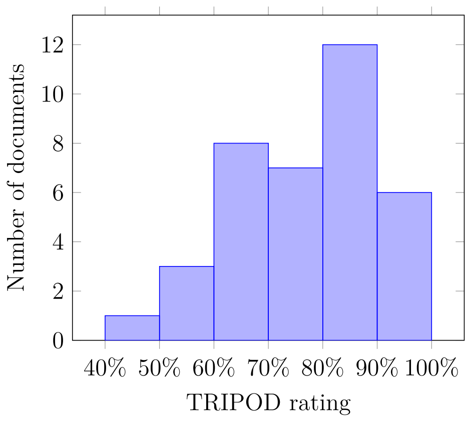


Fig. 7. Distribution of TRIPOD overall ratings across 37 studies.

3.4. *Risk of bias across-studies*

Funnel plots for both the analysis of correlation coefficients and R-squared are shown in Fig. 8. Both analyses present symmetrical funnel plots, which imply low risk of publication bias, but the range of standard errors is low, due to sample limitations, e.g. the lack of results with more subjects. Additional funnel plots, including results without low RoB are shown in Appendix F.

3.5. *Additional analyses*

We additionally analyzed the relationship between the expected effect size and training set size. Due to the small number of results pertaining to R-squared, this analysis was performed only for the correlation coefficient. Fig. 9 shows the expected correlation coefficient between predicted values and true labels as a function of approximate training set size. This comparison is qualitative, and does not take into account confounders, but it is also expected that such procedures are more robust in larger sample sizes. Compare with Fig. 5a, which includes only studies with low RoB and low concerns regarding applicability.

4. Discussion and conclusion

Here, we systematically reviewed available studies on the application of ML to the prediction of human intelligence using MRI data. Most of these studies were published very recently. See Fig. 2. Namely, two-thirds were published in 2019 and 2020. This attests the high and growing interest over this question in the literature.

It is also very clear from Fig. 3 that some highly cited studies exert a

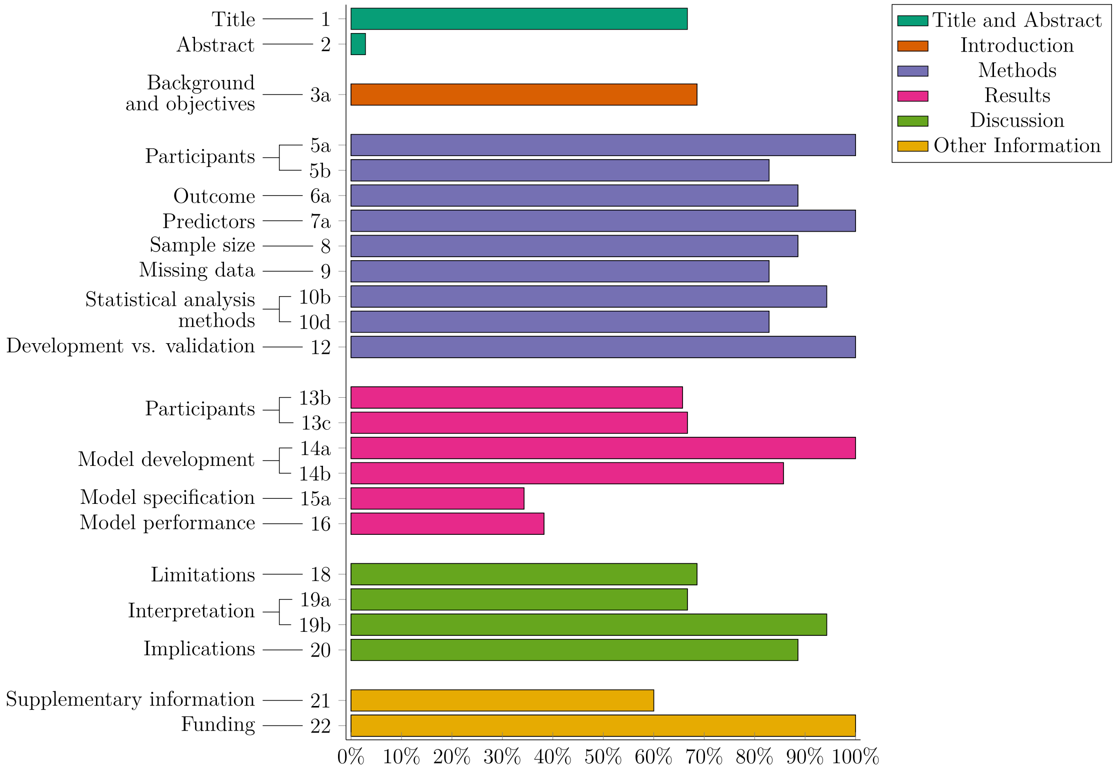
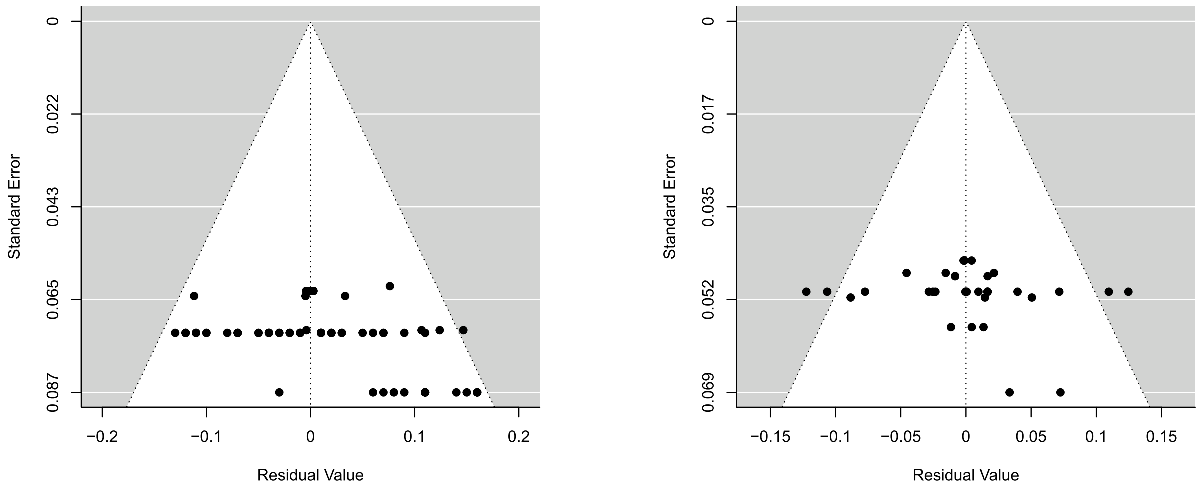


Fig. 6. Overall results from the TRIPOD assessment of reporting quality. Bars represent average scores across studies. Items are nested into topics which are nested within sections, following the specification in TRIPOD. Sections and topics are shown, while items can be inspected in more detail in Table D.5 or Moons et al. (2015), Heus et al. (2019). Items 7a, 10b and 15a were adjusted following Wang et al. (2015). Table D.5 reflects these adjustments.



(a) Correlation coefficient

(b) R-squared

Fig. 8. Funnel plots for (a) the 66 results pertaining to the correlation coefficient and (b) the 34 results pertaining to R-squared meta-analyses.

larger influence in the literature. Later works were highly influenced by these and, in a way, the current state of the literature reflects those earlier successes. A few studies do not cite other earlier studies in Fig. 3, likely because not every document focused exclusively on individualized prediction and/or intelligence. That should be taken into account when examining most results, especially TRIPOD ratings. See the TRIPOD checklist (Heus et al., 2019).

In the case of T-fMRI, results are largely compatible across datasets, but not across tasks. Greene et al. (2018), Gao et al. (2019), Jiang, Zuo, Ford, et al. (2020) show that FC derived from tasks are stronger predictors of G_F than RSFC. The gambling and the working-memory tasks demonstrate higher predictive power in all three studies. Wu et al. (2020), Sripada et al. (2020) also found that the working-memory task is highly discriminative of G, this time using statistical spatial maps.

While some of the studies presented results on more than one MRI modality, only one study presented a model that learns from multimodal data. Jiang, Calhoun, Cui, et al. (2020) presented results on both vertexwise cortical thickness and region of interest (ROI)-based RSFC. They show that a model that uses both modalities at once attains significantly higher predictive accuracy for intelligence compared to single-modality models. Choi et al. (2008) “neurometric model” includes both cortical thickness and T-fMRI statistical maps as inputs, but each part of the model was learned in isolation.

The HCP (Glasser et al., 2016; Van Essen et al., 2013) is the most employed dataset, appearing in 68% of the sample. Dating its first releases back to 2013, it began being employed for the prediction of intelligence as early as 2015 (Finn et al., 2015).

The majority, encompassing 76% of eligible studies, employed linear modeling for regression to some extent. Linear modeling is a strong baseline, also appearing in studies employing non-linear models. The most popular linear approaches include CPM and penalized linear models, each appearing in 36% and 42% of studies using linear models, respectively. CPM (Shen et al., 2017) is a very streamlined approach to predictive modeling. It is based on building linear models to predict outputs from aggregate measures of correlation between inputs and outputs, after thresholding based on significance. Features that are kept are then divided into positive-feature and negative-feature networks (Finn et al., 2015). Features in each network are summarized, e.g., summed or averaged, for each sample. Then, linear regression is used to predict outputs from these aggregate features, either separately or jointly for the positive-feature and negative-feature networks. After its introduction by Finn et al. (2015), albeit not yet named CPM, it became a staple of neuroimaging-based predictive modeling. Even though “connectome” appears in its name, the same principle can also be extended to other domains such as morphometry (Jiang, Calhoun, Cui, et al., 2020). Penalized linear modeling, on the other hand, does not aggregate features. Often, univariate filtering based on significance thresholding is used, akin to CPM. Then, however, remaining features are used as they are, without any additional transformation. The rationale for it is that penalization of coefficients can resolve commonalities and differences in features, and effectively attenuates overfitting. Also, aggregation by summing connectivity values amounts to weighting all features equally, which might be suboptimal.

Non-linear regression modeling appears in only a few studies, 19% of the sample. This might be due to the intrinsic high dimensionality of neuroimaging data, particularly evident for fMRI. At such high dimensionality, overfitting becomes a greater concern for more flexible models. The only non-linear model appearing more than once is KRR, a kernelized penalized linear regression. It is a very flexible approach given that a similarity measure between samples can be derived. Instead of using the base features in the model, features are expanded to higher (potentially infinite) dimensionalities. The kernel is the dot product between samples in this high dimensional space, which allows for efficient computation of models, bypassing the need of explicitly computing features in the new basis. In the sample, the correlation and the Dice overlap kernels were used in different studies. Due to the implicit high dimensionality, penalization is used very often, such as the ridge penalty, in the case of KRR.

Across studies, prediction is usually performed in aggregate measures of the data. Abrol et al. (2021) systematically shows that deep neural networks when trained on raw data outperform classical linear and non-linear ML models in the prediction of age, gender and Mini Mental State Examination scores. They also show that embeddings obtained from deep neural networks provide strong features for classical ML. This suggests that the choice of features in the literature has the potential of negatively biasing the performance of deep neural networks. Deep neural networks allow for using structured data, due to their inductive biases, present in architectures such as CNNs for image data or RNNs for sequence data. Only a few studies use deep neural networks for the prediction of intelligence using neuroimaging. He et al. (2020) modeled G_F based on RSFC with three deep neural networks. Fan et al. (2020) modeled G_F and G_C based on dynFC with RNNs. Vieira, Dubois, Calhoun, and Salmon (2021, not in this review) implements prediction of G also with RNNs, but based on RS-fMRI time series.

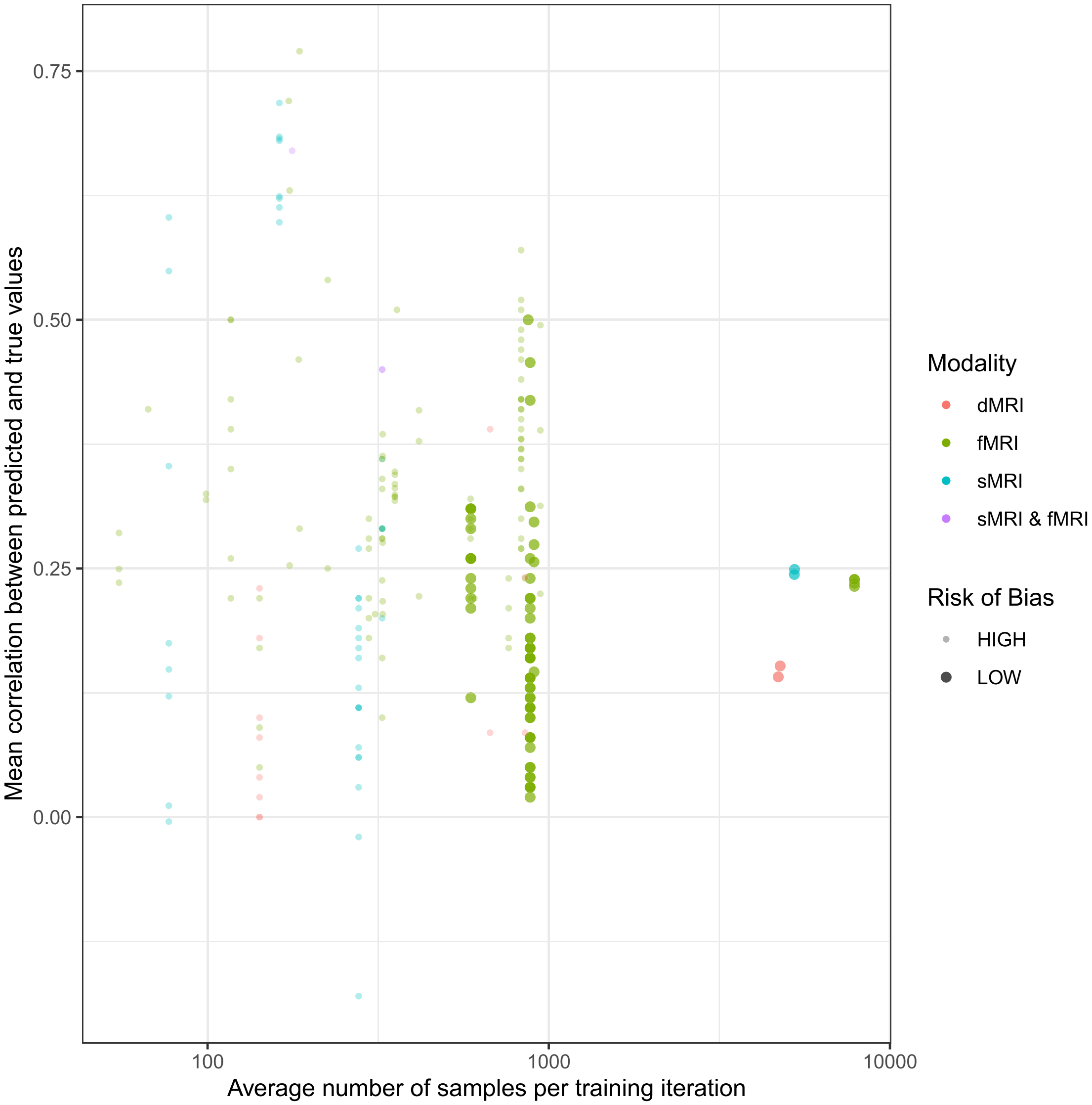


Fig. 9. The expected correlation coefficient according to the approximate training size data employed across studies. With the exception of holdout-based studies, where the actual training set size is known, the approximate training size data was estimated as the total number of data available for training times by $(K-1)/K$, where K is the number of groupings. Low risk of bias refers to studies that were rated with low RoB and low concern regarding applicability in Table 3. Modality refers to the imaging modality of each individual result.

4.1. *Limitations across studies and recommendations*

We must first state that limitations found in the analyzed studies have to be examined under the light of the current review's question, i.e. what current literature regarding the ML-based prediction of intelligence using neuroimaging looks like. Many studies did not focus primarily on the prediction of intelligence, even though they included such results. Studies proposing or benchmarking modeling choices, i.e. preprocessing, ML, and imaging methods, will often include intelligence among their results. A common occurrence in these studies, that include several outcomes, is that they will not give results in text format. When results are shown only graphically, we decided to not use inferred numbers. Also, when assessing the TRIPOD checklist, we only scored items that were clearly within the scope of the document. For example, studies not primarily concerned with prediction were not penalized by not mentioning prediction in their title, i.e. item 1.ii in TRIPOD.

4.1.1. *On the measurement of intelligence and its quality*

For both correlation and R-squared results, G-based results are significantly higher than G_F-based ones. This alludes to Gignac and Bates (2017), who showed that higher measurement quality moderates the observed correlation between intelligence and brain volume. In our assessment in Table 2, G derived from 10-tests in the NIHTB and PennCNB was rated as excellent, while G_F or "cognitive ability" obtained from a single test was rated as fair. A few results on G_F on He et al. (2020) employ the 2-min, 13-item test from the UK Biobank, which quality is unclear but would probably be rated as "poor" in Gignac and Bates (2017). Furthermore, Dubois, Galdi, Paul, and Adolphs (2018), Dubois, Galdi, Paul, & Adolphs, 2018 used the same predictor data based on RS-fMRI, but obtained very disparate results using the HCP. The authors reported $r = 0.263$ and $R^2 = 0.047$, when predicting PMAT-based G_F, versus $r = 0.457$ and $R^2 = 0.206$, when predicting G based on the factor analysis of 10 tests in the PennCNB and NIHTB.

It might be the case that performance would improve with better measurement of G_F, although Bilker et al. (2012) shows high correlations (> 0.9) between the original 60-item RSPM and abbreviated versions with as few as 9 items. For a comparison, the PMAT in the HCP is 24-items long, and in the PNC it is either 18- or 24-items long, which should guarantee a very high correlation with the 60-item RSPM. No study employed the full 60-item computerized RSPM, which should take 17 min to administer, on average (Williams & McCord, 2006). G_F in the UK Biobank, on the other hand, is measured by a very brief verbal and numerical reasoning test tailored to that cohort, with no comparable reference test according to Fawns-Ritchie and Deary (2020). The time restriction adds a substantial working memory load to the test scores (Chuderski, 2015). It remains to be investigated if improvements to the test used in the UK Biobank would lead to measurable improvements in performance.

The measurement of G and G_F may incur risks of bias compromising proper estimation of intelligence. The use of a single-domain test, such as inductive reasoning in RPM, would evaluate an isolated skill and not measure adequately intelligence, which is by definition a set of different cognitive skills. Furthermore, a test that assesses different skills needs to cover more than one cognitive domain, e.g., verbal, visual or spatial, to obtain a complete measurement (Gignac & Bates, 2017).

Another bias in interpreting results can occur due to the omission of information related to the measurement of intelligence. Studies must present a construct, e.g. intelligence, G, or G_F, and the psychological test used to measure it, so that it is possible to verify whether the test is adequate to measure the function contained in the specific construct. However, a psychological test suitable for the construct is not necessarily suitable for the population studied. It is essential to ensure tests are adequately validated for the population under study.

Despite the solid empirical basis of the concepts of G, G_F, and G_C, there are still concerns regarding cognitive abilities associated with G (Kent, 2017). New research on the neuroimaging-based prediction of intelligence should bring more specifications when evaluating cognitive constructs, such as the psychological instrument, validity, and application range.

The current conceptualization of the intelligence construct does not encompass only G_C or G_F. It covers adaptability and problem-solving in real life, considering emotional intelligence factors, decision making (Stankov, 2017), and personality (Kent, 2017). The interaction of these cognitive processes in an integrated way configures a complex multidimensional construct (McGrew, 2009). Due to this characteristic, it is recommended to use as many specifications as possible when performing the intelligence measurement.

The best model for the development of psychological instruments in intelligence evaluation is the Cattell-Horn-Carroll (CHC), seen as the best psychometric evidence for human aptitudes (Abu-Hamour & Al-Hmouz, 2016; Hurks & Bakker, 2016; James, Jacobs, & Roodenburg, 2015; Lecerf, Reverte, Coleaux, Favez, & Rossier, 2010; Wechsler & de Cassia Nakano, 2016). CHC theory consists of a hierarchical multidimensional model with ten factors of cognitive functioning: Fluid intelligence (Gf), Quantitative knowledge (Gq), Crystallized intelligence (Gc), Reading and writing (Grw), Short-term memory (Gsm), Visual processing (Gv), Auditory Processing (Ga), Long-term memory storage and retrieval ability (Glr), Processing Speed (Gs) and Decision speed (Gt). However, there is criticism over its weak explanatory capacity, its failure to make testable predictions, and its enmeshment to the Woodcock-Johnson battery (Wasserman, 2019). TheWoodcock-Johnson battery of tests (Woodcock, McGrew, & Mather, 2001) was designed to be more aligned to the CHC theory. However, there is evidence against this alignment and the lack of support for interpreting most of the scores suggested by its scoring system (Dombrowski, Beaujean, McGill, Benson, & Schneider, 2019). To date, no psychological test measures the broad cognitive abilities established in the CHC model which are contained in intelligence. For an adequate measurement, one should make use of instruments that are most related to the CHC theory, e.g., WAIS or Woodcock-Johnson Tests.

4.1.2. *The prevalence of gF*

The preponderance of G_F has three probable causes: (1) early success, as reported in Finn et al. (2015), which is cited by 27 out of 33 possible studies, as can be seen in Fig. 3; (2) ease of estimation, since it is often taken to comprise the score of a single test; and (3) availability, which compounds with the last reason, since RPM scores are available from the HCP, UK Biobank, BGSP and PNC.

The prevalence of G_F presents some challenges regarding the validity of results. The RPM can be considered a good score to include for the estimation of G and G_F. Current studies show that G_F and G have a strong correlation and are often statistically indistinguishable (Caemmerer, Keith, & Reynolds, 2020). In isolation, however, according to the criteria published in Gignac and Bates (2017), the RPM would be considered at best a “fair” measure of G. Similarly, although it is correlated with G_F, it does not appear to be remarkable in comparison with other tests that measure G_FGignac (2015). These findings point to the necessity of investigating what the models are learning through the RPM, and how much of it is shared between G, G_F and test specific variance. This would better clarify how much the prediction of RPM correlates with prediction of G_F.

Lohman and Lakin (2012) argue that G_F consists of three components: sequential reasoning, quantitative reasoning and inductive reasoning. The latter is the core of RPM. For this reason, Gignac (2015) argues that RPM can be considered an imperfect measure of G_F. This is due to its narrower scope, consisting exclusively of figural type items. All studies that predicted G_F employed the RPM to some extent. Most, 19 out of 20, used solely the RPM, with the remaining one employing both the RPM and NIHTB’s fluid composite score. For this reason, their results necessitate further consideration.

4.1.3. *Modeling approaches*

Schulz et al. (2020, not in this review) show that non-linear models do not show performance advantages in the prediction of G_F in large scale datasets. He et al. (2020), on the other hand, argues that a non-linear traditional machine learning model outperforms deep learning in the prediction of G_F in large scale datasets. Abrol et al. (2021) argues that it might be that processed features, used in both works, discard task-specific information. Representation learning with deep learning could extract more informative features tailored for the task at hand from minimally processed data. We can conclude that ROI-level summarization favors traditional ML, and particularly linear models. On the other hand, less- or minimally-processed and structured data would favor non-linear approaches based on representation learning, as has been shown in Fan et al. (2020) and Vieira et al. (2021).

Linear models are, in fact, more abundant than non-linear ones among the references reviewed. CPM, in its original formulation, averaged features either positively- or negative-aligned with the target variable. Wei et al. (2020) mentions that this gives equal weighting to all selected features, which might be suboptimal for the task, and would then explain the lower performance of CPM compared to linear models that do not equally weight predictors (Dubois, Galdi, Han, et al., 2018; Gao et al., 2019; Wei et al., 2020). While the original CPM is more interpretable, among the reviewed articles more optimized choices have been demonstrated.

4.1.4. *The level of expected evidence*

The literature constructs a clear picture regarding the level of expected evidence: correlations between brain imaging data and intelligence are substantial, albeit reliably low. It hovers around between 0.12 and 0.25 in large sample-size studies based on the UK Biobank (Cox et al., 2019; He et al., 2020), shown in Fig. 9. According to our quantitative analysis, the confidence interval covers between 0.35 and 0.50 for G and 0.13 and 0.17 for G_F based on fMRI data only. A possible explanation for this is that, in fact, the current data only affords such a level of performance. This also means that unexplained components of intelligence could be potentially learned in other spatial and temporal resolutions and imaging contrast mechanisms. Another, more problematic hypothesis is that ML is capturing relationships with other behaviors and demographics that correlate with intelligence, but not intelligence itself. This “shortcut learning" (Geirhos et al., 2020) is a major challenge for ML generalizability and interpretability. Possible shortcuts could include attention and arousal, but can go much deeper, to include substance abuse, malnutrition or socioeconomic status. Population modeling is one alternative to estimate how much brain data contributes to prediction of mental traits, i.e. Dadi et al. (2021, not in this review) demonstrates that, despite statistical significance, multimodal brain data contributes little to the prediction of G_F compared with sociodemographics.

4.1.5. Proper assessment of performance

We noted a large diversity in the methods for the assessment of performance. Techniques such as LOOCV and CV are widely employed. Proper inner loops of validation are often reported for performance tuning. However, the excessive use of the same datasets for internal validation leads to the risk of overfitting across studies, where differences between performance is due to random chance and not systematic differences between modeling approaches. Assessment of performance in an independent sample is often favored in the ML literature for that reason. Several studies implemented external validation, often across datasets (Cox et al., 2019; Elliott et al., 2019; Gao et al., 2019; Greene et al., 2018; Jiang, Calhoun, Fan, et al., 2020; Li et al., 2020). Li et al. (2020) also implements validation across time, demonstrating the stability of predictions. The usage of proper validation strategies is fundamental to assess meaningful differences between models.

On the choice of performance metrics, we see that Pearson correlation coefficient and R-squared are the most common in the literature. This is due to their scale invariance and perceived ease of interpretation. Despite their popularity, both are prone to biases. The correlation coefficient represents the linear association between predictions and true outcomes. Its formulation does not involve actual residuals, so models with arbitrarily large errors can still achieve perfect unitary correlation. Since R-squared involves a ratio, the denominator that represents the variance of true values can arbitrarily reduce or augment it. In other words, too small (or too large) variance of intelligence in the sample can lead to small (or large) R-squared, even under the same model (Alexander, Tropsha, & Winkler, 2015). This means that comparisons between studies, specially when their outcomes and/or populations differ, is at elevated risk of bias. A different choice of population incurs different characteristics of the outcome variance, possibly compromising the comparison. Model comparison on the same data could be performed under a well-behaved metric, such as the MSE or MAE.

4.1.6. Publication bias and selective reporting

We detected censoring for studies with high ROB and small sample sizes, as can be seen in Fig. 9. Their variability and the frequency of negative results diminish with models trained on less than 300 subjects. This is a qualitative indicator of publication bias, but also of selective reporting, since most studies report comparisons with multiple models. This selective reporting can be a result of the issue described in Hosseini et al. (2020), where authors perform optimization of their models on the same data that performance is measured, leading to inflated performance estimates due to overfitting to the test set and leakage.

4.1.7. Lack of diversity of data samples

The diversity of populations under study across studies is skewed towards a select group of countries. The 14 datasets identified can be grouped accordingly into United States (HCP, NKI, PNC, BGSP, COBRE), New Zealand (DMHDS), United Kingdom (UK Biobank), China (UESTC, SLIM, SXMU), South Korea (NRI, KAIST) and North America/Europe (ABIDE-I). Earlier releases of ABIDE were for the most part based on United States populations as well (New York University (NYU), Kennedy Krieger Institute, Stanford, Oregon Health & Science University, University of California, Los Angeles as in Wang et al. (2015)), and the only study using the ADHD-200 dataset employs solely NYU data. This limitation stems from economic factors that affect countries differently. While some datasets sampled highly-educated young adult populations, several others are matched samples from the local general population, which diminishes risks of biases. The prediction of G_F from the HCP, specially that assessed by the RPM, is very predominant in the literature. Albeit large datasets are often employed, the homogeneities across studies raise concerns regarding generalizability to other populations. Future works could perform validation analyses of trained models on new datasets, taking special care of differences in imaging acquisition and pre-processing.

4.1.8. Neuroscientific value

While earlier association works helped to foment new theories on intelligence, current ML-based works have not yet contributed substantially to this endeavor. This comes from the fact that the majority of the works do not try to extract explanatory value from the trained models. Few works test the leverage of different features and how these fit within or without theories such as Parieto Frontal Integration Theory (P-FIT), Network Neuroscience Theory (NNT) and the Multiple-Demand (MD) system (Duncan, 2010). Future works and possibly meta-analyses can solidify these findings, providing support for existing or new theories.

Li et al. (2020) discuss their findings from white-matter RSFC in light of P-FIT. This includes the importance of FC from the superior longitudinal fasciculus, which is central to P-FIT, but also other networks not included in the classical P-FIT. Dryburgh et al. (2020) concludes that many regions included in their modeling approaches, for both autism and neurotypical samples, coincide with P-FIT predictions. Hilger et al. (2020) concludes that the preference for frontal and parietal regions in their study aligns with P-FIT and MD. Sripada et al. (2020) highlight that a fronto-parietal network (FPN) and related executive regions are implicated in both P-FIT and MD. Jiang, Calhoun, Cui, et al. (2020) concludes that several regions included in their model, comprising (DMN), executive control network and a subcortical network, conform to P-FIT, with the exception of the cingulate, which they argue might be omitted due to heterogeneity in the sample. Cox et al. (2019) extensively discuss their results, including non-predictive ones, regarding P-FIT. Their work highlights that cortical and subcortical gray-matter volumes are more predictive of G than white-matter dMRI properties. Volumes of orbitofrontal, subcallosal, central, precentral, insular and precuneus had high associated effects over the prediction of G, but were not entirely encapsulated in classical P-FIT. Greene et al. (2018) concludes that the regions included in their model relating T-fMRI FC to G_F are consistent with P-FIT. Dubois, Galdi, Han, et al. (2018) report that predictive RSFC edges are distributed across the cortex, but connections encompassed by the FPN, DMN, cingulo-opercular network and the visual network are highlighted in their analysis, which align with P-FIT. Yang et al. (2013) concludes that their results based on sMRI morphometrics highlights some of the regions implicated in P-FIT, with the exception of frontal regions, a divergence they attribute to using different methods in the assessment of brain-intelligence associations. Jiang, Calhoun, Fan, et al. (2020)Hilger et al. (2020)Li et al. (2020) cite the NNT, but none discuss their results in the context of this theory.

Another concern lies in the fact that the current theories were mostly derived from univariate association studies. Omissions in the predictive models of features implicated in theories do not, necessarily, contradict said theories. These could be simply caused by the fact that models account for the correlation between predictors. ML-based studies are fundamental to update theories to conform to multivariate associations.

4.1.9. Factors that increase risk of bias

A common occurrence in the assessment of PROBAST was that studies did not take into account the optimistic bias of confounders received high RoB ratings for “Analysis” Table 3. A notorious confounder which should be taken into account is kinship, in datasets like the HCP (Dubois, Galdi, Paul, & Adolphs, 2018). This requires adaptations, such as LOFOCV. Other, more pervasive ones, include movement and brain volume, but also sex and age. Two common approaches in the literature are removing the effect of confounders using linear models and stratifying data in a way to minimize bias due to confounders, the latter a very common approach when dealing with family structure. While our work cannot determine optimal strategies for treatment of confounders, low RoB studies were expected to recognize their effects and account for it in results.

Another common factor leading to high RoB was small sample size. It is a well-known fact from the literature that ML-based studies suffer spurious correlation induced by small samples (Varoquaux, 2018). The few studies that report the standard error of the mean cross-validated performance also likely underestimate it (Varoquaux, 2018). Recognizing the negative impact of small sample sizes, having fewer than 500 subjects was considered as an indicator of possible RoB in the assessment of Table 3.

4.1.10. Recommendations

A number of recommendations can be made based on the limitations we identified: (1) evidence points to the fact that constructs with higher measurement quality are easier to predict, thus, it is desirable to have the highest measurement quality as possible; (2) confounders have to be carefully selected and controlled for. Confounders can inflate performance estimates if not representative of the population (Rao et al., 2017), and in the case of explainability, can lead to erroneous conclusions, e.g., due to mutual causation. In general, age, sex and intracranial volume or other measures of head size have probable causal effects on both neuroimaging- and cognition-derived variables, making them candidates as confounders. Movement has a probable common cause to intelligence, making both correlated (Siegel et al., 2017). Its effect on images, however, is mostly non-neuronal (Li et al., 2019; Siegel et al., 2017), where ML models could in principle learn that subjects that have less motion-related artifacts are likely to be more intelligent, which is a correct conclusion, but not of interest when one is searching for brain-based markers of intelligence. Therefore, the risk of “shortcut learning" makes it a potential candidate as a confounder. Years of education, socioeconomic status and related variables have causal effects to intelligence, but improbable direct effects on brain-derived variables. This in turn makes both not candidates as confounders in the general population, especially if effects on health, particularly vascular and mental health, can be ruled out. Kinship is a potential confounder because both brain- and cognition-derived variables demonstrate a degree of heritability. The model is evaluated as if it would be applied to new subjects not related to the ones in the training set. Also, under non-representative sampling, any variable has the potential to become a confounder. Chyzhyk, Varoquaux, Thirion, and Milham (2018) presents an overview of strategies to control confounders which can be explored by future studies, including a novel strategy based on anti mutual-information; (3) the use of scale- and location-invariant performance metrics, such as Pearson correlation or R-squared, while attractive for its purported interpretability, can lead to erroneous conclusions. Reporting of other metrics such as MSE is desirable, but due to normalization, standardization and possibly residualization of confounders comparison between models and datasets is not straightforward, and authors should be cognizant of those caveats; (4) proper use of CV and other procedures to estimate generalizability performance. This includes using stratification when it makes sense, and also nested validation for selecting best hyperparameters/models, since the “best” combination can be due to random chance. This also includes being aware of possible sources of leakage; (5) studies could probe predictions made by theories, *e.g* P-FIT and NNT, and obtain neuroscientific insights from the ML models; (6) studies that explore data with structured temporal, e.g. time series or dynFC, and/or spatial, e.g. minimally processed sMRI, are likely to benefit from the use of deep learning. However, studies interested in ROI-level associations with intelligence will probably not benefit from deep learning models.

Journals could adopt a guideline for self assessment of reporting quality and RoB by authors, e.g., in the format TRIPOD and PROBAST. Reviewers could then assess the adherence to the guidelines using checklists, in final stages of peer-review. It is not clear, however, how much guidelines actually help to improve quality (Zamanipoor Najafabadi et al., 2020).

4.2. Limitations of the review

Some possible limitations can be identified in our review methodology.

4.2.1. Systematic searches and data retrieval

Searching for manuscripts on predictive modeling on neuroimaging is particularly challenging. In the early literature, the term “predict” would often be used to refer to studies on correlations and associations. For this reason, we had to use a search strategy based on domain-knowledge. This choice, however, incurs the risk of selection bias due to missing documents. Since we successfully retrieved a reasonable number of documents, we believe that we minimized this risk and also obtained a representative sample. It is however expected that our selection missed documents, but we believe that this number should be small.

The large diversity of use of terms and also the fact that intelligence is not the main object of study in several studies makes systematic search more difficult, leading to omissions (for example Marc-Andre Schulz et al., 2020, Avery et al., 2019, Pervaiz, Vidaurre, Woolrich, & Smith, 2020, not in this review). Narrowing down the search to full-articles misses other forms of publications, such as conference papers (for example Mihalik et al., 2019, not in this review). Lastly, the end date of the search will also lead to the omission of new articles (for example Vieira et al., 2021, Dadi et al., 2021, Dhamala, Jamison, Jaywant, Dennis, & Kuceyeski, 2021, Feilong, Guntupalli, & Haxby, 2021, Schulz et al., 2022, Cai et al., 2021, Frith et al., 2021, Sen & Parhi, 2021, not in this review). The authors were made aware of some of these references after the protocol registration, while others were noted by an anonymous review during peer-review. We offer this list in the hope it might be useful for future studies.

The fact that most studies either did not focus solely on intelligence or were not primarily about individualized prediction makes data retrieval difficult. For this reason, in several instances constructs and instruments are not readily identified in searchable text. We thoroughly searched for information in actual figures and supplementary materials. We did not follow citations or other sources to infer this information, since the construct should ideally be stated by authors.

Another source of variance is the fact that terminology is flexible. Studies will often use terms like cognitive ability or others with ambiguous meaning. For example, in Elliott et al. (2019) “cognitive ability” refers to both FSIQ and G_F, while in Sripada et al. (2020) “general cognitive ability” names a measurement that is identified with G in other studies. Some works will refer to a G-like construct as general intelligence, others will refrain from using the term intelligence altogether. We tried to disambiguate authors’ choices with the coherence of the review in mind. This is particularly evident in Table 2, where we tried to unify terminology.

4.2.2. Use and adaptation of standardized tools

We adopted the TRIPOD adherence assessment form (Heus et al., 2019) to evaluate reporting quality. That benefits objectivity in this analysis. Measuring adherence to a specific reporting guideline has the disadvantage of potentially misrepresenting studies. This guideline is not enforced by journals, reviewers or the authors themselves in this research area. This form has been similarly applied to documents published prior to TRIPOD (Zamanipoor Najafabadi et al., 2020). Due to the generality of TRIPOD items, we believe that the risk of bias is low regarding the assessment of reporting quality. Several studies achieved high ratings, as can be seen in Fig. 7.

We employed PROBAST to assess RoB and applicability. PROBAST is a tool designed primarily for studies in health and medicine, but its items are still very applicable to our review question. Another benefit is that the use of standardized tools minimizes biases when compared with an alternative created by authors. This was a post-hoc adaptation from the protocol in Vieira et al. (2021), but, with aforementioned justification, we also consider that the risk of inducing bias is low.

We employed the PRISMA checklist as a reporting guideline. PRISMA was designed for studies that evaluate healthcare interventions, but most items can be applied to our review question. We believe that this choice offers no additional risk of bias for our review.

The quality of measurement of intelligence was evaluated by the first three criteria of the essential guide for categorizing the quality of general intelligence measurement (Gignac & Bates, 2017). Although the guide was proposed for G, we also used it to assess the quality of measurement of G_F.

4.2.3. *Lack of data for further inferences*

The number of studies using modalities other than fMRI with low RoB and low concerns regarding applicability was insufficient for quantitative analysis. For this reason, we only obtained meta-estimates of correlation and R-squared from fMRI. Fig. 9 seems to point towards an approximately unique ceiling in performance, but the small number of studies, especially truly multimodal ones, makes that inference inconclusive.

4.3. *Future work*

Future work could explore other imaging techniques, such as PET, EEG and MEG. These imaging techniques probe different functional aspects from fMRI. PET allows the study of slow metabolic dynamics in the brain and was fundamental for the definition of the P-FIT, being employed in the study of metabolic response differences under cognitively demanding tasks (Jung & Haier, 2007). EEG and MEG, on the other hand, probe fast electrical cerebral dynamics, and their importance was also acknowledged in P-FIT, albeit neither was part of its experimental foundation. We are aware of at least one study that employs EEG to the prediction of G_F (Hakim, Awh, Vogel, and Rosenberg, 2021, not in this review). In addition to other imaging techniques, multimodality presents an avenue for future research. It is currently not possible to establish whether information learned from different modalities overlap due to the lack of large numbers of multimodal models. Studies employing two or more techniques or modalities at once can better disambiguate the predictive power exclusive to each. This type of study is, however, becoming more widespread. Jiang, Calhoun, Cui, et al. (2020) model anatomical and RSFC data jointly, Dhamala et al. (2021, not in this review) use dMRI structural connectivity and RSFC, and Dadi et al. (2021, not in this review) includes joint modeling based on RSFC, dMRI diffusion measurements, and sMRI global and regional volumes.

Most works employ ROI-level features. Although this “summarization” makes ML more amenable, since it diminishes the dimensionality of data, this level of spatial abstraction can discard useful intra-regional information. Feilong et al. (2021, not in this review) systematically demonstrates that accounting for fine-grained, intra-ROI task and resting-state FC differences lead to improvements in the prediction of G and other intelligence measurements. Schulz et al. (2022, not in this review) shows that the performance ceiling for the prediction of G_F from RS-fMRI-, dMRI- and sMRI-extracted predictors has not been reached yet on the largest dataset available today, with linear and non-linear models performing on par with each other. It can then be argued that the true level of association between brain-derived predictors and intelligence cannot be probed with the current paradigm with the available datasets at present. Future developments on data-efficient ML models that can robustly learn from minimally preprocessed data have the potential of resolving this abstraction and discovering relationships hidden by summarization.

Other ML algorithmic developments can improve prediction accuracy and validity in the future. In particular, interpretable and explainable models can further corroborate, falsify and augment current theories on the biological bases of intelligence, which were, for the most part, developed based on coarse-grained spatial attributes of brain anatomy and function.

Refinements of psychometric and neuroscientific theories of intelligence will also lead to a demand for future work. Intelligence differences do not occur in isolation, being permeated by other human behaviors and environmental factors. The extended P-FIT (ExtPFIT) was formulated in Gur et al. (2021), and its generalizability can be tested in a ML-based framework. Other neuroscientific theories and extensions will probably emerge in the future.

Finally, larger scale datasets will diminish small sample-size biases in predictive models (Varoquaux, 2018). Jointly learning across different datasets and discarding confounding information efficiently can boost predictive accuracy. Future works can help answer if the patterns observed in current models generalize across different populations, socio-economic environments, languages and cultures.

4.4. *Conclusions*

Half of the identified studies include linear modeling to predict RPM-based G_F from HCP fMRI. This fact attests the significance and reliability of fMRI-based prediction studies. It also alludes to possible new avenues of research that have been studied infrequently if at all.

By pointing out salient results across studies and limitations, we hope that this work contributes to further developments in this area of research. While predictive modeling “best-practices” are abound, the literature currently lacks reporting guidelines, which could be fulfilled to ease literature search. Some gaps that can be filled by future studies include: extending and validating the current models in new populations, developing models using other spatiotemporal resolutions, other modalities, and imaging techniques, and disambiguating the contribution of neuronal phenomena to the predictions.

Disclosure statement

Authors declare the following that could be perceived as a conflict of interest: authors B.H.V. and C.E.G.S. maintain active collaboration with J.D. (Dubois, Galdi, Han, et al., 2018; Dubois, Galdi, Paul, & Adolphs, 2018) and V.D.C. (Jiang, Calhoun, Cui, et al., 2020; Jiang, Calhoun, Fan, et al., 2020; Jiang, Zuo, Ford, et al., 2020; Xiao et al., 2019; Xiao et al., 2020). J.D. and V.D.C. were not consulted or informed in any form regarding the current study.

Acknowledgements

Authors acknowledge comments provided by anonymous reviewers during peer-review, which substantially improved the paper structure and clarity.

Authors acknowledge funding received through grants #2018/11881-1, São Paulo Research Foundation (FAPESP) (B.H.V - PhD scholarship) and [10001C_197480], Swiss National Science Foundation at the University of Zurich (B.H.V. - postdoctoral position).

Appendix A. SCOPUS search string

We used the following query string for systematic search in SCOPUS:

(TITLE-ABS-KEY (("cortical thickness" OR "functional connectivity" OR "structural connectivity" OR "effective connectivity" OR mri OR fmri OR morphometry) AND (predict* OR "multivariate pattern analysis" OR bases OR cpm OR variability OR mvpa OR "machine learning")) AND (TITLE (intell* OR behav* OR "cognitive ability" OR iq)))

This query string was adapted from an initial search string, defined during preregistration:

(TITLE-ABS-KEY (("cortical thickness" OR "functional connectivity" OR "structural connectivity" OR "effective connectivity" OR mri OR fmri OR morphometry) AND (prediction OR predict OR cpm OR "multivariate pattern analysis" OR bases OR variability OR mvpa)) AND (TITLE (intelligence OR behavioral OR behavior OR "cognitive ability")))

Appendix B. PRISMA 2009 checklist

Table B.4
From: Moher et al. (2009).

Section/topic	#	Checklist item	Reported on page #
Title			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	page 1
Abstract			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	page 1
Introduction			
Rationale	3	Describe the rationale for the review in the context of what is already known.	page 2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	page 2
Methods			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	page 2
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	page 2
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	page 2
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	page 2
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	page 2
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	page 5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	page 5
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	page 5
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	page 6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I2) for each meta-analysis.	page 6
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	page 6
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	page 6
Results			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	page 6
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	page 7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	page 8
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	page 10
Synthesis of results	21	Present the main results of the review. If meta-analyses are done, include for each, confidence intervals and measures of consistency	page 10
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	page 11
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression see Item 16]).	page 12
Discussion			

(continued on next page)

Table B.4 (*continued*)

Section/topic	#	Checklist item	Reported on page #
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	page 12
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	page 13
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	page 18
Funding			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	page 18

Appendix C. Prediction performance metrics

Given a set of true labels $\boldsymbol{y}$, and a set of predictions $\hat{\boldsymbol{y}}$, several performance metrics can be defined.

C.1. Continuous valued labels

The Pearson correlation coefficient, defined as

$$\mathrm{r}(\boldsymbol{y},\hat{\boldsymbol{y}}) = \frac{\sum_i^N (\hat{y}_i - E[\hat{\boldsymbol{y}}])(y_i - E[\boldsymbol{y}])}{\sqrt{\sum_i^N (\hat{y}_i - E[\hat{\boldsymbol{y}}])^2}\sqrt{\sum_i^N (y_i - E[\boldsymbol{y}])^2}},$$

is the most popular performance metric for regression of continuous valued labels.

Other metrics include the MSE,

$$\mathrm{MSE}(\boldsymbol{y},\hat{\boldsymbol{y}}) = E\left[(y-\hat{y})^2\right] = \frac{\sum_i^N (y_i - \hat{y}_i)^2}{N},$$

the MAE,

$$\mathrm{MAE}(\boldsymbol{y},\hat{\boldsymbol{y}}) = E[|y-\hat{y}|] = \frac{\sum_i^N \left|y_i - \hat{y}_i\right|}{N},$$

and Spearman rank correlation coefficient,

$$\rho(\boldsymbol{y},\hat{\boldsymbol{y}}) = \mathrm{r}(R(y), R(\hat{y})),$$

defined in terms of Pearson's, but based on ranks instead of values, as denoted by the rank function $R(\cdot)$.

A few more metrics are linked to the MSE. These include the coefficient of determination, or squared deviance, R-squared,

$$R^2(\boldsymbol{y},\hat{\boldsymbol{y}}) = 1 - \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - E[y_i])^2},$$

the RMSE,

$$\mathrm{RMSE}(\boldsymbol{y},\hat{\boldsymbol{y}}) = \sqrt{\mathrm{MSE}(\boldsymbol{y},\hat{\boldsymbol{y}})},$$

the NRMSE,

$$\mathrm{NRMSE}(\boldsymbol{y},\hat{\boldsymbol{y}}) = \frac{\mathrm{RMSE}(\boldsymbol{y},\hat{\boldsymbol{y}})}{E[\boldsymbol{y}]},$$

the nRMSD,

$$\mathrm{nRMSD}(\boldsymbol{y},\hat{\boldsymbol{y}}) = \frac{\mathrm{RMSE}(\boldsymbol{y},\hat{\boldsymbol{y}})}{\sqrt{\sum_i^N (y_i - E[\boldsymbol{y}])^2}} = \sqrt{1-R^2},$$

and the mean percentage error (as used in Park et al., 2016),

$$\text{Percentage error}(\boldsymbol{y},\hat{\boldsymbol{y}}) = \frac{\sum_i^N |y_i - \hat{y}_i| / y_i}{N}.$$

C.2. Binary valued labels

In our sample, the only reported performance metric for binary valued labels $y_i \in \{0, 1\}$ were the AUC and accuracy.

AUC is defined mathematically as:

$$\text{AUROC}(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = \frac{\sum_i^N \sum_j^N \left(y_j - y_i\right)^2 \mathbf{1}_{\widehat{y_i} > \widehat{y_j}}}{\sum_i^N y_i \sum_i^N (1 - y_i)}.$$

Notice that $(y_j - y_i)^2 \equiv 1$ only when $y_i \neq y_j$, being 0 otherwise.

Likewise, accuracy is:

$$\text{Accuracy}(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = \frac{\sum_i^N \mathbf{1}_{\widehat{y_i} = y_i}}{N}.$$

Appendix D. Adjusted TRIPOD checklist

Table D.5
Adjusted TRIPOD checklist for reporting quality assessment.

Section/topic	Item	Checklist item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	Describe eligibility criteria for participants.	
	5c	Give details of treatments received, if relevant.	
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	Adjusted 7a	Clearly define all predictors used in developing or validating the ML model, including how and when they were measured.	
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	Explain how the study size was arrived at.	
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	
	Adjusted 10b	Specify type of model, all model-building procedures (including any predictor selection, hyperparameter selection if needed), and method for internal validation.	
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
Risk groups	11	Provide details on how risk groups were created, if done.	
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
Model development	14a	Specify the number of participants and outcome events in each analysis.	
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	
Model specification	Adjusted 15a	Present the full prediction model to allow predictions for individuals (i.e. links to the final model online (coding of predictors, codeand final parameters/coefficients, and with the architecture described in full in the article)).	
	15b	Explain how to the use the prediction model.	
Model performance	16	Report performance measures (with CIs) for the prediction model.	
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	
Implications	20	Discuss the potential clinical use of the model and implications for future research.	

(*continued on next page*)

Table D.5 (*continued*)

Section/topic	Item	Checklist item	Page
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
Funding	22	Give the source of funding and the role of the funders for the present study.	

Appendix E. Synthesis of additional meta-analytic results

Forest plots using all identified results are depicted in Fig. E.10, for R-squared, and Fig. E.11, for the correlation coefficient. Contrast with the ones in Fig. 5. Imaging modalities with less than 10 results and datasets with less than 4 results were excluded from these analyses.

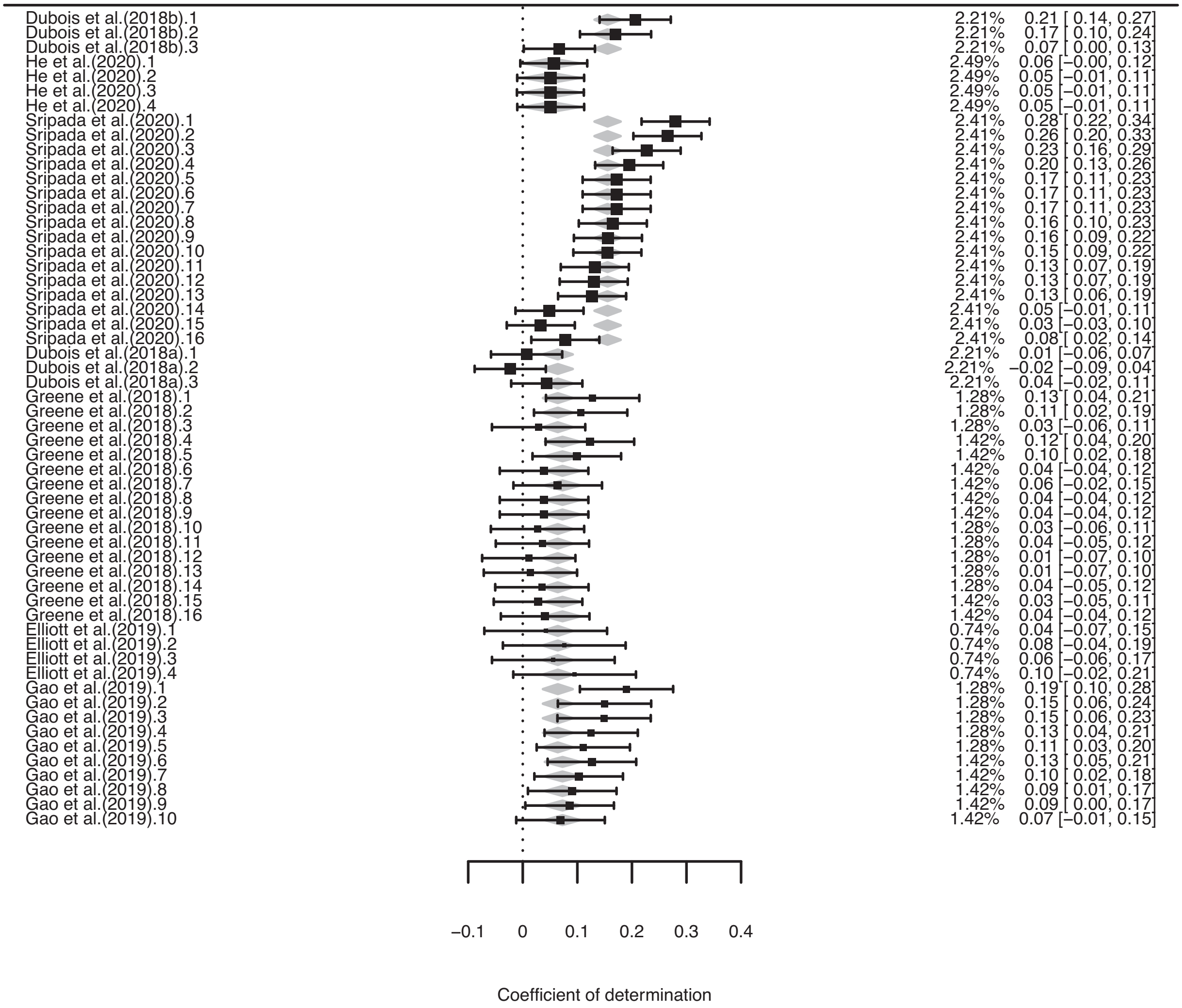


Fig. E.10. Forest plot for the R-squared meta-analysis. Outcome, either G, G_F or G_C, and dataset, either HCP, PNC or the UK Biobank, were included as moderators. All results were obtained using fMRI.

In Fig. E.10, no significant effect of dataset, either one of HCP, PNC, or UK Biobank was found (ANOVA $p = 0.80$). The expected R-squared was estimated as 0.155 ($CI_{95\%} = [0.131, 0.179]$, $p < 0.001$) for G and 0.065 ($CI_{95\%} = [0.045, 0.084]$, $p < 0.001$) for G_F, in a model without dataset moderators. The difference between G and G_F was estimated at 0.09 ($CI_{95\%} = [0.060, 0.121]$, $p < 0.001$). While the result for G is compatible with the one in Fig. 5, for G_F results without low RoB lead to an inflation in expected R-squared.

In Fig. E.11, significant effects were found for the outcome, dataset and imaging modality moderators (ANOVA $p < 0.001$ in all three cases). The expected correlation was estimated as 0.322 ($CI_{95\%} = [0.2602, 0.3835]$, $p < 0.001$) for G in HCP. The expected correlation was estimated as 0.187 ($CI_{95\%} = [0.166, 0.208]$, $p < 0.001$) and 0.159 ($CI_{95\%} = [0.086, 0.232]$, $p < 0.001$) for G_F in HCP and the UK Biobank, respectively, where no difference between datasets was found (ANOVA $p = 0.449$). Fig. 5, on the other hand, estimated higher G_F correlation in the UK Biobank and higher G correlation in the HCP.

The inclusion of a RoB (High/Unclear vs Low) moderator led to significant differences in both analysis.

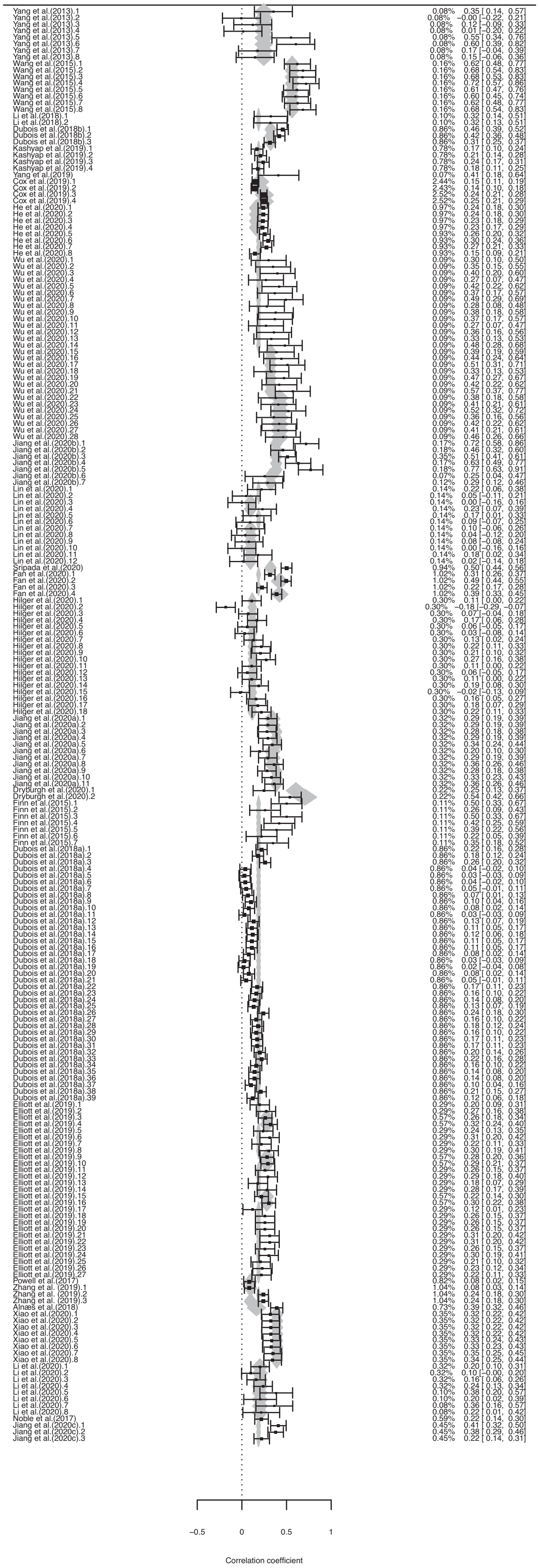

Yang et al.(2013).1 0.08% 0.35 [0.14, 0.57]
Yang et al.(2013).2 0.08% −0.00 [−0.22, 0.21]
Yang et al.(2013).3 0.08% 0.12 [−0.09, 0.33]
Yang et al.(2013).4 0.08% 0.01 [−0.20, 0.22]
Yang et al.(2013).5 0.08% 0.55 [0.34, 0.76]
Yang et al.(2013).6 0.08% 0.60 [0.39, 0.82]
Yang et al.(2013).7 0.08% 0.17 [−0.04, 0.39]
Yang et al.(2013).8 0.08% 0.15 [−0.06, 0.36]
Wang et al.(2015).1 0.16% 0.62 [0.48, 0.77]
Wang et al.(2015).2 0.16% 0.68 [0.54, 0.83]
Wang et al.(2015).3 0.16% 0.68 [0.53, 0.83]
Wang et al.(2015).4 0.16% 0.72 [0.57, 0.86]
Wang et al.(2015).5 0.16% 0.61 [0.47, 0.76]
Wang et al.(2015).6 0.16% 0.60 [0.45, 0.74]
Wang et al.(2015).7 0.16% 0.62 [0.48, 0.77]
Wang et al.(2015).8 0.16% 0.68 [0.54, 0.83]
Li et al.(2018).1 0.10% 0.32 [0.14, 0.51]
Li et al.(2018).2 0.10% 0.32 [0.13, 0.51]
Dubois et al.(2018b).1 0.86% 0.46 [0.39, 0.52]
Dubois et al.(2018b).2 0.86% 0.42 [0.36, 0.48]
Dubois et al.(2018b).3 0.86% 0.31 [0.25, 0.37]
Kashyap et al.(2019).1 0.78% 0.17 [0.10, 0.24]
Kashyap et al.(2019).2 0.78% 0.21 [0.14, 0.28]
Kashyap et al.(2019).3 0.78% 0.24 [0.17, 0.31]
Kashyap et al.(2019).4 0.78% 0.18 [0.11, 0.25]
Yang et al.(2019) 0.07% 0.41 [0.18, 0.64]
Cox et al.(2019).1 2.44% 0.15 [0.11, 0.19]
Cox et al.(2019).2 2.43% 0.14 [0.10, 0.18]
Cox et al.(2019).3 2.52% 0.24 [0.21, 0.28]
Cox et al.(2019).4 2.52% 0.25 [0.21, 0.29]
He et al.(2020).1 0.97% 0.24 [0.18, 0.30]
He et al.(2020).2 0.97% 0.24 [0.18, 0.30]
He et al.(2020).3 0.97% 0.23 [0.18, 0.29]
He et al.(2020).4 0.97% 0.23 [0.17, 0.29]
He et al.(2020).5 0.93% 0.26 [0.20, 0.32]
He et al.(2020).6 0.93% 0.30 [0.24, 0.36]
He et al.(2020).7 0.93% 0.27 [0.21, 0.33]
He et al.(2020).8 0.93% 0.15 [0.09, 0.21]
Wu et al.(2020).1 0.09% 0.30 [0.10, 0.50]
Wu et al.(2020).2 0.09% 0.35 [0.15, 0.55]
Wu et al.(2020).3 0.09% 0.40 [0.20, 0.60]
Wu et al.(2020).4 0.09% 0.27 [0.07, 0.47]
Wu et al.(2020).5 0.09% 0.42 [0.22, 0.62]
Wu et al.(2020).6 0.09% 0.37 [0.17, 0.57]
Wu et al.(2020).7 0.09% 0.49 [0.29, 0.69]
Wu et al.(2020).8 0.09% 0.28 [0.08, 0.48]
Wu et al.(2020).9 0.09% 0.38 [0.18, 0.58]
Wu et al.(2020).10 0.09% 0.37 [0.17, 0.57]
Wu et al.(2020).11 0.09% 0.27 [0.07, 0.47]
Wu et al.(2020).12 0.09% 0.36 [0.16, 0.56]
Wu et al.(2020).13 0.09% 0.33 [0.13, 0.53]
Wu et al.(2020).14 0.09% 0.48 [0.28, 0.68]
Wu et al.(2020).15 0.09% 0.39 [0.19, 0.59]
Wu et al.(2020).16 0.09% 0.44 [0.24, 0.64]
Wu et al.(2020).17 0.09% 0.51 [0.31, 0.71]
Wu et al.(2020).18 0.09% 0.33 [0.13, 0.53]
Wu et al.(2020).19 0.09% 0.47 [0.27, 0.67]
Wu et al.(2020).20 0.09% 0.42 [0.22, 0.62]
Wu et al.(2020).21 0.09% 0.57 [0.37, 0.77]
Wu et al.(2020).22 0.09% 0.38 [0.18, 0.58]
Wu et al.(2020).23 0.09% 0.41 [0.21, 0.61]
Wu et al.(2020).24 0.09% 0.52 [0.32, 0.72]
Wu et al.(2020).25 0.09% 0.36 [0.16, 0.56]
Wu et al.(2020).26 0.09% 0.42 [0.22, 0.62]
Wu et al.(2020).27 0.09% 0.41 [0.21, 0.61]
Wu et al.(2020).28 0.09% 0.46 [0.26, 0.66]
Jiang et al.(2020b).1 0.17% 0.72 [0.58, 0.86]
Jiang et al.(2020b).2 0.18% 0.46 [0.32, 0.60]
Jiang et al.(2020b).3 0.35% 0.51 [0.41, 0.61]
Jiang et al.(2020b).4 0.17% 0.63 [0.49, 0.77]
Jiang et al.(2020b).5 0.18% 0.77 [0.63, 0.91]
Jiang et al.(2020b).6 0.07% 0.25 [0.04, 0.47]
Jiang et al.(2020b).7 0.12% 0.29 [0.12, 0.46]
Lin et al.(2020).1 0.14% 0.22 [0.06, 0.38]
Lin et al.(2020).2 0.14% 0.05 [−0.11, 0.21]
Lin et al.(2020).3 0.14% 0.00 [−0.16, 0.16]
Lin et al.(2020).4 0.14% 0.23 [0.07, 0.39]
Lin et al.(2020).5 0.14% 0.17 [0.01, 0.33]
Lin et al.(2020).6 0.14% 0.09 [−0.07, 0.25]
Lin et al.(2020).7 0.14% 0.10 [−0.06, 0.26]
Lin et al.(2020).8 0.14% 0.04 [−0.12, 0.20]
Lin et al.(2020).9 0.14% 0.08 [−0.08, 0.24]
Lin et al.(2020).10 0.14% 0.00 [−0.16, 0.16]
Lin et al.(2020).11 0.14% 0.18 [0.02, 0.34]
Lin et al.(2020).12 0.14% 0.02 [−0.14, 0.18]
Sripada et al.(2020) 0.94% 0.50 [0.44, 0.56]
Fan et al.(2020).1 1.02% 0.31 [0.26, 0.37]
Fan et al.(2020).2 1.02% 0.49 [0.44, 0.55]
Fan et al.(2020).3 1.02% 0.22 [0.17, 0.28]
Fan et al.(2020).4 1.02% 0.39 [0.33, 0.45]
Hilger et al.(2020).1 0.30% 0.11 [0.00, 0.22]
Hilger et al.(2020).2 0.30% −0.18 [−0.29, −0.07]
Hilger et al.(2020).3 0.30% 0.07 [−0.04, 0.18]
Hilger et al.(2020).4 0.30% 0.17 [0.06, 0.28]
Hilger et al.(2020).5 0.30% 0.06 [−0.05, 0.17]
Hilger et al.(2020).6 0.30% 0.03 [−0.08, 0.14]
Hilger et al.(2020).7 0.30% 0.13 [0.02, 0.24]
Hilger et al.(2020).8 0.30% 0.22 [0.11, 0.33]
Hilger et al.(2020).9 0.30% 0.21 [0.10, 0.32]
Hilger et al.(2020).10 0.30% 0.27 [0.16, 0.38]
Hilger et al.(2020).11 0.30% 0.11 [0.00, 0.22]
Hilger et al.(2020).12 0.30% 0.06 [−0.05, 0.17]
Hilger et al.(2020).13 0.30% 0.11 [0.00, 0.22]
Hilger et al.(2020).14 0.30% 0.19 [0.08, 0.30]
Hilger et al.(2020).15 0.30% −0.02 [−0.13, 0.09]
Hilger et al.(2020).16 0.30% 0.16 [0.05, 0.27]
Hilger et al.(2020).17 0.30% 0.18 [0.07, 0.29]
Hilger et al.(2020).18 0.30% 0.22 [0.11, 0.33]
Jiang et al.(2020a).1 0.32% 0.29 [0.19, 0.39]
Jiang et al.(2020a).2 0.32% 0.29 [0.19, 0.39]
Jiang et al.(2020a).3 0.32% 0.28 [0.18, 0.38]
Jiang et al.(2020a).4 0.32% 0.29 [0.19, 0.39]
Jiang et al.(2020a).5 0.32% 0.34 [0.24, 0.44]
Jiang et al.(2020a).6 0.32% 0.20 [0.10, 0.30]
Jiang et al.(2020a).7 0.32% 0.29 [0.19, 0.39]
Jiang et al.(2020a).8 0.32% 0.36 [0.26, 0.46]
Jiang et al.(2020a).9 0.32% 0.28 [0.18, 0.38]
Jiang et al.(2020a).10 0.32% 0.33 [0.23, 0.43]
Jiang et al.(2020a).11 0.32% 0.36 [0.26, 0.46]
Dryburgh et al.(2020).1 0.22% 0.25 [0.13, 0.37]
Dryburgh et al.(2020).2 0.22% 0.54 [0.42, 0.66]
Finn et al.(2015).1 0.11% 0.50 [0.33, 0.67]
Finn et al.(2015).2 0.11% 0.26 [0.09, 0.43]
Finn et al.(2015).3 0.11% 0.50 [0.33, 0.67]
Finn et al.(2015).4 0.11% 0.42 [0.25, 0.59]
Finn et al.(2015).5 0.11% 0.39 [0.22, 0.56]
Finn et al.(2015).6 0.11% 0.22 [0.05, 0.39]
Finn et al.(2015).7 0.11% 0.35 [0.18, 0.52]
Dubois et al.(2018a).1 0.86% 0.22 [0.16, 0.28]
Dubois et al.(2018a).2 0.86% 0.18 [0.12, 0.24]
Dubois et al.(2018a).3 0.86% 0.26 [0.20, 0.32]
Dubois et al.(2018a).4 0.86% 0.04 [−0.02, 0.10]
Dubois et al.(2018a).5 0.86% 0.03 [−0.03, 0.09]
Dubois et al.(2018a).6 0.86% 0.04 [−0.02, 0.10]
Dubois et al.(2018a).7 0.86% 0.05 [−0.01, 0.11]
Dubois et al.(2018a).8 0.86% 0.07 [0.01, 0.13]
Dubois et al.(2018a).9 0.86% 0.10 [0.04, 0.16]
Dubois et al.(2018a).10 0.86% 0.08 [0.02, 0.14]
Dubois et al.(2018a).11 0.86% 0.03 [−0.03, 0.09]
Dubois et al.(2018a).12 0.86% 0.13 [0.07, 0.19]
Dubois et al.(2018a).13 0.86% 0.11 [0.05, 0.17]
Dubois et al.(2018a).14 0.86% 0.12 [0.06, 0.18]
Dubois et al.(2018a).15 0.86% 0.11 [0.05, 0.17]
Dubois et al.(2018a).16 0.86% 0.11 [0.05, 0.17]
Dubois et al.(2018a).17 0.86% 0.08 [0.02, 0.14]
Dubois et al.(2018a).18 0.86% 0.03 [−0.03, 0.09]
Dubois et al.(2018a).19 0.86% 0.02 [−0.04, 0.08]
Dubois et al.(2018a).20 0.86% 0.08 [0.02, 0.14]
Dubois et al.(2018a).21 0.86% 0.05 [−0.01, 0.11]
Dubois et al.(2018a).22 0.86% 0.17 [0.11, 0.23]
Dubois et al.(2018a).23 0.86% 0.16 [0.10, 0.22]
Dubois et al.(2018a).24 0.86% 0.14 [0.08, 0.20]
Dubois et al.(2018a).25 0.86% 0.13 [0.07, 0.19]
Dubois et al.(2018a).26 0.86% 0.24 [0.18, 0.30]
Dubois et al.(2018a).27 0.86% 0.16 [0.10, 0.22]
Dubois et al.(2018a).28 0.86% 0.18 [0.12, 0.24]
Dubois et al.(2018a).29 0.86% 0.16 [0.10, 0.22]
Dubois et al.(2018a).30 0.86% 0.17 [0.11, 0.23]
Dubois et al.(2018a).31 0.86% 0.17 [0.11, 0.23]
Dubois et al.(2018a).32 0.86% 0.20 [0.14, 0.26]
Dubois et al.(2018a).33 0.86% 0.22 [0.16, 0.28]
Dubois et al.(2018a).34 0.86% 0.16 [0.10, 0.22]
Dubois et al.(2018a).35 0.86% 0.14 [0.08, 0.20]
Dubois et al.(2018a).36 0.86% 0.14 [0.08, 0.20]
Dubois et al.(2018a).37 0.86% 0.10 [0.04, 0.16]
Dubois et al.(2018a).38 0.86% 0.21 [0.15, 0.27]
Dubois et al.(2018a).39 0.86% 0.12 [0.06, 0.18]
Elliott et al.(2019).1 0.29% 0.20 [0.09, 0.31]
Elliott et al.(2019).2 0.29% 0.27 [0.16, 0.38]
Elliott et al.(2019).3 0.57% 0.26 [0.18, 0.34]
Elliott et al.(2019).4 0.57% 0.32 [0.24, 0.40]
Elliott et al.(2019).5 0.29% 0.24 [0.13, 0.35]
Elliott et al.(2019).6 0.29% 0.31 [0.20, 0.42]
Elliott et al.(2019).7 0.29% 0.22 [0.11, 0.33]
Elliott et al.(2019).8 0.29% 0.30 [0.19, 0.41]
Elliott et al.(2019).9 0.57% 0.28 [0.20, 0.36]
Elliott et al.(2019).10 0.57% 0.29 [0.21, 0.37]
Elliott et al.(2019).11 0.29% 0.26 [0.15, 0.37]
Elliott et al.(2019).12 0.29% 0.29 [0.18, 0.40]
Elliott et al.(2019).13 0.29% 0.18 [0.07, 0.29]
Elliott et al.(2019).14 0.29% 0.28 [0.17, 0.39]
Elliott et al.(2019).15 0.57% 0.22 [0.14, 0.30]
Elliott et al.(2019).16 0.57% 0.30 [0.22, 0.38]
Elliott et al.(2019).17 0.29% 0.12 [0.01, 0.23]
Elliott et al.(2019).18 0.29% 0.26 [0.15, 0.37]
Elliott et al.(2019).19 0.29% 0.26 [0.15, 0.37]
Elliott et al.(2019).20 0.29% 0.26 [0.15, 0.37]
Elliott et al.(2019).21 0.29% 0.31 [0.20, 0.42]
Elliott et al.(2019).22 0.29% 0.31 [0.20, 0.42]
Elliott et al.(2019).23 0.29% 0.26 [0.15, 0.37]
Elliott et al.(2019).24 0.29% 0.30 [0.19, 0.41]
Elliott et al.(2019).25 0.29% 0.21 [0.10, 0.32]
Elliott et al.(2019).26 0.29% 0.23 [0.12, 0.34]
Elliott et al.(2019).27 0.29% 0.22 [0.11, 0.33]
Powell et al.(2017) 0.82% 0.08 [0.02, 0.15]
Zhang et al. (2019).1 1.04% 0.08 [0.03, 0.14]
Zhang et al. (2019).2 1.04% 0.24 [0.18, 0.30]
Zhang et al. (2019).3 1.04% 0.24 [0.18, 0.30]
Alnæs et al.(2018) 0.73% 0.39 [0.32, 0.46]
Xiao et al.(2020).1 0.35% 0.32 [0.22, 0.42]
Xiao et al.(2020).2 0.35% 0.32 [0.22, 0.42]
Xiao et al.(2020).3 0.35% 0.32 [0.22, 0.42]
Xiao et al.(2020).4 0.35% 0.32 [0.22, 0.42]
Xiao et al.(2020).5 0.35% 0.33 [0.24, 0.43]
Xiao et al.(2020).6 0.35% 0.33 [0.23, 0.43]
Xiao et al.(2020).7 0.35% 0.35 [0.25, 0.45]
Xiao et al.(2020).8 0.35% 0.34 [0.25, 0.44]
Li et al.(2020).1 0.32% 0.20 [0.10, 0.31]
Li et al.(2020).2 0.32% 0.10 [−0.00, 0.20]
Li et al.(2020).3 0.32% 0.16 [0.06, 0.26]
Li et al.(2020).4 0.32% 0.24 [0.13, 0.34]
Li et al.(2020).5 0.10% 0.38 [0.20, 0.57]
Li et al.(2020).6 0.10% 0.20 [0.02, 0.39]
Li et al.(2020).7 0.08% 0.36 [0.16, 0.57]
Li et al.(2020).8 0.08% 0.22 [0.01, 0.42]
Noble et al.(2017) 0.59% 0.22 [0.14, 0.30]
Jiang et al.(2020c).1 0.45% 0.41 [0.32, 0.50]
Jiang et al.(2020c).2 0.45% 0.38 [0.29, 0.46]
Jiang et al.(2020c).3 0.45% 0.22 [0.14, 0.31]
−0.5
0
0.5
1
Correlation coefficient

Fig. E.11. Forest plot for the correlation coefficient meta-analysis. Imaging modality, either fMRI, sMRI or dMRI, outcome, either G, G_F or G_C, and dataset, either ABIDE, ADHD-200, DMHDS, HCP, NKI, NRI, NRI/KAIST, PNC, SLIM, SXMU, UESTC or the UK Biobank, were included as moderators.

Appendix F. Additional analysis of risk of bias across studies

Funnel plots using all identified results are depicted in Fig. F.12, for the correlation coefficient, and Fig. F.13, for R-squared. Contrast with the ones in Fig. 8.

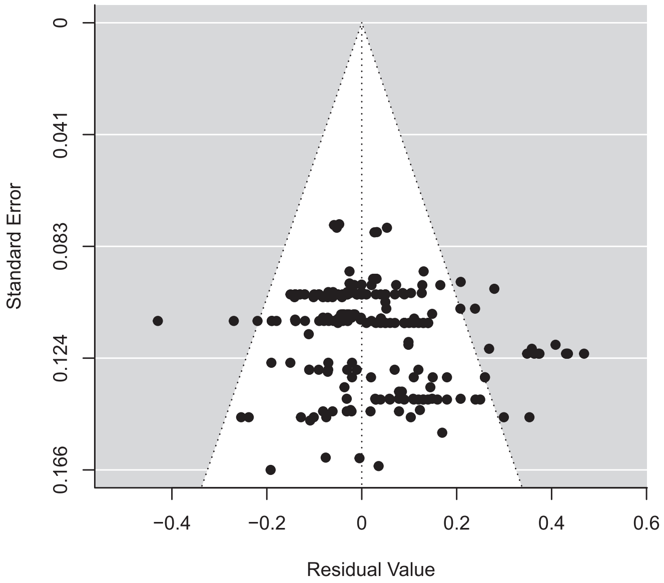


Fig. F.12. Funnel plot for the correlation coefficient meta-analysis. 223 results are depicted.

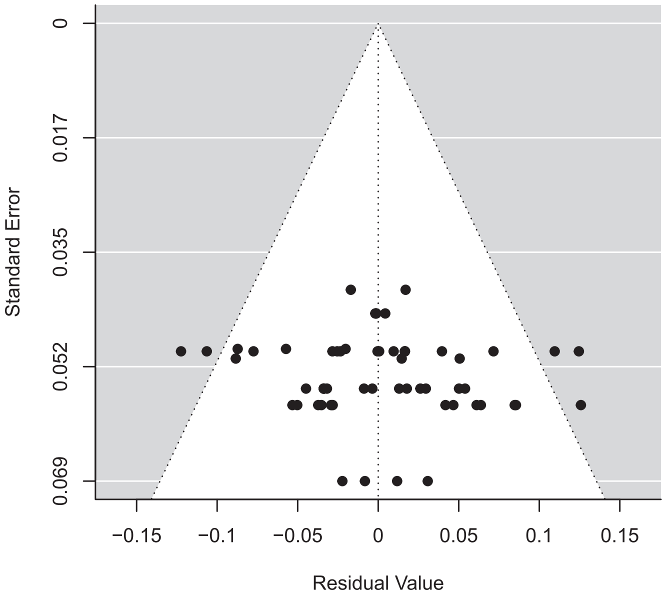


Fig. F.13. Funnel plot for the R-squared meta-analysis. 58 results are depicted.

References

Abrol, A., Zening, F., Salman, M., Silva, R., Yuhui, D., Plis, S., & Calhoun, V. D. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature Communications, 12*(1), 1–17. https://doi.org/10.1038/s41467-020-20655-6. ISSN 20411723.

Abu-Hamour, B., & Al-Hmouz, H. (July 2016). Prevalence and pattern of learning difficulties in primary school students in Jordan. *Australian Journal of Learning Difficulties, 21*(2), 99–113. https://doi.org/10.1080/19404158.2017.1287104

Alexander, D. L. J., Tropsha, A., & Winkler, D. A. (July 2015). Beware of R^2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of Chemical Information and Modeling, 55*(7), 1316–1322. https://doi.org/10.1021/acs.jcim.5b00206. ISSN 1549-9596.

Alnæs, D., Kaufmann, T., Doan, N. T., Córdova-Palomera, A., Wang, Y., Bettella, F., … Westlye, L. T. (2018). Association of heritable cognitive ability and psychopathology with white matter properties in children and adolescents. *JAMA Psychiatry, 75*(3), 287–295. https://doi.org/10.1001/jamapsychiatry.2017.4277. ISSN 2168622X.

Avery, E. W., Yoo, K., Rosenberg, M. D., Greene, A. S., Gao, S., Na, D. L., … Chun, M. M. (2019). Distributed patterns of functional connectivity predict working memory performance in novel healthy and memory-impaired individuals. *Journal of Cognitive Neuroscience, 32*(2), 241–255. https://doi.org/10.1162/jocn_a_01487. ISSN 15308898.

Barbey, A. K. (2018). Network neuroscience theory of human intelligence. *Trends in Cognitive Sciences, 22*(1), 8–20. https://doi.org/10.1016/j.tics.2017.10.001. ISSN 1879307X.

Basten, U., & Fiebach, C. J. (May 2021). Functional brain imaging of intelligence. In *The Cambridge Handbook of Intelligence and Cognitive Neuroscience, number June* (pp. 235–260). Cambridge University Press. https://doi.org/10.1017/9781108635462.016. ISBN 9781108635462. URL https://www.cambridge.org/core/product/identifier/9781108635462{%}23CN-bp-12/type/book{_}part.

Basten, U., Hilger, K., & Fiebach, C. J. (2015). Where smart brains are different: A quantitative meta-analysis of functional and structural brain imaging studies on intelligence. *Intelligence, 51*, 10–27. https://doi.org/10.1016/j.intell.2015.04.009. ISSN 01602896.

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the raven€TMs standard progressive matrices test. *Assessment, 19*(3), 354–369. https://doi.org/10.1177/1073191112446655 (PMID: 22605785).

Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Frontiers in Neuroscience, 11*(OCT), 1–23. https://doi.org/10.3389/fnins.2017.00543. ISSN 1662453X.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Points of significance: Statistics versus machine learning. *Nature Methods, 15*(4), 233–234. https://doi.org/10.1038/nmeth.4642. ISSN 15487105.

Caemmerer, J. M., Keith, T. Z., & Reynolds, M. R. (March 2020). Beyond individual intelligence tests: Application of cattell-horn-Carroll theory. *Intelligence, 79*, Article 101433. https://doi.org/10.1016/j.intell.2020.101433

Cai, B., Zhang, G., Zhang, A., Xiao, L., Hu, W., Stephen, J. M., … Wang, Y. P. (2021). Functional connectome fingerprinting: Identifying individuals and predicting cognitive functions via autoencoder. *Human Brain Mapping, 42*(9), 2691–2705. https://doi.org/10.1002/hbm.25394. ISSN 10970193.

Carroll, J. B. (1997). Psychometrics, intelligence, and public perception. *Intelligence, 24*(1 SPEC.ISS(l)), 25–52. https://doi.org/10.1016/s0160-2896(97)90012-x. ISSN 01602896.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin, 38*(7), 592.

Cattell, R. B. (1971). *Abilites: Their structure, growth and action*. Boston, Massachusetts, United States: Houghton Mifflin. ISBN 0395042755.

Choi, Y. Y., Shamosh, N. A., Cho, S. H., DeYoung, C. G., Lee, M. J., Lee, J.-M., … Lee, K. H. (2008). Multiple bases of human intelligence revealed by cortical thickness and neural activation. *Journal of Neuroscience, 28*(41), 10323–10329. https://doi.org/10.1523/JNEUROSCI.3259-08.2008. ISSN 0270-6474.

Chuderski, A. (2015). The broad factor of working memory is virtually isomorphic to fluid intelligence tested under time pressure. *Personality and Individual Differences, 85*, 98–104. https://doi.org/10.1016/j.paid.2015.04.046. ISSN 01918869.

Chyzhyk, D., Varoquaux, G., Thirion, B., & Milham, M. (2018). Controlling a confound in predictive models with a test set minimizing its effect. In *2018 International Workshop on Pattern Recognition in Neuroimaging* (p. 2018). PRNI. https://doi.org/10.1109/PRNI.2018.8423961.

Cole, M. W., Yarkoni, T., Repovš, G., Anticevic, A., & Braver, T. S. (June 2012). Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *The Journal of Neuroscience, 32*(26), 8988–8999. https://doi.org/10.1523/JNEUROSCI.0536-12.2012. ISSN 0270-6474.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine, 162*(1), 55–63. https://doi.org/10.7326/M14-0697. ISSN 15393704.

Cox, S. R., Ritchie, S. J., Fawns-Ritchie, C., Tucker-Drob, E. M., & Deary, I. J. (2019). Structural brain imaging correlates of general intelligence in UK biobank. *Intelligence, 76*. https://doi.org/10.1016/j.intell.2019.101376

Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage, 178*, 622–637. https://doi.org/10.1016/j.neuroimage.2018.06.001

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., & Varoquaux, G. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage, 192*(March), 115–134. https://doi.org/10.1016/j.neuroimage.2019.02.062. ISSN 10959572.

Dadi, K., Varoquaux, G., Houenou, J., Bzdok, D., Thirion, B., & Engemann, D. (October 2021). Population modeling with machine learning can enhance measures of mental health. *GigaScience, 10*(10), 1–16. https://doi.org/10.1093/gigascience/giab071. ISSN 2047-217X. URL https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giab071/6396189.

Dhamala, E., Jamison, K. W., Jaywant, A., Dennis, S., & Kuceyeski, A. (2021). Distinct functional and structural connections predict crystallised and fluid cognition in healthy adults. *Human Brain Mapping, 42*(10), 3102–3118. https://doi.org/10.1002/hbm.25420. ISSN 10970193.

Dizaji, A. S., Vieira, B. H., Khodaei, M.-R., Ashrafi, M., Parham, E., Hossein-Zadeh, G.-A., … Soltanian-Zadeh, H. (2021). Linking brain biology to intellectual endowment: a review on the associations between human intelligence and neuroimaging data. *Basic and Clinical Neuroscience*. https://doi.org/10.32598/bcn.12.1.574.1

Dombrowski, S. C., Beaujean, A. A., McGill, R. J., Benson, N. F., & Schneider, W. J. (June 2019). Using exploratory bifactor analysis to understand the latent structure of multidimensional psychological measures: An example featuring the WISC-v. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(6), 847–860. https://doi.org/10.1080/10705511.2019.1622421

Dryburgh, E., McKenna, S., & Rekik, I. (2020). Predicting full-scale and verbal intelligence scores from functional Connectomic data in individuals with autism Spectrum disorder. *Brain Imaging and Behavior, 14*(5), 1769–1778. https://doi.org/10.1007/s11682-019-00111-w. ISSN 19317565. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065290036{&}doi=10.1007{%}2Fs11682-019-00111-w{&}partnerID=40{&}md5=ad34c2439a257ce9aac4bf652a905164.

Dubois, J., Galdi, P., Han, Y., Paul, L. K., & Adolphs, R. (2018). Resting-state functional brain connectivity best predicts the personality dimension of openness to experience. *Personality Neuroscience*. https://doi.org/10.1017/pen.2018.8. ISSN 2513-9886.

Dubois, J., Galdi, P., Paul, L. K., & Adolphs, R. (September 2018). A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 373*(1756), 20170284. https://doi.org/10.1098/rstb.2017.0284. ISSN 0962-8436.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences, 14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004. ISSN 13646613.

Elliott, M. L., Knodt, A. R., Cooke, M., Kim, M. J., Melzer, T. R., Keenan, R., … Hariri, A. R. (2019). General functional connectivity: Shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. *NeuroImage, 189*(January), 516–532. https://doi.org/10.1016/j.neuroimage.2019.01.068. ISSN 10959572.

Euler, M. J., & McKinney, T. L. (2021). Evaluating the weight of the evidence: Cognitive neuroscience theories of intelligence. In A. K. Barbey, S. Karama, & R. J. Haier (Eds.), *The Cambridge handbook of intelligence and cognitive neuroscience* (pp. 85–101). Cambridge University Press. https://doi.org/10.1017/9781108635462.008. ISBN 9781108635462.

Fan, L., Jianpo, S., Qin, J., Hu, D., & Shen, H. (2020). A deep network model on dynamic functional connectivity with applications to gender classification and intelligence prediction. *Frontiers in Neuroscience, 14*(August), 1–14. https://doi.org/10.3389/fnins.2020.00881. ISSN 1662453X.

Fawns-Ritchie, C., & Deary, I. J. (2020). Reliability and validity of the UK biobank cognitive tests. *PLoS One, 15*(4), 1–24. https://doi.org/10.1371/journal.pone.0231627

Feilong, M., Guntupalli, J. S., & Haxby, J. V. (2021). The neural basis of intelligence in fine-grained cortical topographies. *eLife, 10*, 1–33. https://doi.org/10.7554/eLife.64058. ISSN 2050084X.

Ferguson, M. A., Anderson, J. S., & Spreng, R. N. (2017). Fluid and flexible minds: Intelligence reflects synchrony in the brain's intrinsic network architecture. *Network Neuroscience, 1*(2), 192–207. https://doi.org/10.1162/netn_a_00010. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042367354{&}doi=10.1162{%}2Fnetn{_}a{_}00010{&}partnerID=40{&}md5=e92d8ed9b86237d15c4c4a69cfa4f762.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., … R Todd Constable. (November 2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience, 18*(11), 1664–1671. https://doi.org/10.1038/nn.4135. ISSN 1546-1726.

Frith, E., Elbich, D. B., Christensen, A. P., Rosenberg, M. D., Chen, Q., Kane, M. J., … Beaty, R. E. (2021). Intelligence and creativity share a common cognitive and neural basis. *Journal of Experimental Psychology: General, 150*(4), 609–632. https://doi.org/10.1037/xge0000958. ISSN 00963445.

Gao, S., Greene, A. S., Constable, R. T., & Scheinost, D. (2019). Combining multiple connectomes improves predictive modeling of phenotypic measures. *NeuroImage, 201*(May), 116038. https://doi.org/10.1016/j.neuroimage.2019.116038. ISSN 10959572.

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z. ISSN 25225839.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence, 52*, 71–79. https://doi.org/10.1016/j.intell.2015.07.006. ISSN 01602896.

Gignac, G. E., & Bates, T. C. (2017). Brain volume and intelligence: The moderating role of intelligence measurement quality. *Intelligence, 64*(May), 18–29. https://doi.org/10.1016/j.intell.2017.06.004. ISSN 01602896.

Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L. R., Auerbach, E. J., Behrens, T. E. J., … Van Essen, D. C. (September 2016). The human connectome project's neuroimaging approach. *Nature Neuroscience, 19*(9), 1175–1187. https://doi.org/10.1038/nn.4361. URL. ISSN 1097-6256 http://www.nature.com/articles/nn.4361.

Graham, S., Jiang, J., Manning, V., Nejad, A. B., Zhisheng, K., Salleh, S. R., … McKenna, P. J. (2010). IQ-related fMRI differences during cognitive set shifting. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhp130. ISSN 10473211.

Grazioplene, R. G., Ryman, S. G., Gray, J. R., Rustichini, A., Jung, R. E., & Deyoung, C. G (2015). Subcortical intelligence: Caudate volume predicts IQ in healthy adults. *Human Brain Mapping, 36*(4), 1407–1416. https://doi.org/10.1002/hbm.22710. ISSN 10970193.

Greene, A. S., Gao, S., Scheinost, D., & Constable, R. T. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nature Communications, 9*(1). https://doi.org/10.1038/s41467-018-04920-3. ISSN 20411723.

Gur, R. C., Butler, E. R., Moore, T. M., Rosen, A. F. G., Ruparel, K., Satterthwaite, T. D., … Gur, R. E. (February 2021). Structural and functional brain parameters related to cognitive performance across development: replication and extension of the parieto-frontal integration theory in a single sample. *Cerebral cortex (New York, N.Y.: 1991), 31*(3), 1444–1463. https://doi.org/10.1093/cercor/bhaa282. ISSN 1460-2199.

Haier, R. J., Jung, R. E., Yeo, R. A., Head, K., & Alkire, M. T. (2005). The neuroanatomy of general intelligence: Sex matters. *NeuroImage, 25*(1), 320–327. https://doi.org/10.1016/j.neuroimage.2004.11.019. ISSN 10538119.

Hakim, N., Awh, E., Vogel, E. K., & Rosenberg, M. D. (2021). Inter-electrode correlations measured with EEG predict individual differences in cognitive ability. *Current Biology, 31*(22), 4998–5008. ISSN 18790445. e6 https://doi.org/10.1016/j.cub.2021.09.036.

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., … Yeo, B. T. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage, 206*(July 2019), 116276. https://doi.org/10.1016/j.neuroimage.2019.116276. ISSN 10959572.

Heus, P., Damen, J. A. A. G., Pajouheshnia, R., Scholten, R. J. P. M., Reitsma, J. B., Collins, G. S., … Hooft, L. (2019). Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open, 9*(4), e025611. https://doi.org/10.1136/bmjopen-2018-025611. ISSN 2044-6055.

Hilger, K., Winter, N. R., Leenings, R., Sassenhagen, J., Hahn, T., Basten, U., & Fiebach, C. J. (2020). Predicting intelligence from brain gray matter volume. *Brain Structure and Function, 225*(7), 2111–2129. https://doi.org/10.1007/s00429-020-02113-7. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088240962{&}doi=10.1007{%}2Fs00429-020-02113-7{&}partnerID=40{&}md5=7c3e11225a12af8c9fff0d0fe24d6196.

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience and Biobehavioral Reviews, 119*(September), 456–467. https://doi.org/10.1016/j.neubiorev.2020.09.036. ISSN 1873-7528.

Hurks, P. P. M., & Bakker, H. (April 2016). Assessing intelligence in children and youth living in the Netherlands. *International Journal of School and Educational Psychology, 4*(4), 266–275. https://doi.org/10.1080/21683603.2016.1166754

James, L., Jacobs, K. E., & Roodenburg, J. (June 2015). Adoption of the cattell–horn–Carroll model of cognitive abilities by australian psychologists. *Australian Psychologist, 50*(3), 194–202. https://doi.org/10.1111/ap.12110

Jiang, R., Calhoun, V. D., Cui, Y., Qi, S., Zhuo, C., Li, J., Jung, R., Yang, J., Du, Y., Jiang, T., Jiang, T., & Sui, J. (2020). Multimodal data revealed different neurobiological correlates of intelligence between males and females. *Brain Imaging and Behavior, 14*(5), 1979–1993. https://doi.org/10.1007/s11682-019-00146-z

Jiang, R., Calhoun, V. D., Fan, L., Zuo, N., Jung, R., Qi, S., Lin, D., Li, J., Zhuo, C., Song, M., Jiang, T., & Sui, J. (2020). Gender differences in connectome-based predictions of individualized intelligence quotient and sub-domain scores. *Cerebral Cortex, 30*(3), 888–900. https://doi.org/10.1093/cercor/bhz134

Jiang, R., Zuo, N., Ford, J. M. J. M., Qi, S., Zhi, D., Zhuo, C., … Sui, J. (2020). Task-induced brain connectivity promotes the detection of individual differences in brain-behavior relationships. *NeuroImage, 207*(November 2019), 116370. https://doi.org/10.1016/j.neuroimage.2019.116370. ISSN 10959572.

Jung, R. E., & Haier, R. J. (2007). The Parieto-frontal integration theory (P-FIT) of intelligence: Converging neuroimaging evidence. *Behavioral and Brain Sciences, 30*(2), 135–154. https://doi.org/10.1017/S0140525X07001185. ISSN 0140-525X.

Kashyap, R., Kong, R., Bhattacharjee, S., Zhou, J., Li, J., Zhou, J., & Yeo, T. (2019). Individual-specific fMRI-subspaces improve functional connectivity prediction of behavior. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2019.01.069. ISSN 10538119.

Kent, P. (May 2017). Fluid intelligence: A brief history. *Applied Neuropsychology: Child, 6*(3), 193–203. https://doi.org/10.1080/21622965.2017.1317480

Kong, R., Li, J., Orban, C., Sabuncu, M. R., Liu, H., Schaefer, A., … Yeo, B. T. T. (2019). Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cerebral Cortex, 29*(6), 2533–2551. https://doi.org/10.1093/cercor/bhy123. ISSN 14602199.

Lecerf, T., Reverte, I., Coleaux, L., Favez, N., & Rossier, J. (March 2010). Indice d'aptitude général pour le WISC-IV: Normes francophones. *Pratiques Psychologiques, 16*(1), 109–121. https://doi.org/10.1016/j.prps.2009.04.001

Li, C., Yang, G., Li, M., & Li, B. (January 2018). Fluid intelligence relates to the resting state amplitude of low-frequency fluctuation and functional connectivity. *NeuroReport, 29*(1), 8–12. https://doi.org/10.1097/WNR.0000000000000917. ISSN 0959-4965.

Li, J., Biswal, B. B., Meng, Y., Yang, S., Duan, X., Cui, Q., … Liao, W. (2020). A neuromarker of individual general fluid intelligence from the white-matter functional connectome. *Translational Psychiatry, 10*(1), 147. https://doi.org/10.1038/s41398-020-0829-3. ISSN 2158–3188.

Li, J., Kong, R., Liégeois, R., Orban, C., Tan, Y., Sun, N., … Yeo, B. T. T. (2019). Global signal regression strengthens association between resting-state functional connectivity and behavior. *NeuroImage, 196*(April), 126–141. https://doi.org/10.1016/j.neuroimage.2019.04.016. ISSN 10959572.

Lin, Y.-C., Baete, S. H., Wang, X., & Boada, F. E. (2020). Mapping brain behavior networks using functional and structural connectome fingerprinting in the HCP dataset. *Brain and Behavior: A Cognitive Neuroscience Perspective, 10*(6). https://doi.org/10.1002/brb3.1647. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084231068{&}doi=10.1002{%}2Fbrb3.1647{&}partnerID=40{&}md5=b0d1f42dce7001774afe436caed798d9.

Lohman, D. F., & Lakin, J. M. (2012). Intelligence and reasoning. In *The Cambridge Handbook of Intelligence, Chapter 21* (pp. 419–441). https://doi.org/10.1017/cbo9780511977244.022. ISBN 9780511977244.

Luders, E., Narr, K. L., Bilder, R. M., Thompson, P. M., Szeszko, P. R., Hamilton, L., & Toga, A. W. (2007). Positive correlations between corpus callosum thickness and intelligence. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2007.06.028. ISSN 10538119.

Luders, E., Narr, K. L., Thompson, P. M., & Toga, A. W. (2009). Neuroanatomical correlates of intelligence. *Intelligence, 37*(2), 156–163. https://doi.org/10.1016/j.intell.2008.07.002. ISSN 01602896.

Marc-Andre Schulz, B. T., Yeo, T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., … Bzdok, D. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, (1), 11. https://doi.org/10.1038/s41467-020-18037-z. ISSN 20411723.

McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence, 33*(4), 337–346. https://doi.org/10.1016/j.intell.2004.11.005. ISSN 01602896.

McGrew, K. S. (January 2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Mihalik, A., Brudfors, M., Robu, M., Ferreira, F. S., Lin, H., Rau, A., … Oxtoby, N. P. (2019). ABCD neurocognitive prediction challenge 2019: Predicting individual fluid intelligence scores from structural MRI using probabilistic segmentation and Kernel Ridge regression. In *, 11791 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 133–142). https://doi.org/10.1007/978-3-030-31901-4_16

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLoS Medicine, 6*(7), 1–6. https://doi.org/10.1371/journal.pmed.1000097

Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., … Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine, 162*(1), W1–W73. https://doi.org/10.7326/M14-0698. ISSN 15393704.

Moons, K. G. M., De Groot, J. A. H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., … Collins, G. S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. *PLoS Medicine, 11*(10). https://doi.org/10.1371/journal.pmed.1001744. ISSN 15491676.

Noble, S., Spann, M. N., Tokoglu, F., & Shen, X. (2017). R. Todd Constable, and Dustin Scheinost. Influences on the test-retest reliability of functional connectivity MRI and its relationship with behavioral utility. *Cerebral Cortex, 27*(11), 5415–5429. https://doi.org/10.1093/cercor/bhx230. ISSN 14602199.

Pamplona, G. S. P., Santos Neto, G. S., Rosset, S. R. E., Rogers, B. P., & Salmon, C. E. G. (2015). Analyzing the association between functional connectivity of the brain and intellectual performance. *Frontiers in Human Neuroscience, 9*(February), 61. https://doi.org/10.3389/fnhum.2015.00061. ISSN 1662-5161.

Park, B. Y., Hong, J., Lee, S. H., & Park, H. (2016). Functional connectivity of child and adolescent attention deficit hyperactivity disorder patients: Correlation with IQ. *Frontiers in Human Neuroscience, 10*(NOV2016), 1–9. https://doi.org/10.3389/fnhum.2016.00565. ISSN 16625161.

Pervaiz, U., Vidaurre, D., Woolrich, M. W., & Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage, 211*(January), 116604. https://doi.org/10.1016/j.neuroimage.2020.116604. ISSN 1095-9572.

Pietschnig, J., Penke, L., Wicherts, J. M., Zeiler, M., & Voracek, M. (October 2015). Meta-analysis of associations between human brain volume and intelligence differences: How strong are they and what do they mean? *Neuroscience and Biobehavioral Reviews, 57*, 411–432. https://doi.org/10.1016/j.neubiorev.2015.09.017. ISSN 01497634.

Poldrack, R. A., Huckins, G., & Varoquaux, G. (May 2020). Establishment of best practices for evidence for prediction. *JAMA Psychiatry, 77*(5), 534. https://doi.org/10.1001/jamapsychiatry.2019.3671. ISSN 2168-622X. URL.

Powell, M. A., Garcia, J. O., Yeh, F.-C., Vettel, J. M., & Verstynen, T. (March 2017). Local connectome phenotypes predict social, health, and cognitive factors. *Network Neuroscience, 2*(1), 86–105. https://doi.org/10.1162/NETN{_}a{_}00031. ISSN 2472-1751 https://www.mitpressjournals.org/doi/abs/10.1162/NETN{_}a{_}00031.

Rao, A., Monteiro, J. M., Mourao-Miranda, J., & Alzheimer's Disease Initiative. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage, 150*(January), 23–49. https://doi.org/10.1016/j.neuroimage.2017.01.066. ISSN 1095-9572.

Ritchie, S. J., Booth, T., Valdés Hernández, M. D. C., Corley, J., Maniega, S. M., Gow, A. J., … Deary, I. J. (2015). Beyond a bigger brain: Multivariable structural brain imaging and intelligence. *Intelligence, 51*, 47–56. https://doi.org/10.1016/j.intell.2015.05.001. ISSN 0160-2896.

Schelini, P. W. (December 2006). Teoria das inteligências fluida e cristalizada: início e evolução. *Estudos de Psicologia (Natal), 11*(3), 323–332. https://doi.org/10.1590/s1413-294x2006000300010

Schulz, M.-A., Bzdok, D., Haufe, S., Haynes, J.-D., & Ritter, K. (2022). Performance reserves in brain-imaging-based phenotype prediction. *bioRxiv (preprint)*, 1–27. https://doi.org/10.1101/2022.02.23.481601

Sen, B., & Parhi, K. K. (2021). Predicting biological gender and intelligence from fMRI via dynamic functional connectivity. *IEEE Transactions on Biomedical Engineering, 68*(3), 815–825. https://doi.org/10.1109/TBME.2020.3011363. ISSN 15582531.

Shen, X., Finn, E. S., Scheinost, D., Rosenberg, M. D., Chun, M. M., Papademetris, X., & Constable, R. T. (February 2017). Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nature Protocols, 12*(3), 506–518. https://doi.org/10.1038/nprot.2016.178. ISSN 1754-2189.

Siegel, J. S., Mitra, A., Laumann, T. O., Seitzman, B. A., Raichle, M., Corbetta, M., & Snyder, A. Z. (2017). Data quality influences observed links between functional connectivity and behavior. *Cerebral Cortex, 27*(9), 4492–4502. https://doi.org/10.1093/cercor/bhw253. ISSN 14602199.

Song, M., Zhou, Y., Li, J., Liu, Y., Tian, L., Yu, C., & Jiang, T. (2008). Brain spontaneous functional connectivity and intelligence. *NeuroImage, 41*(3), 1168–1176. https://doi.org/10.1016/j.neuroimage.2008.02.036. ISSN 10538119.

Spearman, C. (April 1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology, 15*(2), 201. https://doi.org/10.2307/1412107. ISSN 00029556.

Spearman, C. (1927). *The Abilities of Man; Their Nature and Measurement*. https://doi.org/10.2307/2015168

Sripada, C., Angstadt, M., Rutherford, S., Taxali, A., & Shedden, K. (2020). Toward a treadmill test for cognition: Improved prediction of general cognitive ability from the task activated brain. *Human Brain Mapping, 41*(12), 3186–3197. https://doi.org/10.1002/hbm.25007. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085105364{&}doi=10.1002{%}2Fhbm.25007{&}partnerID=40{&}md5=b5db0c7b21d896d82d47d30025c1f2ed.

Stankov, L. (October 2017). Overemphasized "g". *Journal of Intelligence, 5*(4), 33. https://doi.org/10.3390/jintelligence5040033

Sui, J., Jiang, R., Bustillo, J., & Calhoun, V. D. (2020). Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: Methods and promises. *Biological Psychiatry*, (3), 1–11. https://doi.org/10.1016/j.biopsych.2020.02.016. ISSN 18732402.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, Illinois, United States: University of Chicago Press.

Urbina, S. (2011). *Tests of intelligence* (pp. 20–38). New York, NY: Cambridge University Press.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn human connectome project: An overview. *NeuroImage, 80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041. ISSN 10538119.

Varoquaux, G. (October 2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage, 180*(April), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061. ISSN 10538119.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Vieira, B. H., Dubois, J., Calhoun, V. D., & Salmon, C. E. G. (2021). A deep learning based approach identifies regions more relevant than resting-state networks to the prediction of general intelligence from resting-state fMRI. *Human Brain Mapping*, (August), 1–15. https://doi.org/10.1002/hbm.25656

Vieira, B. H., Fachinello, K., Silva, A. K., Foss, M. P., & Salmon, C. E. G. (2021). *On individualized prediction of intellectual ability from brain imaging: A systematic review*. https://doi.org/10.17605/OSF.IO/QVP9K

Wang, L., Wee, C.-Y., Suk, H.-I., Tang, X., & Shen, D. (March 2015). MRI-based Intelligence Quotient (IQ) estimation with sparse learning. *PLoS One, 10*(3), e0117295. https://doi.org/10.1371/journal.pone.0117295. ISSN 1932-6203.

Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., Rudd, A. G., … Bray, B. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One, 15*(6), 1–16. https://doi.org/10.1371/journal.pone.0234722. ISSN 19326203.

Wasserman, J. D. (June 2019). Deconstructing CHC. *Applied Measurement in Education, 32*(3), 249–268. https://doi.org/10.1080/08957347.2019.1619563

Wechsler, S. M., & de Cassia Nakano, T. (June 2016). Cognitive assessment of brazilian children and youth: Past and present perspectives and challenges. *International Journal of School and Educational Psychology, 4*(4), 215–224. https://doi.org/10.1080/21683603.2016.1163654

Wei, L., Jing, B., & Li, H. (2020). Bootstrapping promotes the RSFC-behavior associations: An application of individual cognitive traits prediction. *Human Brain Mapping, 41*(9), 2302–2316. https://doi.org/10.1002/hbm.24947

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized versions of the raven progressive matrices test. *Computers in Human Behavior, 22*(5), 791–800. https://doi.org/10.1016/j.chb.2004.03.005. ISSN 0747-5632.

Woodcock, R. J., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III examiner's manual. Riverside, Itasca, Illinois, United States*.

Wu, D., Li, X., & Jiang, T. (2020). Reconstruction of behavior-relevant individual brain activity: An individualized fMRI study. *Science China. Life Sciences, 63*(3), 410–418. https://doi.org/10.1007/s11427-019-9556-4. ISSN 18691889.

Xiao, L., Stephen, J. M., Wilson, T. W., Calhoun, V. D., & Wang, Y. P. (2019). Alternating diffusion map based fusion of multimodal brain connectivity networks for iq prediction. *IEEE Transactions on Biomedical Engineering, 66*(8), 2140–2151. https://doi.org/10.1109/TBME.2018.2884129. ISSN 15582531.

Xiao, L., Stephen, J. M., Wilson, T. W., Calhoun, V. D., & Wang, Y. P. (2020). A manifold regularized multi-task learning model for IQ prediction from two fMRI paradigms. *IEEE Transactions on Biomedical Engineering, 67*(3), 796–806. https://doi.org/10.1109/TBME.2019.2921207

Yang, J. J., Yoon, U., Yun, H. J., Im, K., Choi, Y. Y., Lee, K. H., … Lee, J. M. (2013). Prediction for human intelligence using morphometric characteristics of cortical surface: Partial least square analysis. *Neuroscience, 246*, 351–361. https://doi.org/10.1016/j.neuroscience.2013.04.051. ISSN 03064522.

Yang, S., Zhao, Z., Cui, H., Zhang, T., Zhao, L., He, Z., … Jiang, X. (2019). Temporal variability of cortical gyral-sulcal resting state functional activity correlates with fluid intelligence. *Frontiers in Neural Circuits, 13*. https://doi.org/10.3389/fncir.2019.00036. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066983254{&}doi=10.3389{%}2Ffncir.2019.00036{&}partnerID=40{&}md5=3b4b269f2b3517d671d86b7b2149bc66.

Yoo, K., Rosenberg, M. D., Noble, S., & Scheinost, D. (2019). R. Todd Constable, and Marvin M. Chun. Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors. *NeuroImage, 197* (November 2018), 212–223. https://doi.org/10.1016/j.neuroimage.2019.04.060. ISSN 10959572.

Zamanipoor Najafabadi, A. H., Ramspek, C. L., Dekker, F. W., Heus, P., Hooft, L., Moons, K. G. M., … Diepen, M. V. (2020). TRIPOD statement: A preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open, 10*(9), 1–10. https://doi.org/10.1136/bmjopen-2020-041537. ISSN 20446055.

Zhang, Z., Allen, G. I., Zhu, H., & Dunson, D. (2019). Tensor network factorizations: Relationships between brain structural connectomes and traits. *NeuroImage, 197* (January), 330–343. https://doi.org/10.1016/j.neuroimage.2019.04.027. ISSN 10959572.