Biological Psychiatry: CNNI

Archival Report

Classification Accuracy of Neuroimaging Biomarkers in Attention-Deficit/Hyperactivity Disorder: Effects of Sample Size and Circular Analysis

Alfredo A. Pulini, Wesley T. Kerr, Sandra K. Loo, and Agatha Lenartowicz

ABSTRACT

BACKGROUND: Motivated by an inconsistency between reports of high diagnosis-classification accuracies and known heterogeneity in attention-deficit/hyperactivity disorder (ADHD), this study assessed classification accuracy in studies of ADHD as a function of methodological factors that can bias results. We hypothesized that high classification results in ADHD diagnosis are inflated by methodological factors.

METHODS: We reviewed 69 studies (of 95 studies identified) that used neuroimaging features to predict ADHD diagnosis. Based on reported methods, we assessed the prevalence of circular analysis, which inflates classification accuracy, and evaluated the relationship between sample size and accuracy to test if small-sample models tend to report higher classification accuracy, also an indicator of bias.

RESULTS: Circular analysis was detected in 15.9% of ADHD classification studies, lack of independent test set was noted in 13%, and insufficient methodological detail to establish its presence was noted in another 11.6%. Accuracy of classification ranged from 60% to 80% in the 59.4% of reviewed studies that met criteria for independence of feature selection, model construction, and test datasets. Moreover, there was a negative relationship between accuracy and sample size, implying additional bias contributing to reported accuracies at lower sample sizes.

CONCLUSIONS: High classification accuracies in neuroimaging studies of ADHD appear to be inflated by circular analysis and small sample size. Accuracies on independent datasets were consistent with known heterogeneity of the disorder. Steps to resolve these issues, and a shift toward accounting for sample heterogeneity and prediction of future outcomes, will be crucial in future classification studies in ADHD.

Keywords: ADHD, Bias, Circular analysis, Classification, Machine learning, Sample size

https://doi.org/10.1016/j.bpsc.2018.06.003

A significant challenge in the assessment and treatment of neuropsychiatric disorders is that diagnosis is typically based on subjective behavioral criteria, a process that is time-consuming and requires considerable expertise and training. The need for objective diagnostic indicators has fueled efforts to define neuropsychiatric biomarkers, particularly based on structural and functional features of the brain, and with increasing deployment of machine learning methods. Results of these efforts have been variable; recent reviews indicate that classification accuracy is distributed broadly between chance and near 100% (1–3). Such variability can lead to puzzling outcomes, as is evident in the case of attention-deficit/hyperactivity disorder (ADHD). On the one hand, reports of accuracies in excess of 90% (4–17) have culminated in the electroencephalography-based theta/beta ratio (TBR) metric (18) gaining Food and Drug Administration support as an adjunct to clinical assessment of ADHD (19,20). On the other hand, the variability echoes increasing awareness of heterogeneity in ADHD in symptom presentation (21), neurocognitive impairment, (22,23) persistence (24–26), treatment response (27,28), and putative mechanistic pathways (29–31), and supports the existence of independent subgroups within ADHD (32–37). The incompatibility between such heterogeneity and a diagnostic tool validated by existing ADHD diagnosis has contributed to discussion over the utility of neuroimaging in diagnosis of ADHD (38–40). It also raises a conceptual question: if current diagnosis of ADHD is too clinically variable for classification, how are high classification accuracies achieved? The answer to this question is important if it lies in methodological limitations, which may continue to be a concern in future studies. Thus, we examine this question using ADHD as an exemplar given the large existing literature base on neuroimaging classifiers of diagnosis.

Potential pitfalls of applying classification approaches to neuropsychiatric data have been discussed extensively (1,3,41,42). Two that are particularly relevant include circular analysis and sample size. First, to evaluate its role in clinical medicine, a machine learning classifier needs to have good

 ISSN: 2451-9022

generalizability: defined by good performance on patients not included in the study (i.e., new patients). In the experimental setting, this is assessed by cross-validation, whereby a subset of a dataset is not included in construction of the classification model ("training") and subsequently used to assess the performance of the model ("testing"). However, the testing accuracy can be inflated owing to a common error of including all data when selecting features to be used for classification (i.e., before training). For instance, a *t* test may be performed on all subjects' data, before cross-validation, to identify brain regions that are the most discriminative of two groups. This step is typically performed to reduce the number of features (e.g., brain regions) that are included in the model. However, including all subjects' data in feature selection (rather than performing this step on the training subset only) creates circularity, or "peeking," in the training model that can inflate reported test accuracy (43). Simulations suggest that accuracy inflation can reach 40% depending on model parameters (3,44,45) (also see the Supplement for simulation results). In 2008, a reported 42% of high-impact journal functional magnetic resonance imaging (MRI) studies were subject to circular analysis, with another 14% lacking methodological detail to reach judgment (43,46), suggesting that such practice is not uncommon. A second concern is small sample size, as it can drastically increase both accuracy and variability of cross-validation accuracy (41,42,47). Simulations show that accuracy estimates, in models designed for neuropsychiatric diagnostics, can become unstable when total sample size is less than 100 to 150 (41,47–50), and the problem is most severe when combined with circular analysis (45).

The objective of this study was to review neuroimaging-based studies on ADHD classification to assess the contribution of circular analysis and sample size to classification accuracy, thereby testing for accuracy-inflating effects of these two factors and whether these effects have changed over time. The results reveal a more accurate portrayal of classification accuracies in ADHD, revealing methodological weaknesses that should be addressed in future studies, and that generalize to studies of any neuropsychiatric disorder.

METHODS AND MATERIALS

We performed a literature search using multiple databases (PubMed, Web of Science) and search engines (Google Scholar), with key words including *ADHD*, *ADD*, *classification*, *machine learning*, *classifier*, *prediction*, and *accuracy*, retaining publications that explicitly described a classification framework to distinguish between ADHD and comparison groups (N = 95 studies) based on neuroimaging features. Studies were excluded if 1) no control group was examined (ADHD only or ADHD vs. other disorder groups) (n = 5); 2) sample size per class or age group was not specified (n = 5); 3) total sample was <6, limiting within-group variance (n = 2); 4) accuracy was shown graphically but not reported in the text (n = 3); 5) the model did not use neuroimaging features (n = 9); and 6) classification was not performed based on original ADHD diagnostic labels (n = 1). One study was excluded due to a retraction. This exclusion protocol yielded a final total of 69 studies (Table 1; see Supplement for list of excluded studies).

Study Characteristics

For each study we identified sample size, population (adult, pediatric), feature type, and classifier model. We used a cutoff of 18 years of age for classifying studies as adult versus child populations. For simplicity, studies with participants up to and including 18 years of age were labeled as child studies and studies with participants over and including 18 years of age were labeled as adult studies. An exception was the 2017 study of Duffy *et al.* (51), who used a range of 2 to 22 years of age, which was labeled as "children" in Table 1 for simplicity. If studies performed separate analyses for adults and for children, we report the study twice, treating each group as a separate population.

Frequency of Circular Analysis

To assess the frequency of circular analysis, we evaluated the methods section of each study. We identified procedures for feature selection and those for classification, with the goal being to identify if the same dataset was used for feature selection and in the testing of the classification model. If this was unambiguously the case, the study was labeled as non-independent (see Table 1) with respect to model testing. In many instances, there was ambiguity regarding nonindependence given the methods description and/or presented workflow. Such studies were labeled as unknown, with respect to nonindependence. For all such studies, we contacted the primary author to seek additional details to reduce the size of the unknown category. Some studies presented rationale for including all subjects' data in model training because the algorithm of feature selection analysis was independent from the analysis of the classifier, and thus should not affect classifier performance (14,51). However, as true independence in such cases can be difficult to ascertain (43), we included such studies in the nonindependent category. Therefore, we adopted a rather strict criterion of requiring a completely different set of subjects to be used for feature selection versus testing, to label a study as free of circular analysis. This definition subsumes cases where features were defined based on prior knowledge (i.e., prior studies de facto use independent data to define the features). It also implies that for studies that use an iterative cross-validation scheme, feature selection must be either based on prior knowledge or performed within the training set of every iteration for the classifier to be guaranteed free of circular analysis. Finally, we also identified studies in which no test set was defined (all data were used in feature selection and model construction) and thus no cross-validation was performed. Such studies may suggest potentially useful features but have no test of model generalizability. At the other extreme, we also identify studies that identified an additional completely independent testing dataset (which we refer to as the validation set to distinguish it from the test set), not involved in feature selection, which provides an additional objective, external validation of model generalizability.

Sample Size and Accuracy

For each study, we obtained the total sample size and classifier specificity, sensitivity, and accuracy. Where multiple models were examined, we took the best-performing model. Where accuracy was unreported, we calculated accuracy from

Table 1. Neuroimaging Classification Studies of ADHD

Study	Diagnosis, *n* (Age Group)	*N*	Independent Set	Features	Classifier	$Performance_{Test\ Set}$	CV	Notes
EEG								
Duffy *et al.*, 2017 (51)	347 ADHD, 619 TD (CH)	966	No	CA	DFA	$ACCest_{CV}$: 88%[a] SEN: 86.8%, SPE: 88.5%	Twofold RS-CV (10 iterations)	NI
Chabot *et al.*, 1996 (74)	407 ADD/ADHD, 242 SDLD, 310 TD (CH)	959	No	PSF, CA, MAL	DFA	$ACCest_{CV}$: 75% Three-class correctly classified percentages: 66% TD, 81% ADHD, 61% SDLD	HS	UN
Chabot and Serfontein, 1996 (12)	407 ADHD, 310 TD (CH)	717	No	PSF	DFA	$ACCest_{CV}$: 91% SEN: 94%, SPE: 88%	HS	UN
Helgadottir *et al.*, 2015 (75)	310 ADHD, 351 TD (CH)	661	36 ADHD, 36 TD Total: 72	PSF, CA	SVM	ACC_{IS}: 76%	IS	
Smith *et al.*, 2003 (76)	50 ADHD-C, 50 ADHD-I, 50 TD (CH)	150	No	LM, AM (ERP)	DFA	ACC_{CV}: 73% (8-12 years) ACC_{CV}: 59% (13-18 years)	LOO CV	
Mueller *et al.*, 2011 (4)	75 ADHD, 75 TD (AD)	150	17 ADHD	LM, AM (ERP)	SVM	ACC_{IS}: 94%	IS	Only ADHD in IS
Mueller *et al.*, 2010 (5)	74 ADHD, 74 TD (AD)	148	No	LM (ERP)	RBF-SVM	ACC_{CV}: 92% SEN: 90%, SPE: 94%	10-fold CV	UN
Kemner *et al.*, 1999 (77)	43 ADHD, 43 TD, 50 AUT, 30 DYS, 16 MCDD (CH)	182	No	AM (ERP)	DFA	ACC_{CV}: 46% (5-classes)	HS	
Biederman *et al.*, 2017 (78)	34 ADHD, 29 TD (AD)	63	No	BN/PSF (ERP)	Linear SVM	AUC_{CV}: 92% SEN: 86%, SPE: 95%	10-fold CV (10 iterations)	UN
Ghassemi *et al.*, 2012 (8)	10 ADHD, 40 TD (AD)	50	No	EN (WAV), LE	K-NNC	ACC_{CV}: 96 %	LOO CV	
Allahverdi *et al.*, 2011 (79)	29 ADHD, 20 TD (CH)	49	No	LE, FD	MLP-NN	ACC_{CV}: All electrodes 69%, central 62%, parietal 61%, occipital 56%, frontal 86%	80/20 RS	
Robaey *et al.*, 1992 (80)	12 ADHD, 12 TD (CH)	24	No	AM, LM (ERP)	DFA	ACC: 79%	No CV	NI
Tenev *et al.*, 2014 (81)	67 ADHD, 50 TD (AD)	117	No	PSF	SVMs	ACC_{CV}: 82%	10-fold CV	UN
Lenartowicz *et al.*, 2014 (82)	52 ADHD, 47 TD (CH)	99	No	PSF, P2	LR	ACC_{CV}: 70%	Sixfold CV	NI
Mohammadi *et al.*, 2016 (9)	30 ADHD, 30 TD (CH)	60	No	FD, EN, FE	MLP-NN	ACC_{CV}: 94%	70/10/20 RS	NI
Ahmadlou and Adeli, 2010 (7)	47 ADHD, 7 TD (CH)	54	No	SLM (WAV)	RBF-NN	ACC_{CV}: 96%	90/10 RS	NI
Mann *et al.*, 1992 (83)	25 ADHD, 27 TD (CH)	52	No	PSF	DFA	ACCest: 77% SEN: 80%, SPE: 74%	No CV	NI
Poil *et al.*, 2014 (84)	22 ADHD, 27 TD (AD) 19 ADHD, 22 TD (CH)	90	No	PSF	RBF-SVM	$ACCest_{CV\text{-}AD}$: 76% SEN: 67%, SPE: 83% $ACCest_{CV\text{-}CH}$: 64% SEN: 56%, SPE: 70%	Twofold CV	NI
Tcheslavski and Beex, 2006 (85)	6 ADHD, 7 TD (CH)	13	No	CA	EDC	ACC_{CV}: 63%	LOO CV	NI
Lubar *et al.*, 1985 (11)	69 ADD, 34 TD (CH)	103	No	PSF	DFA	ACC: 98%	No CV	NI

Table 1. Continued

Study	Diagnosis, *n* (Age Group)	*N*	Independent Set	Features	Classifier	$Performance_{Test\ Set}$	CV	Notes
Alba-Sanchez *et al.*, 2010 (86)	28 ADHD, 10 TD (CH)	38	No	PSF	NN	ACC_{CV}: 70%	Sixfold CV	
Nazhvani *et al.*, 2013 (6)	12 TD, 12 ADHD, 12 BMD (CH, AD)	36	No	LM, AM (VEP)	1-NNC	$ACC_{CV\ ADHD\ vs.\ TD}$: 95%	LOO-CV	
Kim *et al.*, 2015 (87)	53 ADHD, 44 TD (CH)	97	No	TGC	ROC	AUC: 71%	No CV	
Kovatchev *et al.*, 2001 (88)	33 ADHD, 34 TD (CH, AD)	67	No	Cindex	LR	ACC: 80% SEN: 82%, SPE: 77%	No CV	
Magee *et al.*, 2005 (89)	253 ADHD, 67 TD (CH)	320	No	PSF	LR	ACC: 74% SEN: 85%, SPE: 42%	No CV	
EEG (TBR)								
Snyder *et al.*, 2008 (18)	97 ADHD, 62 TD (CH)	159	See main text	TBR	DIS	ACC: 89% SEN: 87%, SPE: 94%	See main text	See main text
Monastra *et al.*, 1999 (55)	221 ADHD-C, 176 ADHD-I, 85 TD (CH, AD)	482	No	TBR	DIS	ACCest: 88% SEN: 86%, SPE: 98%	No CV	UN (SPE)
Sangal and Sangal, 2015 (58)	58 ADHD, 28 TD (CH)	86	No	TBR PSF	ROC	ACCest: 66% SEN_{TBR}: 78%, SPE_{TBR}: 43% ACCest: 77% SEN_{PSF}: 86%, SPE_{PSF}: 57%	No CV	
Monastra *et al.*, 2001 (10)	37 ADHD-I, 59 ADHD-C, 33 TD (CH, AD)	129	See main text	TBR	DIS	ACCest: 91% SEN: 90%, SPE: 94%	See main text	See main text
Ogrim *et al.*, 2012 (56)	62 ADHD, 39 TD (CH)	101	No	TBR, TP	LR	ACC_{TBR}: 58% ACC_{TP}: 62%	No CV	
Liechti *et al.*, 2013 (57)	32 ADHD, 30 TD (CH)	62	No	PSF, TBR	DFA	ACC_{TBR}: 44% SEN: 47%, SPE: 40% ACC_{PSF}: 71% SEN: 69%, SPE: 73%	No CV	
Snyder *et al.*, 2015 (20)	275 (CH)	275	See main text	TBR + clinical evaluation	HA	ACC: 88%	See main text	See main text
MRI								
Ghiassian *et al.*, 2013 (90)	279 ADHD, 790 TD (CH, AD; ADHD200)	1069	77 ADHD, 94 TD Total: 171	SF	RBF-SVM	ACC_{IS}: 63%	IS	
Dai *et al.*, 2012 (91)	285 ADHD, 491 TD (ADHD200, CH)	776	(Estimated) 75 ADHD, 94 TD Total: 169	FF/FC, SF	RBF-SVM	ACC_{IS}: 59% SEN: 44%, SPE: 71%	IS	
Eloyan *et al.*, 2012 (92)	285 ADHD, 491 TD (CH, AD, ADHD200)	776	65 ADHD, 128 TD Total: 193	SF, FF/FC	Various methods	ACC_{IS}: 61% SEN: 21% SPE: 91%	IS	
Ghiassian *et al.*, 2016 (93)	285 ADHD, 491 TD (CH, AD, ADHD200)	776	77 ADHD, 94 TD Total: 171	SF, PD	RBF-SVM	ACC_{IS}: 70%	IS	
Colby *et al.*, 2012 (94)	285 ADHD, 491 TD (CH, AD; ADHD200)	776	51 ADHD-C, 26 ADHD-I, 94 TD Total: 171	SF, FF/FC, DD	RBF-SVM	ACC_{IS}: 55% SEN: 33%, SPE: 80%	IS	

Table 1. Continued

Study	Diagnosis, *n* (Age Group)	*N*	Independent Set	Features	Classifier	Performance$_{\text{Test Set}}$	CV	Notes
Dey *et al.*, 2012 (95)	156 ADHD-C, 99 ADHD-I, 11 ADHD-H, 468 TD (CH, AD, ADHD200)	734	50 ADHD-C, 26 ADHD-I, 2 ADHD-H, 93 TD Total: 171	NF	PCA-LDA	ACC_{IS}: 70% SEN: 87%, SPE: 49%	IS	
Sidhu *et al.*, 2012 (96)	141 ADHD-C, 98 ADHD-I, 429 TD (CH, AD, ADHD200)	668	51 ADHD-C, 26 ADHD-I, 94 TD Total: 171	PD/ PD, PSF	SVM	$ACC_{IS\text{-}PD}$: 71% $ACC_{IS\text{-}PD,\ PSF}$: 67%	IS	
Fair *et al.*, 2013 (35)	112 ADHD-C, 80 ADHD-I, 455 TD (CH)	647	No	FC	SVM-based MVPA	$ACC_{CV\ ADHD\text{-}C\ vs.\ TD}$: 77% SEN: 75%, SPE: 77% $ACC_{CV\ ADHD\text{-}I\ vs.\ TD}$: 83% SEN: 79%, SPE: 87%	LOO CV	
dos Santos Siqueira *et al.*, 2014 (97)	269 ADHD, 340 TD (CH, ADHD200)	609	No	FC	Linear SVM	ACC_{CV}: 73% SEN: 63%, SPE: 83%	LOO CV	
Chang *et al.*, 2012 (98)	210 ADHD, 226 TD (CH, AD; ADHD200)	436	No	SF	Linear SVM	ACC_{CV}: 70%	10-fold CV	
Tan *et al.*, 2017 (99)	215 ADHD, 98 TD (CH, ADHD200)	313	No	FF, PD	Linear SVM	ACC_{CV}: 69% SEN: 78%, SPE: 57%	10-fold CV	
Wolfers *et al.*, 2016 (100)	184 ADHD, 128 TD (CH, AD)	312	No	FF	GPC	AUC_{CV}: 64%	LOO CV	
Cheng *et al.*, 2012 (101)	98 ADHD, 141 TD (CH)	239	No	fALFF, ReHo, FF/FC, SF	RBF-SVM	ACC_{CV}: 76% SEN: 63%, SPE: 85%	LOO CV	
Jie *et al.*, 2016 (102)	118 ADHD, 98 TD (CH, ADHD200)	216	No	FC	SVM	ACC_{CV}: 83% SEN: 84%, SPE: 82%	LOO CV	
Du *et al.*, 2016 (103)	118 ADHD, 98 TD (CH, ADHD200)	216	No	FC	SVM	ACC_{CV}: 95% SEN: 93%, SPE: 97%	10-fold CV	
Qureshi *et al.*, 2017 (104)	53 ADHD-C, 53 ADHD-I, 53 TD (CH, ADHD200)	159	14 ADHD-C, 14 ADHD-I, 14 TD Total: 42	SF	ELM	$ACC_{IS\ ADHD\text{-}C\ vs.\ TD}$: 89% $ACC_{IS\ ADHD\text{-}I\ vs.\ TD}$: 93%	IS	
Qureshi *et al.*, 2016 (105)	53 ADHD-C, 53 ADHD-I, 53 TD (CH, ADHD200)	159	No	SF	H-ELM	$ACC_{CV\ ADHD\text{-}C\ vs.\ TD}$: 78% $ACC_{CV\ ADHD\text{-}I\ vs.\ TD}$: 80%	10-fold CV	
Qureshi and Boreom, 2016 (106)	30 ADHD-C, 30 ADHD-I, 30 TD (CH, ADHD200)	90	No	SF	Linear SVM	$ACC_{CV\ ADHD\text{-}C\ vs.\ TD}$: 79% $ACC_{CV\ ADHD\text{-}I\ vs.\ TD}$: 70%	10-fold CV	
Abibullaev and An, 2012 (17)	7 ADHD, 3 TD (CH)	10	No	PSF	RBF-SVM	ACC_{CV}: 97%	Fivefold CV	UN
Fu *et al.*, 2013 (15)	21 ADHD, 27 TD (AD) Total: 48	48	No	PSF	SVM	ACC_{CV}: 96%	HS	UN
Kurtek *et al.*, 2011 (107)	19 ADHD, 15 TD (AD) Total: 34	34	No	SF	LOO-NNC	ACC_{CV}: 91%	Fivefold CV	
Hammer *et al.*, 2015 (14)	20 ADHD, 20 TD (CH) Total: 40	40	No	FF	LR	ACC_{CV}: 93%	LOO-CV	NI
Iannaccone *et al.*, 2015 (108)	18 ADHD, 18 TD (CH) Total: 36	36	No	FF, SF	Linear SVM	ACC_{CV}: 78% SEN: 78%, SPE: 78%	LOO-CV	UN
Deshpande *et al.*, 2015 (13)	260 ADHD-C, 173 ADHD-I, 744 TD (CH, AD; ADHD200)	1177	No	FF/FC	FCC-ANN	$ACC_{CV\ ADHD\text{-}C\ or\ ADHD\text{-}I\ vs.\ TD}$: >90% (precise values unclear)	LOO-CV	
Peng *et al.*, 2013 (109)	55 ADHD, 55 TD (CH)	110	No	SF	ELM	ACC_{CV}: 90%	LOO-CV	NI

Table 1. Continued

Study	Diagnosis, *n* (Age Group)	*N*	Independent Set	Features	Classifier	Performance$_{\text{Test Set}}$	CV	Notes
Igual *et al.*, 2012 (110)	39 ADHD, 39 TD (CH)	78	No	SF	SVM (Adaboost)	ACC_{CV}: 72% SEN: 60%, SPE: 86%	Fivefold CV	
Johnston *et al.*, 2014 (111)	34 ADHD, 34 TD (CH)	68	No	SF	RBF-SVM	ACC_{CV}: 93%	LOO-CV	
Dey *et al.*, 2014 (112)	15 ADHD-C, 1 ADHD-H, 12 ADHD-I, 38 TD (CH, ADHD200, OHSU database)	66	5 ADHD-C, 1 ADHD-H, 1 ADHD-I, 27 TD Total: 34	FC	Polynomial SVM	ACC_{IS}: 82% SEN: 89%, SPE: 50%	IS	
Hart *et al.*, 2014 (113)	30 ADHD, 30 TD (CH)	60	No	FF	GPC	ACC_{CV}: 77% SEN: 90%, SPE: 63%	LOO-CV	
Lim *et al.*, 2013 (114)	29 ADHD, 29 TD (CH)	58	No	SF	GPC	ACC_{CV}: 79% SEN: 76%, SPE: 83%	LOO-CV	
Wang *et al.*, 2013 (115)	23 ADHD, 23 TD (AD)	46	No	ReHo	SVM	ACC_{CV}: 80% SEN: 87%, SPE: 74%	LOO-CV	
Hart *et al.*, 2014 (116)	20 ADHD, 20 TD (CH)	40	No	FF	GPC	ACC_{CV}: 75% SEN: 80%, SPE: 70%	LOO-CV	
Zhu *et al.*, 2008 (117)	9 ADHD, 11 TD (CH)	20	No	ReHo	PCA-FDA	ACC_{CV}: 85% SEN: 78%, SPE: 91%	LOO-CV	
Zhu *et al.*, 2005 (118)	9 ADHD, 1 TD (CH)	10	No	ReHo	FDA	ACC_{CV}: 85% SEN: 78%,SPE: 91%	LOO-CV	
Semrud-Clikeman *et al.*, 1996 (119)	10 ADHD-C, 10 DYS, 10 TD (CH)	30	No	SF, PD	DFA	ACC_{ADHD-C}: 80% ACC_{DYS} : 90% ACC_{TD}: 90% (three-class model)	No CV	
MEG								
Gomez *et al.*, 2013 (120)	14 ADHD, 14 TD (CH)	28	No	FE	ROC	ACC_{CV}: 82%	LOO-CV	
fNIRS								
Monden *et al.*, 2015 (16)	30 ADHD, 30 TD (CH)	60	No	AM	ROC	AUC: 85% SEN: 90%, SPE: 70%	No CV	

ACC, accuracy; ACCest, estimated accuracy; AD, adults; ADD, attention-deficit disorder; ADHD, attention-deficit/hyperactivity disorder; ADHD200, ADHD-200 Global Competition; ADHD-C, attention-deficit/hyperactivity disorder combined type; ADHD-H, attention-deficit/hyperactivity disorder hyperactive type; ADHD-I, attention-deficit/hyperactivity disorder inattentive type; AM, amplitude measure; ANN, artificial neural network; AUC, area under the curve; AUT, autistic children; BMD, bipolar mood disorder; BN, brain network; CA, coherence analysis; CH, children; Cindex, consistency index; CV, cross-validation; DD, demographic data; DFA, discriminant function analysis; DIS, distribution; DYS, dyslexia; EDC, Euclidean distance–based classifier; EEG, electroencephalography; ELM, extreme learning machine; EN, entropy; ERP, event-related potential; FALFF, fractional amplitude of low-frequency fluctuation; FC, functional connectivity; FCC-ANN, fully connected cascade artificial neural network; FD, fractal dimension; FDA, Fisher discriminative analysis; FE, fuzzy entropy; FF, functional feature; fNIRS, functional near-infrared spectroscopy; GPC, Gaussian process classifier; h-ELM, hierarchical extreme learning machine; HA, human assessment; HS, half split; IS, independent set; LDA, linear discriminant analysis; LE, Lyapunov exponent; LM, latency measures; LOO, leave-one-out; LR, logistic regression; MAL, maturational lag; MCDD, multiple complex developmental disorder; MEG, magnetoencephalography; MLP, multilayer perceptron; MRI, magnetic resonance imaging; MVPA, multivariate pattern analysis; NF, network features; NI, nonindependent test set; NN, neural network; NNC, nearest neighbor classifier; OHSU, Oregon Health and Science University; PCA, principal component analysis; PD, personal data; PSF, power spectra features; RBF, radial basis function; ReHo, regional homogeneity; ROC, receiver-operating characteristic; RS, random split; SEN, sensitivity; SDLD, specific developmental learning disorder; SF, structural feature; SLM, synchronization likelihood method; SPE, specificity; SVM, support vector machine; TBR, theta/beta ratio; TD, typically developing; TGC, theta-phase gamma-amplitude coupling; TP, theta power; UN, unclear; WAV, wavelet; VEP, visual evoked potential.

[a]When the study did not provide the accuracy, we estimated it given the sample size, specificity and sensitivity.

specificity, sensitivity, and sample size. We tested whether accuracy varies with sample size using a logistic regression model with accuracy treated as probability of a binary outcome (i.e., corresponding to correct/incorrect prediction) and sample size as predictor. This model assumes that classification accuracy follows a binomial distribution (41,42,52,53). Influential observations were identified using Cook's D statistic exceeding $4 / n - k - 1$ (n = sample size, k = number of observations) and, if present, were excluded from final model fit.

Time Analysis

Finally, we sought to establish whether the methodological factors of concern (small sample size and circular analysis) are current problems, or whether their presence (if established) is restricted to older studies, preceding awareness of these issues in the field. To do so, we analyzed 1) an analogous logistic regression model with accuracy as a probability of a binary outcome and year of publication as a predictor; 2) a linear regression model with sample size as the dependent variable and year of publication as a predictor; and 3) contingency tables for presence of circular analysis (yes/no/unknown) and time windows constructed by binning years of publication by median split (<2013, ≥2013), and in a second analysis, also the top and bottom 33rd percentiles (≤2011, >2014).

RESULTS

Study Set Characteristics

Of the 69 studies reviewed (Table 1, Figure 1), 32 (46.4%) used electroencephalographic (EEG) features, 35 (50.7%) used functional or structural MRI features, and 2 (2.9%) used magnetoencephalography or functional near-infrared spectroscopy. Sample size varied from 10 to over 1177. Of these studies, 47 (68.1%) included children only, 14 (20.3%) included both adults and children, and 8 (11.6%) were of adults only. Classifier model parameters varied highly across studies. Almost no studies used the exact same set of features, with the exception of studies of TBR. Among algorithms chosen, support vector machines were the most common, used in 26 (37.6%) studies, followed by discriminant linear analysis (13 studies, 18.8%), neural networks (8 studies, 11.6%), and logistic regression (5 studies, 7.3%). Four studies employed receiver-operating characteristic curve analysis (5.8%) to draw conclusions regarding ability of features to discriminate between groups.

Prevalence of Circular Analysis

A total of 15.9% (11 of 69) presented methods that were consistent with circular analysis, whereby feature selection was performed on the full dataset including the test data. Nine studies (13.0%) did not employ any cross-validation. Hence, the reported accuracies were untested with respect to generalizability. In 8 of 69 studies (11.6%), independence was unclear (unknown). That is, the methods provided insufficient information to determine if circular analysis was present. For example, some studies used linear discriminant analysis trained on half the dataset, but *t* tests were used to determine which features were considered by the linear discriminant analysis. Importantly, it was not specified which data were used to perform the *t* tests (training sample only or full sample). We note that before active author inquiry, we encountered a total of 17 studies (24.6%, 17 of 69) with methodological detail insufficient to make a determination regarding feature selection.

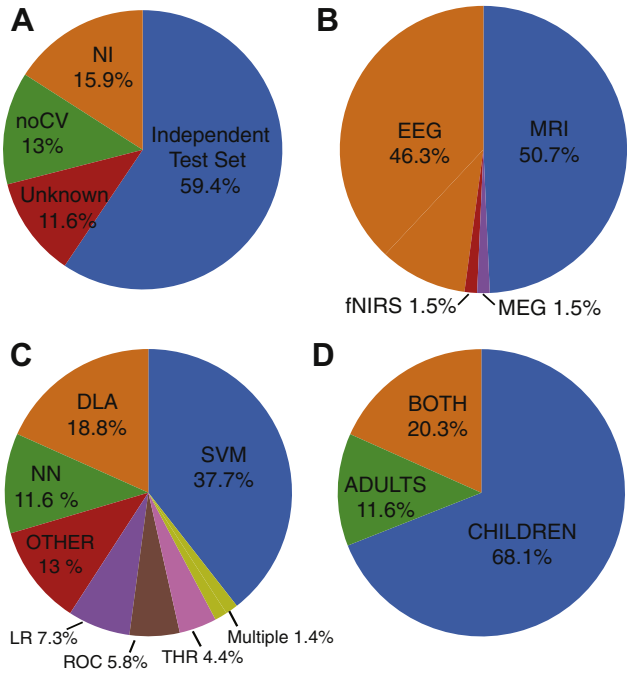

Figure 1. Study characteristics. **(A)** Of the reviewed studies, 28.9% did not meet independence criteria due to nonindependence (NI) or lack of cross-validation (noCV), with another 11.6% lacking clarity to rule out circular analysis. **(B)** Most studies used features derived from electroencephalography (EEG) and magnetic resonance imaging (MRI)–related signals. **(C)** Support vector machine (SVM) and discriminant linear analysis (DLA) were the most common algorithms. **(D)** The majority of ADHD classification studies included pediatric populations. fNIRS, functional near-infrared spectroscopy; LR, logistic regression; MEG, magnetoencephalography; NN, neural network; ROC, receiver-operating characteristic; THR, threshold-based classification.

In sum, we identified 41 studies (59.4%) that met our criteria for independence of the test set relative to training and feature selection. Of these, most (29 of 41 or 70.7%) were studies using functional MRI features [25 as part of the ADHD-200 competition (54)]. Only 26.8% (11 of 41) used EEG features. Thus, where an assessment could be performed, circular analysis was more prevalent in EEG studies than MRI, $\chi^2_{1,\ n\ =\ 51} = 8.52$, $p < .004$.

Sample Size and Classifier Accuracy

In studies that met independence criteria, the relationship between sample size and accuracy was significant (Wald $\chi^2 = 18.9$, $p < .001$; odds ratio = 0.9987; 95% confidence interval = 0.9983–0.9993) (Figure 2); for a one-unit increase in sample size, the odds of correct classification decreased by 0.12%. This translates into a predicted drop of approximately 5.9% in classifier accuracy when increasing a sample from $n = 10$ to $n = 300$, or 25.4% when increasing a sample from $n = 10$ to $n = 1000$. A sample size–accuracy relationship was not significant

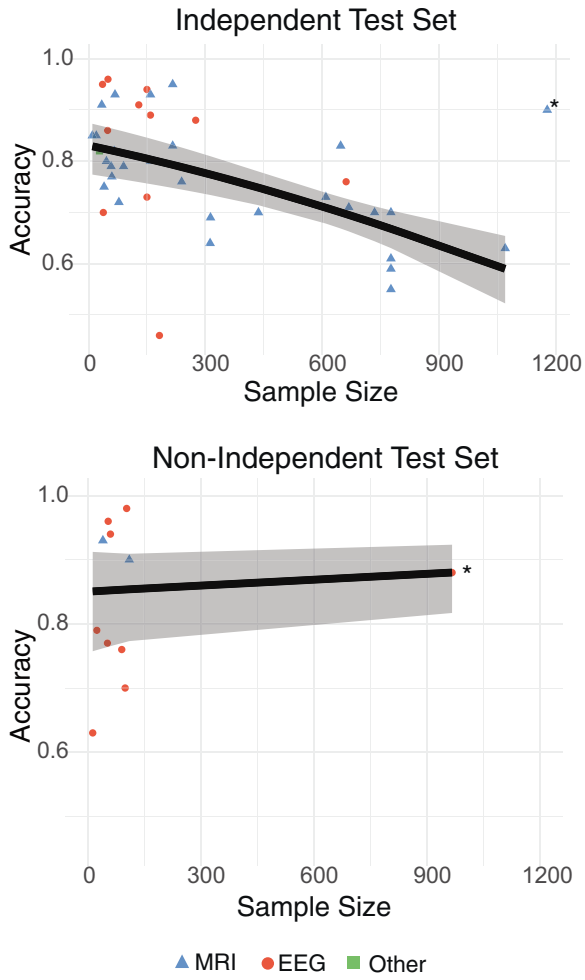

Figure 2. Sample size vs. classification accuracy. A negative relationship between classifier accuracy and sample size was evident in studies that met test set independence criteria (top panel). This group was dominated by magnetic resonance imaging (MRI) studies. In contrast, studies that did not meet independence criteria (bottom panel) were dominated by studies that used electroencephalographic (EEG) features. Shading indicates 95% confidence interval. Starred observations were found as influential by Cook's D, and were excluded from final model fits.

for studies that failed to meet independence criteria (Wald χ^2 = 0.03, p = .88) (Figure 2), possibly because of inflated accuracy across sample sizes. Confirming these effects, the mean accuracy of the 25% largest independent test set studies was significantly lower than the mean accuracy of the 25% smallest studies ($mean_{largest}$ = 68.1%, $mean_{smallest}$ = 84.5% [t_{18} = 4.4, $p < .0001$]), and also significantly lower than the nonindependent studies ($mean_{nonindependent}$ = 83.6% [t_{18} = 3.3, $p < .005$]).

As a larger portion of MRI than EEG studies used independent testing, we repeated the analysis for each modality to test whether this relationship is largely driven by MRI studies. As expected, for MRI studies, the negative association of sample size and classification accuracy was significant (Wald χ^2 = 17.0, $p < .001$; odds ratio = 0.9988, 95% confidence interval = 0.9983–0.9995). For EEG studies, the relationship was not significant (Wald χ^2 = 0.01, p = .91).

EEG-Based Theta/Beta Ratio

Our analysis included seven studies (9.9%) that classified ADHD based on the EEG-signal TBR (10,18,20,55–58). These studies are considered separately because they did not uniformly conform to the above assessment of circular analysis, and also because of their significance as an Food and Drug Administration–approved adjunct to clinical assessment (19,20). Of these, the studies of Ogrim *et al.* (56), Liechti *et al.* (57), and Sangal and Sangal (58) used analyses that did not include cross-validation. The remaining four studies [Snyder

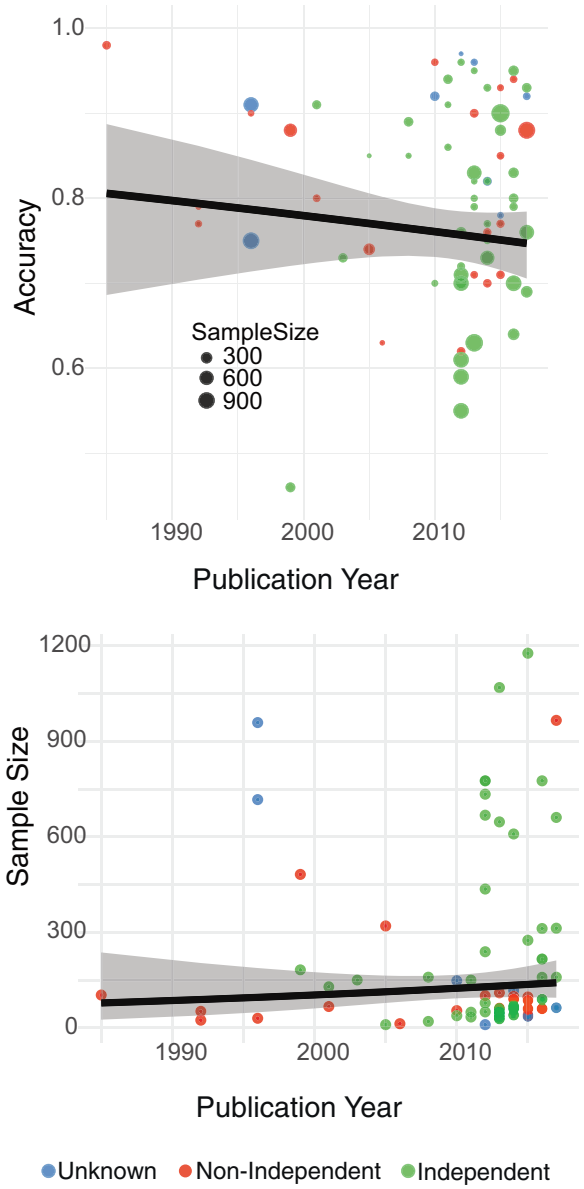

Figure 3. Classification across publication year. Neither accuracy (top panel) nor sample size (bottom panel) could be predicted from publication year. The relationship between the two (Figure 2) was also significant with publication year as a covariate (see text). Frequency of circular analysis also did not vary by year. Shading indicates 95% confidence interval.

et al. (18,20) and Monastra *et al.* (10,55)] used a distribution-based classification scheme. They predicted ADHD diagnosis based on a TBR threshold defined as 1.5 standard deviations greater than the mean of a normative control population (55). In the Monastra *et al.* (10,55) studies, the 1999 study (55) identified the threshold, whereas the 2001 study (10) provided the cross-validation result using new participants. In the Snyder *et al.* (18,20) studies, the thresholds were defined based on an external database in the 2008 study (18), and based on the 2008 result in the 2015 study (20). Thus, the Snyder *et al.* (18,20) studies and the 2001 Monastra *et al.* study (10) can be considered independent cross-validation and by this definition do not fall under circular analysis. However, these studies had limitations with respect to estimation of specificity. The non-ADHD comparison sample size averaged 16 individuals per age group [i.e., n = 7, 11, and 15 per tested age group in Monastra *et al.* (10); n = 9, 20, and 33 per tested age group in Snyder *et al.* (18)]. Finally, in the 2015 study of Snyder *et al.* (20), accuracy based on TBR alone was not reported. In all, test results are either lacking or underpowered for effective assessment of TBR classification generalizability.

Time Effects

As shown in Figure 3, year of publication predicted neither accuracy (Wald χ^2 = 0.77, p = .38) nor sample size (for a linear fit [$F_{1,67}$ = 0.22, p = .64], for an exponential fit [$F_{1,67}$ = 0.75, $p < .39$]). Given a median split (based on year published) of studies into those published in or post-2013 (n = 36) versus pre-2013 (n = 33), there was no difference in proportion of studies that met independence criteria (pre-2013 = 19, during 2013/post-2013 = 22), failed the independence criteria (pre-2013 = 10, during 2013/post-2013 = 10), or were unclassified (pre-2013 = 4, during 2013/post-2013 = 4) (χ^2 = 0.09, p = .97). A similar result was obtained comparing the bottom third (oldest) versus the top third (newest) of studies. The relationship between accuracy and sample size reported in the previous section remained significant with the inclusion of publication date as a covariate (Wald χ^2 = 25.5, $p <$ .001; odds ratio = 0.9988, 95% confidence interval = 0.9983–0.9992).

DISCUSSION

The aim of our study was to assess the contribution of circular analysis and small-sample bias to accuracy of diagnostic classification studies in ADHD using neuroimaging biomarkers. We found circular analysis in 15.9% of ADHD classification studies, lack of cross-validation in 13%, and insufficient methodological detail to establish its presence in another 11.6%. Our results reveal that accuracy of classification is 60% to 80% in the 59.4% of studies that met our criteria for independence of feature selection, model construction, and test datasets. There was a negative relationship between accuracy and sample size even in the presence of independent testing, suggesting that small-sample accuracies may be subject to bias.

Methodological Factors and Classification Accuracy

A key conclusion from our analysis is that in 28.9% of the studies reviewed, reported accuracy was likely inflated owing to presence of circular analysis or lack of internal validation (test set). In some cases, the use of a full dataset for feature selection was justified by using an analysis thought to be independent from the contrast of ADHD patients versus control subjects [e.g., mean effect across all subjects within a condition (14), principal component analysis (51)]. However, the independence of such approaches is difficult to guarantee and can still contribute to bias during testing (43), and therefore should be avoided. External validation, an even stronger test on generalizability, was absent in 55 studies (79.7%), suggesting that our estimates of true accuracy in classification of ADHD may be optimistic still. Time analyses did not support the conclusion that rates of circular analysis are decreasing across publication year. However, our estimate of 15.9% of studies reviewed is nearly a third of that reported in 2008, when 42% of high impact–journal functional MRI studies were subject to circular analysis (43,46), supporting an awareness of these methodological issues in the community. Nevertheless, the frequency of the lack of sufficient methodological detail (24.6% before author inquiry, 11.6% after author inquiry) was high and highlights a need for systematicity in review criteria of classification studies. There are now a number of excellent reviews, many specifically targeting biomarker studies in neuropsychiatry, that provide such guidance (1–3,41,42).

Replicating recent review findings of Varoquaux (41) and Varoquaux *et al.* (42) (classification using MRI and magnetoencephalography), Arbabshirani *et al.* (1) (classification across brain disorders, using functional and structural MRI features), and Schnack and Kahn (48) (classification in studies of schizophrenia using structural MRI), we show, in the context of ADHD, that accuracy in classification studies of neuroimaging data decreased with sample size. This suggests that bias is at play in small-sample studies, particularly given that, in unbiased analyses, accuracy is known to increase with sample size (47,49,59). Sources of this bias likely include publication bias, with small-sample studies that fail to obtain high classification accuracy being unlikely to be published, leading to underestimation of accuracy variance in small-sample studies. In classification of psychiatric conditions, such as ADHD, a pertinent source of bias may be sample homogeneity in small samples that is not representative of the broader population (48). An important caveat to our observations, the interaction between sample size and accuracy may be additionally affected by choice of cross-validation scheme (e.g., k-fold vs. leave one out), data preprocessing (e.g., control for motion artifacts), and classifier. An exhaustive analysis of these factors fell outside of the scope of the current study, owing to variability in these factors among studies, but a preliminary analysis did not reveal differences in choice of classifier or cross-validation scheme across sample size (see Supplement). It is notable that accuracy did not appear to decrease across year of publication, whereas sophistication in machine learning has certainly improved. The decrease in accuracy with sample size that we observed appears robust to these alternative methodological choices.

Critically, the solution to small-sample problems lies in rigorous statistical assessment of classifier accuracy. This can be achieved using the binomial test (for two-class problems) and permutation testing (50). Permutation testing, in particular,

is a reliable, flexible, and readily available tool to assess the significance and variability of a given accuracy (53,60). Reporting of both significance and an estimate of variability, such as confidence intervals, is perhaps the most important recommendation because, independent of availability of larger samples, such reporting continues to be done inconsistently based on a 2017 review of 237 classification studies across brain disorders (1). Finally, although difficult to quantify, it is inherent that the amount of data per subject varies from study to study, and thus the reliability varies depending on the neuroimaging measure employed. This fact underscores further the importance of data quality in addition to data quantity in predictive modeling.

Value of Biomarkers in ADHD Diagnosis and Beyond

This study was motivated by an apparent inconsistency between reports of high classification accuracies and known heterogeneity in ADHD. We found that the subset of studies with independent test sets reported an accuracy in the range of 60% to 80%. The fact that these values were significantly above 50% suggests that neuroimaging-based biomarkers were associated with ADHD and therefore have some value. However, these accuracies are too low to be used without other supporting information in clinical practice because they would result in substantial false positive and false negative rates [also see Loo and Barkley (61)]. We also note that the test set was difficult to define in the studies of TBR (10,18,20,55), which is significant because TBR is a Food and Drug Administration–approved adjunct to clinical assessment (19,20). These studies also did not include large control samples to accurately estimate the standard error, which could mean that the specificity of the TBR has been overestimated. Such a conclusion is consistent both with reported variability in the group-difference effect size of TBR (38,62,63) and, in particular, with the observation that decreasing effect sizes of TBR across studies appears to be correlated with a change in TBR in the control sample rather than the ADHD sample (62).

However, the low and variable accuracies are consistent with the inherent heterogeneity of ADHD, documented in ADHD in symptom presentation (21), neurocognitive impairment (22,23), persistence (24–26), treatment response (27,28), and putative mechanistic pathways (29–31). Given a heterogeneous population, classification models will learn to accurately identify those individuals with features that are shared among subpopulations, but will be less successful in identifying individuals who have features specific to a subpopulation. However, as argued by Schnack and Kahn (48), a drop in accuracy in a new testing sample in this context carries information about the mutual homogeneity of the sample and may help to identify shared versus nonshared features.

What is the future of biomarkers in ADHD? Echoing recent reviews, we suggest that the primary goals within ADHD ought to include parsing of heterogeneity and prediction of future outcomes, rather than diagnosis. Addressing heterogeneity, dimensional analysis approaches [e.g., Research Domain Criterion initiative (64,65)] seek to identify novel subgroups, based on shared neuroimaging (and other feature) profiles. A promising example of this approach is that of Bansal *et al.* (66), who developed an automated routine to first discover natural groupings based on brain morphology. Using these novel groupings, they achieved classification sensitivity of 93.6% and specificity of 88.5% on an independent testing set including children with ADHD and control subjects. In complement, a shift toward using machine learning and biomarkers to predict future outcomes—development and aging, education, learning, criminality, health-related behaviors, response to treatments—is likely to have a greater impact than prediction of diagnosis on personalized clinical practices than can directly improve patients' lives (67–69). For instance, brain network connectivity associated with sustained attention performance has been shown to predict ADHD symptoms in an independent sample (70–73), defining a potential tool for diagnosis-independent assessment of attentional integrity.

Conclusions

In this study, we found that unbiased classification accuracy in ADHD diagnosis in the range of 60% to 80%, too low to be viewed as an independently useful biomarker of disease, is consistent with known heterogeneity in this disorder. These data are also consistent with contributions of circular analysis and small-sample bias to inflation of higher accuracies, thus accounting for the discrepancy. We conclude that steps to resolve these issues, as well as a shift toward accounting for sample heterogeneity and prediction of future outcomes, will be crucial in increasing the utility of classification in ADHD.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by National Institutes of Health Grant Nos. MH101282 and NS97484 (to SKL), UCLA-California Institute of Technology Medical Scientist Training Program (National Institutes of Health Grant No. T32 GM08042), and the Eisenhower Medical Center Department of Internal Medicine.

We would like to thank Francisco Pereira, Scott Makeig, and Iman Rezazadeh for helpful discussions.

The authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Department of Psychiatry and Biobehavioral Sciences (WTK, SKL, AL) and Department of Biomathematics (WTK), University of California, Los Angeles, Los Angeles; and Department of Internal Medicine (WTK), Eisenhower Medical Center, Rancho Mirage, California; and the Paris Descartes University (AAP), Paris, France.

Address correspondence to Agatha Lenartowicz, Ph.D., Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, 760 Westwood Plaza, Suite A7-457, Los Angeles, CA 90095; E-mail: alenarto@g.ucla.edu.

Received May 4, 2018; revised Jun 15, 2018; accepted Jun 18, 2018.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.bpsc.2018.06.003.

REFERENCES

1. Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017): Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage 145:137–165.
2. Woo CW, Chang LJ, Lindquist MA, Wager TD (2017): Building better biomarkers: Brain models in translational neuroimaging. Nat Neurosci 20:365–377.
3. Kassraian-Fard P, Matthis C, Balsters JH, Maathuis MH, Wenderoth N (2016): Promises, pitfalls, and basic guidelines for

applying machine learning classifiers to psychiatric imaging data, with autism as an example. Front Psychiatry 7:177.
4. Mueller A, Candrian G, Grane VA, Kropotov JD, Ponomarev VA, Baschera GM (2011): Discriminating between ADHD adults and controls using independent ERP components and a support vector machine: A validation study. Nonlinear Biomed Phys 5:5.
5. Mueller A, Candrian G, Kropotov JD, Ponomarev VA, Baschera GM (2010): Classification of ADHD patients on the basis of independent ERP components using a machine learning system. Nonlinear Biomed Phys 4(suppl 1):S1.
6. Nazhvani AD, Boostani R, Afrasiabi S, Sadatnezhad K (2013): Classification of ADHD and BMD patients using visual evoked potential. Clin Neurol Neurosurg 115:2329–2335.
7. Ahmadlou M, Adeli H (2010): Wavelet-synchronization methodology: A new approach for EEG-based diagnosis of ADHD. Clin EEG Neurosci 41:1–10.
8. Ghassemi F, Moradi MH, Tehrani-Doost M, Abootalebi V (2012): Using non-linear features of EEG for ADHD/normal participants' classification. Procedia Soc Behav Sci 32:148–152.
9. Mohammadi MR, Khaleghi A, Nasrabadi AM, Rafieivand S, Begol M, Zarafshan H (2016): EEG classification of ADHD and normal children using non-linear features and neural networks. Biomed Eng Lett 6:66–73.
10. Monastra VJ, Lubar JF, Linden M (2001): The development of a quantitative electroencephalographic scanning process for attention deficit-hyperactivity disorder: Reliability and validity studies. Neuropsychology 15:136–144.
11. Lubar JF, Bianchini KJ, Calhoun WH, Lambert EW, Brody ZH, Shabsin HS (1985): Spectral analysis of EEG differences between children with and without learning disabilities. J Learn Disabil 18:403–408.
12. Chabot RJ, Serfontein G (1996): Quantitative electroencephalographic profiles of children with attention deficit disorder. Biol Psychiatry 40:951–963.
13. Deshpande G, Wang P, Rangaprakash D, Wilamowski B (2015): Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. IEEE Trans Cybern 45:2668–2679.
14. Hammer R, Cooke GE, Stein MA, Booth JR (2015): Functional neuroimaging of visuospatial working memory tasks enables accurate detection of attention deficit and hyperactivity disorder. Neuroimage Clin 9:244–252.
15. Fu J, Liu Z, Gao X (2013): Investigtion of low frequency drift in attention deficit hyperactivity disorder fMRI signal. Presented at the ICIST 2013-The 3rd International Conference on Information Science and Technology, March 23–25, Yangzhou, China.
16. Monden Y, Dan I, Nagashima M, Dan H, Uga M, Ikeda T, *et al.* (2015): Individual classification of ADHD children by right prefrontal hemodynamic responses during a go/no-go task as assessed by fNIRS. Neuroimage Clin 9:1–12.
17. Abibullaev B, An J (2012): Decision support algorithm for diagnosis of ADHD using electroencephalograms. J Med Syst 36:2675–2688.
18. Snyder SM, Quintana H, Sexson SB, Knott P, Haque AF, Reynolds DA (2008): Blinded, multi-center validation of EEG and rating scales in identifying ADHD within a clinical sample. Psychiatry Res 159:346–358.
19. Food and Drug Administration (2013): De novo classification request for neuropsychiatric EEG-based assessment aid for ADHD (NEBA) system. Submission no. K112711. Available at: https://www.accessdata.fda.gov/cdrh_docs/reviews/K112711.pdf. Accessed April 27, 2018.
20. Snyder SM, Rugino TA, Hornig M, Stein MA (2015): Integration of an EEG biomarker with a clinician's ADHD evaluation. Brain Behav 5:e00330.
21. Willcutt EG, Nigg JT, Pennington BF, Solanto MV, Rohde LA, Tannock R, *et al.* (2012): Validity of DSM-IV attention deficit/hyperactivity disorder symptom dimensions and subtypes. J Abnorm Psychol 121:991–1010.
22. Sonuga-Barke EJS, Castellanos FX (2005): A common core dysfunction in attention-deficit/hyperactivity disorder: A scientific red herring? Behav Brain Sci 28:443–444.
23. Nigg JT, Willcutt EG, Doyle AE, Sonuga-Barke EJ (2005): Causal heterogeneity in attention-deficit/hyperactivity disorder: Do we need neuropsychologically impaired subtypes? Biol Psychiatry 57:1224–1230.
24. Lara C, Fayyad J, de Graaf R, Kessler RC, Aguilar-Gaxiola S, Angermeyer M, *et al.* (2009): Childhood predictors of adult attention-deficit/hyperactivity disorder: Results from the World Health Organization World Mental Health Survey Initiative. Biol Psychiatry 65:46–54.
25. Biederman J, Petty CR, Evans M, Small J, Faraone SV (2010): How persistent is ADHD? A controlled 10-year follow-up study of boys with ADHD. Psychiatry Res 177:299–304.
26. Biederman J, Petty CR, Clarke A, Lomedico A, Faraone SV (2011): Predictors of persistent ADHD: An 11-year follow-up study. J Psychiatr Res 45:150–155.
27. Swanson J, Arnold LE, Kraemer H, Hechtman L, Molina B, Hinshaw S, *et al.* (2008): Evidence, interpretation, and qualification from multiple reports of long-term outcomes in the Multimodal Treatment study of Children With ADHD (MTA): Part I: Executive summary. J Atten Disord 12:4–14.
28. Hinshaw SP, Arnold LE, Group MTAC (2015): Attention-deficit hyperactivity disorder, multimodal treatment, and longitudinal outcome: Evidence, paradox, and challenge. Wiley Interdiscip Rev Cogn Sci 6:39–52.
29. Castellanos FX, Tannock R (2002): Neuroscience of attention-deficit/hyperactivity disorder: The search for endophenotypes. Nat Rev Neurosci 3:617–628.
30. Rubia K, Alegria A, Brinson H (2014): Imaging the ADHD brain: Disorder-specificity, medication effects and clinical translation. Expert Rev Neurother 14:519–538.
31. Arnsten AF (2006): Fundamentals of attention-deficit/hyperactivity disorder: Circuits and pathways. J Clin Psychiatry 67(suppl 8):7–12.
32. Loo SK, McGough JJ, McCracken JT, Smalley SL (2017): Parsing heterogeneity in attention-deficit hyperactivity disorder using EEG-based subgroups. J Child Psychol Psychiatry 59:223–231.
33. Gates KM, Molenaar PC, Iyer SP, Nigg JT, Fair DA (2014): Organizing heterogeneous samples using community detection of GIMME-derived resting state functional networks. PLoS One 9:e91322.
34. Clarke AR, Barry RJ, Dupuy FE, Heckel LD, McCarthy R, Selikowitz M, *et al.* (2011): Behavioural differences between EEG-defined subgroups of children with Attention-Deficit/Hyperactivity Disorder. Clin Neurophysiol 122:1333–1341.
35. Fair DA, Nigg JT, Iyer S, Bathula D, Mills KL, Dosenbach NU, *et al.* (2012): Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data. Front Syst Neurosci 6:80.
36. Sonuga-Barke E, Bitsakou P, Thompson M (2010): Beyond the dual pathway model: Evidence for the dissociation of timing, inhibitory, and delay-related impairments in attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 49:345–355.
37. van Hulst BM, de Zeeuw P, Durston S (2015): Distinct neuropsychological profiles within ADHD: A latent class analysis of cognitive control, reward sensitivity and timing. Psychol Med 45:735–745.
38. Loo SK, Arns M (2015): Should the EEG-based theta to beta ratio be used to diagnose ADHD? ADHD Report 23:8–13.
39. Lenartowicz A, Loo SK (2014): Use of EEG to diagnose ADHD. Curr Psychiatry Rep 16:498.
40. Loo SK, Makeig S (2012): Clinical utility of EEG in attention-deficit/hyperactivity disorder: A research update. Neurotherapeutics 9:569–587.
41. Varoquaux G (2018): Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 180(Pt A):68–77.
42. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B (2017): Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. Neuroimage 145:166–179.
43. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009): Circular analysis in systems neuroscience: The dangers of double dipping. Nat Neurosci 12:535–540.

44. Cawley G, Talbot NLC (2010): On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079–2107.
45. Simon R (2003): Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). ACM SIGKDD Explorations Newsletter 5:31.
46. Vul E, Pashler H (2012): Voodoo and circularity errors. Neuroimage 62:945–948.
47. Honorio J (2014): Classification on brain functional magnetic resonance imaging: Dimensionality, sample size, subject variability and noise. In: Chen CH, editor. Frontiers of Medical Imaging. Dartmouth, MA: University of Massachusetts Dartmouth, 153–165.
48. Schnack HG, Kahn RS (2016): Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. Front Psychiatry 7:50.
49. Nieuwenhuis M, van Haren NE, Hulshoff Pol HE, Cahn W, Kahn RS, Schnack HG (2012): Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. Neuroimage 61:606–612.
50. Combrisson E, Jerbi K (2015): Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci Methods 250:126–136.
51. Duffy FH, Shankardass A, McAnulty GB, Als H (2017): A unique pattern of cortical connectivity characterizes patients with attention deficit disorders: A large electroencephalographic coherence study. BMC Med 15:51.
52. Pereira F, Botvinick M (2011): Information mapping with pattern classifiers: A comparative study. Neuroimage 56:476–496.
53. Pereira F, Mitchell T, Botvinick M (2009): Machine learning classifiers and fMRI: A tutorial overview. Neuroimage 45:S199–S209.
54. ADHD Consortium (2012): The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience. Front Syst Neurosci 6:62.
55. Monastra VJ, Lubar JF, Linden M, VanDeusen P, Green G, Wing W, *et al.* (1999): Assessing attention deficit hyperactivity disorder via quantitative electroencephalography: An initial validation study. Neuropsychology 13:424–433.
56. Ogrim G, Kropotov J, Hestad K (2012): The quantitative EEG theta/beta ratio in attention deficit/hyperactivity disorder and normal controls: Sensitivity, specificity, and behavioral correlates. Psychiat Res 198:482–488.
57. Liechti MD, Valko L, Muller UC, Dohnert M, Drechsler R, Steinhausen HC, *et al.* (2013): Diagnostic value of resting electroencephalogram in attention-deficit/hyperactivity disorder across the lifespan. Brain Topogr 26:135–151.
58. Sangal RB, Sangal JM (2015): Use of EEG beta-1 power and theta/beta ratio over Broca's area to confirm diagnosis of attention deficit/hyperactivity disorder in children. Clin EEG Neurosci 46:177–182.
59. Chu C, Hsu AL, Chou KH, Bandettini P, Lin C, , Alzheimer's Disease Neuroimaging Initiative. (2012): Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage 60:59–70.
60. Noirhomme Q, Lesenfants D, Gomez F, Soddu A, Schrouff J, Garraux G, *et al.* (2014): Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. Neuroimage Clin 4:687–694.
61. Loo SK, Barkley RA (2005): Clinical utility of EEG in attention deficit hyperactivity disorder. Appl Neuropsychol 12:64–76.
62. Arns M, Conners CK, Kraemer HC (2013): A decade of EEG theta/beta ratio research in ADHD: A meta-analysis. J Atten Disord 17:374–383.
63. Loo SK, Cho A, Hale TS, McGough J, McCracken J, Smalley SL (2013): Characterization of the theta to beta ratio in ADHD: Identifying potential sources of heterogeneity. J Atten Disord 17:384–392.
64. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, *et al.* (2010): Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. Am J Psychiatry 167:748–751.
65. Cuthbert BN, Insel TR (2010): Toward new approaches to psychotic disorders: The NIMH Research Domain Criteria project. Schizophr Bull 36:1061–1062.
66. Bansal R, Staib LH, Laine AF, Hao X, Xu D, Liu J, *et al.* (2012): Anatomical brain images alone can accurately diagnose chronic neuropsychiatric illnesses. PLoS One 7:e50698.
67. Gabrieli JD, Ghosh SS, Whitfield-Gabrieli S (2015): Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron 85:11–26.
68. Rosenberg MD, Casey BJ, Holmes AJ (2018): Prediction complements explanation in understanding the developing brain. Nat Commun 9:589.
69. Kessler D, Angstadt M, Sripada C (2016): Growth charting of brain connectivity networks and the identification of attention impairment in youth. JAMA Psychiatry 73:481–489.
70. O'Halloran L, Cao Z, Ruddy K, Jollans L, Albaugh MD, Aleni A, *et al.* (2018): Neural circuitry underlying sustained attention in healthy adolescents and in ADHD symptomatology. Neuroimage 169:395–406.
71. Rosenberg MD, Finn ES, Scheinost D, Papademetris X, Shen X, Constable RT, *et al.* (2016): A neuromarker of sustained attention from whole-brain functional connectivity. Nat Neurosci 19:165–171.
72. Rosenberg MD, Finn ES, Scheinost D, Constable RT, Chun MM (2017): Characterizing attention with predictive network models. Trends Cogn Sci 21:290–302.
73. Yoo K, Rosenberg MD, Hsu WT, Zhang S, Li CR, Scheinost D, *et al.* (2018): Connectome-based predictive modeling of attention: Comparing different functional connectivity features and prediction methods across datasets. Neuroimage 167:11–22.
74. Chabot RJ, Merkin H, Wood LM, Davenport TL, Serfontein G (1996): Sensitivity and specificity of QEEG in children with attention deficit or specific developmental learning disorders. Clin Electroencephalogr 27:26–34.
75. Helgadottir H, Gudmundsson OO, Baldursson G, Magnusson P, Blin N, Brynjolfsdottir B, *et al.* (2015): Electroencephalography as a clinical tool for diagnosing and monitoring attention deficit hyperactivity disorder: A cross-sectional study. BMJ Open 5:e005500.
76. Smith JL, Johnstone SJ, Barry RJ (2003): Aiding diagnosis of attention-deficit/hyperactivity disorder and its subtypes: Discriminant function analysis of event-related potential data. J Child Psychol Psychiatry 44:1067–1075.
77. Kemner C, van der Gaag RJ, Verbaten M, van Engeland H (1999): ERP differences among subtypes of pervasive developmental disorders. Biol Psychiatry 46:781–789.
78. Biederman J, Hammerness P, Sadeh B, Peremen Z, Amit A, Or-Ly H, *et al.* (2017): Diagnostic utility of brain activity flow patterns analysis in attention deficit hyperactivity disorder. Psychol Med 47:1259–1270.
79. Allahverdi A, Nasrabadi AM, Mohammad M (2011): Detecting ADHD children using symbolic dynamic of nonlinear features of EEG. Presented at the 19th Iranian Confernece on Electrical Engineering, May 17–19, Tehran, Iran.
80. Robaey P, Breton F, Dugas M, Renault B (1992): An event-related potential study of controlled and automatic processes in 6-8-year-old boys with attention deficit hyperactivity disorder. Electroencephalogr Clin Neurophysiol 82:330–340.
81. Tenev A, Markovska-Simoska S, Kocarev L, Pop-Jordanov J, Muller A, Candrian G (2014): Machine learning approach for classification of ADHD adults. Int J Psychophysiol 93:162–166.
82. Lenartowicz A, Delorme A, Walshaw PD, Cho AL, Bilder RM, McGough JJ, *et al.* (2014): Electroencephalography correlates of spatial working memory deficits in attention-deficit/hyperactivity disorder: Vigilance, encoding, and maintenance. J Neurosci 34:1171–1182.
83. Mann CA, Lubar JF, Zimmerman AW, Miller CA, Muenchen RA (1992): Quantitative-analysis of EEG in boys with attention-deficit-hyperactivity disorder: Controlled study with clinical implications. Pediatr Neurol 8:30–36.
84. Poil SS, Bollmann S, Ghisleni C, O'Gorman RL, Klaver P, Ball J, *et al.* (2014): Age dependent electroencephalographic changes in

attention-deficit/hyperactivity disorder (ADHD). Clin Neurophysiol 125:1626–1638.
85. Tcheslavski GV, Beex AA (2006): Phase synchrony and coherence analyses of EEG as tools to discriminate between children with and without attention deficit disorder. Biomed Signal Process Control 1:151–161.
86. Alba-Sanchez F, Yanez-Suarez O, Brust-Carmona H (2010): Assisted diagnosis of attention-deficit hyperactivity disorder through EEG bandpower clustering with self-organizing maps. Conf Proc IEEE Eng Med Biol Soc 2010:2447–2450.
87. Kim JW, Lee J, Kim BN, Kang T, Min KJ, Han DH, *et al.* (2015): Theta-phase gamma-amplitude coupling as a neurophysiological marker of attention deficit/hyperactivity disorder in children. Neurosci Lett 603:25–30.
88. Kovatchev B, Cox D, Hill R, Reeve R, Robeva R, Loboschefski T (2001): A psychophysiological marker of attention deficit/hyperactivity disorder (ADHD)–defining the EEG consistency index. Appl Psychophysiol Biofeedback 26:127–140.
89. Magee CA, Clarke AR, Barry RJ, McCarthy R, Selikowitz M (2005): Examining the diagnostic utility of EEG power measures in children with attention deficit/hyperactivity disorder. Clin Neurophysiol 116:1033–1040.
90. Ghiassian S, Greiner R, Jin P, Brown MRG (2013): Learning to classify psychiatric disorders based on fMR images: Autism vs healthy and ADHD vs healthy. Presented at the NIPS 2013 Workshop on Machine Learning and Interpretation in Neuroimaging, December 9–10, Lake Tahoe, Nevada.
91. Dai D, Wang J, Hua J, He H (2012): Classification of ADHD children through multimodal magnetic resonance imaging. Front Syst Neurosci 6:63.
92. Eloyan A, Muschelli J, Nebel MB, Liu H, Han F, Zhao T, *et al.* (2012): Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. Front Syst Neurosci 6:61.
93. Ghiassian S, Greiner R, Jin P, Brown MR (2016): Using functional or structural magnetic resonance images and personal characteristic data to identify ADHD and autism. PLoS One 11:e0166934.
94. Colby JB, Rudie JD, Brown JA, Douglas PK, Cohen MS, Shehzad Z (2012): Insights into multimodal imaging classification of ADHD. Front Syst Neurosci 6:59.
95. Dey S, Rao AR, Shah M (2012): Exploiting the brain's network structure in identifying ADHD subjects. Front Syst Neurosci 6:75.
96. Sidhu GS, Asgarian N, Greiner R, Brown MR (2012): Kernel principal component analysis for dimensionality reduction in fMRI-based diagnosis of ADHD. Front Syst Neurosci 6:74.
97. dos Santos Siqueira A, Biazoli Junior CE, Comfort WE, Rohde LA, Sato JR (2014): Abnormal functional resting-state networks in ADHD: Graph theory and pattern recognition analysis of fMRI data. Biomed Res Int 2014:380531.
98. Chang CW, Ho CC, Chen JH (2012): ADHD classification by a texture analysis of anatomical brain MRI data. Front Syst Neurosci 6:66.
99. Tan L, Guo X, Ren S, Epstein JN, Lu LJ (2017): A computational model for the automatic diagnosis of attention deficit hyperactivity disorder based on functional brain volume. Front Comput Neurosci 11:75.
100. Wolfers T, van Rooij D, Oosterlaan J, Heslenfeld D, Hartman CA, Hoekstra PJ, *et al.* (2016): Quantifying patterns of brain activity: Distinguishing unaffected siblings from participants with ADHD and healthy individuals. Neuroimage Clin 12:227–233.
101. Cheng W, Ji X, Zhang J, Feng J (2012): Individual classification of ADHD patients by integrating multiscale neuroimaging markers and advanced pattern recognition techniques. Front Syst Neurosci 6:58.
102. Jie B, Wee CY, Shen D, Zhang D (2016): Hyper-connectivity of functional networks for brain disease diagnosis. Med Image Anal 32:84–100.
103. Du J, Wang L, Jie B, Zhang D (2016): Network-based classification of ADHD patients using discriminative subnetwork selection and graph kernel PCA. Comput Med Imaging Graph 52:82–88.
104. Qureshi MNI, Oh J, Min B, Jo HJ, Lee B (2017): Multi-modal, multi-measure, and multi-class discrimination of ADHD with hierarchical feature extraction and extreme learning machine using structural and functional brain MRI. Front Hum Neurosci 11:157.
105. Qureshi MN, Min B, Jo HJ, Lee B (2016): Multiclass classification for the differential diagnosis on the ADHD subtypes using recursive feature elimination and hierarchical extreme learning machine: Structural MRI study. PLoS One 11:e0160697.
106. Qureshi MN, Boreom L (2016): Classification of ADHD subgroup with recursive feature elimination for structural brain MRI. Conf Proc IEEE Eng Med Biol Soc 2016:5929–5932.
107. Kurtek S, Klassen E, Ding Z, Jacobson SW, Jacobson JL, Avison MJ, *et al.* (2011): Parameterization-invariant shape comparisons of anatomical surfaces. IEEE Trans Med Imaging 30:849–858.
108. Iannaccone R, Hauser TU, Ball J, Brandeis D, Walitza S, Brem S (2015): Classifying adolescent attention-deficit/hyperactivity disorder (ADHD) based on functional and structural imaging. Eur Child Adolesc Psychiatry 24:1279–1289.
109. Peng X, Lin P, Zhang T, Wang J (2013): Extreme learning machine-based classification of ADHD using brain structural MRI data. PLoS One 8:e79476.
110. Igual L, Soliva JC, Escalera S, Gimeno R, Vilarroya O, Radeva P (2012): Automatic brain caudate nuclei segmentation and classification in diagnostic of Attention-Deficit/Hyperactivity Disorder. Comput Med Imaging Graph 36:591–600.
111. Johnston BA, Mwangi B, Matthews K, Coghill D, Konrad K, Steele JD (2014): Brainstem abnormalities in attention deficit hyperactivity disorder support high accuracy individual diagnostic classification. Hum Brain Mapp 35:5179–5189.
112. Dey S, Rao AR, Shah M (2014): Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects. Front Neural Circuits 8:64.
113. Hart H, Chantiluke K, Cubillo AI, Smith AB, Simmons A, Brammer MJ, *et al.* (2014): Pattern classification of response inhibition in ADHD: Toward the development of neurobiological markers for ADHD. Hum Brain Mapp 35:3083–3094.
114. Lim L, Marquand A, Cubillo AA, Smith AB, Chantiluke K, Simmons A, *et al.* (2013): Disorder-specific predictive classification of adolescents with attention deficit hyperactivity disorder (ADHD) relative to autism using structural magnetic resonance imaging. PLoS One 8:e63660.
115. Wang X, Jiao Y, Tang T, Wang H, Lu Z (2013): Altered regional homogeneity patterns in adults with attention-deficit hyperactivity disorder. Eur J Radiol 82:1552–1557.
116. Hart H, Marquand AF, Smith A, Cubillo A, Simmons A, Brammer M, *et al.* (2014): Predictive neurofunctional markers of attention-deficit/hyperactivity disorder based on pattern classification of temporal processing. J Am Acad Child Adolesc Psychiatry 53:569–578 e561.
117. Zhu CZ, Zang YF, Cao QJ, Yan CG, He Y, Jiang TZ, *et al.* (2008): Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. Neuroimage 40:110–120.
118. Zhu CZ, Zang YF, Liang M, Tian LX, He Y, Li XB, *et al.* (2005): Discriminative analysis of brain function at resting-state for attention-deficit/hyperactivity disorder. Med Image Comput Comput Assist Interv 8:468–475.
119. Semrud-Clikeman M, Hooper SR, Hynd GW, Hern K, Presley R, Watson T (1996): Prediction of group membership in developmental dyslexia, attention deficit hyperactivity disorder, and normal controls using brain morphometric analysis of magnetic resonance imaging. Arch Clin Neuropsychol 11:521–528.
120. Gomez C, Poza J, Fernandez A, Bachiller A, Gomez J, Hornero R (2013): Entropy analysis of MEG background activity in attention-deficit/hyperactivity disorder. Conf Proc IEEE Eng Med Biol Soc 2013:5057–5060.