The World Journal of Biological Psychiatry

ISSN: 1562-2975 (Print) 1814-1412 (Online) Journal homepage: www.tandfonline.com/journals/iwbp20

EEG/ERP-based biomarker/neuroalgorithms in adults with ADHD: Development, reliability, and application in clinical practice

Andreas Müller, Sarah Vetsch, Ilia Pershin, Gian Candrian, Gian-Marco Baschera, Juri D. Kropotov, Johannes Kasper, Hossam Abdel Rehim & Dominique Eich

To cite this article: Andreas Müller, Sarah Vetsch, Ilia Pershin, Gian Candrian, Gian-Marco Baschera, Juri D. Kropotov, Johannes Kasper, Hossam Abdel Rehim & Dominique Eich (2020) EEG/ERP-based biomarker/neuroalgorithms in adults with ADHD: Development, reliability, and application in clinical practice, The World Journal of Biological Psychiatry, 21:3, 172-182, DOI: 10.1080/15622975.2019.1605198

To link to this article: https://doi.org/10.1080/15622975.2019.1605198

THE WORLD JOURNAL OF BIOLOGICAL PSYCHIATRY
2020, VOL. 21, NO. 3, 172–182
https://doi.org/10.1080/15622975.2019.1605198

ORIGINAL INVESTIGATION

EEG/ERP-based biomarker/neuroalgorithms in adults with ADHD: Development, reliability, and application in clinical practice

Andreas Müller[a], Sarah Vetsch[a], Ilia Pershin[a], Gian Candrian[a], Gian-Marco Baschera[a], Juri D. Kropotov[b], Johannes Kasper[c], Hossam Abdel Rehim[d] and Dominique Eich[e]

[a]Brain and Trauma Foundation Grisons/Switzerland, Chur, Switzerland; [b]N.P. Bechtereva Institute of the Human Brain, Russian Academy of Sciences, St. Petersburg, Russia; [c]Praxisgemeinschaft für Psychiatrie und Psychotherapie, Lucerne, Switzerland; [d]Psychiatrie und Psychotherapie Rapperswil, Rapperswil, Switzerland; [e]Department of Psychiatry, Psychotherapy and Psychosomatics, University of Zurich, Zurich, Switzerland

ABSTRACT

Objectives: The electrophysiological characteristics of attention-deficit/hyperactivity disorder (ADHD) and recent machine-learning methods promise easy-to-use approaches that can complement existing diagnostic tools when sufficiently large samples are used. Neuroalgorithms are models of multidimensional brain networks by means of which ADHD patient data can be separated from healthy control data.

Methods: Spontaneous electroencephalographic and event-related potential (ERP) data were collected three times over the course of 2 years from a multicentre sample of adults comprising 181 patients with ADHD and 147 healthy controls. Spectral power and ERP amplitude and latency measures were used as input data for a semi-automatic machine-learning framework.

Results: ADHD patients and healthy controls could be classified with a sensitivity ranging from 75% to 83% and specificity values of 71% to 77%. In the analysis of the repeated measurements, sensitivity values of the selected logistic regression model remained high (72% and 76%), while specificity values slightly decreased over time (64% and 67%).

Conclusions: Implementation of the system in clinical practice requires facilities to track affected networks, as well as expertise in neuropathophysiology. Therefore, the use of neuroalgorithms can enhance the diagnostic process by making it less subjective and more reliable and linking it to the underlying pathology.

ARTICLE HISTORY
Received 6 September 2018
Revised 13 February 2019
Accepted 3 April 2019

KEYWORDS
ADHD; biomarker; neuroalgorithm; machine learning; EEG/ERP

Introduction

Psychiatric disorders are highly complex because genetic, biological, and mental factors interactively generate behaviour, emotions and cognition in specific cultural fields. Concurrently, the organism interacts with the biosocial, physical und structural environment in terms of transactional patterns and adaptation processes (Guntern 1982). Moreover, using questionnaires and information from patients and their relatives, clinical diagnoses of mental disorders rely on subjective descriptions and external observations of patient behaviour. Due to the complexity of mental disorders and in view of the inherent subjectivity, diagnoses are error-prone, even though the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases provide accurate descriptions of the various conditions.

From biomarkers to neuroalgorithms

In this light, great efforts have been made to search for and develop biological markers for mental disorders (Ritsner 2009; Mueller et al. 2010, 2011a, 2011b; McLoughlin et al. 2014; Helgadóttir et al. 2015; Kropotov 2016). Most such biomarkers are genetic, pharmacogenetic, biochemical, or epigenetic constellations (Bonvicini et al. 2016; Wang et al. 2018), blood, plasma, or serum measures (Cubero-Millán et al. 2017; Wang et al. 2018), or hormonal states (Pauli-Pott et al. 2017). A number of so-called electroencephalography (EEG)/evoked potential, magnetic resonance imaging (MRI)/functional MRI and diffusion tensor imaging biomarker candidates have also been established (Rubia et al. 2014; Kropotov 2016).

However, patient groups and healthy subjects have complex characteristics and thus are hard to

CONTACT Andreas Müller andreas_mueller@hin.ch Brain and Trauma Foundation Grisons/Switzerland, Poststrasse 22, 7000 Chur, Switzerland
Supplemental data for this article can be accessed at https://doi.org/10.1080/15622975.2019.1605198.

distinguish using single markers. Symptoms can be driven by different neurobiological pathways, the consideration of which is important for treatment response (Cuthbert and Insel 2013; Insel 2014; Insel and Cuthbert 2015). In view of the multitude of processes involved in the bio-psychosocial interplay of mental disorders such as attention deficit/hyperactivity disorder (ADHD), there is an urgent need for novel solutions towards multimodal and multilevel diagnostic systems in psychiatry, which would then result in precise treatment decisions (see Figure 1).

Summarising, in addition to the significant subjectivity in making diagnoses, the current constriction to a subdomain of the total of elements involved in the genesis of psychiatric diseases may also lead to misdiagnoses of certain conditions. Accordingly, for many researchers, there is a desire to consider pathophysiological processes and the use of more objective methods in the prediction, recognition and treatment of mental illnesses, and of ADHD in particular (Castellanos and Tannock 2002; Thome et al. 2012; Casey et al. 2013; Faraone et al. 2014).

ADHD and EEG/event-related potential (ERP) biomarkers

In this paper, we focus on EEG/ERP measures. The advantages of ERP studies include the possibility to follow cognitive processes non-invasively at the millisecond scale (Banaschewski and Brandeis 2007). Several studies have already shown that children, adolescents and adults with ADHD display alterations in brain functions measured by means of EEG/ERP attributes (Lenartowicz and Loo 2014; Snyder et al. 2015; Yang et al. 2015; Kakuszi et al. 2016).

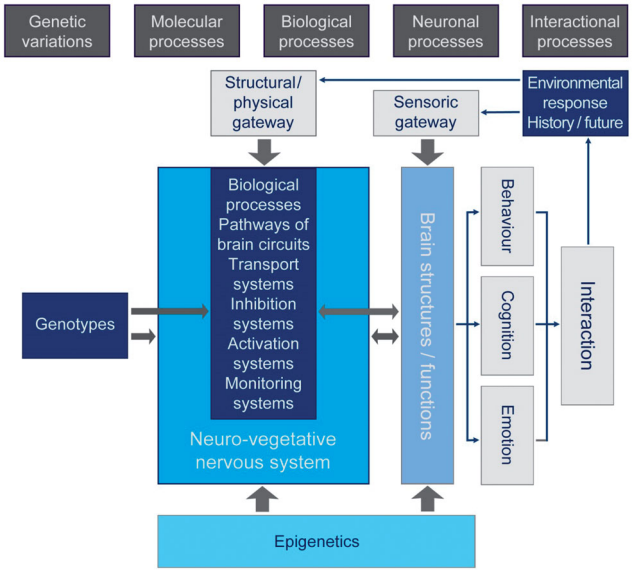

Figure 1. Elements of multiparameter diagnoses. Elements and their interaction in a multi-parameter diagnosis of mental disorders. Ultimately, all levels should be considered in a diagnosis. Currently, only descriptions of emotion, cognition, behaviour and their interactions flow in a diagnosis. Neuroalgorithms as defined in this paper are located in the 'Brain functions' box. In order to develop a precise diagnosis, all levels must be considered in the future.

Big data approaches in psychiatry

The use of big data approaches has increased significantly over the last 10 years in the healthcare sector (El Aboudi and Benhlima 2018; Islam et al. 2018; Khalifa 2018; Meskó et al. 2018). This corresponds to the idea that there are combinations of markers that can distinguish datasets better than individual variables do. However, in the field of psychiatry, the use of big data approaches has been limited to date (Jollans and Whelan 2018). Jollans and Whelan (2018) have described various barriers to the use of neural markers in applied psychiatry. These include the still insufficient knowledge regarding statistical approaches to handle the enormous amounts of data, the need for larger samples, a lack of understanding of the psychiatric regularity of neuropathology, and the fear of turning away from the previous diagnostic criteria used in psychiatry.

EEG/ERP-based big data approaches in ADHD

The first studies using big data techniques to classify patients with ADHD and healthy individuals based on EEG/ERPs were published in 2010 and 2011. Dealing with an adult sample, they achieved classification accuracies of 90% using Support Vector Machine (SVM) (Mueller et al. 2010, 2011a). Using the same methodology, Tenev et al. (2014) classified individuals based on spectral analyses. The authors analysed EEG frequency bands at rest (eyes open/closed) and during two continuous performance tasks (VCPT (visual continuous performance task) and ECPT (emotional continuous performance task)). The classification performance ranged from 69.2% (VCPT only) to 82.3% (combination of all conditions). Öztoprak et al. (2017) classified boys as ADHD and 'non-ADHD' based on time-frequency band analysis of evoked potentials recorded during a Stroop test. The Time-Frequency Hermite-Atomizer technique was used to extract the high-resolution time-frequency features. SVM-recursive feature elimination provided the most distinctive features. The allocation to the different groups was 100% correct for the testing dataset. Nazhvani et al. (2013)

reported a classification accuracy of 92.85% in a study of healthy individuals and patients with ADHD and bipolar mood disorder, where they used evoked potentials during a visual stimulation task. They subsequently analysed the data using nonlinear machine learning. Helgadóttir et al. (2015) examined a larger sample (310 patients with ADHD and 351 healthy subjects aged 5.8–14 years) using statistical pattern recognition and achieved 76%/81% (cross-validation) accuracy for age-specific classification and 76%/73% (cross-validation) accuracy for age-independent classification. The above study indicates that chronological age is an important classification feature. Heinrich et al. (2014) investigated neural mechanisms of motor control using evoked potentials in combination with MRI and achieved a classification rate of 90% in a linear discrimination analysis. The above study indicates that both cognitive and motor inhibition must be understood as fundamental difficulties in children with ADHD. In summary, various studies have reported the good classification capability of EEG and evoked potentials. Non-linear methods of classification often achieve results with accuracies of 90%–100%, although sensitivity and specificity are often not specified in such studies. Linear methods can be used to achieve accuracy values in the range of 70%–90%. The disadvantage of using non-linear methods is the lack of traceability of the results. With the exception of one study (Helgadóttir et al. 2015), the sample sizes studied are often small.

This study aims at developing neuroalgorithms for the optimal separation of healthy people and adults with ADHD using a large sample, at testing the reliability of the algorithms after 1 and 2 years and at drafting the conditions for the application in practice.

Materials and methods

Sample

The sample consisted of 328 adult participants 18 to 60 years of age at study enrolment. One-hundred and eighty-one participants were diagnosed with ADHD based on the DSM-5, and 147 constituted the control group. Participants with ADHD were recruited by advertisements in the local media and by informing local psychiatrists and ADHD associations. Controls were recruited by advertisements in the local media, schools, companies and associations.

Before the testing sessions, medicated subjects were asked to refrain from taking psychotropic drugs that could be stopped upon short notice. The intake of methylphenidates was discontinued at least 24 h before testing. The exclusion criteria were an intelligence quotient (IQ) <80, neuropsychological performance quotient <75, history of brain injury requiring rehabilitation, epilepsy, a primary mental disorder other than ADHD and insufficient knowledge of the German or French languages. Control participants with psychiatric diagnoses or histories of psychotropic medication intake were not included.

Procedure

The examinations were carried out in the framework of a Brain and Trauma Foundation Grisons/Switzerland study named 'Biomarker-oriented diagnostics of ADHD and comorbidity – children, adolescents and adults'. The study was approved by the cantonal ethics committee of Zurich (LeitEKZH_2013-0327/EKNZ_2014_160). Data were collected between July 2014 and July 2017 at five different locations in Switzerland (Zurich, Chur, Lausanne, Lucerne and Rapperswil). Persons interested in participation were asked to fill in and return an ADHD screening questionnaire before the first appointment. At the initial consultation, informed consent was provided and ADHD diagnosis based on DSM-5 criteria was made or verified by a psychiatric specialist. In the subsequent first testing session, questionnaires were filled, an interview (SKID (Wittchen et al. 1997)) and an IQ test (Formann et al. 2011) were conducted, and a series of neuropsychological tests (candit.com) was performed. In a second session, EEG data were recorded. During the course of the study, a blood sample was obtained by an external laboratory. The present work describes the analysis of the EEG/ERP data.

Over a period of 2 years, EEG/ERP data were collected from the participants with ADHD at five measurement time points separated by 6 months and from the controls at three measurement time points separated by 1 year. The present report only includes data from the first assessment and the 1- and 2-year follow-ups. The first examination provided the basis for the formation of the classification models, and the follow-ups were intended to provide information regarding the reliability of the models.

EEG and ERP tasks

EEG data were recorded over 4 min with eyes closed and eyes opened in the resting condition as well as during a VCPT lasting approximately 22 min. This test was a cued go/no-go task and primarily served to assess the executive function of suppressing an action.

The 400 trials, each of which consisted of a pair of consecutively presented visual stimuli, were grouped into four categories. In go trials, a picture of an animal is followed by a picture of an animal and the participant is asked to press a button as fast as possible. In no-go trials, a picture of an animal is followed by a picture of a plant and the participant is asked to withhold from pressing the button. In ignore (the picture of a plant is followed by a picture of a plant) and novelty (the picture of a plant is followed by a picture of a human being, the latter being presented along with a novel sound) trials, no action is asked from the participants. A detailed description of the task is provided in a previous publication by the authors (Mueller et al. 2010).

Electrophysiological recording and processing

EEGs were recorded using the NeuroAmp x23 (BEE Medic GmbH, Switzerland), which is a personal computer-controlled 23-channel digital EEG system with direct current coupling and 24-bit resolution. The input signals were bandpass-filtered between 0.5 and 50 Hz and sampled at a rate of 500 Hz. The montage was changed from linked-earlobes to common average reference before data processing. Electrodes were placed according to the International 10-20 system using a fitting electrode cap with tin electrodes (Electro-cap International, USA). Impedance was kept below 5 kΩ for all electrodes.

Raw EEGs recorded using ERPrec software (BEE Medic GmbH, Switzerland) were processed and analysed using Matlab-based in-house software. Eye-blinks and horizontal eye movements were detected using independent component analysis (ICA) decomposition, and removed from EEGs by zeroing the activation of the respective components. Remaining artefacts were removed by rejecting filtered EEG segments with amplitudes higher than 100 μV and/or excessive activity in the 0–3 and 20–50 Hz frequency bands (threshold = channel *z* score of 6).

Feature extraction

The features of interest comprised the eyes-closed, eyes-open, and VCPT signal power in a series of different frequency bands, as well as ERP peak amplitudes and latencies. The feature set also included peak amplitudes and latencies of independent ERP components obtained by decomposing the multivariate ERP signal using group ICA (Mueller et al. 2010). ERP quantification was performed by an ERP expert using the peak amplitude approach with guidance from a Matlab-based tool. The point on the ERP waveform at which the wave reached the maximum (or the minimum) was determined within a time window that characterised the particular ERP component. The time window size was fixed at 80% of the time distance from the component peak to the nearer adjacent peak on the grand average curve, and the centre of the window was equal to the peak of the component. The ERP features identified by the tool were displayed using a graphical user interface. The respective points on the curves were either adopted or adjusted by the expert, the latter occurring, for example, if the proposed maximum did not correspond to a local extremum. Previous to peak selection, the ERP waves were baseline-corrected using the 100-ms pre-stimulus period.

Statistical evaluations in the machine-learning framework

Two principal issues concerning the application of machine-learning algorithms in neurophysiological studies are lack of data and difficulties in the interpretation of measured features. The first issue leads to poor signal-to-noise ratio, while the second issue leads to difficulties in operationalisation and, thus, wide feature sets. In order to handle these problems in the present study, we developed a semi-automatic framework allowing us to prepare data, train models, perform model/feature selection and evaluate performance. The grid search allowed us to find the best model from a range of many possible options, while the nested cross-validation ensured the absence of overfitting during parameter tuning and model selection.

Model selection

Grid search is a default hyperparameter search technique in machine learning. It performs an exhaustive search over a set of specified hyperparameters called the parameter grid to find the best set of hyperparameters. The key feature of the grid search in the framework is that it allows one to search over different classification algorithms including hyperparameters of the whole pipeline of preprocessing transforms and the final estimator (classification algorithm).

In the setup of the present study, the parameter grid specified 158 models with different hyperparameters. Each of these models represented the whole pipeline (see Figure 2), including both preprocessing transforms and the final classifier. The preprocessing

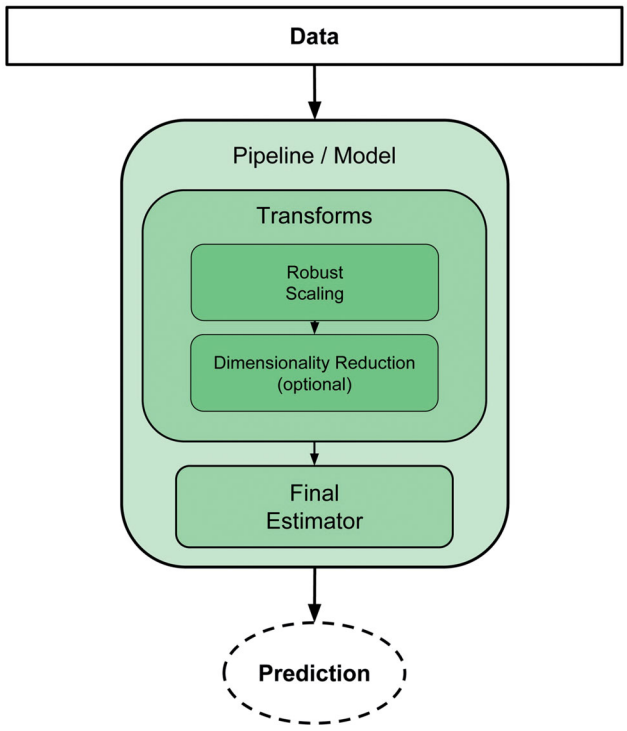

Figure 2. Pipeline of the models in the setup of the study.

steps included robust scaling and principal component analysis (PCA) dimensionality reduction as an optional transform, while the final estimator was determined using one of five classification algorithms: Regularised Logistic Regression, SVM (linear kernel), SVM (radial basis function kernel), Random Forest and XGBoost. A complete list of the models used and their hyperparameters can be found in the supplementary materials.

Feature selection

To limit the number of irrelevant features, a mixed theory-driven and statistics-driven feature selection approach was chosen. The initial set of features was selected according to theoretical ADHD models. Further feature selection was performed based on coefficients from the best-performing linear models after model selection. The framework was run multiple times to select both models and features. This approach granted more control and flexibility over feature selection by allowing us to test different hypotheses using particular feature sets at the cost of an increased risk of overfitting. A complete list of the features used can be found in the supplementary materials.

Cross-validation

The general methodology of machine learning assumes the use of three sets of data for the three stages of learning: a training set to fit the model, a development set to tune the hyperparameters of the model during model selection and a test set for the final evaluation. Each of these sets must be representative, i.e., contain a sufficient amount of data. A general approach to improve the reliability of the validation with smaller data samples is called cross-validation, and one of the most common cross-validation methods is *K*-fold cross-validation. This method allows one to use the data more efficiently at the cost of a reasonable increase in computational time by performing *K* trials with different non-overlapping training and testing sets and then averaging the obtained scores (Borra and Di Ciaccio 2010).

The framework performed *K*-fold cross-validation twice within two nested loops. This approach is called nested cross-validation. As shown in Figure 3, within the inner loop, the framework iteratively selects *K* development sets and performs model selection using a grid search. In the outer loop, the framework iteratively selects *N* test sets to perform final evaluation of the selected models. Thus, at every stage of processing in the framework, the three datasets remain entirely separated.

In our study, we used 10 folds in both loops. At the very beginning, in the first iteration of the outer loop the framework splits the data into 10 equal parts. One part is retained for testing at the end of the iteration, while the remaining 90% of the data are used in the inner loop. On the first iteration of the inner loop, the framework splits the above 90% of the data again into 10 equal parts. One of these 10 parts is again retained for testing as a development set, while the remaining data are used for training.

Scoring

The ROC AUC (area under the receiver operating characteristic curve) (Hajian-Tilaki 2013) was the main scoring metric for the performance evaluation. This metric represents overall performance of a model by combining the true-positive rate (sensitivity) and the false-positive rate (1 – specificity). For binary classifiers, the ROC AUC value varies from 0.5 to 1, where 1 represents perfect performance of a classifier and 0.5 is a random-guess baseline.

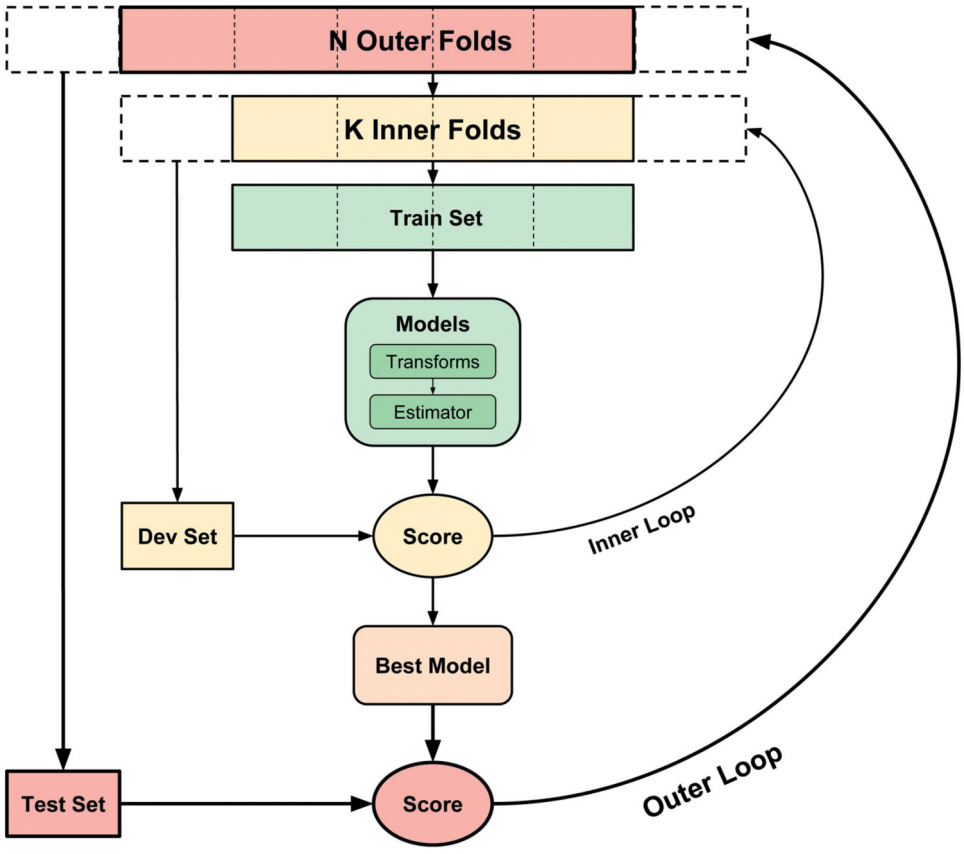

Figure 3. Workflow of the framework with two loops of the nested cross-validation.

Table 1. Mean (M) and standard deviation (SD) values for age and sex.

	Controls (M/SD)	Patients with ADHD (M/SD)	Total	*F*	χ^2	*P*
	$n = 147$	$n = 181$	$N = 328$			
Age (years)	32.10/11.96	34.54/10.16		15.11		0.050
Sex (male/female)	47/100	91/90			11.15	0.001

Group differences in age were assessed using independent-sample *t*-tests (Welch correction) and differences in sex were assessed using the chi-square test (Fisher exact test).

Testing reliability

In addition to training and evaluating different models using the machine-learning framework, one of models was selected to further test the reliability of the model evaluation. The model, which was based on data derived from measurement point 1, was tested using data derived from measurement points 2 (after 1 year) and 3 (after 2 years). To assess the test–retest reliability, we used the Intraclass Correlation Coefficient (ICC) (Koo and Li 2016). A double-mixed model with absolute agreement was selected.

Results

Sample

As shown in Table 1, controls and patients with ADHD did not differ in age. However, the proportion of females and males was different in the control and ADHD groups. There were fewer males in the control group than in the ADHD group.

Classification

We confine our report to the five classification models that achieved the highest scores in the validation procedure as described above. Classification was performed using a selected set of 47 features. Grid search

yielded the following parameter settings for the five best performing models:

1. SVM with linear kernel; penalty parameter of the error term $C = 0.1$; original feature set
2. Logistic Regression with L2 penalty; regularisation strength $C = 0.01$; original feature set
3. Logistic Regression with L2 penalty; regularisation strength $C = 0.1$; PCA components
4. Logistic Regression with L2 penalty; regularisation strength $C = 0.1$; original feature set
5. SVM with linear kernel; penalty parameter of the error term $C = 0.1$; PCA components

Table 2. Statistical performance measures for the five selected models.

Model	TPR	TNR	FPR	FNR	ACC	AUC	*P*
1 SVM	0.83	0.75	0.25	0.17	0.80	0.85	<0.001
2 Logistic regression	0.75	0.77	0.23	0.25	0.76	0.84	<0.001
3 Logistic regression	0.81	0.72	0.28	0.19	0.77	0.84	<0.001
4 Logistic regression	0.82	0.73	0.27	0.18	0.78	0.85	<0.001
5 SVM	0.82	0.71	0.29	0.18	0.77	0.84	<0.001

Differences in performance with the random-guess baseline were tested using one-tailed *t*-tests.
TPR, true-positive rate (sensitivity); TNR, true-negative rate (specificity); FPR, false-positive rate; FNR, false-negative rate; ACC, accuracy; AUC, area under the ROC curve.

The statistical properties of the five models are presented in Table 2. All models performed significantly better than the random-guess baseline. The models exhibited sensitivities between 75% and 83%, and specificities between 71% and 77%. Accuracy ranged between 76% and 80%, and the ROC AUC values were between 0.84 and 0.85. Classifier performance in terms of ROC curves is presented in Figure 4. ROC curves illustrate the true-positive rate (sensitivity) as a function of the false-positive rate (1 – specificity). Figure 5 provides insight into the process of model validation by depicting the cross-validation scores for the five models during the model development and evaluation stages.

Reliability

Due to its relatively high specificity, classification model number 2 was selected for test–retest analysis. When compared to the initial 75% sensitivity and 77% specificity at measurement point one, the sensitivity values at the subsequent measuring points (72% and 76%, respectively) were in a similar range, but the specificity decreased to <70% at measurement points 2 and 3. Table 3 summarises sensitivity and specificity values as well as ROC AUC values for the three measurement points.

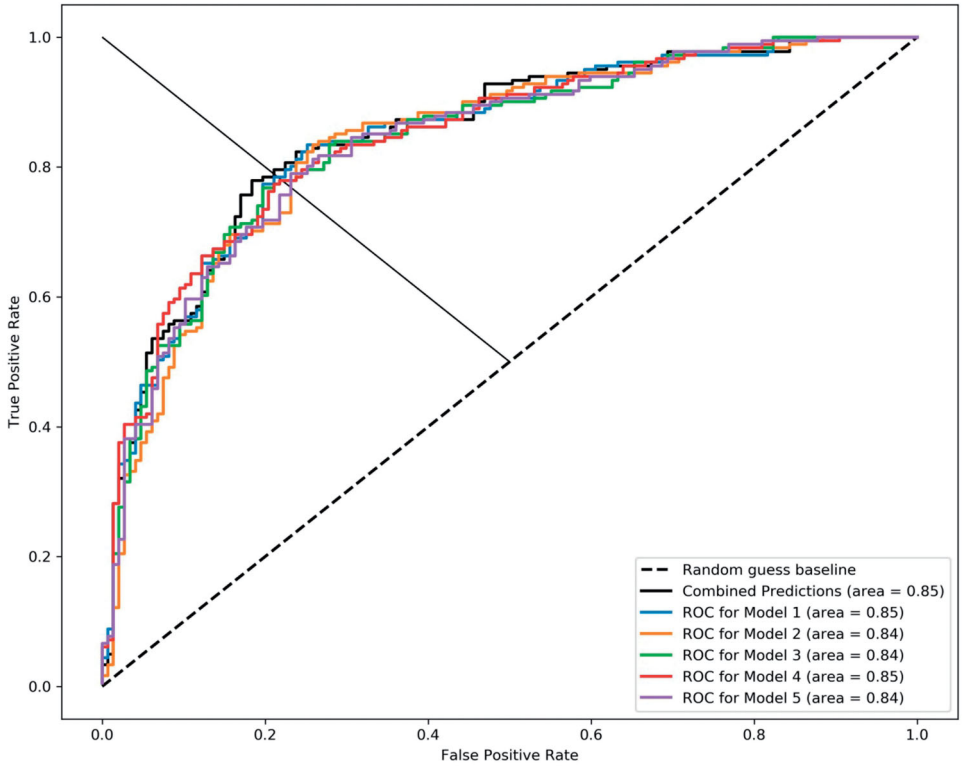

Figure 4. Receiver operating characteristic (ROC) curves corresponding to the five classification models. In ROC curves, the true-positive rate (sensitivity) is plotted against the false-positive rate (1 – specificity) for various threshold settings. The performance of different classification models can be compared using the area under the ROC curve (AUC). Here, the five best performing models exhibit very comparable AUC values in the range of 0.84 and 0.85.

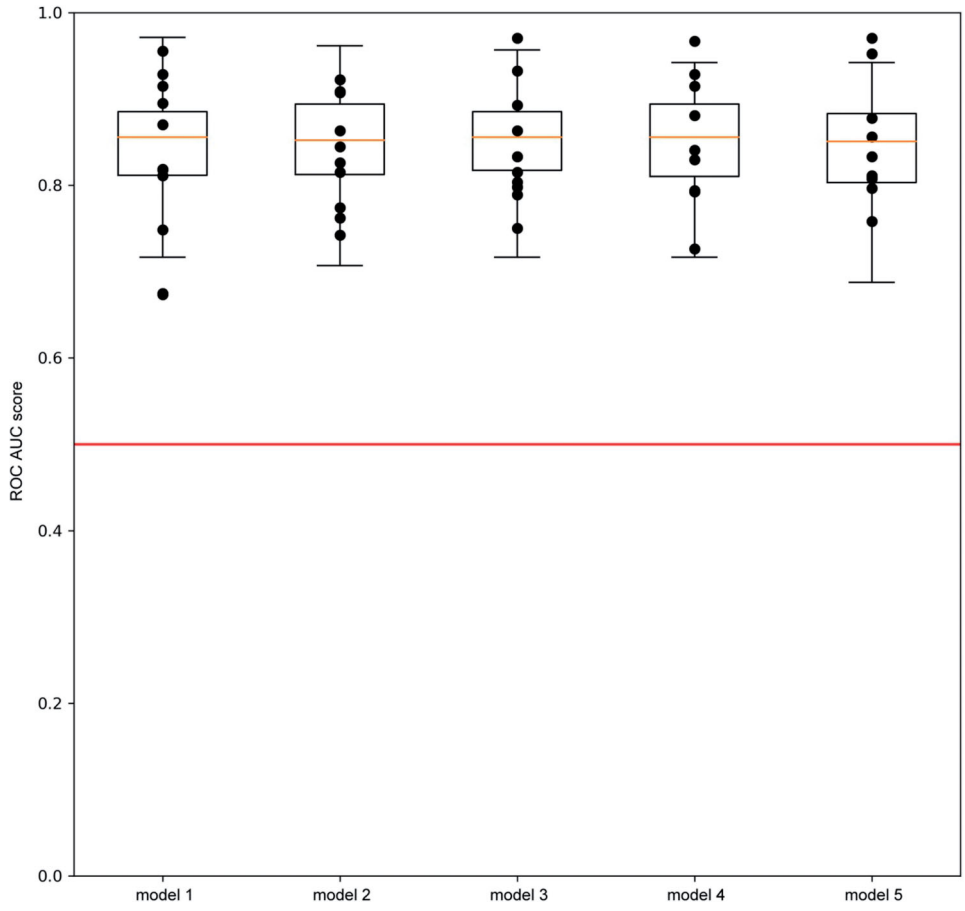

Figure 5. Training and evaluation cross-validation scores for the classification models. The framework performed nested cross-validation using 10 folds in each of two loops. The boxes represent the training stage cross-validation scores using the development sets (inner loop), the black dots show the evaluation cross-validation scores using the test sets (outer loop). The *x*-axis represents the models as listed in Table 2; the *y*-axis represents the ROC AUC scores.

Table 3. Statistical performances of the selected models at different measurement points.

	Measurement 1 (outset)	Measurement 2 (after 12 months)	Measurement 3 (after 24 months)
Sensitivity	0.75	0.72	0.76
Specificity	0.77	0.64	0.67
AUC	0.84	0.68	0.72

Measurement 1: research group $n = 181$, control group $n = 147$; Measurement 2: research group $n = 136$, control group = 137; and Measurement 3: research group $n = 127$, control group $n = 124$.
AUC, area under the ROC curve.

Table 4. Test–retest reliability (intraclass correlation) measures.

	95% confidence interval		*F*-test with true value of 0		
ICC	Lower limit	Upper limit	Value df1	df2	*p*
0.623	0.560	0.683	5.945 243	486	<.001

Model: two-way mixed, absolute agreement. Total $N = 244$ (after exclusion of 84 cases due to missing values). Research group $n = 120$, control group $n = 124$.

Notwithstanding the sensitivity and specificity values, the ICC of 0.623 indicates good consistency of classification performance over time (see Table 4).

Discussion

Here we present a set of classifiers based on supervised machine learning intended to complement conventional psychiatric-psychological diagnoses of ADHD using an objective and reliable measure based in electrophysiology. The vast majority of the presented statistical models exhibited promising sensitivity values higher than 80%. However, the specificity values ranged between 71% and 77%. Thus, the requirements for an ideal ADHD marker, as defined by Thome et al. (2012), with diagnostic sensitivity and specificity values >80%, could not be entirely met in the current study. Nevertheless, our results are promising. A biomarker should not only allow for differentiation between patients with ADHD and healthy controls, but also between patients with ADHD and those with other psychiatric disorders. Since many psychiatric diseases overlap at the aetiological, symptomatic and neuronal levels, this requirement constitutes a great challenge, which nevertheless should be considered in subsequent studies.

Psychiatric diagnoses are not made based on brief snapshots, and classification systems should remain valid over time. Therefore, the reliability of one selected classifier was assessed by including repeated EEG and ERP measurements obtained in the ADHD and control groups in the analysis. The reliability of this classifier was shown to be satisfactory. However, the approach of examining the classifier stability could be reconsidered by accounting for the clinical course of the individuals when assessing them at the follow-ups.

The reported classification performance indicates that moderate to good results in the separation of patients with ADHD and healthy controls are possible with the use of linear models. The choice between linear and non-linear models must be based on required power, robustness, and interpretability in each particular case. However, during the first stage of research, it seems reasonable to test different kinds of models to ensure that classification is possible. Later, if non-linear models are shown to not perform significantly better than the linear models, it would be logical to preferentially choose linear models, as they allow many more avenues for the inspection of the effects of particular features of the model.

A problem associated with machine learning is the dependence of the statistical models on the sample. We addressed this issue by collecting multicentre samples to ensure that the algorithms are not unilaterally associated with professionals making the diagnoses.

The applicability of the method in practice is ultimately the decisive factor. We would like to stress here that integration of all dimensions (including the life system, questionnaires/interviews, and neuropsychological investigation) in the diagnostic process is essential. In the application of a biomarker in practice, a clear statement regarding the probability of belonging to the ADHD group is required on the one hand (Bzdok and Yeo 2017), and the tracing of significant individual variables is required on the other hand. However, the latter requires intensive training of the diagnosing professional.

It is important to execute a standardised procedure in clinical practice so that the classifier can be correctly addressed. To enable the use of the classifier in clinical practice subsequent to the study, it was recreated so as to incorporate all of the feature extraction functions used in the study. This would ensure that features will be extracted in exactly the same way in future patients (see example in the Supplementary Materials).

In individual cases, a classification index can provide useful information regarding the presence of a disorder. However, the application of such an index in practice is much more complex, as the clinician must first be able to understand the neuropathophysiology of the disorder. Ultimately, basic mechanisms beyond the classification output should lead to diagnosis as well as the selection of a comprehensive treatment plan (Cavanagh et al. 2017; Stephan et al. 2017). In order to understand such mechanisms, multimodal data should be integrated into the diagnostic process. If diagnoses are reduced to an index, precise medicine will result in depersonalised psychiatry (Fuchs 2013).

An instrument such as the ADHD index can be misused, comparable to conventional psychological diagnostic tests. For instance, health insurance companies could no longer provide services in case of negative tests. Furthermore, in everyday clinical practice, conclusions could also be drawn prematurely. Therefore, the use of relatively easy-to-use indices always requires in-depth expertise on the part of the user.

In summary, the results are promising and can presumably be enhanced with consideration of a few factors. Most notably, the feature selection procedure must be improved by including it in the inner loop of the nested cross-validation scheme along with model selection to avoid the risk of overfitting. Furthermore, possible age-related effects have to be addressed, for instance by removing such effects from the data by using regression models prior to classification.

We have developed recommendations for future investigations aiming to develop neuroalgorithms in applied neuroscience. These recommendations can be found in the supplementary materials.

Acknowledgments

We would like to thank the supporting foundations: Hirschmann Foundation, Uniscientia Foundation, Hand in Hand Anstalt, Fondation Claude & Giuliana, Propter Homines Foundation, Karl Mayer Foundation, Senta Herrmann Foundation, Maiores Foundation, Unus pro multis Foundation. Without their support, research activity of the kind described above wouldn't have been possible.

Statement of interest

The authors of the publication did not receive direct research support from the above mentioned foundations and no other financial support besides the ordinary salary from Brain and Trauma Foundation. Andreas Müller and Juri Kropotov hold stocks of and serve on the board of directors of HBImed AG.

References

Banaschewski T, Brandeis D. 2007. Annotation: what electrical brain activity tells us about brain function that other techniques cannot tell us - a child psychiatric perspective. J Child Psychol Psychiatry. 48:415–435.

Bonvicini C, Faraone S, Scassellati C. 2016. Attention-deficit hyperactivity disorder in adults: a systematic review and meta-analysis of genetic, pharmacogenetic and biochemical studies. Mol Psychiatry. 21:872.

Borra S, Di Ciaccio A. 2010. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Comput Stat Data Anal. 54:14.

Bzdok D, Yeo BT. 2017. Inference in the age of big data: Future perspectives on neuroscience. Neuroimage. 155: 549–564.

Casey B, Craddock N, Cuthbert BN, Hyman SE, Lee FS, Ressler KJ. 2013. DSM-5 and RDoC: progress in psychiatry research? Nat Rev Neurosci. 14:810.

Castellanos FX, Tannock R. 2002. Neuroscience of attention-deficit/hyperactivity disorder: the search for endophenotypes. Nat Rev Neurosci. 3:617–628.

Cavanagh JF, Napolitano A, Wu C, Mueen A. 2017. The Patient Repository for EEG Data + Computational Tools (PRED + CT). Front Neuroinform. 11:67.

Cubero-Millán I, Ruiz-Ramos M-J, Molina-Carballo A, Martínez-Serrano S, Fernández-López L, Machado-Casas I, Tortosa-Pinto P, Ruiz-López A, Luna-Del-Castillo J-d-D, Uberos J, Muñoz-Hoyos A. 2017. BDNF concentrations and daily fluctuations differ among ADHD children and respond differently to methylphenidate with no relationship with depressive symptomatology. Psychopharmacology. 234:267–279.

Cuthbert BN, Insel TR. 2013. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. 11:126.

El Aboudi N, Benhlima L. 2018. Big data management for healthcare systems: architecture, requirements, and implementation. Adv Bioinformatics. 2018:4059018.

Faraone SV, Bonvicini C, Scassellati C. 2014. Biomarkers in the diagnosis of ADHD-promising directions. Curr Psychiatry Rep. 16:497.

Formann AK, Waldherr K, Piswanger K. 2011. Wiener Matrizen-Test 2. Ein Rasch-skalierter sprachfreier Kurztest zur Erfassung der Intelligenz [A Rasch scaled language-free short test for assessing intelligence]. Göttingen (Germany): Hogrefe.

Fuchs T. 2013. Personalisierte Psychiatrie? Eine Kritik und Gegendarstellung [Personalised psychiatry? A critique and counterstatement]. In: Heinze M, Schlimme JE, Kupke C, editors. Personalisierte Psychiatrie: zur Kritik eines Konzepts. Berlin (Germany): Parados; p. 85–99.

Guntern G. 1982. Auto-Organization in Human Systems. Syst Res. 27:323–337.

Hajian-Tilaki K. 2013. Receiver Operating Characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med. 4:627–635.

Heinrich H, Hoegl T, Moll GH, Kratz O. 2014. A bimodal neurophysiological study of motor control in attention-deficit hyperactivity disorder: a step towards core mechanisms? Brain. 137:1156–1166.

Helgadóttir H, Gudmundsson ÓÓ, Baldursson G, Magnússon P, Blin N, Brynjólfsdóttir B, Emilsdóttir Á, Gudmundsdóttir GB, Lorange M, Newman PK, et al. 2015. Electroencephalography as a clinical tool for diagnosing and monitoring attention deficit hyperactivity disorder: a cross-sectional study. BMJ Open. 5:e005500.

Insel TR. 2014. The NIMH research domain criteria (RDoC) project: precision medicine for psychiatry. Am J Psychiatry. 171:395–397.

Insel TR, Cuthbert BN. 2015. Medicine. Brain disorders? Precisely. Science. 348:499–500.

Islam MS, Hasan MM, Wang X, Germack HD, Noor-E-Alam MA. 2018. A systematic review on healthcare analytics: Application and theoretical perspective of data mining. Healthcare. 6:pii: 54.

Jollans L, Whelan R. 2018. Neuromarkers for mental disorders: Harnessing population neuroscience. Front Psychiatry 9:242.

Kakuszi B, Tombor L, Papp S, Bitter I, Czobor P. 2016. Altered response-preparation in patients with adult ADHD: A high-density ERP study. Psychiatry Res Neuroimaging. 249: 57–66.

Khalifa M. 2018. Health analytics types, functions and levels: A review of literature. Stud Health Technol Inform. 251: 137–140.

Koo TK, Li MY. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 15:155–163.

Kropotov JD. 2016. Functional neuromarkers for psychiatry: Applications for diagnosis and treatment. Amsterdam (Netherlands): Academic Press.

Lenartowicz A, Loo SK. 2014. Use of EEG to diagnose ADHD. Curr Psychiatry Rep. 16:498.

McLoughlin G, Makeig S, Tsuang MT. 2014. In search of biomarkers in psychiatry: EEG-based measures of brain function. Am J Med Genet. 165:111–121.

Meskó B, Hetényi G, Győrffy Z. 2018. Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv Res. 18:545.

Mueller A, Candrian G, Kropotov J. 2011b. ADHS-Neurodiagnostik in der Praxis [Neurodiagnostics of ADHD in practice]. Berlin (Germany): Springer.

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.

Mueller A, Candrian G, Grane VA, Kropotov JD, Ponomarev VA, Baschera GM. 2011a. Discriminating between ADHD adults and controls using independent ERP components and a support vector machine: a validation study. Nonlinear Biomed Phys. 5:5.

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.

Öztoprak H, Toycan M, Alp YK, Arıkan O, Doğutepe E, Karakaş S. 2017. Machine-based classification of ADHD and nonADHD participants using time/frequency features of event-related neuroelectric activity. Clin Neurophysiol. 128:2400–2410.

Pauli-Pott U, Schloß S, Ruhl I, Skoluda N, Nater UM, Becker K. 2017. Hair cortisol concentration in preschoolers with attention-deficit/hyperactivity symptoms-Roles of gender and family adversity. Psychoneuroendocrino. 86:25–33.

Ritsner M. 2009. The handbook of neuropsychiatric biomarkers, endophenotypes and genes: volume I: neuropsychological endophenotypes and biomarkers. Dordrecht (Netherlands): Springer Science & Business Media.

Rubia K, Alegría AA, Brinson H. 2014. Brain abnormalities in attention-deficit hyperactivity disorder: a review. Rev Neurol. 58:S3–S16.

Snyder SM, Rugino TA, Hornig M, Stein MA. 2015. Integration of an EEG biomarker with a clinician's ADHD evaluation. Brain Behav. 5:e00330.

Stephan KE, Siemerkus J, Bischof M, Haker H. 2017. Hat computational psychiatry relevanz für die klinische praxis der psychiatrie [Is computational psychiatry of relevance for the clinical practice of psychiatry]? Zeitschrift Für Psychiatrie, Psychologie Und Psychotherapie. 65:11.

Tenev A, Markovska-Simoska S, Kocarev L, Pop-Jordanov J, Müller A, Candrian G. 2014. Machine learning approach for classification of ADHD adults. Int J Psychophysiol. 93: 162–166.

Thome J, Ehlis A-C, Fallgatter AJ, Krauel K, Lange KW, Riederer P, Romanos M, Taurines R, Tucha O, Uzbekov M, Gerlach M. 2012. Biomarkers for attention-deficit/hyperactivity disorder (ADHD). A consensus report of the WFSBP task force on biological markers and the World Federation of ADHD. World J Biol Psychiatry. 13:379–400.

Wang L-J, Li S-C, Lee M-J, Chou M-C, Chou W-J, Lee S-Y, Hsu C-W, Huang L-H, Kuo H-C. 2018. Blood-bourne MicroRNA biomarker evaluation in attention-Deficit/Hyperactivity disorder of han chinese individuals: An exploratory study. Front Psychiatry 9:227.

Wittchen H-U, Zaudig M, Fydrich T. 1997. Strukturiertes Klinisches Interview für DSM-IV [Structured clinical interview for DSM-IV]. Göttingen (Germany): Hogrefe.

Yang M-T, Hsu C-H, Yeh P-W, Lee W-T, Liang J-S, Fu W-M, Lee C-Y. 2015. Attention deficits revealed by passive auditory change detection for pure tones and lexical tones in ADHD children. Front Hum Neurosci.Neurosci. 9:470.