Symptom and performance validity in samples of adults at clinical evaluation of ADHD: a replication study using machine learning algorithms

Oliver Hirsch, Anselm B.M. Fuermaier, Oliver Tucha, Björn Albrecht, Mira-Lynn Chavanon & Hanna Christiansen

To cite this article: Oliver Hirsch, Anselm B.M. Fuermaier, Oliver Tucha, Björn Albrecht, Mira-Lynn Chavanon & Hanna Christiansen (2022) Symptom and performance validity in samples of adults at clinical evaluation of ADHD: a replication study using machine learning algorithms, Journal of Clinical and Experimental Neuropsychology, 44:3, 171-184, DOI: 10.1080/13803395.2022.2105821

To link to this article: https://doi.org/10.1080/13803395.2022.2105821
Symptom and performance validity in samples of adults at clinical evaluation of ADHD: a replication study using machine learning algorithms

Oliver Hirsch¹, Anselm B.M. Fuermaier¹b, Oliver Tucha¹cd, Björn Albrechte, Mira-Lynn Chavanone and Hanna Christiansen⁵

¹Department of Psychology, FOM University of Applied Sciences, Siegen, Germany; ²Department of Clinical and Developmental Neuropsychology, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, The Netherlands; ³Department of Psychiatry and Psychotherapy, University Medical Center Rostock, Rostock, Germany; ⁴Department of Psychology, Maynooth University, National University of Ireland, Maynooth, Ireland; ⁵Department of Psychology, Clinical Child and Adolescent Psychology/Philippus University Marburg, Marburg, Germany

ABSTRACT

Introduction: Research has shown non-trivial base rates of noncredible symptom report and performance in the clinical evaluation of attention-deficit/hyperactivity disorder (ADHD) in adulthood. The goal of this study is to estimate and replicate base rates of symptom and performance validity test failure in the clinical evaluation of adult ADHD and derive prediction models based on routine clinical measures.

Methods: This study reuses data from a previous publication of 196 adults seeking ADHD assessment and replicates the findings on an independent sample of 700 adults recruited in the same referral context. Measures of symptom and performance validity (one SVT, two PVTs) were applied to estimate base rates. Prediction models were developed using machine learning.

Results: Both samples showed substantial rates of noncredible symptom report (one SVT failure: 35.7% – 36.6%), noncredible test performance (one PVT failure: 32.1% – 49.3%; two PVT failures: 18.9% – 27.3%), or both (each one SVT and PVT failure: 13.3% – 22.4%; one SVT and two PVT failures: 9.7% – 13.7%). Machine learning algorithms resulted in generally moderate to weak prediction models, with advantages of the reused sample compared to the independent replication sample. Associations between measures of symptom and performance validity were negligible to small.

Conclusions: This study highlights the necessity to include measures of symptom and performance validity in the clinical evaluation of adult ADHD. Further, this study demonstrates the difficulty to characterize the group failing symptom or performance validity assessment.

Introduction

Invalid symptom report and/or performance is a common phenomenon in the clinical assessment and represents a challenge for clinicians as it complicates, distorts, and biases diagnostic assessment and treatment planning. Base rates of noncredible symptom report or performance depend on the population studied and context of the assessment. With regard to attention-deficit/hyperactivity disorder (ADHD) in adulthood, research in the last two decades has documented various incentives why individuals at clinical assessment of ADHD may intentionally exaggerate or produce symptoms and impairments in order to obtain an (unwarranted) diagnosis of ADHD in adulthood (Harrison & Edwards, 2010; Musso & Gouvier, 2014). For example, nonmedical use of stimulant medication occurs frequently not exclusively but particularly on university campuses, to improve cognitive functions and thereby study performance, for recreational use, or to sell the medication to peers on the black market (Fuermaier et al., 2021; Rabiner, 2013; Rostain, 2006). Furthermore, one may benefit from special accommodation provided by the university (e.g., allowance of different forms of exams, receiving extra time for assignments and exams, getting an own room for taking tests) or at work place (e.g., assisting technology such as noise-canceling headphones, quiet office or positioning) if diagnosed with ADHD (Adamou et al., 2013; Musso & Gouvier, 2014; Tucha et al., 2015).

Finally, recent research also revealed that young adults may seek a diagnosis of ADHD as an excuse for academic failure or misbehavior in the social setting, e.g., being late, forgetful, inattentive, or impulsive (Fuermaier et al., 2021). Unwarranted diagnoses of ADHD result in much needed resources not being available to genuine patients with...
ADHD who are in real need of support, whereas those with a false positive diagnosis are potentially at an increased risk of adverse health effects due to superfluous treatment. On a societal level, illegitimate diagnoses of ADHD may fuel public debates on the existence of ADHD as a serious health problem (Hoelzle et al., 2019).

In order to strengthen quality of clinical data and increase the validity of conclusions drawn from a clinical assessment, there is broad consensus on the relevance of the assessment of symptom and performance validity in the clinical evaluation of ADHD in adulthood (Musso & Gouvier, 2014; Sibley, 2021; Sibley et al., 2018; Tucha et al., 2015). Various approaches and instruments have been considered and studied in this context, including self- and observer-report rating scales (in terms of symptom validity testing, SVT), personality inventories, routine neuropsychological tests (in terms of embedded performance validity testing, PVT), and tests specifically designed to detect non-credible performance (stand-alone PVTs). The increasing number of publications in this field and heightened awareness among clinicians for this issue resulted in a lively debate on the usefulness of the various instruments and approaches, their sensitivity and specificity to detect non-credible symptom report and performance, as well as the consequences of SVT/PVT failure for the individual assessment and treatment plan (Hoelzle et al., 2019; Nelson & Lovett, 2019; P. S. Marshall et al., 2016; Wallace et al., 2019).

Base rate estimations of noncredible symptom reporting and performance between studies may not be fully comparable as research in this field made use of different and diverse measures of symptom and performance validity, ranging from self-report questionnaires, embedded PVTs, stand-alone PVTs (Schmand & Lindeboom, 2005), or a combination of those. Early studies in this field commonly used a single PVT, such as the Word Memory Test (WMT; Green, 2003) to determine noncredible performance of individuals at clinical evaluation of adult ADHD, and reported base rates ranging from 15 to up to 48% (Harrison & Edwards, 2010; J. Suhr et al., 2008; Sullivan et al., 2007). However, empirical evidence from more recent research and current practice standards claim the use of multiple independent PVTs to determine performance validity, and define invalid performance if at least two PVTs are failed (Jennette et al., 2021; Rhoads et al., 2021; Sherman et al., 2020; Soble et al., 2020; Sweet et al., 2021). Based on the stricter criterion of at least two PVT failures on a battery of several PVTs, recent research reported base rates of non-credible performance of 11–19% in the clinical evaluation of adult ADHD (Abramson et al., 2021; Phillips et al., 2022; P. S. Marshall et al., 2016). These figures come close to the estimation of clinically working neuropsychologists in a survey of Martin & Schroeder (Martin & Schroeder, 2020) who reported an estimated base rate of 20% of invalid test performance in the clinical evaluation of ADHD.

In this context, Hirsch and Christiansen (Hirsch & Christiansen, 2018) examined a sample of 196 adults presenting for ADHD diagnostic evaluation, using a comprehensive battery of various clinical and neuropsychological measures, including the Amsterdam Short Term Memory test (ASTM; Schmand & Lindeboom, 2005) as a measure of performance validity. The study revealed an ASTM failure rate of 32.1%, which was in the range of earlier studies reporting indications of non-credible performance on a single PVT (Harrison & Edwards, 2010; J. Suhr et al., 2008; Sullivan et al., 2007). This study further showed that scores of SVT and PVT were only marginally and non-significantly related to each other. Invalid symptom reporting and cognitive performance appear to be dissociable constructs in individuals at neuropsychological evaluation for ADHD, which substantiates claims that an independent assessment of both constructs is warranted and necessary standard of practice (Aase et al., 2021; White et al., 2022).

Next to estimating base rates of non-credible performance, Hirsch and Christiansen (Hirsch & Christiansen, 2018) examined evidence for negative response bias in those who failed the ASTM. No significant differences were found between the groups passing and failing the ASTM with regard to self- and observer ratings, however, patients with ADHD failing the ASTM showed lower attention performance on continuous performance tasks than those passing. The diversity in methodology and partly rather small sample sizes in early research bring into question the robustness of the findings and call for more research on the development and evaluation of SVTs/PVTs for adults with ADHD, and to characterize those individuals failing SVTs/PVTs. Because the motivations behind SVT/PVT failures can be manifold and remain unclear in many cases, a thorough examination of those individuals failing symptom and performance validity assessment appears beneficial. The goal of such a characterization is not to replace SVTs/PVTs, but to learn about the individual and clinical characteristics of those who are more likely failing SVT/PVT assessment. Such a characterization may help to define the groups presenting noncredibly and may guide further clinical assessment and treatment in order to differentiate between and tackle reasons for invalid symptom reporting and performance.
The present study reuses the earlier sample of 196 patients with ADHD of Hirsch and Christiansen (Hirsch & Christiansen, 2018) and additionally considers an independent replication sample of $N = 700$ patients presenting for ADHD assessment in the same context. The goals of the present study are to replicate this study with its original methodology and to extend its analysis with current conceptual standards and advanced statistical methodology. Thus, the goals of this study are to (1) determine the failure rates of individuals fulfilling ADHD diagnosis on measures of symptom validity (Infrequency Index of the CAARS; J. A. Suhr et al., 2011) and performance validity. Performance validity is determined with a single PVT (ASTM) as applied in the original study as well as with the failure on more than one PVT as suggested by current standards. In this context, this study will also add evidence to the association between measures of symptom and performance validity. Moreover, and (2), this study seeks to establish prediction models of those individuals failing performance validity testing based on clinical and neuropsychological measures. In contrast to standard data analysis techniques such as univariate or multivariate group comparisons as employed in the vast majority of previous research in this field, including the original study (Hirsch & Christiansen, 2018), the present study employs machine learning (ML) algorithms for the estimation of prediction models. ML algorithms distinguish between training sets (for establishing the model) and test sets (for validation and application of the model) and are assumed to outperform traditional data analysis techniques in the validity of derived models. ML techniques are promising in this context and have been successfully applied to distinguish between credible and noncredible responses in a variety of clinical settings (Orrù et al., 2020). These techniques are able to handle non-normal, non-linear, and multivariate data, and data from different levels of measurement (Kotsiantis et al., 2006). Findings of this study are expected to be robust because of the large samples and rigorous diagnostic and statistical methodology and will be interpreted regarding their implications for the clinical assessment and treatment planning of adults with ADHD.

**Methods**

**Patients sample from Hirsch and Christiansen (Hirsch & Christiansen, 2018; $n = 196$)**

The sample was drawn from 292 patients who were consecutively seeking diagnostic counseling regarding ADHD from February 2011 to February 2015 at the ADHD outpatient clinic of the Philipps University Marburg, Germany. Fifty-eight (19.8%) of the 292 participants received no ADHD diagnosis, and 38 (13%) discontinued the diagnostic process. Therefore, a study sample of 196 (67%) adult patients with likely ADHD diagnosis resulted in which there were 127 men (65%) with a mean age of 32.1 years ($SD = 10.7$, range = 18–62 years) and 69 women (35%) with a mean age of 33.2 years ($SD = 11.0$, range = 18–59 years). Of these, 97 (49.5%) completed Grammar school, 58 (29.6%) Secondary School, 37 (18.9%) received a Basic Schooling Degree, and 4 (2.0%) had no school degree. All were German residents with German as first language. The diagnostic assessment was carried out by experienced licensed clinical psychologists on the basis of a detailed clinical history, the Wender-Reimherr Interview (WRI), the Conners’ Adult ADHD Rating Scales–Long version/Self-Rating (CAARS-L: S) and Observer-Rating (CAARS-L: O), the WURS-k, the German ADHD Self-Rating Scale (ADHS-SB), the Quantified Behavior Test Plus (Qb+©), and the Test of Attentional Performance (TAP: GO/NOGO, Divided Attention, Sustained Attention). In addition, the German version of the Amsterdam Short Term Memory Test (ASTM) was applied as a measure of performance validity (PVT; Hirsch & Christiansen, 2018).

**Current sample ($n = 700$)**

Participants of the current sample were recruited in the same referral context as the earlier sample at the ADHD outpatient clinic of the University of Marburg, Germany, between June 2015 and December 2019. In the current sample, there were 414 men (59.1%) with a mean age of 32.3 years ($SD = 9.6$, range = 18–60 years) and 286 women (40.9%) with a mean age of 34.4 years ($SD = 11.1$, range = 18–60 years) who fulfilled diagnostic criteria for ADHD based on the following assessments: 1) structured clinical interview, 2) informant ratings, 3) DSM-5 criterion assessment, 4) impairment ratings, 6) assessment of other or comorbid disorders (Sibley, 2021). The outpatient clinic at Marburg University is specialized on adult ADHD and nationally leading; a reason why many patients present for assessment. Further, adult ADHD is still rather neglected in Germany, explaining the large number of adults coming in for first time assessments. We excluded 84 subjects not having received a diagnosis of adult ADHD. In our current sample, 372 (53.1%) completed Grammar school, 192 (27.4%) Secondary School, 118 (16.9%) received a Basic Schooling Degree, and 18 (2.6%) had no school degree. All were German residents with German as first
language. All patients were medication naïve and were examined by experienced, licensed clinical psychologists relying on a detailed clinical history, and the structured diagnostic interview for ADHD in adults (DIVA 2.0), a DSM-5 based clinical interview assessing the ADHD core symptoms in childhood and adulthood, as well as psychological domains often impaired in adult ADHD (https://www.divacenter.eu/Content/VertalingPDFs/German%20DIVA%202.0_FORM.pdf). The diagnosis was based on the DIVA 2.0 results in order to fulfill DSM-5 diagnostic criteria. A standardized clinical interview (SCID) was used to assess potential comorbidity or other primary disorders explaining symptoms. The Conners’ Adult ADHD Rating Scales (CAARS-L: self-and observer-ratings), and the Qb+ (Christiansen et al., 2014; Hirsch & Christiansen, 2017; Ulberstad, 2012) were further used for criterion assessment according to DSM, as were impairment ratings. The Amsterdam Short Term Memory Test (ASTM) was additionally applied as a measure of performance validity (PVT; Schmand & Lindeboom, 2005). Apart from the DSM-based structured clinical interview and the TAP, all assessments are the same as in the sample of Hirsch and Christiansen (2018). We replaced the WRI with the DIVA, as the latter provided DSM-diagnostic assessment for both adult- and childhood and thus direct estimation of existing impairment prior to age 12 years. The TAP was no longer used as the information provided by the QbTest comprises all three ADHD core symptoms and different studies with the QbTest demonstrated satisfactory psychometric properties, outperforming the TAP (Hirsch & Christiansen, 2017).

All participants of both samples gave written informed consent. Our study conforms to the Declaration of Helsinki and was approved by the local ethics committee of the Faculty of Psychology at the Philipps University in Marburg, Germany.

The study was not preregistered. Materials and analysis code for this study are available by emailing the corresponding author.

**Measures**

CAARS-L: S and CAARS-L: O. The German version of the CAARS-L: S assesses ADHD symptoms in adults aged 18 years or older. Symptoms are rated on a Likert-type scale (0 = not at all/never to 3 = very much/very frequently). The long version consists of 66 items, but only 42 items were included in the original factor analysis by Conners, Erhardt, and Sparrow (Conners et al., 1999) due to statistical restrictions made by the authors. Four factors emerged from their analyses: inattention/memory problems, hyperactivity/restlessness, impulsivity/emotional lability, and problems with self-concept. Confirmatory factor analyses of the German version in healthy adults and ADHD patients supported this factor analytic solution (2013, 2011). The four subscales were significantly influenced by age, gender, and the number of years in education. Symptom severity decreased with increasing age, males scored higher than females on hyperactivity and sensation-seeking behavior, and females scored higher than males on problems with self-concept. Overall symptom ratings were higher for individuals who had received less education. Test–retest reliability ranged between .85 and .92; sensitivity and specificity were high for all four subscales. The CAARS-L: S represents a reliable and cross-culturally valid measure of current ADHD symptoms in adults (Christiansen et al., 2012). The same is true for the observer version of the scale (CAARS-L: O), in which a person who has a close relationship to the participant under examination should rate the same items (Christiansen et al., 2014). The hypothesized factor structure could also be supported, and the observer version also possesses good statistical quality criteria. The maximum scores for the subscales are as follows: Inattention/Memory Problems = 36, Hyperactivity/Restlessness = 36, Impulsivity/Emotional Lability = 36, Problems with Self-Concept = 18. Norms based on gender and age are provided as well as T-scores (M = 50, SD = 10), with participants scoring ≥65 representing the 85th percentile and above (Christiansen et al., 2014). We used the subscales Inattention/Memory Problems, Hyperactivity/Restlessness, Impulsivity/Emotional Lability, and Problems with Self-Concept from the self and observer version for prediction.

The Inconsistency Index was created as a validity measure to assess response inconsistency (Conners et al., 1999). The long version of the CAARS contains eight pairs of items with similar content. The absolute value of the differences between each item pair is calculated, and these differences are added together. The other validity index used in this study was introduced by Suhr, Buelow, and Riddle (J. A. Suhr et al., 2011) and is called the Infrequency Index. For this, 12 items from the scales Inattention/Memory Problems, Impulsivity/Emotional Lability, and DSM-IV Hyperactive/Impulsive Symptoms are selected that occur “pretty much,” “often” to “very much, very frequently” by 10% or less of the total participant sample. A score of 21 points or higher is indication of noncredible symptom report.

Quantified Behavior Test Plus (Qb+). The Qb+ is performed in front of a computer screen. The test equipment consists of an infrared camera, a headband with a reflective marker attached to it, and a responder button. The test was approved by the Federal Drug
Administration in 2014 (Dolgin, 2014) and is used as a diagnostic tool for ADHD (Edebol et al., 2012; Lis et al., 2010; Söderström et al., 2014). The Qb+ is a CPT measuring sustained attention with a 1-back working memory task (recall of the same object in shape and color) combined with a simultaneous high-resolution motion tracking system. It separately assesses hyperactivity, inattention, and impulsivity with nine parameters and takes 20 minutes. Presented stimuli are a blue circle, a blue square, a red circle, and a red square. A response key is to be pressed when two identical stimuli are shown in succession. The task requires stimulus information to be maintained in working memory until the next stimulus is presented and a matching process can be done (Lis et al., 2010). The ratio of target to nontarget stimuli is 25:75. During performance of the CPT, the movements of the participant are recorded with an infrared camera tracking a reflective marker attached to a headband worn by the participant. The infrared camera is placed about 1 m away from the participant, who is sitting in front of a computer screen. Participants are seated on a chair with back support but no armrest, to assure that they sit comfortably during testing, but do not adopt a reclining posture. Participants’ activities during the test are recorded by reading the coordinates (X and Y) of the headband marker. The position of the marker is sampled 50 times per second, with a spatial resolution of 1/27 mm per camera unit. The Qb+ reports a total of nine parameters, which can be divided into activity and CPT measures. The reported activity measures include five parameters: (a) Time Active, which reflects the percentage of time the subject has moved more than 1 cm/s; (b) Distance, which reflects the total distance traveled by the reflective headband marker and is measured in meters; (c) Area, measured as the surface covered by the headband reflector during the test and presented in square centimeters; (d) Total Number of Micro events that are small movements of the reflective marker that occur when a position change since the last micro event is greater than 1 mm; and (e) Motion Simplicity, a measure of complexity of the motion pattern, which is reported as a percentage. Four CPT measures are reported: (f) Reaction Time (RT) as the average time of all correct responses. This score indicates latency in information processing and motor response speed; (g) RT Variation (RTVar) is calculated using the standard deviation of the mean of correct response times and is a measure of the participant’s inconsistency in response times; (h) The total number of missed targets is represented in the Omission Errors score, while the total number of false hits is depicted by (i) Commission Errors. Normative data have been gathered from 1,307 individuals between 6 and 60 years of age for both versions of the test (QbTest 6–12 and Qb+) with an even age and gender distribution (Ulberstad, 2012). Q scores are derived for hyperactivity, inattention, and impulsivity. They are interpreted similar to Z scores with a mean of 0 and a standard deviation of 1. We used the Q scores of the listed parameters as predictors in our current study. In addition to its use as a cognitive performance measure, the Qb+ was adopted as an embedded validity indicator (EVI) in this study. There is a large body of evidence advocating the utility of CPTs as validity indicators in the neuropsychological assessment. Numerous studies have been published in this context, both for the assessment of ADHD (Harrison & Armstrong, 2020; Musso & Gouvier, 2014; Ord et al., 2021; Scimeca et al., 2021; Tucha et al., 2015) and patients with brain injury (Erdodi et al., 2014, 2018). Among typical variables of computerized attention tasks, the variability of reaction times received particular interest and was strongly supported for its utility as EVI by several research groups (Czornik et al., 2021; Erdodi et al., 2018; Ord et al., 2021; Stevens et al., 2016). Even though the Qb+ itself has not been studied in this context, converging evidence from other CPTs with similar and comparable characteristics may also apply to the Qb+. Most of the studies used age-adjusted T-scores to determine invalid cognitive performance, with T-scores > 65 indicating invalid cognitive performance (Harrison & Armstrong, 2020; Scimeca et al., 2021). For the present study we adopted the findings of this large empirical evidence base and applied an equivalent cut score (Q score ≥ 1.5) to the RTVar of the Qb+ to determine invalid cognitive performance.

ASTM. The Amsterdam Short Term Memory Test (ASTM) is a performance validity test (PVT) for the assessment of negative response bias and insufficient motivation in psychological examinations (Schmand & Lindeboom, 2005). It is presented as a test of short-term memory and attention. Five semantically related words are shown for 8 s. They should be read aloud and memorized. Then a simple arithmetic problem is given. After this, five words are again presented, three of which were previously shown. Those three should be identified by the participant. Thirty tasks total a maximum of 90 points. The reliability of the test is satisfactory. Internal consistency in different samples was around 90%. In a sample of mixed neurological patients, test–retest correlation was .85 within an interval of 1 to 3 days. The cutoff value for the ASTM is ≤ 84 points (indicating fail). Sensitivity to feigned cognitive dysfunction was 91% (in experimental simulators) and specificity to genuine cognitive dysfunction was 89% (in neurological patients). Healthy controls from age
9 years onwards master this test almost perfectly. Patients with neurological disorders like concussion, brain tumors, multiple sclerosis, or difficult-to-treat epilepsy rarely have difficulties in handling this test, provided that they do not have serious cognitive deficits.

**Statistical analyses**

In addition to SVT (CII) and PVT (ASTM) failure rates as computed in the original sample, we defined a group failing two PVTs (ASTM and RTVar of the Qb+) and groups failing both a SVT (CII) and one or both PVTs (ASTM and/or RTVar of the Qb+). First, we presented SVT/PVT failure rates per group, as well as associations between SVT and PVT measures (Spearman correlation coefficients of raw scores as well as Cohen’s Kappa for classification rates).

We try to classify those failing two PVTs (ASTM-RTVar) versus those with normal scores in both measures or one abnormal score in the AST or in the Qb+ RTVar. We further try to correctly classify those failing both a SVT and a PVT (ASTM-CII) versus those with normal scores in both measures or one abnormal score in the ASTM or in the CII. In preparation of estimating prediction models, metric variables were z-transformed using a robust z score transformation based on the median absolute deviation (Iglewicz & Hoaglin, 1993) or the Q based norm scores of the Qb+ test were entered in the analyses as these were also scaled in this specific range. We also used gender and age as predictor variables for our model.

We applied several machine learning strategies as there is no strategy that fits all analyses (Lantz, 2015). k nearest neighbor (k-NN) is a supervised learning algorithm classifying unlabeled data points by assigning them the class of the most similar labeled data points. A data point is labeled based upon its similarity with respect to its neighbors (Bali & Sarkar, 2016). The parameter k implies the number of nearest neighbors used. The square root of the number of training examples can be used as an empirically derived rule (Lantz, 2015).

We performed holdout cross-validation, 10-fold cross-validation with 50 repetitions, leave-one-out cross-validation, and nested cross-validation with hyperparameter tuning in order to try to compensate for under- or overfitting (Rhys, 2020).

Naïve Bayes Classifier is based on the Bayes theorem and predicts the class membership probability of observations assigning observations to the class with the largest probability score. We used 10-fold cross-validation repeated 50 times to evaluate the performance of our model-building procedure.

Support Vector Machine (SVM) identifies the optimal decision boundary that separates data points from different classes and predicts the class of new data based on this observation boundary. Linear and radial kernel transformations were used. We performed nested cross-validation with hyperparameter tuning using polynomial, radial, sigmoid, and linear kernels as possible candidates.

Linear discriminant analysis results in a discriminant function which is a linear equation of the input variables intended for group separation. Quadratic discriminant analysis can be applied in situations in which classes are best separated by a nonlinear decision boundary. Flexible discriminant analysis (FDA) is an extension of discriminant analysis. FDA is able to model multivariate non-normality and non-linear relationships. The models were cross-validated using 10-fold cross-validation repeated 50 times.

The random forest algorithm applies bagging to decision trees. At each splitting step of the decision tree algorithm, a random sample of predictors is chosen. The results are averaged over the number of trees chosen for the analyses. We applied the boosting algorithm XGBoost and performed nested cross-validation including hyperparameter tuning.

We performed several cross-validation methods in order to try to compensate for under- or overfitting. In holdout cross-validation a stratified random proportion of the data set is used as a test set, and the model is trained on the remaining data. We used two-thirds of the data for training and one-third of the data for testing. In 10-fold cross-validation with 50 repetitions as used in our study the data is split in equal-sized chunks which are called folds. One of the folds is used as the test set and the other folds are used as the training set. This is continued until all folds have been used as the test set. The data is shuffled 50 times around and the process is repeated as described. The results are then averaged. In leave-one-out cross-validation a single observation is regarded as a test case and the model is trained on the rest of the data.

Then the model predicts the class membership of the test case, and this is compared with its actual status. This is done until every observation has been used once as a test case. Machine learning models have associated parameters. Some of them are not estimated from the data but control how a model makes predictions. These are called hyperparameters and can be optimized by a process called hyperparameter tuning. In nested cross-validation hyperparameter tuning is included in the process of cross-validation. In an inner loop, different values of hyperparameters are cross-validated. The best parameters are then used in the outer cross-validation loop for each fold of data as one of the above described cross-validation methods (holdout, k-fold, leave-one-out) are used here. Nested cross-validation approaches are known to produce robust and unbiased performance estimates regardless of sample size (Vabala et al., 2019).
We primarily used the Matthews correlation coefficient (MCC), a special case of the phi coefficient, to evaluate model performance as this method is more reliable and produces a high score only if the prediction obtained good results in all four categories of the confusion matrix proportionally to the size of the positive and of the negative elements (Chicco & Jurman, 2020; Montesinos López, 2022). It is further unaffected by unbalanced datasets, and it is a contingency matrix method of calculating the Pearson product-moment correlation coefficient which can range between −1 and +1. It is superior to other performance measures like accuracy, F1 score, balanced accuracy, kappa coefficient, area under the curve (AUC), bookmaker informedness, and markedness (Chicco & Jurman, 2020; Chicco et al., 2021). The Matthews correlation coefficient was further used to select the most appropriate model after using different cross-validation methods within the same machine learning strategy. We further mention values of accuracy, and of the kappa coefficient (Lantz, 2015). Accuracy is defined as the success rate and is calculated as the number of true positives and true negatives divided by the total number of predictions. The kappa coefficient adjusts accuracy by accounting for the possibility of a correct prediction by chance. A coefficient of less than .20 is regarded as poor agreement, .20 to .40 represents fair agreement, .41 to .60 moderate agreement, .61 to .80 good agreement, and .81 to 1.00 very good agreement (Lantz, 2015). We also discuss the percentages of correctly classified individuals who passed the ASTM and/or the Qb + RTVar and those who passed the ASTM and/or the CII and the percentages of correctly classified patients with a diagnosis of ADHD who failed the ASTM and the Qb + RTVar and those who failed the ASTM and the CII as these values are of direct clinical importance.

For classification purposes multicollinearity is not regarded as a problematic feature (Lantz, 2015). This was apparent as reduced less correlated predictor sets resulted in almost identical classification results in our samples.

It can be expected that the suspicious sub-samples are significantly underrepresented compared to the respective sample classified as normal or failing just one SVT/PVT. This imbalance would in turn lead to an unrealistic correct classification of the majority samples in machine learning. Therefore, the methods of over- and undersampling must be applied here. Both the minority classes were oversampled with replacement and the majority classes were undersampled without replacement (Lunardon et al., 2014).

In predicting class membership in the ASTM-RTVar models (i.e., failing 2 PVTs) we excluded Qb+ variables as predictors and in predicting class membership in the ASTM-CII models (i.e., failing PVT and SVT) we excluded CAARS variables as predictors.

All analyses were performed with R 4.0.2, RStudio Version 1.2.5042, and R packages mlr, tidyrverse, ROSE, caret, caretEnsemble, and xgboost were used.

Results

Table 1 presents base rates of noncredible symptom reporting and performance of the original and current sample of individuals with ADHD, according to the various criteria of SVT and/or PVT failure. In our earlier sample of individuals with ADHD, n = 63 (32.1%) failed the ASTM and n = 70 (35.7%) scored in the noncredible range of the CII as a measure of SVT. While the base rate of noncredible responses on the CII was roughly confirmed in the current sample (36.6%), the proportion of individuals with ADHD failing the ASTM was much larger in the current sample (i.e., 49.3%) than in the original sample. As expected, base-rates were lower when stricter criteria for noncredible performance were applied, i.e., 18.9% (original sample) and 27.3% (current sample) when two PVTs were failed, 13.3% (original sample) and 22.4% (current sample) for individuals failing the ASTM and CII, and 9.7% (original sample) and 13.7% (current sample) for individuals failing the ASTM, RTVar and CII. The measures of SVT (CII) and PVT (ASTM) showed poor association, which is reflected both in correlation analyses (original sample: \( r = -0.068, p = .345 \); current sample: \( r = -0.156, p < .001 \)) and classification correspondence statistics (original sample: Kappa = 0.080, p = 0.264; current sample: Kappa = -0.054, p < .001).

As the classification results with this class imbalance differed as expected from upsampling techniques, we report the results of the combination of over- and undersampling performed by the ROSE library in R (Lunardon et al., 2014). After over- and undersampling in both combined criteria in our earlier sample n = 96 (49%) were classified as having results in the normal range or failing one SVT or one PVT while n = 100 (51%) were classified as having suspicious

<table>
<thead>
<tr>
<th>Model</th>
<th>Original sample (n = 196)</th>
<th>Current sample (n = 700)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASTM (1 PVT)</td>
<td>32.1%</td>
<td>49.3%</td>
</tr>
<tr>
<td>CII (1 SVT)</td>
<td>35.7%</td>
<td>36.6%</td>
</tr>
<tr>
<td>ASTM-RTVar (2 PVTs)</td>
<td>18.9%</td>
<td>27.3%</td>
</tr>
<tr>
<td>ASTM-CII (PVT + SVT)</td>
<td>13.3%</td>
<td>22.4%</td>
</tr>
<tr>
<td>ASTM-RTVar-CII (2 PVTs + SVT)</td>
<td>9.7%</td>
<td>13.7%</td>
</tr>
</tbody>
</table>

Note: ASTM = Amsterdam Short Term Memory Test; CII = CAARS Infrequency Index; RTVar = RT Variation of Qb+
results in both SVT and PVT. In our current sample after over- and undersampling in both combined criteria \( n = 360 \) (51.4\%) were classified as having results in the normal range or failing one SVT or one PVT while \( n = 340 \) (48.6\%) were classified as having suspicious results in both SVT and PVT. These are necessary steps to avoid overcorrect classification of imbalanced majority classes in machine learning.

With regard to the ASTM-RTVar and the ASTM-CII prediction models, reduced models of less correlated variables revealed that in any of the analyses no subset of predictors provided a model with significantly better model evaluation parameters. For better comparisons we therefore report results of the full model.

Using hyperparameter tuning in kNN, Support Vector Machine (SVM) and Random forest resulted in overfitting which is a known issue in machine learning (Domingos, 2012; Hosseini et al., 2020). We applied the tuned models on the historic dataset to the current dataset which represents an independent validation and vice versa. Bad model fits resulted revealing that noisy data was overfitted. We therefore present results of the most realistic models from different machine learning algorithms. Detailed analyses of the remaining models can be found in the Supplementary document. Table 2 displays the results of several discriminant analyses methods regarding the ASTM-RTVar-models.

In our earlier sample, quadratic discriminant analysis resulted in the best classification result regarding the ASTM-RTVar model with a MCC value of .545, a kappa coefficient of .545, and an overall accuracy of .773. In this analysis, 77% of those with a likely diagnosis of ADHD and no abnormal scores in the ASTM and the Qb+ RTVar or one abnormal score in the ASTM or in the Qb+ RTVar were correctly classified, and 78% with a likely diagnosis of ADHD and abnormal scores in the ASTM and in the Qb+ RTVar. For our current sample, also quadratic discriminant analysis resulted in the best classification result regarding the ASTM-RTVar model. A MCC value of .275, a kappa coefficient of also .275, and an overall accuracy of .638 resulted, meaning that overall 63.8% of patients were correctly classified. Of those, 64% were correctly classified as having a likely diagnosis of ADHD passing the ASTM and the CII or having one abnormal score in the ASTM or in the CII, and 64% were correctly classified who had a likely diagnosis of ADHD failing the ASTM and the CII.

In our earlier sample, quadratic discriminant analysis resulted in the best classification result regarding the ASTM-CII model with an MCC value of .570, a kappa coefficient of .564, and an overall accuracy of .783 (Table 3). In this analysis, 71% of those with a likely diagnosis of ADHD and no abnormal scores in the ASTM and the

<table>
<thead>
<tr>
<th>Table 2. Results of different discriminant analyses methods of our ASTM-RTVar-models with 10-fold cross-validations using MCC as a model evaluation parameter.</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear discriminant analysis</td>
</tr>
<tr>
<td>current sample</td>
</tr>
<tr>
<td>.201</td>
</tr>
<tr>
<td>quadratic discriminant analysis</td>
</tr>
<tr>
<td>current sample</td>
</tr>
<tr>
<td>.275</td>
</tr>
<tr>
<td>flexible discriminant analysis</td>
</tr>
<tr>
<td>current sample</td>
</tr>
<tr>
<td>.221</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. Results of different discriminant analyses methods of our ASTM-CII-models with 10-fold cross-validations using MCC as a model evaluation parameter.</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear discriminant analysis</td>
</tr>
<tr>
<td>current sample</td>
</tr>
<tr>
<td>.259</td>
</tr>
<tr>
<td>quadratic discriminant analysis</td>
</tr>
<tr>
<td>current sample</td>
</tr>
<tr>
<td>.329</td>
</tr>
<tr>
<td>flexible discriminant analysis</td>
</tr>
<tr>
<td>current sample</td>
</tr>
<tr>
<td>.296</td>
</tr>
</tbody>
</table>

CII or one abnormal score in the ASTM or in the CII were correctly classified, and 86% with a likely diagnosis of ADHD and abnormal scores in the ASTM and in the CII. In our current sample, after excluding the overfitting hyperparameter tuning in kNN, holdout cross-validation in kNN, holdout cross-validation resulted in the highest relative MCC value of .345, a kappa coefficient of .343, and an overall accuracy of .671. In this analysis, 62% were correctly classified as likely having a diagnosis of ADHD and showing no abnormal scores in the ASTM and in the CII or one abnormal score in the ASTM or in the CII, while 72% were correctly classified who had a likely diagnosis of ADHD and who showed abnormal scores in the ASTM and in the CII.

**Discussion**

The present study confirms high base rates of symptom and performance validity test failure in the clinical evaluation of adult ADHD. The PVT failure rate (as assessed with the ASTM) of the original study (32.1\%) was exceeded in the current sample with 49.3\% of individuals with ADHD failing the ASTM. The higher PVT failure rate in the current compared to the sample from Hirsch and Christiansen (Hirsch & Christiansen, 2018) could reflect a sharp increase in exaggerated or fabricated cognitive deficits across the years in this referral context (Sibley, 2021; Sibley et al., 2018). Alternatively, even though numbers are still largely in the range of previous estimations of single PVT failure (Harrison &
Edwards, 2010; P. Marshall et al., 2010; J. Suhr et al., 2008; Sullivan et al., 2007), various other reasons can be considered for the difference in ASTM failure rate, such as differences in sample characteristics (e.g., slight differences in age and gender), participant selection and recruitment (possible transition in referral over the years), or assessment procedure (e.g., different structured clinical interview). Unstable ASTM failure rates and sensitivity to changes in design and procedure may question its reliability in the assessment of performance validity in the clinical evaluation of adult ADHD, and may give reasons to favor performance validity tests that rely on test principles that resemble more closely core features of ADHD (e.g., symptoms of inattention; Fuermaier et al., 2016). With regard to the assessment of the credibility of symptom reports, as determined by the CII derived from the Conners’ ADHD scales, stable rates of noncredible symptom reporting were observed between both samples, with 35.7% in the earlier and 36.6% in the current sample of patients with ADHD. When applying current standards of at least two SVT/PVT failures indicating noncredible symptom report or performance, our base rates ranging from 9.7% to 27.3% fell largely in the range of earlier studies in this field of up to 20% invalid test performance as suggested by empirical research (Abramson et al., 2021; Phillips et al., 2022; P. S. Marshall et al., 2016) and estimated by clinically working neuropsychologists (Martin & Schroeder, 2020).

The high rates of symptom and performance validity test failures confirm recent consensus studies drawing the attention to substantial percentages of examinees in the clinical evaluation of adult ADHD who feign, exaggerate, or fabricate cognitive deficits during neuropsychological evaluation (Sherman et al., 2020). Of note, reasons for invalid data are manifold and do not necessarily represent an intentional attempt to feign cognitive deficits. For example, one must also consider that invalid symptom report and test performance may be indications of help-seeking behavior (symptom exaggeration as a cry for help), low intelligence, severe cognitive impairments, or motivational issues. It is widely recommended by position reports and consensus papers that SVT/PVT assessment should be an integral part of any clinical assessment (Bush et al., 2005; Sweet et al., 2021), yet, those failing SVT/PVT assessment represent a heterogeneous group that need further attention in subsequent assessment in order to differentiate between the various underlying reasons for SVT/PVT failure.

In further analyses, we computed several prediction models using ML algorithms on the earlier as well as current sample, in order to determine possible characteristics of those failing SVT/PVT assessments. Generally, prediction models based on clinical and neuropsychological variables were moderate to poor and did not identify strong candidates that repeatedly stand out in the various models to reliably predict PVT failure in either of the samples (Hirsch & Christiansen, 2018) after up- and downsampling due to class imbalance, regardless of the applied machine learning algorithm. Without using this technique, it was obvious in our both samples that a high misbalance occurred in favor of predicting the most prominent class. Therefore, it was justified to correct for class imbalance. The algorithms with hyperparameter tuning (kNN, SVM, random forest) each showed the overfitting known in machine learning, which is also referred to as over-hyping (Hosseini et al., 2020). The application of the models to the respective other data set (historical versus current and vice versa) resulted in very poor model fits in each case, although good to very good parameters were achieved on the training data. Initially, the technique of cross-validation and in particular nested cross-validation was seen as a solution against overfitting (Hosseini et al., 2020), although we could not confirm this. However, it is well known that cross-validation is not a comprehensive solution against overfitting (Domingos, 2012). Models with higher complexity, i.e., with more variables, also called flexible models, perform better in terms of accuracy (Montesinos López, 2022). Our models are considered flexible models, although they do not perform well. Recursive feature elimination (REF) is a procedure that also does not prevent overfitting (Hosseini et al., 2020). There is no unique formal scientific procedure available in the literature regarding hyperparameter tuning. Our procedure for checking hyperparameter tuning on an independent dataset is also recommended (Hosseini et al., 2020). We consider the poor model fits as confirmation that there is no overarching pattern in adult ADHD that could predict failing in symptom and performance validity testing. The partially good model fits after hyperparameter tuning could not be confirmed on the respective independent datasets, so that we assume that specific patterns for the prediction of failing in symptom and performance validity tests in adult ADHD do not exist. Because no clinical or neuropsychological characteristics can be identified that reliably predict SVT/PVT failure (including the CII as a measure of SVT), we conclude that further research using different methodology is needed to study the group of SVT/
PVT failure in more detail. Interviews and the consideration of collateral information may be necessary in order to uncover underlying reasons for SVT/PVT failure, that may be of various nature. Further, we conclude that the use of SVT/PVTs in the clinical assessment is pivotal as invalid cognitive performance seemingly cannot be predicted satisfactorily by routine clinical measures. The clinician’s confidence on the occurrence of invalid test performance is increased if multiple, preferably minimally related, measures are applied (Sherman et al., 2020). It is also advised to make use of measures of different and diverse test principles, including SVTs and PVTs, based on embedded and stand-alone testing (Erdodi et al., 2014, 2018), sample validity continuously throughout the assessment and across cognitive domains (Soble et. al., 2020; Rhoads et al., 2021; Sweet et al., 2021), and by using measures that have been proven valid in the context of the clinical evaluation of adult ADHD (Wallace et al., 2019).

Moreover, negligible to weak associations were found between SVT (CII) and PVT (ASTM) measures, with only slight agreement in their categorization. Similar to the extensive body of neuropsychological research showing the discrepancy between self-reports and performance tests of cognitive functioning (Fuermaier et al., 2015; Koerts et al., 2011; Toplak et al., 2013), the findings of the present study support the view that SVT and PVT measures are of different nature indicating different types of noncredible behavior and may not be interpreted interchangeably (Hirsch & Christiansen, 2018). This finding further supports the notion of Sherman et al. (Sherman et al., 2020) stating that a proper assessment of symptom and performance validity should include several approaches that are minimally related to each other.

**Limitations**

This study has to be interpreted in the context of several limitations. First, the comparability of the earlier and current sample may be limited due to unspecific differences in recruitment or selection, sample characteristics, and the applied assessment battery, which complicates the interpretations of findings. Second, comparisons of prediction models between the earlier and current samples are complicated due to class imbalance which was controlled for by up- and downsampling. Third, due to the difference of the structured clinical interview in the assessment battery (WRI vs DIVA) between the earlier and current sample as well as the neuropsychological tasks (TAP + QbTest vs QbTest alone), the prediction models could not make use of all data available but included only those measures that were available in both samples. This led to the exclusion of scores on the test battery of attention performance (TAP) which was administered in the earlier sample only and where significant differences between individuals passing and failing PVT assessment occurred on test variables of divided attention, sustained attention, and inhibitory control (Hirsch & Christiansen, 2018). It could be speculated that the validity of prediction models in the current sample was improved if subtests of the TAP had been included. Finally, and fourth, it must be stressed that there is no consensus yet on the specific measures used to assess noncredible symptom reporting and performance. Even though the use of multiple PVTs, i.e., failure on ≥ 2 independent PVTs, or a multidimensional definition of noncredible symptom report or performance, became standard in practice and research, results may differ depending on the battery used in the respective study.

**Conclusions**

Based on two samples of individuals with ADHD of large size, this study confirms high base rates of noncredible symptom report and performance as defined by one SVT (35.7% − 36.6%), one PVT (32.1% − 49.3%), two PVTs (18.9% − 27.3%), each one SVT and PVT (13.3% − 22.4%), and one SVT and two PVTs (9.7% − 13.7%). The results caution clinicians to uncritically assume validity of clinical data without adequate assessment. The occurrence of noncredible symptom report or test performance suggests that all information derived from self-reports or performance tests in the given clinical assessment may be invalid. Failing to recognize invalid reports or performance may bias, distort, or even invalidate clinical diagnosis, recommendation, and treatment planning, and may result in superfluous and potentially harmful treatment for individuals with an unwarranted diagnosis of ADHD, while at the same time scarce medical resources are occupied and thus not accessible for those who genuinely struggle with ADHD. Treatment programs may be filled with individuals with low treatment motivation and cooperation, which may make the intervention appear ineffective and fuels public’s concerns on the effectiveness of psychological or pharmacological treatments. Position papers and consensus reports uniformly recommend the use of measures of symptom and performance validity in any clinical assessment (Bush et al., 2005; Sweet et al., 2021), including the clinical assessment of adult ADHD (Musso & Gouvier, 2014). Because no clinical or
neuropsychological characteristics served to reliably predict SVT/PVT failure, we conclude that the assessment of the underlying reasons for invalid data is more extensive, and includes further assessment with the individual as well as collateral information (Sherman et al., 2020).

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

The author(s) reported there is no funding associated with the work featured in this article.

**ORCID**

Anselm B.M. Fuermaier [http://orcid.org/0000-0002-2331-0840](http://orcid.org/0000-0002-2331-0840)

Oliver Tucha [http://orcid.org/0000-0001-8427-5279](http://orcid.org/0000-0001-8427-5279)

Björn Albrecht [http://orcid.org/0000-0001-7936-2044](http://orcid.org/0000-0001-7936-2044)

**References**


Edebol, H., Helldin, L., & Norlander, T. (2012). Objective measures of behavior manifestations in adult ADHD and differentiation from participants with bipolar II disorder, borderline personality disorder, participants with disconfirmed ADHD as well as normative participants. *Clinical Practice and Epidemiology in Mental Health, 8*(1), 134–143. [https://doi.org/10.2174/1745017901208010134](https://doi.org/10.2174/1745017901208010134)


Schmand, B., & Lindeboom, J. (2005). Amsterdamer Kurzzeitgedächtnistest (AKGT) [Amsterdam Short Term Memory Test]. PITS B. V.


Ulberstad, F. (2012). *QbTest technical manual*. Qbtech AB.

