ARTICLE   OPEN

Check for updates

FASDetect as a machine learning-based screening app for FASD in youth with ADHD

Lukas Ehrig[1,2,7], Ann-Christin Wagner[2,7], Heike Wolter[2], Christoph U. Correll[2,3,4], Olga Geisel[2,8] and Stefan Konigorski [1,2,5,6,8 ✉]

Fetal alcohol-spectrum disorder (FASD) is underdiagnosed and often misdiagnosed as attention-deficit/hyperactivity disorder (ADHD). Here, we develop a screening tool for FASD in youth with ADHD symptoms. To develop the prediction model, medical record data from a German University outpatient unit are assessed including 275 patients aged 0–19 years old with FASD with or without ADHD and 170 patients with ADHD without FASD aged 0–19 years old. We train 6 machine learning models based on 13 selected variables and evaluate their performance. Random forest models yield the best prediction models with a cross-validated AUC of 0.92 (95% confidence interval [0.84, 0.99]). Follow-up analyses indicate that a random forest model with 6 variables – body length and head circumference at birth, IQ, socially intrusive behaviour, poor memory and sleep disturbance – yields equivalent predictive accuracy. We implement the prediction model in a web-based app called FASDetect – a user-friendly, clinically scalable FASD risk calculator that is freely available at https://fasdetect.dhc-lab.hpi.de.

INTRODUCTION

Fetal alcohol-spectrum disorder (FASD) is an umbrella term for medical conditions caused by prenatal alcohol exposure, including fetal alcohol syndrome (FAS), partial fetal alcohol syndrome (pFAS), alcohol related birth defects (ARBD), and alcohol-related neurodevelopmental disorder (ARND). The global prevalence of FASD is estimated to be between 2–5% of the Western world's population[1]. Despite the prevalence rate, FASD is highly under-diagnosed and many patients miss out on the beneficial effects of an early childhood diagnosis and subsequent early intervention[2–5].

Established diagnostic systems for FASD are based on the manifestation of growth deficiencies, craniofacial dysmorphia, central nervous system damage/dysfunction, and gestational alcohol exposure[6,7]. These neuropsychological impairments can manifest as deficits in intelligence, learning, memory, executive function and academic achievements, language and motor development and attention[8]. People with FASD have a higher risk to develop secondary psychiatric conditions, like conduct disorder, attention-deficit/hyperactivity disorder (ADHD) and sleep disorders, as well as to experience adverse life events[8–11]. Hyperactivity, inattention and impulsivity are characteristically seen both in patients with ADHD and FASD. More than half of FASD patients suffer from comorbid ADHD[11]. These overlapping symptoms of FASD and ADHD complicate the diagnostic process and can lead to misdiagnosis as well as delayed intervention for FASD. In a study conducted in 547 children and adolescent who were adopted or in foster care and who underwent a compre-hensive multidisciplinary diagnostic evaluation to identify FASD, 156 youth met criteria for FASD, but in as many as 80% the FASD diagnosis had been missed and 6% were misdiagnosed within the FASD spectrum. The mental health diagnosis most commonly given to those children upon referral was ADHD[12]. The very high proportion of missed FASD diagnosis and youth receiving a misdiagnosis underscore the importance of evaluating youth diagnosed with ADHD in order to detect any missed FASD diagnosis.

The purpose of the present study is to (i) develop a machine learning algorithm for detection of FASD in patients with ADHD symptoms based on retrospectively gathered out-patient data, and (ii) subsequently use this algorithm to create an easy and fast as well as clinically scalable online screening tool. Based on the analysis of medical record data from a German University outpatient department including 275 patients with FASD with or without ADHD and 170 patients with ADHD without FASD, we identify a random forest model based on 6 variables – body length and head circumference at birth, IQ, socially intrusive behaviour, poor memory and sleep disturbance –that yields sufficient accuracy to differentiate youth with versus without FASD. We implement this algorithm in a screening tool called FASDetect which is easy to use and yields a quick screening result.

RESULTS

Study sample

This study was conducted at the outpatient unit of the department of child and adolescent psychiatry at the Campus Charité Virchow of the Charité Universitätsmedizin Berlin, Germany. The sample for the analysis was selected to allow a comparison of patients with a diagnosis of ADHD with patients with a diagnosis of FASD. More specifically, a group of consecutively assessed patients with a clinical diagnosis of ADHD without FASD and a group of patients with an expert diagnosis of FASD (with or without comorbid ADHD) was compared. Alto-gether, 694 patients with ADHD symptoms were identified consecutively from the general patient pool being potentially

¹Digital Health Center, Hasso Plattner Institute for Digital Engineering, University of Potsdam, Potsdam, Germany. ²Department of Child and Adolescent Psychiatry, Charité Universitätsmedizin Berlin, Berlin, Germany. ³The Zucker Hillside Hospital, Department of Psychiatry, Northwell Health, Glen Oaks, NY, USA. ⁴Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Department of Psychiatry and Molecular Medicine, Hempstead, NY, USA. ⁵Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁶Department of Statistics, Harvard University, Cambridge, MA, USA. ⁷These authors contributed equally: Lukas Ehrig, Ann-Christin Wagner. ⁸These authors jointly supervised this work: Olga Geisel, Stefan Konigorski. ✉email: stefan.konigorski@hpi.de
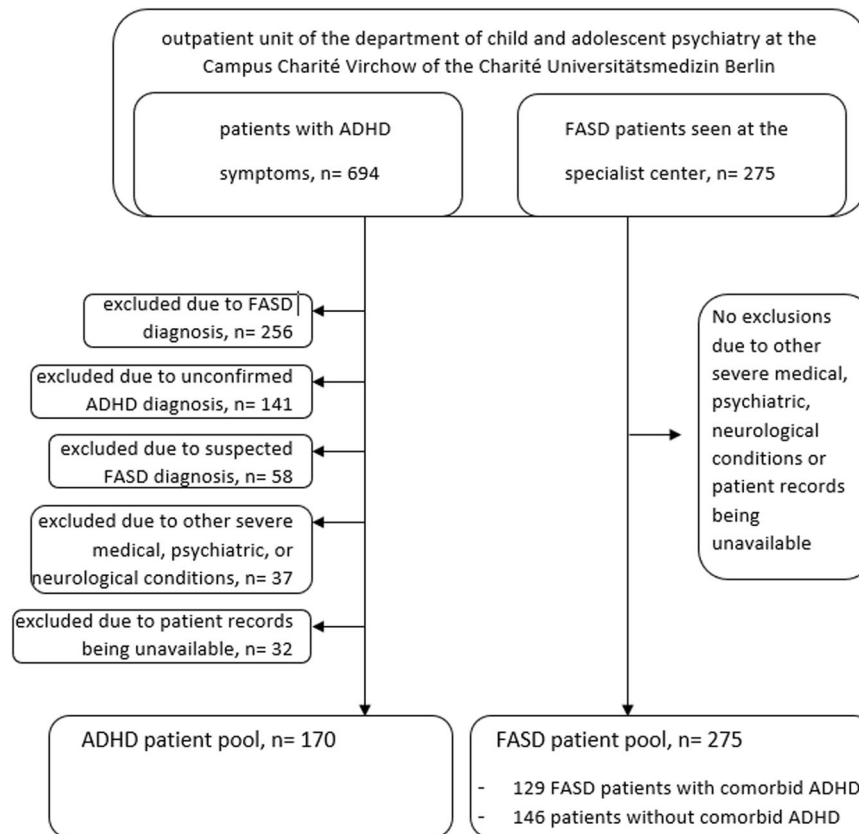
Fig. 1 Flow chart of the ADHD and FASD patient groups included in the study.

eligible for the study. 256 of the 694 ADHD patients had a confirmed FASD diagnosis and therefore were excluded from the ADHD pool. Further, 141 patients were excluded from the ADHD group due to an unconfirmed ADHD diagnosis; 58 because they had a suspected but not confirmed FASD diagnosis; 37 due to other severe medical, psychiatric, or neurological conditions; and 32 patients were excluded because patient records were unavailable. This yielded in total 170 patients in the ADHD group. The consecutively enrolled FASD group was recruited from the specialist center and consisted of 275 youth, including 129 FASD patients with comorbid ADHD and 146 patients without comorbid ADHD diagnosis. These 275 patients included most of the 256 FASD patients from the general patient pool. See also Fig. 1 for an illustration of the two study groups.

Description of the patients' characteristics

Tables 1, 2 give an overview of the main characteristics of the $n = 445$ FASD and ADHD patients, which included 159 female (mean age at initial presentation, 9.6 years [range, 0.2–18.8 years]) and 286 male (mean age at initial presentation, 8.9 years [range, 0.1–19.0 years]) patients. 139 of the FASD patients had a FAS diagnosis, 127 had a pFAS diagnosis and 9 patients were diagnosed with ARND. 170 patients belonged to the ADHD group (31 female; mean age at initial presentation, 8.7 years [range, 3.7–16.8 years]; 139 male; mean age at initial presentation, 8.4 years [range, 2.3–15.7 years]) and 275 patients belonged to the FASD group (128 female; mean age at initial presentation, 9.9 years [range, 0.2–18.8 years]; 147 male; mean age at initial presentation, 9.4 years [range, 0.1–19.0 years]). There were very low pairwise correlations between these variables, with the exception of head circumference and birth length (Pearson correlation coefficient of 0.57).

Table 1. Patient characteristics.

Variable	All patients		FASD patients		ADHD patients	
	Mean	SD	Mean	SD	Mean	SD
Age at first presentation	9.15	4.03	9.60	4.53	8.43	2.94
Gestational age	38.57	2.39	38.58	2.21	38.56	2.68
z-score of birth length	−0.58	1.16	−0.87	1.13	−0.08	1.03
z-score of birth weight	−0.48	1.19	−0.80	1.09	0.05	1.17
z-score of head circumference at birth	−0.45	1.11	−0.70	1.11	0.07	0.93
z-score of length at first presentation	−0.27	1.17	−0.50	1.12	0.27	1.12
z-score of weight at first presentation	−0.02	1.21	−1.20	1.15	0.38	1.24
z-score of head circumference at first presentation	−1.06	1.55	−1.37	1.37	0.23	1.58
Number of mothers pregnancies	2.78	2.14	2.84	2.12	2.53	2.22
Number of mothers births	2.09	1.39	2.21	1.48	1.91	1.23

Mean and standard deviation (SD) of number of mothers pregnancies and births, gestational age, age at first presentation, length, weight, head circumference at time of birth and first presentation for all patients, FASD and ADHD patient groups.

Prediction models to separate FASD and ADHD

The statistical analysis aimed at developing and evaluating a prediction model that would be able to separate FASD from ADHD cases with sufficient accuracy. After data preprocessing and variable selection (see Materials & Methods), we tested the

Table 2. Patient characteristics.

Variable	All patients		FASD patients		ADHD patients	
	Yes	No	Yes	No	Yes	No
Male sex	286 (64.3%)	159 (35.7%)	147 (53.5%)	128 (46.5%)	139 (81.8%)	31 (18.2%)
Intelligence below 85 IQ points	119 (34.4%)	227 (65.6%)	92 (51.4%)	87 (48.6%)	27 (16.2%)	140 (83.8%)
Socially intrusive behaviour	196 (47.6%)	216 (52.4%)	155 (63.5%)	89 (36.5%)	41 (24.4%)	127 (75.6%)
Impairment in memory	272 (66.2%)	139 (33.8%)	208 (86.0%)	34 (14.0%)	64 (37.6%)	105 (62.1%)
Sleep disorder/disturbance	147 (37.1%)	249 (62.9%)	107 (47.3%)	119 (52.7%)	40 (23.5%)	130 (76.5%)
Speech development disorder	221 (54.2%)	187 (45.8%)	145 (60.7%)	94 (39.3%)	76 (45.0%)	93 (55.0%)
Psychiatric comorbidities	289 (72.3%)	111 (27.8%)	185 (80.4%)	45 (19.6%)	104 (61.2%)	66 (38.9%)
Prescribed psycho-tropic medications	186 (47.7%)	204 (52.3%)	64 (29.1%)	156 (70.9%)	122 (71.8%)	48 (28.2%)

Absolute and relative frequencies for sex and cognitive, behavioral and sleep variables, speech development, psychiatric comorbidities and medication in all patients, FASD and ADHD patient groups.

Table 3. Cross-validated evaluation results of prediction with 13 variables.

Model (13 variables)	AUC	Accuracy	Precision	Recall	Brier
Logistic Regression	0.91 [0.83, 0.99]	0.84	0.89	0.84	0.12
Support Vector Machine	0.90 [0.80 0.99]	0.85	0.85	0.92	0.12
Random Forest	**0.92 [0.84, 0.99]**	**0.85**	**0.86**	**0.91**	**0.11**
Gradient Boosting Decision Tree	0.91 [0.82, 0.99]	0.85	0.86	0.91	0.12
kNN Classifier	0.90 [0.81, 0.99]	0.84	0.87	0.88	0.12
Gaussian Process Classifier	0.90 [0.81, 0.99]	0.84	0.86	0.89	0.12

The best model is highlighted in bold.

Table 4. Cross-validated evaluation results of the pipeline with 6 variables.

Model (6 variables)	AUC	Accuracy	Precision	Recall	Brier
Logistic Regression	0.90 [0.81, 0.99]	0.82	0.89	0.81	0.13
Support Vector Machine	0.90 [0.81, 0.99]	0.84	0.85	0.91	0.12
Random Forest	**0.93 [0.85, 1.00]**	**0.85**	**0.87**	**0.91**	**0.11**
Gradient Boosting Decision Tree	0.91 [0.83, 0.99]	0.84	0.87	0.87	0.12
kNN Classifier	0.91 [0.83, 0.99]	0.84	0.87	0.88	0.11
Gaussian Process Classifier	0.91 [0.82, 0.99]	0.83	0.86	0.88	0.12

The best model is highlighted in bold.

performance of 6 machine learning algorithms to predict ADHD or FASD using the 13 remaining variables on our data with nested cross-validation. Table 3 provides an overview of the main results for the prediction model based on the 13 variables number of mother's births, gestational age, z-scores of length, weight and head circumference at birth, z-scores of length and weight at initial presentation, as well as the presence of low IQ, socially intrusive behavior, speech development disorder, poor memory, sleep disturbance and psychiatric comorbidities. When predicting FASD cases among ADHD patients, an AUC of 0.92 (95% confidence interval CI [0.84, 0.99]) was reached by the RF model. 91% of the FASD patients were correctly identified and overall 85% of patients received a correct classification. Of all patients that were classified as FASD cases, 86% were true FASD cases. The kNN and Gaussian Process classifiers both reached an AUC of 0.90 and accuracy of 0.84. The SVM also had a ROC AUC of 0.90 ([0.80, 0.99]), but recognized more positive cases with a sensitivity of 0.92, the highest among all evaluated algorithms. Logistic regression and GBDT both yielded an AUC of 0.91 (95% CI [0.83,

0.99] and 0.91 [0.82, 0.99], respectively). The highest positive predictive value (0.89) was reached by the logistic regression model, however at the cost of the lowest sensitivity (0.84). The RF had a Brier score of 0.11, the other models had a Brier score of 0.12.

In all experiments and cross-validation trials, only 6 of the 13 variables were frequently selected in the ML pipelines. These six variables were: z-scores of body length and head circumference at birth, IQ below 85 IQ points, socially intrusive behaviour, poor memory and sleep disturbance. When using this reduced variable set in our second set of analyses, the RF model had an AUC of 0.93 (95% CI [0.85, 1]) and could on average identify 91% of FASD cases in the test sets, with 85% of patients being classified correctly. Patients that were classified as FASD patients were true cases in 87%. All other algorithms separated the ADHD and the FASD groups similarly well with an AUC of 0.90 or 0.91 (see Table 4). Hence, the various performance metrics of the algorithms were very similar compared to the prediction models using 13 variables.
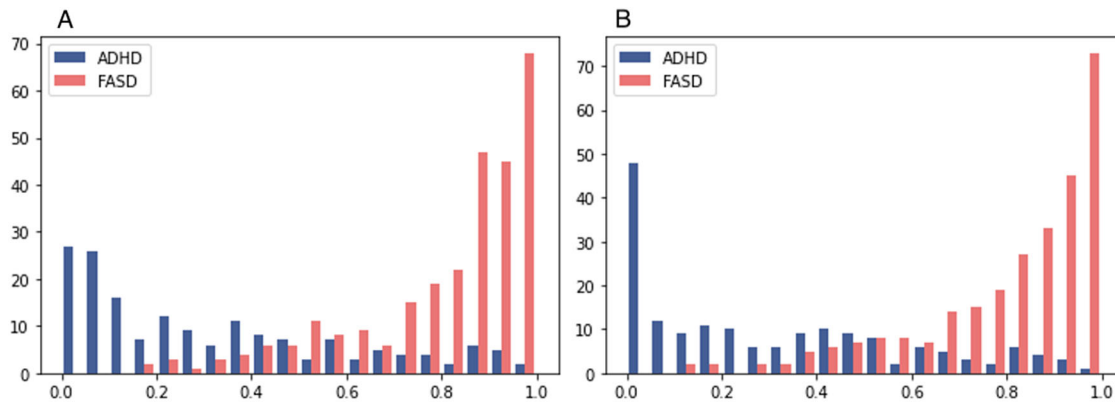
Fig. 2 Distribution of predicted probabilities for the random forest model. The x-axis shows the predicted probability of having FASD and the y-axis the number of actual ADHD/FASD cases for this probability, for the model with 13 variables (**A**, left panel) and with 6 variables (**B**, right panel).

The results of all experiments including ROC curves can be found in Supplementary Figs. 1–6.

For our screening application, we selected the RF model because of its high sensitivity, robustness to changes of the variable set, and its good overall performance. The probability score distributions of the RF model are depicted in Fig. 2 and illustrate that the estimated probabilities of having FASD are generally high for FASD patients and low for ADHD patients. There are only few patients that are assigned a low risk of FASD while having a diagnosis of FASD or that are assigned a high risk despite having an ADHD diagnosis without FASD. The figure also shows the number of true and false classifications at different probability thresholds. For any probability threshold used for the decision whether a patient is assigned to the ADHD or FASD group, ADHD patients right of that threshold (i.e., that were assigned a higher probability by the prediction model) are false positives, FASD patients left of that threshold are false negatives, and all others were classified correctly.

Implementation of machine learning model in FASDetect screening application

In the final step, we developed a screening app for the detection of FASD among ADHD cases based on the RF algorithm. Our focus was on the target user group of medical professionals from different fields (e.g., pediatricians, psychiatrists). Requirements derived for the application included that it should be user-friendly, quick and easy-to-use and that the screening result is immediately visible.

The frontend of the application was built using Vue.js/quasar, the backend using Python/flask. The resulting app consists of three screens and is based on the RF model of 6 variables that can be quickly and appropriately assessed by all possible users. The first screen contains the disclaimer and provides some information about the app. The next screen contains a questionnaire, where information about the 6 variables is obtained. The last screen shows the results and some context of how to interpret the screening results (see Fig. 3). In order to facilitate quick decision-making, the results are visually represented using a traffic light metaphor. A yellow signal is shown in FASDetect when the model estimates the FASD risk to be 50–74% and therefore classifies the patient as a potential FASD case. When the risk exceeds 75%, the red signal is shown, indicating a high risk. The FASDetect app is designed in such a way that if all the variables are known, the data entry and retrieval of the result can be completed very easily in less than 1 min. Currently, the app exists in English and German, but can easily be extended to include more languages. The app is available open-source and free-of-charge at https://fasdetect.dhc-lab.hpi.de.

DISCUSSION

In this study, we developed a screening tool, called FASDetect, based on machine learning models to detect FASD among patients with ADHD symptoms. FASDetect only requires answers to 6 questions to yield a quick screening result. Our motivation was that the diagnosis of FASD is often challenging as well as time-consuming and the most common mental health diagnosis given to FASD patients is ADHD when missing the FASD diagnosis[12]. Also, we were not aware of any tool to screen for the risk of FASD in patients with ADHD. To develop the prediction model, medical record data from a German University outpatient department were assessed including 275 patients with FASD with or without ADHD and 170 patients with ADHD without FASD. We compared different machine learning algorithms and implemented a random forest model in FASDetect, which performed best with a cross-validated AUC of 0.92 (95% confidence interval [0.84, 0.99]).

The high predictive accuracy in our study is similar to previous studies using machine learning in patients with ADHD or FASD, but all prior studies focused on different patient groups and had different objectives than our study. For example, Duda et al.[13] showed that machine learning algorithms are capable of accurately differentiating between patients with ADHD and autism-spectrum disorder with an AUC of 0.96. In another study, Zhang et al.[14] successfully used machine learning to distinguish between FASD patients and healthy controls through use of eye movement, psychometric and neuroimaging data with 85% classification accuracy. Further studies have investigated the use of machine learning models to classify and diagnose FAS[15–17]. For example, Fu et al.[15] developed a transfer learning approach utilizing a network learned on large facial recognition datasets and demonstrated its applicability in an experimental evaluation. Blanck-Lubarsch et al.[16] showed a high accuracy of 90% using decision trees, support vector machine and k-nearest neighbor models to analyze facial 3D scans to differentiate children from the severe end of the FAS spectrum among healthy controls, based on a study sample of 30 patients and 30 controls. Based on similar input data containing 3D facial scans of 149 individuals, Fang et al.[17] developed an automated classification algorithm, which diagnosed up to 91% of FAS patients correctly within their ethnicity group. With similar predictive accuracy but requiring input data that is much easier to collect in clinical practice and with less data privacy challenges, our work provides a scalable and cost-effective screening and diagnostic support tool for classifying FASD among patients with ADHD. FASDetect has the potential to optimize screening and diagnostic procedures that can help improve treatment selection and outcome predictions in clinical psychology and psychiatry[18]. Importantly, FASDetect can be
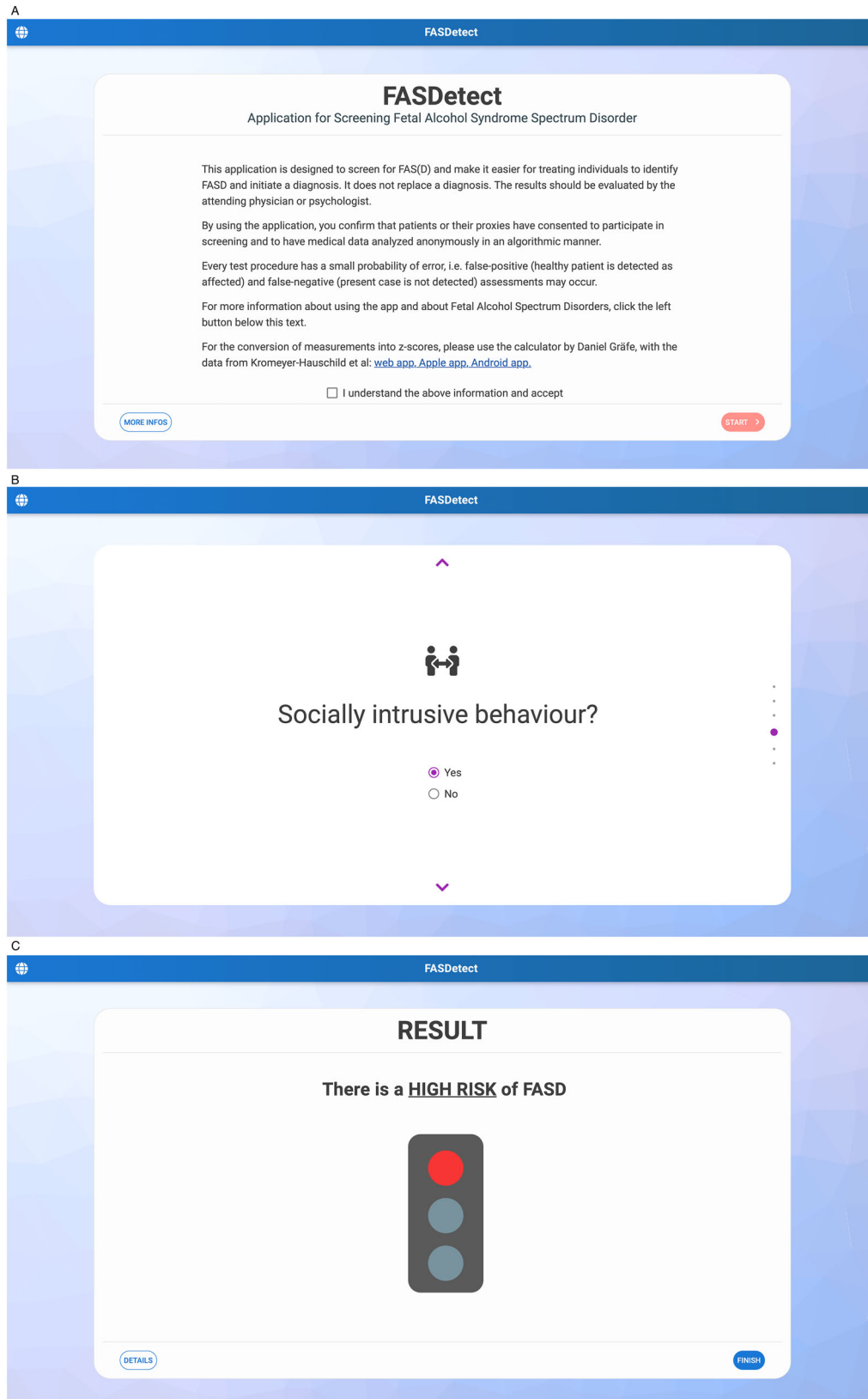
Fig. 3 Illustration of the FASDetect app. The first screen (**A**, left panel) onboards the user, the second screen (**B**, middle panel) contains a questionnaire asking for input to the 6 selected variables in the middle of which we show here one question on socially intrusive behaviour, and the third screen (**C**, right panel) shows the screening result displayed in form of a traffic light.

scaled easily to many patients worldwide, as no extra equipment is needed to utilize it and as it requires little effort. These characteristics are most helpful for a screening application and set our study apart from prior research in this area.

The 6 most important variables that were retained for efficient FASD screening via FASDetect are the z-scores of birth length and birth head circumference, low IQ, social intrusiveness, poor memory and sleep disturbance. All of these variables are known to be medically linked to FASD[19–25]. Previous studies have shown that FASD patients are more likely to be microcephalic and remain to be microcephalic and length growth-restricted throughout life. They also show a lower intelligence than ADHD patients and have been found to suffer from memory problems. Socially intrusive behaviour and sleep disturbance are also often seen in FASD patients. All of this is also shown in our data, which adds face validity to the finding that these predictors were selected during automatic feature selection. Thus, we are optimistic that our results will generalize and can be replicated in other populations.

FASDetect may represent the time-saving clinical screening application for FASD that has been missing until now. Such a tool is urgently needed in clinical practice. In next steps, FASDetect has to be evaluated prospectively and licensed for medical use. Then, we can imagine the following use: If the screening result shows red or yellow, further medical examination is highly recommended. Child psychiatrists who specialise in FASD should examine the patient and investigate the presence of FASD. Experts consider additional information, such as facial dysmorphia or prenatal alcohol exposure that are required to meet official medical diagnostic criteria but were considered inapplicable for a screening tool. For any future implementation of FASDetect in clinical practice, the following considerations are relevant. It is important to prevent any premature diagnosis based on a screening tool. In order to achieve this, every physician or clinical facility using FASDetect should be trained and sensitized to this issue, and there should be a clear protocol established, such as described above, on how to deal with patients with a high screened risk of FASD. It is further important to note that FASDetect is trained and aiming to screen for (and not diagnose) FASD patients among youth patients with ADHD, so we suggest to disregard any use in the general population or in adults, which this tool was not developed for and where results might contain biases. Furthermore, race was not assessed in our study. While facial features can certainly differ considerably between races, we are not aware of studies that have indicated a different presentation of symptoms among races. In our study, we expect that most patients were Caucasian from Western Europe, so any application in other contexts that differ in the race distribution in the target population should be interpreted with caution. Regarding the choice of variables in our prediction model, we aimed to select variables in the final model that are less prone to bias and are more likely to yield accurate and generalizable prediction models. To this end, we had made an expert screening of the assessed variables using expert-generated directed acyclic graphs[26]. As one example, the variable "foster care" was not included in the model since we expected that this variable might have introduced unwanted confounding. Finally, for any application of FASDetect, it should be noted that we used birth length percentiles in our model specific for the German population, which should be evaluated for their application to other populations in follow-up studies, or adjusted to the national norm if it turns out that this may be necessary.

Paediatricians vastly underrecognize FASD and are often unfamiliar with the diagnostic criteria, leading to a higher chance for misdiagnosis and missed diagnosis[27]. The risk of underrecognition and misdiagnosis is at least as high for child and adolescent psychiatrists. FASDetect could enable inexperienced medical staff to screen for FASD and direct patients to specialists. This can help FASD patients to be diagnosed earlier in life and be seen by specialists. Thus, FASDetect could help to reduce the misdiagnosis rates and aide the diagnostic process in busy clinical settings. The successful implementation promises an earlier diagnosis for FASD patients who are currently frequently incorrectly diagnosed with ADHD. Thus, patients who are screened using FASDetect will benefit from earlier treatment, a reduction of secondary conditions and eventually from improved general health.

The results of this study have to be interpreted within its limitations. First, the analysis of archived patient records was limited by the available content of the data. Including further clinical variables might further improve the predictive accuracy of FASDetect. Second, we only examined the discriminatory power and accuracy of the FASDetect app for FASD cases among a sample of patients with a primary diagnosis of ADHD. Further studies are needed that include a broader variety of mental health diagnoses, ideally also oppositional defiant disorder, autism-spectrum disorders and youth with intellectual disability/low IQ who share some other features of FASD than patients with ADHD. The inclusion of further variables that were not available such as reduced eyesight, head circumference at initial presentation and academic achievements are promising predictor candidates for future iterations of the model that are relatively easy to obtain clinically and that should therefore be assessed in future studies.

Third, FASD cases were not distributed evenly within the spectrum (139 FAS, 127 pFAS, 9 ARND), which may have aided the differentiation of the ADHD and FASD groups by the machine learning algorithms. Future research is needed to evaluate how well FASDetect identifies patients across the entire FASD spectrum. Fourth, the study was conducted in a university hospital setting, and testing of generalizability to other clinical settings is further required. Fifth, the patient data for the FASD cases was gathered by psychiatrists specialized in FASD diagnosis. The ADHD cases were diagnosed by outpatient clinicians trained in child and adolescent psychiatry, but without a specific focus on ADHD. The high level of expertise and elaborate testing (e.g. intelligence testing) cannot necessarily be expected of the average user of FASDetect. We adapted the selection of variables that went into final screening tool accordingly. Nevertheless, it is possible that variables seem less distinctive to lesser experienced pediatricians and may be underrecognized when screening with FASDetect.

To our knowledge, this study is the first that developed an empirically-based, machine-learning-derived screening app that robustly differentiates between FASD and ADHD using parameters that can be relatively easily obtained as part of clinical care. The tool, which we call FASDetect, provides a green-yellow-red light rating system on the risk for FASD in ADHD patients calculated from easily obtainable patient data and is an efficient tool for general pediatric practice. The FASDetect is freely available, and we hope that future research with this tool can validate and extend its utility and assess to what degree FASDetect can aide clinical diagnosis and decision-making for subjects with FASD compared to usual care.

MATERIALS AND METHODS
Study population
This study was conducted at the outpatient unit of the department of child and adolescent psychiatry at the Campus Charité Virchow of the Charité Universitätsmedizin Berlin, Germany. For the analysis, a group of consecutively assessed patients with a clinical diagnosis of ADHD without FASD and a group of patients with an expert diagnosis of FASD (with or without comorbid ADHD) was compared. ADHD patients were included from the general pool of patients who were treated at Campus Charité Virchow of the Charité Universitätsmedizin Berlin between January 2019 and September 2020. FASD patients were included

npj

from two sources: from the general pool of ADHD patients described above, as well as from the pool of ambulatory patients of the FASD specialist center at the Campus Charité Virchow of the Charité – Universitätsmedizin Berlin who were treated between January 2019 and September 2020. The two groups were ascertained based on the following inclusion and exclusion criteria.

Inclusion criteria for children and adolescents with ADHD were

a. age between 0 and 19 years,
b. diagnosis of ADHD, combined type of inattentive type, with or without oppositional defiant or conduct disorder according to ICD-10 by child and adolescent psychiatrists at our department of child and adolescent psychiatry at the Campus Charité Virchow of the Charité Universitätsmedizin Berlin,
c. diagnosis of ADHD confirmed during longitudinal assessment and care at our department

Exclusion criteria for children and adolescents with ADHD were

a. severe medical, psychiatric, or neurological conditions (such as microdeletion, microduplication, genetic syndromal diseases, autism-spectrum disorders or hydrocephalus) which can affect the youth's behaviour
b. suspected or confirmed comorbid FASD diagnosis

Inclusion criteria for children and adolescents with FASD (with or without ADHD) were

a. age between 0 and 19 years,
b. diagnosis of FASD according to ICD-10 and the 4 digit code[7]
c. diagnosis of FASD confirmed as part of longitudinal assessment and care at our department

Exclusion criteria for children and adolescents with FASD were severe medical, psychiatric, or neurological conditions.

Of each patient, the following data were extracted retrospectively from medical records: height, weight and head circumference at all available time points; presence or absence of any psychiatric comorbidities, prescribed psychotropic medications yes versus no, fascial dysmorphia and malformation; the results of intelligence tests, whether or not the patient's IQ was below 85 IQ points; as well as pregnancy- and birth-related data such as consumption of alcohol, nicotine and other drugs, number of the mother's pregnancies and births, child's gestational age at first ultrasound and at time of birth, Apgar score[28] and pH of the umbilical cord after birth. The presence or absence of oppositional, hyperactive and impulsive behavior, lack of concentration and attention, developmental disorders, sleep disorders, socially intrusive behavior, and impaired executive function and cognitive flexibility were also assessed clinically. Those symptoms were recorded during clinical assessments, history taking, parent and patient interviews and through behavioral questionnaires such as the child behavior checklist[29] or DISYPS[30]. Assessed symptoms were documented as "present" or "absent", no degree of severity was assessed.

Statistical analysis

The statistical analysis aimed at developing and evaluating a prediction model that would be able to separate FASD from ADHD cases with sufficient accuracy. All machine learning analyses were performed in Python 3.7.3. The code is publicly available at https://github.com/HIAlab/FASDetect. After overall data quality control steps, the training and evaluation of different prediction algorithms was performed in several steps.

In a first overall quality control step, we removed variables with more than 35% missing values for either group (ADHD/FASD). This missing values threshold was chosen in order to include head circumference at birth, which had 35% missing values for ADHD patients, as an indicator for growth deficiencies in FASD patients that is easy to assess and well-suited for use in a screening application. The quality control retained 42 predictive variables, from which we further removed variables with redundant information, such as re-coded duplicates (20 variables), variables that would be too complex to assess for practitioners during a clinical screening visit (5 variables, e.g., executive dysfunction), and variables that might limit generalizability (8 variables). For some variables, multiple reasons for exclusion applied. From the resulting 13 variables, none had more than 23% missing values across both the ADHD and FASD groups. On average, 11% of the variable values were missing for the ADHD group and 12% for the FASD group.

Next, we tested the performance of 6 machine learning algorithms to predict ADHD or FASD using the 13 remaining variables on our data with nested cross-validation (see Fig. 4). To initialize our machine learning pipeline, we randomly split the entire data set into 10 folds (outer split), where each of these folds consisted of 10% of the ADHD cases ($n = 170$) and 10% of the FASD cases ($n = 275$), respectively. We used these outer folds to
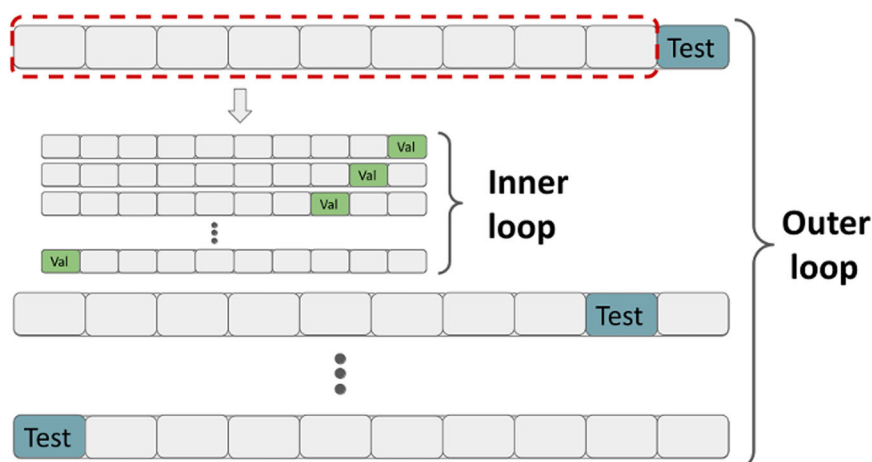

Fig. 4 Overview of the 10-fold nested cross-validation procedure. The data are randomly split into 10 stratified folds where one fold is held out as a test set (blue colour). For each split, the 9 folds are split again into 10 folds, with one fold (green colour) to validate the hyperparameters. The hyperparameters with the best average ROC AUC on the validation sets are used to fit the machine learning pipeline on the complete training set (i.e., the 9 outer folds framed in red colour) and tested against the test set (blue colour), resulting in 10 ROC AUC scores.

perform 10-fold cross-validation (CV) with nine folds for training and the remaining fold for testing. The training data from the outer split with 90% of the data were split again into 10 stratified folds used for training and 10% for validation of the hyperparameters of the pipeline (see below) using a grid search. After the optimized hyperparameter configuration was found in the nested 10-fold CV, the respective model was refit on the complete training data of the outer split (i.e. training and validation data of the inner split) and evaluated against the fold's test set. The nested CV scheme is depicted in Fig. 4.

The training and testing of the different models contained the following steps which are described in more detail below: robust scaling, imputation, feature selection, and model fitting, all embedded in the 10-fold CV. To ensure that the contribution of each variable was similar in the prediction models, we transformed all 13 variables using robust scaling. In robust scaling, the median is subtracted from the value of each variable and each value is then divided by the interquartile range. As a second data processing step, missing values were imputed using k-nearest neighbours (kNN) imputation: each missing value was imputed using the (uniform or distance-weighted) mean value from k_i nearest neighbours found in the training set with non-missing values for the variable, where k_i is a hyperparameter of the pipeline. The distance between two points was measured by Euclidean distance, ignoring variables that were missing for either point. In the next step, we performed a variable selection among the 13 selected variables based on their estimated mutual information with the target variable. Mutual information measures the dependency between two random variables based on entropy and allows to capture also non-linear relationships. Each variable is ranked based on its mutual information with the target variable, and the highest-ranking k_f variables are selected, where k_f is a hyperparameter optimized in the pipeline. Finally, based on these transformed and quality-controlled variables, we trained and evaluated the different machine-learning algorithms. In particular, we tested a logistic regression (LR), support vector machine (SVM), random forest (RF), gradient boosting decision tree (GBDT), kNN classification and Gaussian process classification algorithms. We used the lightgbm package for gradient-boosting decision trees, for all other algorithms, we used the Scikit-learn implementation. Optimized hyperparameters included the number of neighbours used for imputation (k_i), the number of variables to select (k_f) or the decision whether to average values of the neighbours distance-based or uniformly for imputation. Model-specific hyperparameters for the GBDT model included the learning rate, boosting type and number of trees. For random forest models, optimized model-specific hyperparameters were the minimum number of samples required to split an internal node, and the number of trees in the ensemble. For logistic regression, the regularization parameter was optimized. For support vector machines and Gaussian process classifier, the regularization parameter and kernel type were optimized. Hyperparameters optimized for the kNN classifier were the distance metric, the decision whether to average the values of the neighbours either distance-based or uniformly and the number of neighbours.

The main outcome measure for the classification quality of each algorithm was the area under the receiver operating characteristic (ROC) curve (AUC), which was averaged across the 10 test datasets. The reported confidence intervals for ROC AUC scores are the average interval boundaries of confidence intervals calculated for each CV fold according to DeLong[31]. In addition, we assessed the accuracy, precision, recall, and the calibration measured through the Brier score of each model. Lower Brier scores indicate better calibration[32].

In a follow-up analysis, our aim was to evaluate the performance of a most parsimonious prediction model using fewer variables, which is easier to apply in practice. To this aim, the pipeline was run again with a modified variable selection step, where only variables were selected that had been selected by at least half of the different machine learning models in at least 9 of the 10 CV trials. As described above, a variable was selected in a CV trial of an experiment with a classifier when the estimated mutual information with the target was among the k_f highest ranking features on the training set and the classifier with the best hyperparameters (including the number of variables, validated on the validation sets of this CV trial) used this variable. Six variables satisfied these criteria and were used to train the machine learning pipelines a second time.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

REFERENCES

1. May, P. A. et al. Prevalence and epidemiologic characteristics of FASD from various research methods with an emphasis on recent in-school studies. *Dev. Disabil. Res Rev.* **15**, 176–192 (2009).
2. Alex, K. & Feldmann, R. Children and adolescents with fetal alcohol syndrome (FAS): better social and emotional integration after early diagnosis. *Klin. Padiatr.* **224**, 66–71 (2012).
3. Paley, B. & O'Connor, M. J. Behavioral interventions for children and adolescents with fetal alcohol spectrum disorders. *Alcohol Res. Health* **34**, 64–75 (2011).
4. Peadon, E., Rhys-Jones, B., Bower, C. & Elliott, E. J. Systematic review of interventions for children with Fetal Alcohol Spectrum Disorders. *BMC Pediatr.* **9**, 35 (2009).
5. Patrenko, C. L., Tahir, N., Mahoney, E. C. & Chin, N. P. A qualitative assessment of program characteristics for preventing secondary conditions in individuals with fetal alcohol spectrum disorders. *J. Popul Ther. Clin. Pharm.* **21**, e246–e259 (2014).
6. Landgraf, M. N., Nothacker, M. & Heinen, F. Diagnosis of fetal alcohol syndrome (FAS): German guideline version 2013. *Eur. J. Paediatr. Neurol.* **17**, 437–446 (2013).
7. Astley, S. J. & Clarren, S. K. Diagnosing the full spectrum of fetal alcohol-exposed individuals: introducing the 4-digit diagnostic code. *Alcohol* **35**, 400–410 (2000).
8. Mattson, S. N., Bernes, G. A. & Doyle, L. R. Fetal Alcohol Spectrum Disorders: A Review of the Neurobehavioral Deficits Associated With Prenatal Alcohol Exposure. *Alcohol Clin. Exp. Res.* **43**, 1046–1062. (2019).
9. Streissguth, A. P. et al. Risk factors for adverse life outcomes in fetal alcohol syndrome and fetal alcohol effects. *J. Dev. Behav. Pediatr.* **25**, 228–238 (2004).
10. Freunscht, I. & Feldmann, R. Young adults with Fetal Alcohol Syndrome (FAS): social, emotional and occupational development. *Klin. Padiatr.* **223**, 33–37 (2011).
11. Popova, S. et al. Comorbidity of fetal alcohol spectrum disorder: a systematic review and meta-analysis. *Lancet* **387**, 978–987. (2016).
12. Chasnoff, I. J., Wells, A. M. & King, L. Misdiagnosis and missed diagnoses in foster and adopted children with prenatal alcohol exposure. *Pediatr.* **135**, 264–270 (2015).
13. Duda, M., Haber, N., Daniels, J. & Wall, D. P. Crowdsourced validation of a machine-learning classification system for autism and ADHD. *Transl. Psychiatry.* **7**, e1133 (2017).
14. Zhang, C. et al. Detection of Children/Youth With Fetal Alcohol Spectrum Disorder Through Eye Movement, Psychometric, and Neuroimaging Data. *Front Neurol.* **10**, 80 (2019).
15. Fu, Z., Jiao, J., Suttie, M. & Noble, J. A. Facial Anatomical Landmark Detection Using Regularized Transfer Learning With Application to Fetal Alcohol Syndrome Recognition. *IEEE J. Biomed. Health Inf.* **26**, 1591–1601 (2022).

16. Blanck-Lubarsch, M., Dirksen, D., Feldmann, R., Bormann, E. & Hohoff, A. Simplifying Diagnosis of Fetal Alcohol Syndrome Using Machine Learning Methods. *Front Pediatr.* **9**, 707566 (2021).

17. Fang, S. et al. Automated diagnosis of fetal alcohol syndrome using 3D facial image analysis. *Orthod. Craniofac Res.* **11**, 162–171 (2008).

18. Dwyer, D. B., Falkai, P. & Koutsouleris, N. Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annu Rev. Clin. Psychol.* **14**, 91–118 (2018).

19. Streissguth, A. P. et al. Fetal alcohol syndrome in adolescents and adults. *JAMA* **265**, 1961–1967 (1991).

20. Peadon, E. & Elliott, E. J. Distinguishing between attention-deficit hyperactivity and fetal alcohol spectrum disorders in children: clinical guidelines. *Neuropsychiatr. Dis. Treat.* **6**, 509–515 (2010).

21. Lewis, C. E. et al. Verbal learning and memory impairment in children with fetal alcohol spectrum disorders. *Alcohol Clin. Exp. Res* **39**, 724–732 (2015).

22. Kalberg, W. O. & Buckley, D. FASD: what types of intervention and rehabilitation are useful? *Neurosci. Biobehav Rev.* **31**, 278–285 (2007).

23. Spohr, H. L., Willms, J. & Steinhausen, H. C. Fetal alcohol spectrum disorders in young adulthood. *J. Pediatr.* **150**, 175–179, 9.e1 (2007).

24. Dylag, K. A. et al. Sleep problems among children with Fetal Alcohol Spectrum Disorders (FASD)- an explorative study. *Ital. J. Pediatr.* **47**, 113 (2021).

25. Jan J. E. et al. Sleep Health Issues for Children with FASD: Clinical Considerations. Int J Pediatr. 2010;2010:639048.

26. Piccininni, M., Konigorski, S., Rohmann, J. L. & Kurth, T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. *BMC Med. Res. Methodol.* **20**, 179 (2020).

27. Rojmahamongkol, P., Cheema-Hasan, A. & Weitzman, C. Do pediatricians recognize fetal alcohol spectrum disorders in children with developmental and behavioral problems? *J. Dev. Behav. Pediatr.* **36**, 197–202 (2015).

28. Apgar, V. A proposal for a new method of evaluation of the newborn infant. *Curr. Res Anesth. Analg.* **32**, 260–267 (1953).

29. Döpfner M. P., Julia; Kinnen, Claudia. Deutsche Schulalter-Formen der Child Behavior Checklist von Thomas M. Achenbach Elternfragebogen über das Verhalten von Kindern und Jugendlichen (CBCL/6-18R), Lehrerfragebogen über das Verhalten von Kindern und Jugendlichen (TRF/6-18R), Fragebogen für Jugendliche (YSR/11-18R).

30. Döpfner MG-D, Anja; Lehmkuhl, Gerd. Diagnostik-System für psychische Störungen nach ICD-10 und DSM-5 für Kinder und Jugendliche - II. 2008.

31. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* **44**, 837–845 (1988).

32. Bella A., Ferri, C., Hernández-Orallo, J. and Ramírez-Quintana, M. J. Calibration of machine learning models. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. 2010. 128–46.

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS
Conceptualization and methodology, O.G., S.K. and H.W. statistical analysis and implementation, L.E. and A.W. data generation, A.W., H.W., O.G. writing of first draft, L.E. and A.W. Supervision, S.K., O.G., H.W., C.U.C. Critical revision of the manuscript, all authors. All authors read and agreed on the final version of the manuscript. Lukas Ehrig and Ann-Christin Wagner contributed equally to this work as joint first authors. S.K. and O.G. contributed equally to this work as joint senior authors supervising the work.

FUNDING

COMPETING INTERESTS
C.U.C. has been a consultant and/or advisor to or has received honoraria from: AbbVie, Acadia, Alkermes, Allergan, Angelini, Aristo, Boehringer-Ingelheim, Cardio Diagnostics, Cerevel, CNX Therapeutics, Compass Pathways, Damitsa, Denovo, Gedeon Richter, Hikma, Holmusk, IntraCellular Therapies, Janssen/J&J, Karuna, LB Pharma, Lundbeck, MedAvante-ProPhase, MedInCell, Merck, Mindpax, Mitsubishi Tanabe Pharma, Mylan, Neurocrine, Newron, Noven, Otsuka, Pharmabrain, PPD Biotech, Recordati, Relmada, Reviva, Rovi, Seqirus, SK Life Science, Sunovion, Sun Pharma, Supernus, Takeda, Teva, and Viatris. He provided expert testimony for Janssen and Otsuka. He served on a Data Safety Monitoring Board for Compass Pathways, Lundbeck, Relmada, Reviva, Rovi, Supernus, and Teva. He has received grant support from Janssen and Takeda. He received royalties from UpToDate and is also a stock option holder of Cardio Diagnostics, Mindpax, LB Pharma and Quantic. O.G. received honoraria from Takeda and Novartis. C.U.C. and O.G. declare no competing non-financial interests. The remaining authors declare no competing financial or non-financial interests.

ETHICAL STATEMENT
This study was conducted in accordance with the principles of the Declaration of Helsinki and Good Clinical Practice and approved by the local ethics committee at Charité – Universitätsmedizin Berlin (EA2/053/20). Consent/assent requirements were waived for this retrospective chart review study.

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-023-00864-1.

Correspondence and requests for materials should be addressed to Stefan Konigorski.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.