OPEN

Citation: Transl Psychiatry (2015) 5, e514; doi:10.1038/tp.2015.7

www.nature.com/tp

ORIGINAL ARTICLE

Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning

JA Kosmicki[1,2], V Sochat[3], M Duda[4] and DP Wall[4]

Although the prevalence of autism spectrum disorder (ASD) has risen sharply in the last few years reaching 1 in 68, the average age of diagnosis in the United States remains close to 4—well past the developmental window when early intervention has the largest gains. This emphasizes the importance of developing accurate methods to detect risk faster than the current standards of care. In the present study, we used machine learning to evaluate one of the best and most widely used instruments for clinical assessment of ASD, the Autism Diagnostic Observation Schedule (ADOS) to test whether only a subset of behaviors can differentiate between children on and off the autism spectrum. ADOS relies on behavioral observation in a clinical setting and consists of four modules, with module 2 reserved for individuals with some vocabulary and module 3 for higher levels of cognitive functioning. We ran eight machine learning algorithms using stepwise backward feature selection on score sheets from modules 2 and 3 from 4540 individuals. We found that 9 of the 28 behaviors captured by items from module 2, and 12 of the 28 behaviors captured by module 3 are sufficient to detect ASD risk with 98.27% and 97.66% accuracy, respectively. A greater than 55% reduction in the number of behaviorals with negligible loss of accuracy across both modules suggests a role for computational and statistical methods to streamline ASD risk detection and screening. These results may help enable development of mobile and parent-directed methods for preliminary risk evaluation and/or clinical triage that reach a larger percentage of the population and help to lower the average age of detection and diagnosis.

Translational Psychiatry (2015) **5,** e514; doi:10.1038/tp.2015.7; published online 24 February 2015

INTRODUCTION

Rates of autism spectrum disorder (ASD) continue to climb, now impacting 1 in 68 individuals in the United States.[1] Despite important progress in understanding the genetics of ASD,[2,3] ASD remains diagnosed through behavioral examination. The diagnosis of ASD is currently made using instruments designed to measure impairments in the two core domains of ASD, as defined by the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V): (1) communication and social interaction and (2) restricted interests and repetitive behaviors. The Autism Diagnostic Observation Schedule (ADOS)[4] is one of the most widely used instruments to assist in ASD diagnosis. The ADOS consists of a series of semi-structured activities designed to elicit specific behaviors of social interaction, communication, imaginative use of objects, restricted interests and repetitive behaviors. The diagnostic test is split into four modules, each tailored to specific individuals based on their language and developmental level to ensure coverage of a diverse set of behavioral manifestations.[4] A certified professional at a clinical facility first administers the ADOS examination and then scores the individual based on his or her observations to determine the final diagnosis. The initial assessment can take between 30 and 60 minutes, and the scoring increases the total time to between 60 and 90 minutes. Due to variance in inter-rater reliability, additional professionals may re-score the individual, further increasing the time between testing and receipt of the official clinical diagnosis.[4]

Even ignoring the geographic and logistical hurdles in finding a certified professional to administer the ADOS, the time required for the exam and the rise in the number of children at risk for ASD have contributed to increasing bottlenecks in the healthcare system.[5] The average age of diagnosis in the United States hovers stubbornly around 4 years,[5] and families may wait as long as 13 months for the diagnosis after the initial screening,[6] and even longer if they are from a minority population or are of lower socioeconomic status.[7] Such delays impede early intervention speech and behavioral therapies that provide substantial benefits to children.[8,9] For the estimated 27% of individuals undiagnosed at 8 years of age,[5] opportunities for therapeutic intervention have dissipated. Therefore, risk assessment and triage tools that can reach families earlier and enable them to receive the care they need are badly needed.

Given the promising findings from our previous work on the first module of the ADOS[10,11] and the ADI-R,[12] we postulated that we might obtain similar results when examining records from the other two modules of the ADOS, which apply to a large portion of the population suspected of having an ASD.[13] Improving upon our previous work, here we utilized the best-estimate clinical diagnosis when possible and incorporated stepwise backward feature selection into our machine learning pipeline to quantitatively select the optimal set of significant behavioral features that can accurately detect ASD risk in a large population of individuals. We assembled a collection of ADOS evaluations for 4540 individuals and developed a classifier for each module that exhibited optimal

[1]Department of Medicine, Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA; [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; [3]Graduate Program in Biomedical Informatics, Stanford University, Stanford, CA, USA and [4]Division of Systems Medicine, Department of Pediatrics and Psychiatry (by courtesy), Stanford University, Stanford, CA, USA. Correspondence: Dr DP Wall, Division of Systems Medicine, Department of Pediatrics and Psychiatry (by courtesy), Stanford University, Medical School Office Building, 1265 Welch Road, Stanford, CA 94305, USA.
E-mail: dpwall@stanford.edu
Received 6 October 2014; revised 8 December 2014; accepted 19 December 2014

performance in classification of individuals both on and off the spectrum. Each classifier was trained on over 600 individuals and tested independently on more than 1000 individuals. The resulting classifiers contained fewer items than the ADOS-2 (ref. 14) and pinpointed several behaviors that could help guide future efforts focused on expeditious observation-based screening both in and out of clinical settings.

MATERIALS AND METHODS

Data sets

Data for modules 2 and 3 came from five separate repositories: Boston Autism Consortium (AC), Simons Simplex Collection v14 (SSC),[15] Autism Genetic Resource Exchange (AGRE),[16] National Database of Autism Research (NDAR)[17] and the Simons Variation in Individuals Project (SVIP)[18] (Table 1). The ADOS examination classified individuals into three discrete categories (autism, autism spectrum, and non-spectrum) by summing the scores from a subset of items from the ADOS and cross-referencing this total score with the thresholds for autism, autism spectrum and non-spectrum. ADOS scores for each item fall on an integer scale of 0–3, with scores of 7 or 8 reserved for behaviors not exhibited during the test. In a preprocessing step, the ADOS algorithm recodes scores of 3 to 2 and scores of 7 or 8 to 0 to improve reliability and validity.[14] For our analyses, we recoded scores of 7 and 8 as 0, but elected to leave scores of 2 and 3 as distinct answer codes to increase granularity in the classification. In addition, we grouped strict autism and autism spectrum categories together into one autism spectrum cohort, leaving only two classes for machine learning, an autism spectrum class and a non-spectrum class.

Recruitment varied by study. Individuals in AC, AGRE, NDAR and SSC were recruited with a suspicion of having ASD, and individuals in the SVIP were required to have or be related to an individual with the 16p11.2 duplication/deletion.[19] Gender remained consistent across both modules; males comprised 82–86% of individuals with ASD and 61–63% of individuals without ASD. The intelligence quotient (IQ) was consistent across both modules and between individuals with and without ASD (Table 2). Due to the diverse phenotypic effects of the 16p11.2 duplication/deletion, individuals in SVIP were enriched for comorbidities, including ADHD, developmental coordination disorder, phonological disorder and others. Thus, the individuals in the SVIP proved useful for testing the specificity of the algorithms (i.e., differentiating between ASD and other behavioral disorders and developmental delays). A complete description of the phenotypic diversity of the samples used is provided in Supplementary Table S1.

Different versions of the ADOS were used in each data set, namely ADOS Version 1 (ref. 4) (AC, SSC, AGRE and NDAR), and ADOS version 2 (ref. 14) (SVIP). To ensure consistency across data sets, we computed the ADOS-2 diagnosis for all individuals in AC, AGRE, NDAR and SSC using the ADOS-2 algorithms. We elected to do this because the ADOS-2 incorporates repetitive and restrictive behaviors, and it has been shown to more accurately identify cases from non-spectrum controls in lower-functioning populations.[14] Not all individuals had either a clinician's diagnosis or the best-estimate clinical diagnosis. Specifically, 76% of the ASD cases and 46% of the non-autism controls had a recorded clinician's diagnosis or the best-estimate clinical diagnosis. Therefore, we elected to use the diagnosis provided by the ADOS-2 algorithm for our classifier labels in the training processes.

Table 1. Training and testing data description

	Module 2			*Module 3*		
	Autism	*Autism spectrum*	*Non-spectrum*	*Autism*	*Autism spectrum*	*Non-spectrum*
AC	111	16	10	164	33	60
AGRE[a]	314	28	23	454	56	93
NDAR[b]	315	47	282	109	21	27
SSC	575	27	0	1333	233	0
SVIP	14	4	33	21	10	127
Total	1329	122	348	2081	353	307

Abbreviations: AC, Autism Consortium; ADOS, Autism Diagnostic Observation Schedule; AGRE, Autism Genetic Resource Exchange; NDAR, National Database of Autism Research; SSC, Simons Simplex Collection; SVIP, Simons Variance in Individuals Project. Total number of individuals given a diagnosis of autism, autism spectrum or non-spectrum from the ADOS-2. [a]AGRE was used for training the module 3 classifiers. [b]NDAR data set was used for training the module 2 classifiers.

Machine learning

We used machine learning to develop two classifiers: one derived from ADOS module 2 and the other from ADOS module 3. For each module, our strategy involved training eight different machine learning algorithms (Table 3) using stepwise backward feature selection, and testing the final classifier on four independent data sets. We chose stepwise backward feature selection over stepwise forward feature selection to allow for interactions between features.[20] We used each module's items as features, and the individuals' ADOS-2 diagnoses as our prediction class. All machine learning analyses were performed in R and Weka[21] (version 3-7-9). As the number of individuals with ASD outnumbered those without in both module 2 (~4:1) and module 3 (~5:1) across all data sets, we selected the data set with the highest number of individuals without ASD as our training set. Module 2 classifiers were trained from an NDAR collection of 362 with ASD and 282 individuals without ASD. Module 3 classifiers were trained on AGRE, with 510 individuals with ASD and 93 individuals without ASD (Table 1).

The 28 features for each of module 2 and module 3 were ranked using a support vector machine (SVM) based on their ability to differentiate between individuals with and without ASD. We used stepwise backward feature selection with 10-fold cross-validation in all eight machine learning algorithms. This feature selection procedure determined the optimal number of features by first training a classifier with all 28 features, iteratively removing the lowest-ranked feature, and building a new model using 90% of the data for training and the remaining 10% for testing. The process ended once a single feature remained, yielding a final set of 28 classifiers, which could each be assessed for their sensitivity and specificity. By plotting the sensitivity, specificity and accuracy of each classifier versus the number of features, the best classifier was identified as the one with the highest performance and smallest number of features (Figure 1). We aimed to maximize specificity (the true negative rate) over sensitivity (the true positive rate) because of the large class imbalance (Table 1).

Validation

After finding the optimal classifiers for modules 2 and 3, we validated these classifiers on the remaining four data sets not used for training. The module 2 classifier was tested on AC, AGRE, SSC and SVIP, totaling 1089 individuals with ASD and 66 individuals without ASD (Table 1). The module 3 classifier was tested on AC, NDAR, SSC and SVIP, totaling 1924 individuals with ASD and 214 individuals without ASD (Table 1).

RESULTS

Module 2 results

Two algorithms using the same nine features displayed optimal performance on the NDAR training data (98.90% sensitivity, 98.58% specificity and 98.76% accuracy), a logistic regression[22] and a logistic model tree (LMT)[23] (Table 3; Figure 1). LMTs combine decision trees with logistic regression, thereby allowing the incorporation of nonlinear patterns into the model. When such nonlinear patterns exist and help explain additional variance in the data, LMTs outperform logistic regression.[23] However, in our data, no such patterns were detected and the nine-feature LMT consisted of just the root node with a logistic regression model. Thus we chose logistic regression over LMT for use in further testing and validation.

For independent validation of the nine-feature logistic regression classifier, we collated score sheets for module 2 from the AC, AGRE, SSC and SVIP (Table 1) to determine whether the classifier could recapitulate the sensitivity and specificity of training data on held-out test data. Across our four test sets, the logistic regression classifier misclassified 13 out of 1089 individuals with ASD (98.81% sensitivity) and 7 out of 66 individuals without ASD (89.39% specificity), resulting in 98.27% accuracy (Supplementary Table

S2). Of the 13 misclassified individuals with autism, 6 had a clinical diagnosis of autism, 3 had a clinical diagnosis of pervasive developmental disorder–not otherwise specified and 1 had a best-estimate clinical diagnosis of non-spectrum. For the seven misclassified individuals without autism, three had a non-spectrum clinical diagnosis, three had an autism best-estimate clinical diagnosis and one individual had a clinical diagnosis of broad spectrum. For a subset of individuals, their best-estimate clinical diagnosis was available (autism $N=618$, non-spectrum $N=35$). When independently predicting the best-estimate clinical diagnosis, the sensitivity and specificity of the nine-feature logistic regression model was 98.38% and 88.57%, respectively.

Although the ADOS-2 module 2 uses different algorithms for individuals based on their age, our nine-feature logistic regression classifier does not.[14] Because age and the log-odds of the prediction were significantly correlated ($r=0.45$; $P<2.2\times10^{-16}$), we hypothesized that adding age as a covariate to the regression might explain additional variance in the outcome. However, the effect of age on the classifier was negligible (β 0.015, odds ratio 1.055), and adding it to the model slightly decreased sensitivity (−0.28%) and accuracy (−0.16%). Therefore, we elected not to incorporate age into the regression. IQ measures were also significantly correlated after controlling for gender, including full-scale IQ ($r=-0.37$; $P<2.2\times10^{-16}$), verbal IQ ($r=-0.42$; $P<2.2\times10^{-16}$) and nonverbal IQ ($r=-0.27$; $P<3.8\times10^{-15}$).

The behaviors tested assessed by the module 2 classifier segregated into the two domains associated with ASD: (1) social communication and social interactions and (2) restricted interests and repetitive behaviors. Feature A5 (stereotyped/idiosyncratic use of words or phrases), A8 (descriptive, conventional, instrumental or informational gestures), B1 (unusual eye contact), B3 (shared enjoyment in interaction), B6 (spontaneous initiation of joint attention), B8 (quality of social overtures) and B10 (amount of reciprocal social communication) correspond to the domain of social communication and interaction. D2 (hand and finger and other complex mannerisms) and D4 (unusual repetitive interests or stereotyped behaviors) stem from the domain of restricted interests and repetitive behaviors.

Module 3 results

Of the eight machine learning algorithms trained for module 3, the radial kernel SVM[24] performed best overall on the AGRE training data (100% sensitivity, 98.92% specificity and 99.83% accuracy) (Table 3; Figure 2) and contained 12 behavioral features. This 12-feature SVM classifier was tested on the four data sets not

Table 2. Sample description

		Module 2					*Module 3*				
DX		*N*	*IQ1*	*IQ2*	*IQ3*	*Range*	*N*	*IQ1*	*IQ2*	*IQ3*	*Range*
Autism	Age	1451	52	68	98	12–490	2434	88	111	141	38–559
	Full IQ	710	61	77	90	25–130	1782	83	96	108	26–167
	VIQ	702	57	74	87	19–129	1784	82	95	108	19–167
	NVIQ	705	68	83	95	26–139	1787	85	97	108	26–161
Non-spectrum	Age	348	34	37	48	13–183	307	80	108	130	35–207
	Full IQ	42	76	88	105	54–132	181	87	100	110	63–160
	VIQ	42	75	90	105	54–123	181	89	100	109	49–135
	NVIQ	43	78	88	103	54–137	182	89	98	108	54–169

Abbreviations: DX, ADOS-2 diagnosis; IQ1, first quartile; IQ2, second quartile (median); IQ3, third quartile; VIQ, verbal IQ; NVIQ, nonverbal IQ. All ages are in months.

Table 3. Machine learning algorithms used in training

		Module 2			*Module 3*		
Classifier	*Description*	*Sensitivity*	*Specificity*	*Features*	*Sensitivity*	*Specificity*	*Features*
ADTree	ADTree is based on boosting and combines multiple types of decision trees.	0.967	0.982	10/28	0.988	0.871	9/28
Functional tree	Functional trees use linear/logistic regression at decision nodes and linear models at leaf nodes.	0.981	0.986	12/28	0.994	0.978	14/28
LibSVM*	SVMs search for the highest dimensional plane that separates the classes by the largest margin.	0.997	0.979	14/28	1	0.989	12/28
LMT	Logistic model trees use decision trees with logistic regression models at leaf nodes.	0.989	0.986	9/28	0.998	0.967	15/28
Logistic regression*	Predicts a categorical outcome based on a series of predictor features.	0.989	0.986	9/28	0.996	0.978	19/28
Naive Bayes	Naive Bayes is a probabilistic classifier based on Bayes' theorem.	0.981	0.975	14/28	0.961	0.957	14/28
NBTree	Naive Bayes trees are decision trees that use naive Bayes classifiers at leaf nodes.	0.970	0.979	8/28	0.980	0.925	14/28
Random forest	Random forest trains multiple decision trees returning the most common class.	0.981	0.965	20/28	0.990	0.981	11/28

Abbreviations: ADTree, alternating decision tree; LMT, logistic model trees; NBTree, Naive Bayes Tree; SVM, support vector machine. *Logistic regression and LibSVM were the top-performing algorithms for module 2 and module 3 with respect to sensitivity, specificity and number of features. Description of the eight machine learning algorithms used in training to determine the best algorithm and optimal number of features. Sensitivity, specificity and number of features used over the total number of features in the best-performing iteration of each algorithm for modules 2 and 3 are listed.

used in training: AC, NDAR, SSC and SVIP. Across the four test sets, our classifier misclassified 44 out of 1924 individuals with ASD and 6 out of 214 individuals without ASD (97.71% sensitivity, 97.20% specificity and 97.66% accuracy) (Supplementary Table S3). Of the 44 individuals with ASD who were misclassified, clinical diagnoses were available for 30. Six had a confirmed autism diagnosis, six had Asperger's disorder and the remaining 18 had pervasive developmental disorder–not otherwise specified. For the six

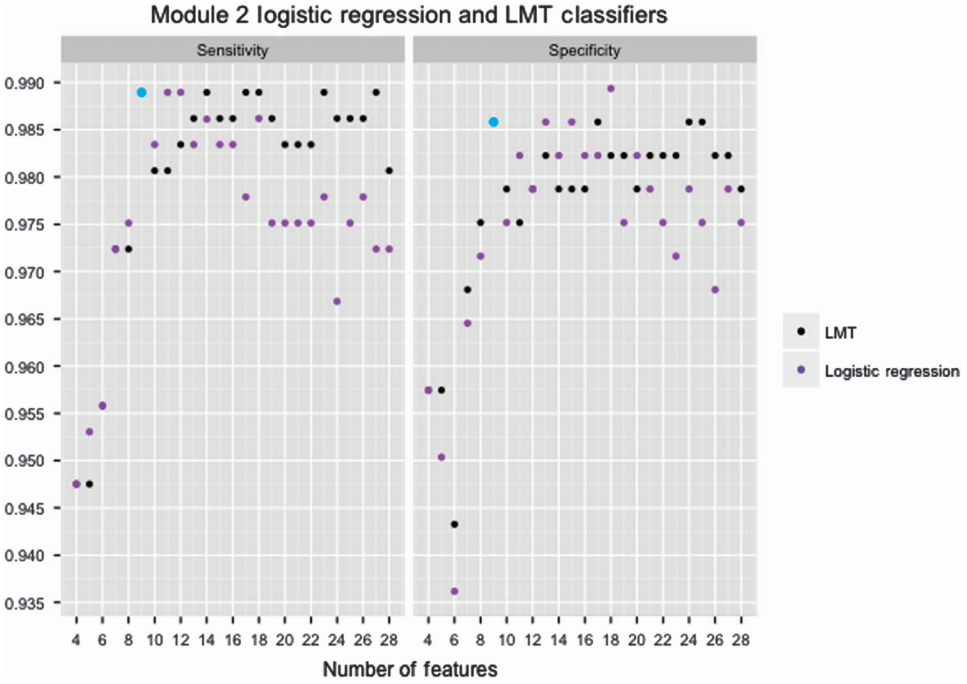

Figure 1. Module 2 logistic regression and logistic model tree (LMT) training results. Sensitivity and specificity of the module 2 logistic regression and LMT classifiers based on the number of features used during training on the National Database of Autism Research are provided in Table 1. The nine-feature logistic regression classifier (blue dot) was used in testing.

Figure 2. Module 3 SVM training results. Sensitivity and specificity of the module 3 SVM classifier based on the number of features used during training on Autism Genetic Resource Exchange are provided in Table 1. The 12-feature SVM classifier was used in testing. SVM, support vector machine.

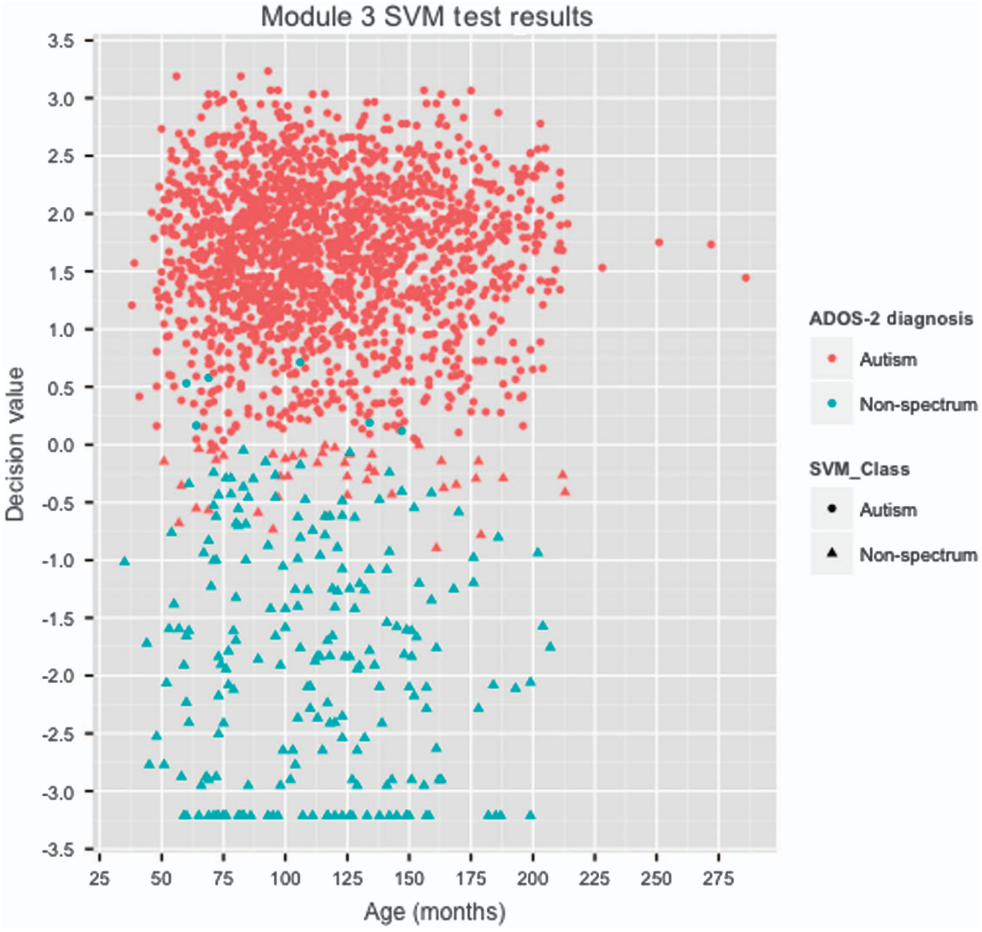

Figure 3. Module 3 SVM test results. The 12-feature SVM decision values from testing data for the two classes: autism (red) and non-spectrum (blue). Forty-four misclassified individuals with autism (red triangles), and six individuals without autism (blue circles) contributed to 97.71% sensitivity and 97.20% specificity. ADOS, Autism Diagnostic Observation Schedule; SVM, support vector machine.

individuals without autism that were misclassified, three had a non-spectrum clinical diagnosis, and the remaining three individuals had no recorded clinical or best-estimate clinical diagnosis. For the individuals for whom a best-estimate clinical diagnosis was available (autism $N = 1568$; non-spectrum $N = 175$), the 12-feature SVM displayed 99.11% sensitivity and 70.86% specificity (Figure 3).

Similar to the module 2 classifier, the features in the module 3 SVM classifier aligned with the two core domains of ASD. Feature A7 (reporting of events), A8 (conversation), A9 (descriptive, conventional, instrumental or informational gestures), B1 (unusual eye contact), B2 (facial expressions directed to others), B7 (quality of social overtures), B8 (quality of social response) and B9 (amount of reciprocal social interaction) correspond to the domain of social communication and interaction. A4 (stereotyped/idiosyncratic use of words or phrases), D1 (unusual sensory interest in play material/person), D2 (hand and finger and other complex mannerisms) and D4 (excessive interest in unusual or highly specific topics or objects) stem from the domain of restricted interests and repetitive behaviors.

DISCUSSION

Despite significant evidence for the genetic heritability of ASD,[25] it remains diagnosed through behavior. Although use of standard instruments for ASD diagnosis has been effective, the practice remains difficult to scale and time intensive, contributing to the growing waiting times between initial warning signs and diagnosis. Machine learning techniques have been previously applied by our group and others to test whether ASD[10–12] and ADHD[26] detection can be achieved with smaller numbers of behavioral measurements. Here, we sought to expand upon our previous work to a wider range of ages and levels of vocabulary by applying machine learning techniques to recorded clinical evaluations of individuals using modules 2 and 3 of the ADOS. We implemented stepwise backward feature selection with eight machine learning algorithms to create small but robust classifiers that retained levels of sensitivity and specificity similar to those of the full ADOS. The logistic regression algorithm produced the top-performing classifier for module 2 using nine features that exhibited 98.81% sensitivity and 89.39% specificity when tested across 1089 individuals with ASD and 66 individuals without ASD. A SVM consisting of 12 behavioral items showed the optimal performance when run on score sheets from module 3, exhibiting 97.71% sensitivity and 97.20% specificity when tested across 1924 individuals with ASD and 214 individuals without ASD.

Both the module 2 and module 3 classifiers contained a large number of items found on the ADOS-2 algorithms, suggesting that our abbreviated classifiers preserve much of the diagnostic validity of the original algorithm. However, we cannot discount the inherent bias in features used in the ADOS-2 algorithm, as those features are used in forming the diagnosis. Despite this, several features in both the ADOS-2 module 2 and module 3 algorithms ranked low in their classification ability. In module 2, A7 and B11 were ranked 13th and 25th, whereas in module 3, B4 and B10 ranked 13th and 14th, respectively, out of the 28 features. The low ranking of these features can be explained by lack of variation in

Table 4. Module 2 activities

Activity	*Required for exam?*
Construction task	Yes
Response to name	No
Make-believe play	No
Joint interactive play	Yes
Conversation	No
Response to joint attention	No
Demonstration task	Yes
Description of a picture	Yes
Telling a story from a book	Yes
Free play	Yes
Birthday party	Yes
Snack	Yes
Anticipation of a routine with objects	Yes
Bubble play	Yes

Abbreviation: ADOS, Autism Diagnostic Observation Schedule. List of the 14 observational activities administered in module 2 of the ADOS-2. Of the 14, only 10 are needed to measure the behaviors used by the Logistic Regression classifier (Supplemental Discussion).

Table 5. Module 3 activities

Activity	*Required for exam?*
Construction task	No
Make-believe play	No
Joint interactive play	Yes
Demonstration task	No
Description of a picture	Yes
Telling a story from a book	Yes
Cartoons	No
Conversation and reporting	Yes
Emotions	No
Social difficulties and annoyance	Yes
Break	Yes
Friends and marriage	Yes
Loneliness	Yes
Creating a story	No

Abbreviation: ADOS, Autism Diagnostic Observation Schedule. List of the 14 observational activities administered in module 3 of the ADOS-2. Of the 14, only 8 are needed to measure the behaviors needed by the support vector machine classifier (Supplemental Discussion).

responses between individuals with and without ASD. Of the 9 and 12 features used in the module 2 and 3 classifiers, five behaviors overlapped between the two machine learning classifiers identified in our study, namely unusual eye contact, quality of social overtures, amount of reciprocal social interaction, descriptive, conventional, instrumental and informational gestures, and hand, finger and other complex mannerisms. Since each module of ADOS is designed for a specific level of developmental ability, the inclusion of these five features in both classifiers may reflect their relative importance to the classification of ASD independent of the language and developmental level of the individual.

When performing a clinical evaluation of an individual with ASD using ADOS modules 2 and 3, the clinician uses 14 prescribed activities designed to elicit specific behaviors by the subject under evaluation. It is possible that the smaller number of behaviors represented in our classifiers may correspond to a compensatory reduction in the number of activities needed for an ASD risk assessment. For example, 3 of the 14 activities in module 2 (Table 4) and 6 of the 14 activities in module 3 (Table 5) would no longer be required to measure the behaviors used in the classifiers (Supplemental Discussion). Further examination and testing of this possibility is certainly needed, but it supports the possiblity that use of fewer behaviors may translate to shorter timeframes for observation. We have previously tested the potential for detection of risk for ASD in short home videos,[27] and we hope in future studies to test whether the behaviors used in the classifiers presented here may also be adequately measured in short home video clips.

Lastly, the output of the module 2 logistic regression classifier provides a quantitative score of the log-odds of the confidence in the classification. Borderline log-odds indicate lower confidence, and therefore need for more testing, before arriving at a risk score and/or diagnosis. The ability to quantitatively measure risk provides another dimension to understand the prediction from the classifier itself. Disagreements among diagnostic exams are not uncommon.[28] By providing the probability of the classification, the module 2 logistic regression classifier could assist in instances of uncertainty. In additionally, if such a scoring system could be used as a pre-clinical screening method, it may be possible to prioritize individuals based on the log-odds of the classification—enabling brief appointments for individuals with clear risk, and longer appointments for individuals that prove clinically challenging.

Limitations

Given that our study focused on analysis of archival records, we were limited by the content of these preexisting data sets. Due to the nature of recruitment, there was a large imbalance in favor of individuals with ASD versus those who tested negative for ASD in AC, AGRE, NDAR and SSC (Table 1). Although the AC and SSC were family-based studies and collected detailed phenotype data for all family members, the ADOS and the ADI-R were administered only to the child with risk for ASD and not to the parents ($N = 2760$) or unaffected siblings ($N = 2278$). Therefore, the individuals without a confirmed ASD in this study were all at least initially suspected of having ASD and administered an ADOS. As such, these non-spectrum individuals served as valuable controls for our study, helping to support the possibility that our classifiers can distinguish between individuals with ASD and those with other developmental delays (Supplementary Table S1). To further measure the specificity of such classification tools, more effort is needed both to balance the number of individuals with and without ASD and to recruit individuals confirmed to have other developmental and/or learning delays.

Defining an appropriate 'truth set' for classifier construction and validation is an important challenge in the field. For ASD, the choice of the truth set is typically among the ADOS, ADI-R, the clinician's diagnosis and the best-estimate clinical diagnosis or some combination thereof.[14] However, none of the potential truth sets are truly independent, as the ADOS and ADI-R can (and often should) influence the clinician's diagnosis and all three can contribute to the best-estimate clinical diagnosis.[28] In the present study, we used the ADOS-2 diagnosis for our truth set during the machine learning trainining processes, given the class imbalance and the fact that 54% of the individuals who tested negative for ASD by the ADOS-2 were missing both the clinician's and best-estimate clinical diagnosis. Yet in our independent validation procedures, we tested the classifiers' performance against all available best-estimate clinical diagnoses. Both analyses provided encouraging results, suggesting that measurement of fewer behaviors can achieve results similar to a full ADOS exam and/or a clinical decision. Nevertheless, it is important to note that the high performance exhibited by the classifiers is based on a truth set that contains subjective observations, and therefore potential biases.[14]

CONCLUSION

Time-intensive behavioral examinations and questionnaires are currently the primary methods used in the diagnosis of ASD. Using machine learning, we created classifiers from two modules of one of the most universally administered behavioral tests, the ADOS. The logistic regression classifier based on analysis of archival records from ADOS module 2 consisted of nine items, 67.86% fewer than the complete ADOS module 2, and performed with 98.81% sensitivity and 89.39% specificity in independent testing. The SVM module 3 classifier based on analysis of archived ADOS module 3 records consisted of 12 items, 57.14% fewer than the complete ADOS module 3, and performed with more than 97% sensitivity and specificity in testing. These results support the notion that fewer behaviors when measured using machine learning tools can achieve high levels of accuracy in autism risk prediction. Furthermore, these results may help encourage future efforts to develop screening-based instruments for ASD detection and mobile health approaches that ultimately enable individuals to receive more expedient care than is possible under the current paradigms.

CONFLICT OF INTEREST

DPW is the scientific founder of Cognoa Inc. (cognoa.com), a digital health company focused on mobile solutions for detection, monitoring and treatment of developmental delay, including ASD. The remaining authors declare no conflict of interest.

ACKNOWLEDGMENTS

We thank all the members of the Wall lab and the Analytical and Translational Genetics Unit (ATGU) for critical input on study design and results interpretation, especially Mark Daly, Elise Robinson, Todd DeLuca and Elaine Lim for their time and assistance. Additionally, we would like to thank the reviewers whose comments substantially improved the quality of the manuscript. Finally, we thank all the families who enrolled in the AC, AGRE, NDAR, SSC and SVIP projects. The work was supported in part by funds to DPW from the Simons Foundation, Nancy Lurie Marks Family Foundation, the Harvard Catalyst Program and grant 1R01MH090611-01A1 from the National Institutes of Health.

REFERENCES

1 Centers for Disease Control and Prevention. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *MMWR Surveill Summ* 2014; **63**: 1–21.
2 Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D *et al.* The contribution of de novo coding mutations to autism spectrum disorder. Nature 2014 11. */13* **515**: 216–221.
3 De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 2014; **515**: 209–215.
4 Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC *et al.* The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord* 2000; **30**: 205–223.
5 Shattuck PT, Durkin M, Maenner M, Newschaffer C, Mandell DS, Wiggins L *et al.* Timing of identification among children with an autism spectrum disorder: findings from a population-based surveillance study. *J Am Acad Child Adolesc Psychiatry* 2009; **48**: 474–483.
6 Wiggins LD, Baio J, Rice C. Examination of the time between first evaluation and first autism spectrum diagnosis in a population-based sample. *J Dev Behav Pediatr* 2006; **27**: 79–87.
7 Bernier R, Mao A, Yen J. Psychopathology, families, and culture: autism. *Child Adolesc Psychiatr Clin N Am* 2010; **19**: 855–867.
8 Howlin P. Children with autism and Asperger syndrome: a guide for practitioners and parents. John Wiley & Sons, Ltd: Chichester, UK, 1998.
9 Pinto-Martin JA, Young LM, Mandell DS, Poghosyan L, Giarelli E, Levy SE. Screening strategies for autism spectrum disorders in pediatric primary care. *J Dev Behav Pediatr* 2008; **29**: 345–350.
10 Wall DP, Kosmicki J, Deluca TF, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Transl Psychiatry* 2012; **2**: e100.
11 Duda M, Kosmicki JA, Wall DP. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Transl Psychiatry* 2014; **4**: e424.
12 Wall DP, Dally R, Luyster R, Jung J-Y, Deluca TF. Use of artificial intelligence to shorten the behavioral diagnosis of autism. PLoS One. 2012; **7**: e43855.
13 Gotham K, Pickles A, Lord C. Trajectories of autism severity in children using standardized ADOS scores. *Pediatrics* 2012; **130**: 1278–1284.
14 Gotham K, Risi S, Pickles A, Lord C. The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. *J Autism Dev Disord* 2007; **37**: 613–627.
15 Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* 2010; **68**: 192–195.
16 Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P *et al.* The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am J Hum Genet* 2001; **69**: 463–466.
17 Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing heterogeneous data: the national database for autism research. *Neuroinformatics* 2012; **10**: 331–339.
18 The Simons Vip Consortium. Simons Variation in Individuals Project (Simons VIP): a genetics-first approach to studying autism spectrum and related neurodevelopmental disorders. *Neuron* 2012. **73**: 1063–1067.
19 Shinawi M, Liu P, Kang S-HL, Shen J, Belmont JW, Scott DA *et al.* Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J Med Genet* 2010; **47**: 332–341.
20 John GH Kohavi R Pfleger K. (eds). Irrelevant features and the subset selection problem. *Machine Learning: Proceedings of the Eleventh International Conference*. San Francisco, CA, USA: Morgan Kaufmann Publisher, 1994. 121–129.
21 Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *SIGKDD Explor Newsl* 2009; **11**: 10–18.
22 Le Cessie S, Van Houwelingen J. Ridge estimators in logistic regression. *Appl Statist* 1992; **41**: 191–201.
23 Landwehr N, Hall M, Frank E. Logistic model trees. *Mach Learn* 2005; **59**: 161–205.
24 Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011; **2**: 27.
25 Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch Gen Psychiatry* 2011; **68**: 1095–1102.
26 Tenev A, Markovska-Simoska S, Kocarev L, Pop-Jordanov J, Muller A, Candrian G. Machine learning approach for classification of ADHD adults. *Int J Psychophysiol* 2014; **93**: 162–166.
27 Fusaro VA, Daniels J, Duda M, DeLuca TF, D'Angelo O, Tamburello J *et al.* The potential of accelerating early detection of autism through content analysis of YouTube videos. *PLoS One* 2014; **9**: 4.
28 Lord C, Risi S, DiLavore PS, Shulman C, Thurm A, Pickles A. Autism from 2 to 9 years of age. *Arch Gen Psychiatry* 2006; **63**: 694–701.

Supplementary Information accompanies the paper on the Translational Psychiatry website (http://www.nature.com/tp)