Use of Machine Learning to Shorten Observation-based Screening and Diagnosis of Autism

Citation: Transl Psychiatry (2012) 2, e100, doi:10.1038/tp.2012.10

www.nature.com/tp

Use of machine learning to shorten observation-based screening and diagnosis of autism

DP Wall[1,2], J Kosmicki[1], TF DeLuca[1], E Harstad[3] and VA Fusaro[1]

The Autism Diagnostic Observation Schedule-Generic (ADOS) is one of the most widely used instruments for behavioral evaluation of autism spectrum disorders. It is composed of four modules, each tailored for a specific group of individuals based on their language and developmental level. On average, a module takes between 30 and 60 min to deliver. We used a series of machine-learning algorithms to study the complete set of scores from Module 1 of the ADOS available at the Autism Genetic Resource Exchange (AGRE) for 612 individuals with a classification of autism and 15 non-spectrum individuals from both AGRE and the Boston Autism Consortium (AC). Our analysis indicated that 8 of the 29 items contained in Module 1 of the ADOS were sufficient to classify autism with 100% accuracy. We further validated the accuracy of this eight-item classifier against complete sets of scores from two independent sources, a collection of 110 individuals with autism from AC and a collection of 336 individuals with autism from the Simons Foundation. In both cases, our classifier performed with nearly 100% sensitivity, correctly classifying all but two of the individuals from these two resources with a diagnosis of autism, and with 94% specificity on a collection of observed and simulated non-spectrum controls. The classifier contained several elements found in the ADOS algorithm, demonstrating high test validity, and also resulted in a quantitative score that measures classification confidence and extremeness of the phenotype. With incidence rates rising, the ability to classify autism effectively and quickly requires careful design of assessment and diagnostic tools. Given the brevity, accuracy and quantitative nature of the classifier, results from this study may prove valuable in the development of mobile tools for preliminary evaluation and clinical prioritization—in particular those focused on assessment of short home videos of children—that speed the pace of initial evaluation and broaden the reach to a significantly larger percentage of the population at risk.
Translational Psychiatry (2012) **2,** e100; doi:10.1038/tp.2012.10; published online 10 April 2012

Introduction

Although autism has a significant genetic component,[1] it is primarily diagnosed through behavioral characteristics. Diagnosing autism has been formalized with instruments carefully designed to measure impairments indicative of autism in three developmental areas: language and communication, reciprocal social interactions and restricted or stereotypical interests and activities. One of the most widely used instruments is the Autism Diagnostic Observation Schedule-Generic (ADOS).[2] The ADOS consists of a variety of semi-structured activities designed to measure social interaction, communication, play and imaginative use of materials. The exam is divided into four modules, each geared towards a specific group of individuals based on their language and developmental level, ensuring coverage for a wide variety of behavioral manifestations. Module 1 contains 10 activities and 29 items, is focused on individuals with little or no language and is therefore most typical for assessment of younger children. The ADOS observation is run by a certified professional in a clinical environment and its duration can range from 30 to 60 min. Following the observation period, the administrator will then score the individual to determine their ADOS-based diagnosis, increasing the total time from observation through scoring to between 60 and 90 min in length.

The long length of the ADOS exam and the need for administration in a clinical facility by a trained professional both contribute to delays in diagnosis and an imbalance in coverage of the population needing attention.[3] The clinical facilities and trained clinical professionals tend to be geographically clustered in major metropolitan areas and far outnumbered by the individuals in need of clinical evaluation. Families may wait as long as 13 months between initial screening and diagnosis,[4] and even longer if part of a minority population or lower socioeconomic status.[5] These delays directly translate into delays in the delivery of speech and behavioral therapies that have significant positive impacts on a child's development, especially when delivered early.[6,7] Thus, a large percentage of the population is diagnosed after developmental windows in which behavioral therapy would have had maximal impact on future development and quality of life. The average age of diagnosis in the United States is 5.7 years and an estimated 27% remain undiagnosed at

[1]Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA; [2]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA, USA and [3]Division of Developmental Medicine, Children's Hospital Boston, Boston, MA, USA
Correspondence: Dr DP Wall, Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115-6030, USA.
E-mail: dpwall@hms.harvard.edu
Keywords: autism classification; autism classifier; algorithm to classify autism; autism diagnostic observation schedule; autism spectrum disorders; machine learning; rapid detection of autism
Received 13 Septemeber 2011; revised 27 December 2011; accepted 8 January 2012

8 years of age.[3] At these late stages in development, many of the opportunities to intervene with therapy have evaporated.

Significant attention has been paid to the design of abbreviated screening examinations that are meant to foster more rapid diagnosis, including the Autism Screening Questionnaire (designed to discriminate between pervasive developmental disorder and non-pervasive developmental disorder diagnoses[8]), the Modified Checklist for Autism in Toddlers[9] and the Social Communication Questionnaire,[10] to name a few. Although these have widespread use and value, the ADOS, because of its high degree of clinical utility and diagnostic validity, remains one of the dominant behavioral tools for finalizing a clinical diagnosis. Research has focused on manual selection of preferred questions from the full ADOS for use in scoring following the observation period, work that has led to critical advances in diagnostic validity and steps toward a reliable measure of severity of the autism phenotype.[11] Our aim in this research study was similarly minded, but specifically focused on testing whether statistical and data-driven selection of the ADOS questions could result in an abbreviated and accurate instrument for classification of autism.

With this goal, we sought to statistically identify a subset of elements from the full ADOS Module 1 that could enable faster screening both in and out of clinical settings without compromising the diagnostic validity of the ADOS. As a valuable by-product of the widespread adoption and use of ADOS, research efforts have banked large collections of score sheets from ADOS together with the clinical diagnosis that can be utilized to address this aim directly. Leveraging these databases, we assembled a collection of complete ADOS evaluations for over 1050 children, focusing on Module 1 data alone to gain insight into the development of shorter approaches for early detection. Through the application of machine-learning methods, we were able to construct classifiers and objectively measure the sensitivity and specificity of each with respect to diagnostic validity and similarity to the original[2] and revised[11] ADOS algorithms. We developed a classifier using decision tree learning that performed optimally for classification of a wide range of individuals both on and off the spectrum. This classifier was substantially shorter than the standard ADOS and pinpointed several behavioral patterns that could guide future methods for expeditious observation-based screening and diagnosis in and out of clinical settings.

Materials and methods

Constructing a classifier. We used ADOS Module 1 data from the Autism Genetic Resource Exchange (AGRE)[12] repository of families with at least one child diagnosed with autism as our input for machine-learning classification. The ADOS examination classified individuals into categories of autism or autism spectrum based on the ADOS diagnostic algorithm. This algorithm added the answers from a subset of items extracted from the full exam for classification on or off the autism spectrum according to a threshold score. Those individuals who did not meet the required threshold were classified as non-spectrum and were used as controls in our study. For the purposes of our study, we restricted the analyses to individuals with the classification of autism. Any individuals with a majority (50% or more) of missing answers in the ADOS exam were excluded. The final data matrix contained 612 individuals with a classification of autism and 11 individuals with a classification of non-spectrum (Table 1).

We constructed 16 alternative classifiers by performing a series of machine-learning analyses (performed using Weka[13]) on the 29 ADOS Module 1 items to differentiate individuals with a classification of autism from those with a classification of non-spectrum. For each algorithm, we performed 10-fold cross-validation, utilizing 90% for training and the remaining 10% for testing to construct the classifiers and measure their sensitivity, specificity and accuracy. This level of cross-validation has been shown previously to perform optimally for structured, labeled data while reducing bias in the resulting classifier.[14] We then plotted the specificity of the classifiers against its sensitivity to visualize the performance and selected the classifier with the best sensitivity, specificity and accuracy (Table 2).

Validating the classifier. In addition to the 10-fold cross-validation, we validated our classifier by testing it on independently collected ADOS data from other individuals with autism in the Boston Autism Consortium (AC) and the Simons Simplex Collection[15] (SSC). The AC data contained 110 individuals who met criteria on the Module 1 ADOS algorithm for autism and an additional four individuals given the non-spectrum classification. The SSC data comprised 336 individuals who met Module 1 cutoffs for autism but lacked ADOS data for non-spectrum individuals.

Table 1 Summary of the data used for both construction and validation of the autism diagnostic classifier

	AGRE		*AC*		*Simons*	
	Autism	***Non-spectrum***	***Autism***	***Non-spectrum***	***Autism***	***Non-spectrum***
Sample size	612	11	110	4	336	0
Q1	4.7375	2.99	3.6875	2.771	5.167	0
Median	6.64	4.57	5.625	3.083	6.75	0
Q3	8.86	6.93	8.4167	6.729	10	0
IQR	4.1225	3.94	4.7292	3.958	4.833	0

Abbreviations: AC, autism consortium; AGRE, autism screening questionnaire; Simons, Simons Foundation.
We acquired complete sets of answers to the Autism Diagnostic Observation Schedule-Generic (ADOS) Module 1 evaluation from the AGRE, the Boston AC and the Simons Foundation. The table lists the total numbers of individuals classified as having autism and individuals classified as non-spectrum represented in each of the three data sets as well as a breakdown of age using the interquartile range.

Table 2 The 16 machine-learning algorithms used for constructing classifiers from the ADOS Module 1 data

Classifier name	*Description*	*FPR*	*TPR*	*Accuracy*
ADTree	An ADTree combines decision trees, voted decision trees and voted decision stumps. The algorithm is based on boosting, which yields accurate predictions by combining a series of 'weak' learners that together can classify accurately.[16]	0.000	1.000	1.000
BFTree	The top node of the decision tree splits the data, so the maximum reduction of impurity (misclassified data) is achieved. This is called the 'best' node and it is expanded upon first (unlike in a C4.5 tree, for example, where nodes are expanded upon according to the depth first).[18]	0.600	0.993	0.979
Decision Stump	A Decision Stump classifier is a single-level decision tree with one node. Terminal nodes extend directly off of this node, thus classification is made based on a single attribute.[19]	1.000	1.000	
FT	Functional trees are classification trees that can use multiple linear regression or multiple logistic regression at decision nodes and linear models at leaf nodes.[17]	0.000	1.000	1.000
J48	J48 is a Java implementation of the C4.5 algorithm; it generates either pruned or an unpruned or C4.5 decision tree. C4.5 build trees from training data using the concept of information entropy.[20]	0.200	0.998	0.994
J48graft	This class generates a grafted C4.5 decision tree that can either be pruned or unpruned. Grafting adds nodes to already created decision trees to improve accuracy.[21]	0.333	1.000	0.992
Jrip	This classifier is an optimized version of the Incremental Reduced Error Pruning, implementing a propositional learner RIPPER (Repeated Incremental Pruning to Produce Error Reduction).[22]	0.333	0.995	0.987
LADTree	LADTree produces a multi-class ADTree. It has the capability to have more than two class inputs. It performs additive logistic regression using the LogitBoost strategy.[23]	0.133	0.997	0.994
LMT	Logistic model trees combine decision trees with logistic regression models. LMTs are generated by creating a logistic model at the root using LogitBoost. The tree is extended at child nodes by using LogitBoost. Nodes are split until no additional split can be found.[24]	0.133	1.000	0.997
Nnge	Nearest neighbor algorithms define a distance function to separate classes. By using generalized exemplars, it reduces the role of the distance function (relying too heavily on the distance function can produce inaccurate results) by grouping classes together.[25]	0.200	0.998	0.994
OneR	This algorithm finds association rules. It finds the one attribute that classifies instances so as to reduce prediction errors.[26]	0.400	0.993	0.984
PART	A set of rules is generated using the 'divide-and-conquer' strategy. From here, all instances in the training data that are covered by this rule get removed and this process is repeated until no instances remain.[27]	0.200	1.000	0.995
Random Tree	The Random Tree classifier draws trees at random from a set of possible trees with *k* random features at each node and performs no pruning.[28]	0.400	0.987	0.978
REPTree	An REPTree is a fast decision tree learner that constructs a decision/regression tree using information gain for splitting, and prunes the tree using reduced-error pruning with backfitting.[29]	0.467	0.998	0.987
Ridor	This classifier is an implementation of a Ripple-Down Rule Learner. An example of this is when the classifier picks a default rule (based on the least weighted error) and creates exception cases stemming from this one.[30]	0.267	0.997	0.990
Simple Cart	Classification and regression trees are used to construct prediction models for data. They are made by partitioning the data and fitting models to each partition.[31]	0.667	0.992	0.976

Abbreviations: ADTree, alternating decision tree; FPR, false positive rate; FT, functional tree; TPR, true positive rate.
The FPR and TPR are provided along with the overall accuracy. The ADTree and the FT, both performed with 100% accuracy. The ADTree contained fewer items (eight in ADTree compared with nine in the FT) and was selected for further analysis in our study.

Balancing classes through simulation. Because machine-learning algorithms maximize performance criteria that place equal weight on each data point without regard to class distinctions, we elected to simulate controls to increase the number of score sheets that would correspond to an ADOS classification of non-spectrum. This enabled us to test whether the imbalance in the classes autism and non-spectrum inadvertently introduced biases that would skew downstream results and interpretation. To create a simulated control, we randomly sampled scores from the existing set of 15 controls, that is, the total number of individuals who did not meet the criteria for a classification of autism or autism spectrum in all the three studies. We did this for each of the 29 items in the ADOS Module 1 by randomly drawing from the set of recorded scores for that item. This guaranteed that the simulated scores were drawn from the same distribution of observed scores. This process was repeated 1000 times to create artificial controls that were subsequently used to further challenge the specificity of the classifier, that is, its ability to correctly categorize individuals with atypical development or apparent risk of neurodevelopmental delay but not on the autism spectrum. We also utilized the simulated controls to recreate a classifier based on completely balanced data, 612 observed ADOS score sheets for individuals categorized as having autism and 612 individuals (15 observed + 597 simulated) not meeting ADOS autism or autism spectrum cutoffs. Additionally, we simulated controls based on the full set of answers that would correspond to a classification of non-spectrum rather than restricting to the observed distribution alone. These simulated controls yielded the same results as those above and thus we elected to use the former simulated controls for imbalance

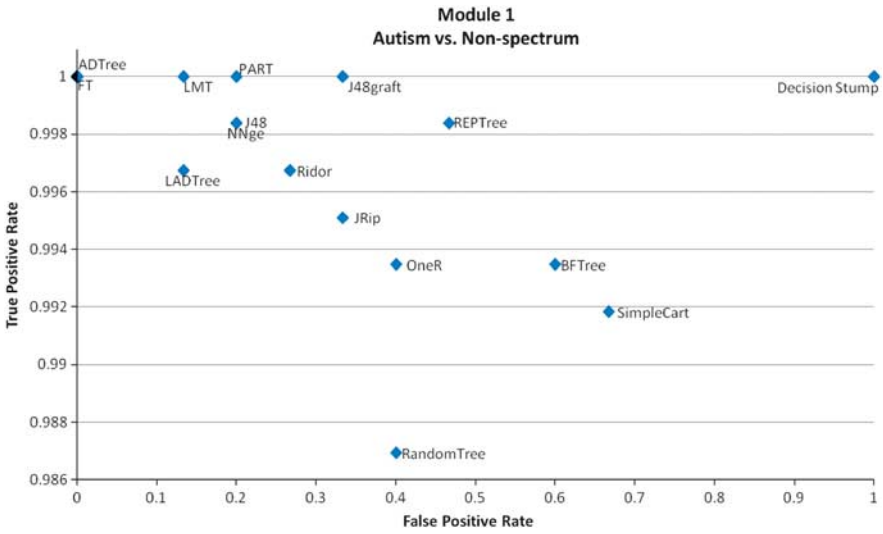

Figure 1 Receiver operating characteristic curves mapping sensitivity versus specificity for the 16 different machine-learning algorithms tested on the ADOS Module 1 training data. We identified the best classifiers as those closest to the point (1, 0) on the graph indicating perfect sensitivity (true positive rate) and one specificity (false positive rate). The best performing models were the ADTree and functional tree (FT). The ADTree was chosen over the FT because it used less items. See Table 2 for a summary of the 16 machine-learning algorithms used in our analysis.

class analysis and for measurements of the specificity of the classifier.

Results

Because the AGRE data contained only 11 controls for Module 1, we included all other Module 1 individuals with a classification of non-spectrum from the Boston AC in the analysis to bring the total number of controls to 15. We found that this improved the accuracy of our classifier when compared with the accuracy of only using the 11 controls alone. We then tested the performance of 16 different machine-learning algorithms on the 29 items in Module 1 (Table 2). We chose the best algorithm by comparing the sensitivity, specificity and accuracy (Figure 1). Two algorithms, the alternating decision tree (ADTree)[16] and the functional tree,[17] operated with perfect sensitivity, specificity and accuracy, resulting in classifiers with 8 and 9 questions, respectively. Because it was our goal to shorten the exam without appreciable loss of accuracy, we selected the ADTree as the optimum algorithm for further analysis and validation. The ADTree classifier correctly classified all 612 individuals from the AGRE who previously received a diagnosis of autism by the ADOS Module 1 algorithm, as well as all 15 individuals from the AGRE and AC who were given a classification of non-spectrum. The ADTree classifier consisted of 8 items out of the 29 used in the analysis and included A2, B1, B2, B5, B9, B10, C1 and C2 (Table 3).

These eight items segregated into two of the three main functional domains associated with autism, language/communication and social interactions, both important indicators of autism. Item A2 (Frequency of Vocalization Directed to Others) corresponded to the language and communication domain. Items B1 (Unusual Eye Contact), B2 (Responsive Social Smile), B5 (Shared Enjoyment in Interaction), B9 (Showing) and B10 (Spontaneous Initiation of Joint Attention) corresponded to the domain of social interaction. Items C1 (Functional Play with Objects) and C2 (Imagination/Creativity) were designed to assess how the subject plays with objects. The eight items formed the elements of a decision tree that enabled classification of either autism or non-spectrum (Figure 2). Two items appeared more than once in the tree (B9 and B10), indicating the possibility that these items have a relatively more important role in classification of autism and that the domain of social interaction may have more utility in observation-based screening and diagnosis of autism. Each item in the tree either increased or decreased a running total statistic known as the ADTree score. A negative score indicated a classification of autism, whereas a positive score yielded the classification of non-spectrum. Importantly, the

Table 3 The eight items found in the ADTree classifier

Question code	*Question subject*	*Core domain*
A2*	Frequency of vocalization directed to others	Communication
B1*	Unusual eye contact	Social interaction
B2	Responsive social smile	Social interaction
B5*	Shared enjoyment in interaction	Social interaction
B9*	Showing	Social interaction
B10*	Spontaneous initiation of joint attention	Social interaction
C1	Functional play with objects	Play
C2	Imagination/creativity	Play

Abbreviation: ADTree, alternating decision tree.
Listed are the question code used by the Autism Genetic Research Exchange (AGRE), a brief description of the question, and the domain to which the question belongs. Five of the items in the ADTree classifier (*) are found on the Autism Diagnostic Observation Schedule-Generic (ADOS) revised algorithm (Gotham *et al*[11]), an algorithm containing 14 total items and demonstrating high diagnostic validity.

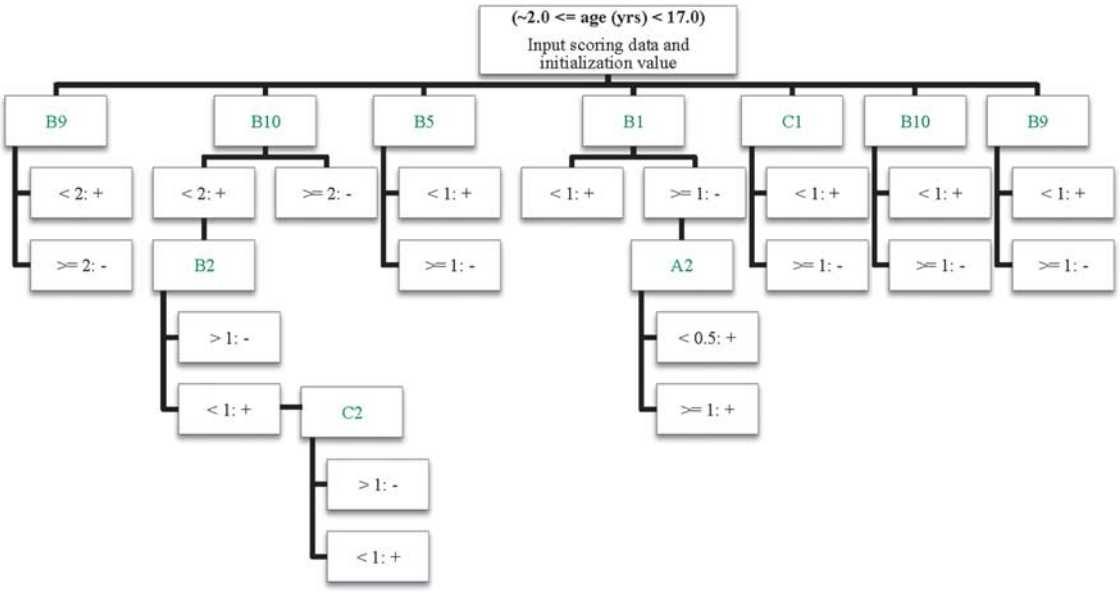

Figure 2 Diagrammatic representation of the classifier generated by the ADTree algorithm. The ADTree was found to perform best out of the 16 different machine-learning approaches (Figure 1, Table 2). The resulting tree enables one to follow each path originating from the top node and increment (+) or decrement (−) prediction variables accordingly. In our case, variables with a negative sign yielded the classification of autism, whereas those with a positive sign resulted in a classification of non-spectrum. The magnitude of the score corresponded to confidence in the class prediction.

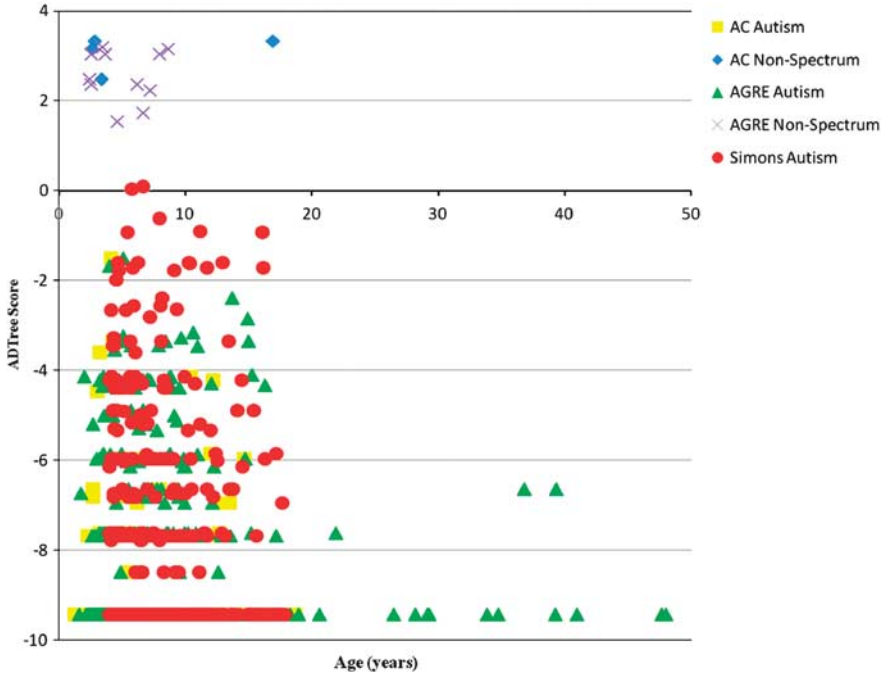

Figure 3 The ADTree scores of individuals in the AGRE, Boston AC and SSC data sets plotted against age in years (range from 13 months to 49 years). A majority of the ADTree scores were large, indicating confidence in the class predictions, and uncorrelated with the ages of the individuals.

amplitude of the score provided a measure of confidence in the classification outcome, with larger absolute values indicating higher confidence overall, as previously indicated in Freund and Mason.[16] In our study, the vast majority of the scores were away from the borderline for both the case and control classes (Figure 3), demonstrating that a majority of the predictions made by the classifier were robust and unambiguous.

For independent validation of our eight-question classifier, we collated score sheets for Module 1 from the Boston AC and

SSC. Here the objective was to determine if the classifier could correctly recapitulate the classification, i.e., autism versus non-spectrum, provided by the ADOS assessments of the individuals recruited to these two independent studies. The classifier correctly classified all 110 individuals previously meeting cutoffs for autism in AC. The classifier also performed with high accuracy on the SSC dataset misclassifying only 2 of the 336 individuals given a classification of autism in the original SSC (99.7% accuracy). Upon further examination of the two misclassified individuals from SSC, we learned that their ADTree scores were approximately zero, at 0.1 and 0.039. The low scores, corresponding to low statistical confidence in the classifications, suggested inadequate classifier power and the potential presence of non-spectrum behaviors in the misclassified subjects themselves.

Because of the limited number of controls who received any ADOS Module, we elected to simulate 1000 controls by randomly sampling from the group of observed answers in the 15 individuals classified as non-spectrum. This procedure enabled us to construct a series of artificial score sheets for the ADOS Module 1 that were within the bounds of answers likely to be provided by prospectively recruited individuals who would not receive a diagnosis of autism following an ADOS exam. The classifier correctly classified 944 out of the 1000 simulated controls (94.4% accuracy). Upon closer inspection of the 56 simulated individuals misclassified with autism, we found that all but 6 had ADTree scores less than one unit away from the classification of non-spectrum (Figure 3).

Because of the small number of controls and the imbalance between the numbers of cases and controls, we elected to perform a machine-learning procedure called upsampling to assess and rule out biases in the original classifier. Upsampling balances the numbers of cases and controls by progressive sampling from the population of observed data. We constructed a classifier using the ADTree algorithm with the 612 individuals with a classification of autism from the AGRE and 612 individuals with a classification of non-spectrum, of which 11 were from the AGRE, 4 were from the AC and the remaining 597 were from the simulated controls. The resulting classifier correctly classified 609 out of the 612 individuals with autism and all 612 individuals with a classification of non-spectrum (99.8% accuracy). The resulting ADTree consisted of seven items, six of which were also in the original classifier derived from the imbalanced data. Additionally, the ensuing ADTree remained largely unchanged from the original (data not shown), lending further support to the robustness of our classifier and supporting the notion that the imbalance of classes did not bias our results.

Discussion

Current practices for the behavioral diagnosis of autism can be effective but in many cases overly prohibitive and time consuming. One of the most trusted and widely used instruments in the field of autism spectrum disorders is the ADOS, an exam broken up into four modules to accommodate varying developmental level and language ability. We used machine-learning techniques to determine if we could achieve high classification accuracy with a small selection of items from the exam. In our case, several alternative machine-learning strategies yielded classifiers with near perfect accuracy and low rates of false positives. The top-performing ADTree algorithm resulted in an eight-item classifier with 99.7% sensitivity and 94% specificity when tested across 1058 individuals with autism and a collection of 1000 simulated and 15 observed non-spectrum controls. The ADTree algorithm resulted in a simple decision tree (Figure 2) with potential value for use in screening and/or clinical diagnostic settings.

The ADTree classifier contains five questions also found on the ADOS revised algorithm[11] (Table 3), suggesting that our classifier retains at least some of the diagnositic validity of this 14-item algorithm. Additionally, the classifier results in a quantitative score that is a direct measurement of both classification confidence as well as severity (or extremeness) of phenotype. Therefore, this score represents an empirical measure of confidence in the classification that can flag borderline cases warranting closer inspection and further behavioral assessment. The ADTree score may also be integrated with other instruments, for example, Social Responsiveness Scale, to enrich content while keeping diagnosis time frames short. In addition, as a quantitative measure of phenotype, the ADTree score could be integrated with genetic data to improve our understanding of the genotype–phenotype map for autism over a diversity of subjects.

The statistical reduction in the number of items from the ADOS Module 1 suggests that a compatible reduction in the activities associated with the exam is possible. Module 1 contains 10 activities (Table 4), each designed to elicit specific behaviors and responses that are coded in the 29 items. Considering only the 8 items in our classifier, 2 of the 10 activities, namely 'response to name' and 'response to joint attention,' could be removed because neither is required for the eight-question classifier (Table 4). How this or other alterations could have an impact on the observation process overall remains an open research question, but as our clinical and research databases expand together with our abilities to refine machine-learning approaches like the one described here, it is conceivable that further statistical reductions that

Table 4 The 10 activities used in an observation of a subject to answer the 29 items found on the ADOS Module 1

Activity	*Needed for classifier?*
Free play	Yes
Response to name	No
Response to joint attention	No
Bubble play	Yes
Anticipation of a routine with objects	Yes
Responsive social smile	Yes
Anticipation of a social routine	Yes
Functional and symbolic imitation	Yes
Birthday party	Yes
Snack	Yes

Abbreviation: ADOS, Autism Diagnostic Observation Schedule-Generic.
Our work resulted in an accurate classifier containing only eight items from the full test. In all, 2 of the 10 activities would not be needed to use this classifier in an evaluation of a subject.

enable rapid detection with high accuracy will be discovered. In a similar vein, we anticipate that our classifier and potentially others realized through similar studies on different instruments and databases (clinical and research) will inform the development of mobile tools for preliminary evaluation and clinical prioritization—in particular those focused on assessment of short home videos of children (for example, http://vid.autworks.hms.harvard.edu)—that speed the pace of initial evaluation and broaden the reach to a significantly larger percentage of the population at risk.

Limitations. Our study was limited by the content of existing repositories that, for reasons related to the recruitment processes of those studies, contain very few individuals who did not meet the criteria for an autism classification. In a prospective design for a study like ours, we would include equal numbers of cases and controls for optimal calculations of sensitivity and specificity of the classifier. Going forward, we hope to expand our work through the inclusion of new ADOS Module 1 (and other modules) data from both individuals with autism spectrum disorders and individuals without autism, particularly non-spectrum individuals with learning delays and neurodevelopmental conditions, to appropriately challenge the specificity and better reflect the population of cases seen in clinical environments.

Again because of limitations in available data, our classifier was trained only on non-spectrum individuals and those with classic autism. Therefore, we were not able to test whether our classifier could accurately distinguish between autism, Asperger's syndrome and pervasive developmental disorder-not otherwise specified. Nevertheless, those individuals not meeting the formal criteria for autism diagnosis were generally recruited to the study as high-risk individuals or as siblings of an individual with autism. Thus, these controls may have milder neurodevelopmental abnormalities that correspond to other categories outside of classic autism. Given that our classifier generally performed well at distinguishing these individuals from those with classic autism supports the possibility that our classifier already has inherent sensitivity to behavioral variation within and outside of the autism spectrum. Additional ADOS data from a range of individuals with autism spectrum disorders and importantly non-spectrum individuals with other learning and developmental delays would enable us to measure the value beyond that of classic autism, as well as enable us to retrain the classifier to improve both sensitivity and specificity.

Conclusions

Currently, autism spectrum disorder is diagnosed through behavioral exams and questionnaires that require significant time investment for both parents and clinicians. In our study, we performed a data-driven approach to select a reduced set of questions from one of the most widely used instruments for behavioral diagnosis, the ADOS. Using machine-learning algorithms, we found the ADTree to perform with almost perfect sensitivity, specificity and accuracy in distinguishing individuals with autism from individuals without autism. The ADTree classifier consisted of eight questions, 72.4% less than the complete ADOS Module 1, and performed with $>99\%$ accuracy when applied to independent populations of individuals with autism, misclassifying only 2 out of 446 cases. Given this reduction in the number of items without appreciable loss in accuracy, our findings may help to guide future efforts, chiefly including mobile health approaches, to shorten the evaluation and diagnosis process overall such that families can receive care earlier than under current diagnostic modalities.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements. We would like to thank members of the Tonellato and Wall labs for key input on study design and results interpretation, including Rebecca Dally for help with data management and Tristan Nelson for assistance with database queries. We thank the AGRE, SSC and AC projects for access to data, as well as the families enrolled in these projects for their invaluable contributions. We thank Vlad Kustanovich for assistance with downloading and handling the AGRE data. We also thank Rhiannon Luyster for comments on an earlier draft.

1. Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E *et al.* Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med* 1995; **25**: 63–77.
2. Lord C, Risi S, Lambrecht L, Cook Jr 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.
3. 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.
4. 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**(2 Suppl): S79–S87.
5. Bernier R, Mao A, Yen J. Psychopathology, families, and culture: autism. *Child Adolesc Psychiatr Clin N Am* 2010; **19**: 855–867.
6. Howlin P. *Children with Autism and Asperger's Syndrome: A Guide for Practitioners and Parents*. Wiley: Chichester, UK, 1998.
7. 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.
8. Berument SK, Rutter M, Lord C, Pickles A, Bailey A. Autism screening questionnaire: diagnostic validity. *Br J Psychiatry* 1999; **175**: 444–451.
9. Robins DL, Fein D, Barton ML, Green JA. The Modified Checklist for Autism in Toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. *J Autism Dev Disord* 2001; **31**: 131–144.
10. Eaves LC, Wingert HD, Ho HH, Mickelson EC. Screening for autism spectrum disorders with the social communication questionnaire. *J Dev Behav Pediatr* 2006; **27**(2 Suppl): S95–S103.
11. 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.
12. 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.
13. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: an update. *SIGKDD Explorations* 2009; **11**: 1.
14. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Mellish CS (ed). In *Proceedings IJCAI-95: 1995 Montreal*. Morgan Kaufmann: Los Altos, CA, 1995, pp 1137–1143.
15. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* 2010; **68**: 192–195.
16. Freund Y, Mason L. The alternating decision tree learning algorithm. In *Machine Learning: Proceedings of the Sixteenth International Conference*. Morgan Kaufmann Publishers Inc., 1999, pp 124–133.
17. Gama J. Functional Trees. *Machine Learning* 2004; **55**: 219–250.
18. Shi H. Best-first Decision Tree Learning. In *Master Thesis, The University of Waikato*, Hamilton, NZ, 2007.
19. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kaufmann: San Francisco, 1996, pp 148–156.
20. Quinlan R. *C4.5*. Morgan Kaufmann Publishers: San Mateo, 1993.

21. Webb GI. Decision tree grafting. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann: Nagoya, Japan, 1997, pp 846–851.
22. Cohen WW. Fast Effective Rule Induction. In *Machine Learning*: *The 12th International Conference*, Lake Tahoe. Morgan Kaufmann: CA, 1995, pp 115–123.
23. Holmes G, Pfahringer B, Kirkby R, Frank E, Hall M. Multiclass alternating decision trees. *ECML* 2001; 161–172.
24. Landwehr N, Hall M, Frank E. Logistic Model Trees. *Machine Learning* 2005; **59**: 161–205.
25. Martin B. *Instance-Based learning: Nearest Neighbor With Generalization*. University of Waikato: Hamilton, New Zealand, 1995.
26. Holte RC. Very simple classification rules perform well on most commonly used datasets. *Machine Learning: Proceedings of the Sixteenth International Conference* 1993; **11**: 63–91.
27. Frank E, Witten IH. Generating Accurate Rule Sets Without Global Optimization. In: *Machine Learning: Proceedings of the Fifteenth International Conference*. Morgan Kaufmann Publishers: San Francisco CA, 1998.
28. Brieman L. Random Forest. *Machine Learning* 2001; **45**: 5–32.
29. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Amsterdam: Morgan Kaufman, 2005.
30. Gaines BR, Compton P. Induction of Ripple-Down Rules Applied to Modeling Large Databases. *J Intell Inf Syst* 1995; **5**: 211–228.
31. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont: California, 1984.