Journal of the American Medical Informatics Association, 25(8), 2018, 1000–1007
doi: 10.1093/jamia/ocy039
Advance Access Publication Date: 7 May 2018
Research and Applications

Research and Applications

Machine learning approach for early detection of autism by combining questionnaire and home video screening

Halim Abbas,[1] Ford Garberson,[1] Eric Glover,[2] and Dennis P Wall[1,3,4]

[1]Cognoa Inc., Palo Alto, CA, USA www.linkedin.com/in/halimabbas, [2]eric_g@ericglover.com, [3]Department of Pediatrics, Stanford University, Stanford, CA, USA, [4]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

Correspondence to: Cognoa Inc., Palo Alto, CA, USA; halim@cognoa.com

Received 19 September 2017; Revised 16 March 2018; Editorial Decision 25 March 2018; Accepted 2 April 2018

ABSTRACT

Background: Existing screening tools for early detection of autism are expensive, cumbersome, time- intensive, and sometimes fall short in predictive value. In this work, we sought to apply Machine Learning (ML) to gold standard clinical data obtained across thousands of children at-risk for autism spectrum disorder to create a low-cost, quick, and easy to apply autism screening tool.
Methods: Two algorithms are trained to identify autism, one based on short, structured parent-reported questionnaires and the other on tagging key behaviors from short, semi-structured home videos of children. A combination algorithm is then used to combine the results into a single assessment of higher accuracy. To overcome the scarcity, sparsity, and imbalance of training data, we apply novel feature selection, feature engineering, and feature encoding techniques. We allow for inconclusive determination where appropriate in order to boost screening accuracy when conclusive. The performance is then validated in a controlled clinical study.
Results: A multi-center clinical study of $n = 162$ children is performed to ascertain the performance of these algorithms and their combination. We demonstrate a significant accuracy improvement over standard screening tools in measurements of AUC, sensitivity, and specificity.
Conclusion: These findings suggest that a mobile, machine learning process is a reliable method for detection of autism outside of clinical settings. A variety of confounding factors in the clinical analysis are discussed along with the solutions engineered into the algorithms. Final results are statistically limited and will benefit from future clinical studies to extend the sample size.

Key words: supervised machine learning, autism spectrum disorder, diagnostic techniques and procedures, mobile applications

INTRODUCTION

Diagnosis within the first few years of life dramatically improves the outlook of children with autism, as it allows for treatment while the child's brain is still rapidly developing.[1,2] Unfortunately, autism is typically not diagnosed earlier than age 4 in the United States, with approximately 27% of cases remaining undiagnosed at age 8.[3] This delay in diagnosis is driven primarily by a lack of effective screening tools and a shortage of specialists to evaluate at-risk children. The use of higher accuracy screening tools to prioritize children to be seen by specialists is therefore essential.

Most autism screeners in use today are based on questions for the parent or the medical practitioner, that produce results by comparing summed answer scores to predetermined thresholds. Notable examples are the Modified Checklist for Autism in Toddlers, Revised (M-CHAT),[4] a checklist-based screening tool for autism that is intended to be administered during developmental screenings for children between the ages of 16 and 30 months, and the Child Behavior Checklist (CBCL).[5] Both are parent-completed screening tools. For both instruments, responses to each question are summed with each question given equal weighting, and if the total is above a pre-determined threshold the child is considered to be at high risk of

autism. In the case of CBCL there are multiple scales based upon different sets of questions corresponding to different conditions. The "Autism Spectrum Problems" scale of CBCL is used when comparing its performance to the performances of our algorithms in this paper.

In this paper, we present two new machine learning screeners that are reliable, cost-effective, short enough to be completed in minutes, and achieve higher accuracy than existing screeners on the same age span as existing screeners. One is based on a short questionnaire about the child, which is answered by the parent. The other is based on identification of specific behaviors by trained analysts after watching two or three short videos of the child within their natural environment that are captured by parents using a mobile device.

The parent questionnaire screener keys on behavioral patterns similar to those probed by a standard autism diagnostic instrument, the Autism Diagnostic Interview – Revised (ADI-R).[6] This clinical tool consists of an interview of the parent with 93 multi-part questions with multiple choice and numeric responses which are delivered by a trained professional in a clinical setting. While this instrument is considered a gold-standard, and gives consistent results across examiners, the cost and time to administer it can be prohibitive in a primary care setting. In this paper, we present our approach to using clinical ADI-R instrument data to create a screener based on a short questionnaire presented directly to parents without supervision.

The video screener keys on behavioral patterns similar to those probed in another diagnostic tool, the Autism Diagnostic Observation Schedule (ADOS).[7] ADOS is widely considered a gold standard and is one of the most common behavioral instruments used to aid in the diagnosis of autism.[8] It consists of an interactive and structured examination of the child by trained clinicians in a tightly controlled setting. ADOS is a multi-modular diagnostic instrument, with different modules for subjects at different levels of cognitive development. In this paper, we present our approach to mining ADOS clinical records, with a focus on younger developmental age, to create a video-based screener that relies on an analyst evaluating short videos of children filmed by their parents at home.

The use of behavioral patterns commonly probed in ADI-R and ADOS scoresheets as inputs to train autism screening classifiers was introduced, studied, and clinically validated in previous work.[9–12] There are several new aspects in this paper. First, the algorithms detailed in the present study have been designed to be more accurate and more robust against confounding biases between training and application data. Next, this paper focuses considerable attention on the impact of confounding factors on machine learning algorithms in this context. Examples of these confounding biases will be discussed below and highlighted in Table 2. Labeled data usually originates from tightly controlled clinical environments and is, hence, clean but sparse, unbalanced, and of a different context to the data available when applying the screening techniques in a less formal environment. This paper also presents a combination between the algorithms for a more powerful single screener. Lastly, this paper generalizes the algorithms to be non-binary, sometimes resulting in an "inconclusive" determination when presented with data from more challenging cases. This allows higher screening accuracy for those children who do receive a conclusive screening, while still presenting a clinically actionable inconclusive outcome in the more challenging cases.

These classifiers of this paper were applied to screen children in a clinical study using the Cognoa[13] App. To date, Cognoa has been used by over 250 000 parents in the US and internationally. The majority of Cognoa users are parents of young children between 18 and 30 months. The clinical study consisted of 162 at-risk children who had undergone full clinical examination and received a clinical diagnosis at a center specialized in neurodevelopmental disorders.

METHODS

It is not feasible to amass large training sets of children who have been evaluated by the mobile screeners and who also have received a professional medical diagnosis. Our approach is to start with historical medical instrument records of previously diagnosed subjects, and use those as training data for screeners that will rely on information acquired outside the clinical setting. Expected performance degradation from applying the algorithms into a less controlled setting would result in inaccurate screeners if conventional machine learning methods were used. Much of this paper outlines the details of creative machine learning methods designed to overcome this challenge and create reliable screeners in this setting.

Training data were compiled from multiple repositories of ADOS and ADI-R score-sheets of children between 18 and 84 months of age including Boston Autism Consortium, Autism Genetic Resource Exchange, Autism Treatment Network, Simons Simplex Collection, and Vanderbilt Medical Center. Since such repositories are highly imbalanced with very few non-autistic patients, the controls across the datasets were supplemented with balancing data obtained by conducting ADI-R interviews by a trained clinician on a random sample of children deemed at low risk for autism from Cognoa's user base. For both algorithms a smaller set of optimal features was selected using methods that will be discussed below. Details about the final selected features are given in the Supplementary Material.

The clinical validation sample consists of 230 children who presented to one of three autism centers in the United States between 18 and 72 months of age. All participants were referred through the clinics' typical referral program process, and only those with English-speaking parents were considered for the study. The three clinical centers were approved on a multisite IRB (project number 2202803). Every child received an ADOS as well as standard screeners like M-CHAT and CBCL as appropriate, and a diagnosis was ultimately ascertained by a licensed health care provider. For 162 of those children, the parents also used their mobile devices to complete the short parental questionnaire and submit the short videos required for the screeners discussed in this paper. The sample breakdown by age group and diagnosis for both the training and clinical validation datasets is shown in Table 1.

Approach

We trained two independent ML classifiers and combined their outputs into a single screening assessment. The parent questionnaire classifier was trained using data from historical item-level ADI-R score-sheets with labels corresponding to established clinical diagnoses. The video classifier was trained using ADOS instrument scoresheets and diagnostic labels. In each case, progressive sampling was used to verify sufficient training volume as detailed in the Supplementary Materials. Multiple machine learning algorithms were evaluated including ensemble techniques on the training data. A number of algorithms performed well. Random Forests were chosen because of robustness against overfitting.

ADI-R and ADOS instruments are designed to be administered by trained professionals in highly standardized clinical settings and typically take hours. In contrast, our screening methods are deliberately

Table 1. Dataset Breakdown by Age Group and Condition Type for Each of the Sources of Training Data and for the Clinical Validation Sample. The Negative Class Label Includes Normally Developing (i.e. neurotypical) Children as Well as Children with Developmental Delays and Conditions other than Autism

Age (years)	Condition	Classification type Questionnaire	Number of samples		
			training	Video training	Clinical validation
< 4	Autism	+	414	1445	84
< 4	Other condition	−	133	231	18
< 4	Neurotypical	−	74	308	3
≥ 4	Autism	+	1885	1865	37
≥ 4	Other condition	−	154	133	11
≥ 4	Neurotypical	−	26	277	9

Table 2. Differences Between Training and Application Environments. These Differences are Expected to Cause Bias that Cannot be Captured by Cross-validation Studies

Aspect	Training Setting	Application Setting
Source	ADI-R and ADOS instrument administered by trained professionals during clinical eval-uations	Short parent questionnaires displayed on smartphone, and behavior tagging by ana-lysts after observing two or three 1-minute home videos uploaded by parents
Proctor	Highly trained medical professionals	Parents answering the questionnaires are un-trained, and the analysts evaluating the home videos are only minimally trained. As a result, their answers may not be as consistent, objective, or reliable
Setting	Clinic setting with highly standardized and semi-structured interactions	At home. Not possible to recreate the structured clinical environment, resulting in an undesired variability of the output signals. Subjects might also behave differently at the clinic than at home, further amplifying the bias
Duration	The ADI-R can take up to 4 hours to complete; The ADOS can take up to 45 minutes of direct observation by trained professionals	Under 10 minutes to complete the parent questionnaire, and a few minutes of home video. As a result, some symptoms and behavioral patterns might be present but not observed. Also causes big uncertainty about the severity and frequency of observed symptoms
Questionnaires	Sophisticated language involving psychological concepts, terms, and subtleties unfamiliar to nonexperts	Simplified questions and answer choices result in less nuanced, noisier inputs

designed to be administered at home by parents without expert supervision, and to take only minutes to complete. This change of environment causes significant data degradation and biases resulting in an expected loss of screening accuracy. For each classifier, we present mindful adjustments to ML methodology to mitigate these issues. These biases and efforts to mitigate them are discussed below.

Differences Between Training and Application Environments

The screeners are trained on historical patient records that correspond to controlled, lengthy clinical examinations, but applied via web or mobile app aimed at unsupervised parents at home. Table 2 details the various mechanisms by which confounding biases may consequently creep into the application data. Note that inaccuracies introduced by such biases cannot be probed by cross- validation or similar analysis of the training data alone.

Hyperparameter Optimization

For each parental questionnaire and video model that will be discussed below, model hyperparameters were tuned with a bootstrapped grid search. In all cases, class labels were used to stratify the folds, and (age, label) pairs were used to weight-balance the samples. More details can be found in the Supplementary Materials.

Parent Questionnaire

Multiple model variants representing incremental improvements over a generic ML classification approach are discussed below.

Generic ML Baseline Variant

A random forest was trained over the ADI-R instrument data. Each of the instrument's 155 data columns was treated as a categorical variable and one-hot encoded. The subject's age and gender were included as features as well. Of the resulting set of features, the top 20 were selected using feature-importance ranking in the decision forest.

Robust Feature Selection Variant

Due to the small size and sparsity of the training dataset, generic feature selection was not robust, and the selected features (along with the performance of the resulting model) fluctuated from run to run due to the stochastic nature of the learner's underlying bagging approach. Many ADI-R questions are highly correlated, leading to multiple competing sets of feature selection choices that were seemingly equally powerful during training, but which had different performance characteristics when the underlying sampling bias was exposed via full bootstrapped cross-validation. This resulted in a wide performance range of the variant of the Generic ML baseline method as shown in Table 3.

Table 3. Performance of Increasingly Effective Classifier Variants Based on the Training Data for the Parent Questionnaire. Results in the Top Table are Based on Cross-validated Training Performance. Results in the BottomTable (which are only available for variants using the optimally selected features) are Based on Actual Clinical Results

	AUC			Sensitivity			Specificity		
	All ages	< 4 years	>= 4 years	All ages	< 4 years	>= 4 years	All ages	< 4 years	>= 4 years
Training scenario									
Generic ML baseline	0.932 to 0.950	0.928 to 0.953	0.928 to 0.953	0.976 to 0.982	0.975 to 0.984	0.975 to 0.984	0.628 to 0.645	0.625 to 0.648	0.625 to 0.648
Robust feature selection variant	0.958	0.958	0.958	0.982	0.982	0.982	0.624	0.624	0.624
Age silo variant	0.953	0.939	0.961	0.962	0.939	0.977	0.777	0.774	0.779
Severity-level feature encoding variant	0.965	0.950	0.974	0.962	0.912	0.993	0.748	0.833	0.692
Aggregate features variant	0.972	0.987	0.963	0.992	0.988	0.994	0.754	0.894	0.661
With inconclusive allowance [up to 25\%]	0.991	0.997	0.983	1.000	1.000	1.000	0.939	0.977	0.881
Application scenario									
Age silo variant	0.62	0.68	0.54	0.65	0.62	0.52	0.48	0.46	0.24
Severity-level feature encoding variant	0.67	0.69	0.64	0.64	0.62	0.58	0.48	0.46	0.33
Aggregate features variant	0.68	0.73	0.68	0.68	0.69	0.65	0.57	0.62	0.48
With inconclusive allowance [up to 25\%]	0.72	0.72	0.73	0.70	0.72	0.67	0.67	0.71	0.53

Robust feature selection overcame that limitation using a two-step approach. First, a 100-count bootstrapped feature selection was run, with a weight balanced 90% random sample selected in each iteration. The top 20 features were selected each time, and a rank-invariant tally was kept for the number of times each feature made it to a top-20 list. Next, the top 30 features in the tally were kept as candidates and all other features were discarded. A final feature-selection run was used to pick the best subset of these candidate features. This approach was found to be more robust to statistical fluctuations, usually selecting the same set of features when run multiple times. A minimal subset of maximally performant features was chosen and locked for clinical validation, totaling 17 features for the young children and 21 features for the old. Details about these selected features are available in the Supplementary Material.

Age Silo Variant

This variant built upon the improvements of the robust feature selection method, by exploiting of the dichotomy between pre-phrasal and fully-phrasal language capability in at-risk children. Language development is significant in this domain as it is known to affect the nature in which autism presents, and consequently the kinds of behavioral clues to look for in order to screen for it.

This variant achieved better performance by training separate classifiers for children in the younger and older age groups of Table 1. The age dichotomy of $<4, \geq 4$ was chosen to serve as the best proxy for language ability. Feature selection, model parameter-tuning, and cross-validation were run independently for each age group classifier. Before siloing by age group, the classifier was limited to selecting features that work well across children of both developmental stages. Siloing enabled the classifiers to specialize on features that are most developmentally appropriate within each age group.

Severity-level Feature Encoding Variant

Building upon the method including age siloing above, this variant achieved better performance by replacing one-hot feature encoding with a more context-appropriate technique. One-hot encoding does not distinguish between values that correspond to increasing levels of severity of a behavioral symptom, and values that do not convey a clear concept of severity. This is especially troublesome since a typical ADI-R instrument question includes answer choices from both types of values. For example, ADI-R question 37, which focuses on the child's tendency to confuse and mix up pronouns, allows for answer codes 0, 1, 2, 3, 7, 8, and 9. Among those choices, 0 through 3 denote increasing degrees of severity in pronominal confusion, while 7 denotes any other type of pronominal confusion not covered in 0-3 regardless of severity. Codes 8 and 9 denote the non-applicability of the question (for example, to a child still incapable of phrasal speech) or the lack of an answer (for example, if the question was skipped) respectively. When coding the answers to such questions, generic one-hot encoding would allow for non-symptomatic answer codes to be selected as screening features based on phantom correlations present in the dataset.

Severity-level encoding converts all answer codes that do not convey a relevant semantic concept to a common value, thereby reducing the chance of useless feature selection, and reducing the number of features to choose from. In addition, severity-level encoding condenses the signal according to increasing ranges of severity. For example, the encoding of ADI-R question 37 would map its responses to new features with 1s in the following cases (all other new features would be zero): (0 →"=0," 1→ "1," 2→ ["1," "2"], 3→ ["1," "2," "3"], 7→ "= 7," 8, 9→*None*). This more closely resembles the way medical practitioners interpret such answer choices, and helps alleviate the problem of sparsity over each of the one-hot encoded features in the dataset.

Aggregate Features Variant

Building upon the method including severity level encoding above, this variant achieved better performance by incorporating aggregate

features such as the minimum, maximum, and average severity level, as well as number of answer choices by severity level across the questions corresponding to the 20 selected features. These new features were especially helpful due to the sparse, shallow, and wide nature of the training set, whereupon any semantically meaningful condensation of the signal can be useful to the trained classifier.

Inconclusive Results Variant

Children with more complex symptom presentation are known to pose challenges to developmental screening. These children often screen as false positives or false negatives, resulting in an overall degradation of screening accuracy that is observed by all standard methods and has become acceptable in the industry. Given that our low-cost instruments do not rely on sophisticated observations to differentiate complex symptom cases, our approach was to avoid assessing them altogether, and to try instead to spot and label them as "inconclusive."

Building upon the method including feature engineering, two methods to implement this strategy were devised. The first was to train a binary classifier with a continuous output score, then replace the cutoff threshold with a cutoff range, with values within the cutoff range considered inconclusive. A grid search was used to determine the optimal cutoff range representing a tradeoff between inconclusive determination rate and accuracy over conclusive subjects. The second approach was to train and cross-validate a simple binary classifier, label the correctly and incorrectly predicted samples as conclusive or inconclusive respectively, and then build a second classifier to predict whether a subject would be incorrectly classified by the first classifier. At runtime, the second classifier was used to spot and label inconclusives. The conclusives were sent for classification by a third, binary classifier trained over the conclusive samples only. Both methods for labeling inconclusive results yielded similar performance. Therefore, the simpler method of using a threshold range in the machine learning output was used to report inconclusive results for this paper.

The inconclusive rate is a configurable model parameter that controls the tradeoff between coverage and accuracy. Throughout this paper, the inconclusive rate for this variant was set to 25%.

Video

The second of our two-method approach to autism screening is an ML classifier that uses input answers about the presence and severity of target behaviors among subjects. This information was provided by an analyst upon viewing two or three 1-minute home videos of children in semi-structured settings that are taken by parents on their mobile phones. The classifier was trained on item-level data from two of the ADOS modules (module 1: preverbal, module 2: phrased speech) and corresponding clinical diagnosis.

Two decision forest ML classifiers were trained corresponding to each ADOS module. For each classifier, 10 questions were selected using the same robust feature selection method, and the same allowance for inconclusive outcomes was made as for the parental questionnaire classifier. Each model was independently parameter-tuned with a bootstrapped grid search. Class labels were used to stratify the cross-validation folds, and (age, label) pairs were used to weight-balance the samples.

Problems related to the change of environment from training to application are especially significant in the case of video screening because ADOS involves a 45 minute direct observation of the child by experts, whereas our screening was based on unsupervised short home videos. Specifically, we expect the likelihood of inconclusive or unobserved behaviors and symptoms to be much higher in the application than in the training data, and the assessed level of severity or frequency of observed symptoms to be less reliable in the application than in the training data. The following improvements were designed to help overcome these limitations.

Presence of Behavior Encoding

To minimize potential bias from a video analyst misreading the severity of a symptom in a short cell phone video, this encoding scheme improves feature reliability at the expense of feature information content by collapsing all severity gradations of a question into one binary value representing the presence vs absence of the behavior or symptom in question. Importantly, a value of 1 denotes the presence of behavior, regardless of whether the behavior is indicative of autism or of normalcy. This rule ensures that a value of 1 corresponds to a reliable observation, whereas a 0 does not necessarily indicate the absence of a symptom but possibly the failure to observe the symptom within the short window of observation.

Missing Value Injection to Balance the Nonpresence of Features for the Video Screener Training Data

While collapsing severity gradations into a single category overcomes noisy severity assessment, it does not help with the problem of a symptom not present or unnoticeable in a short home video. For this reason, it is important that the learning algorithm treat a value of 1 as semantically meaningful, and a value of 0 as inconsequential. To this end, we augmented the training set with duplicate samples that had some feature values flipped from 1 to 0. The injection of 0s was randomly performed with probabilities such that the sample-weighted ratio of positive to negative samples for which the value of any particular feature is 0 is about 50%. Such ratios ensure that the trees in a random forest will be much less likely to draw conclusions from the absence of a feature.

Combination

It is desirable to combine the questionnaire and video screeners to achieve higher accuracy. However, the needed overlapping training set was not available. Instead, the clinical validation dataset itself was used to train the combination model.

The numerical responses of each of the parent questionnaire and video classifiers were combined using L2-regularized logistic regression, which has the advantage of reducing the concern of overfitting, particularly given the logistic model has only three free parameters. Bootstrapping and cross -validation studies showed that any overfitting that may be present from this procedure is not detectable within statistical limitations. Since each of the individual methods was siloed by age, separate combination algorithms were trained per age group silo. For each combination algorithm, optimal inconclusive output criteria were chosen using the logistic regression response, using the same techniques as for the parental questionnaire and video classifiers. The performance characteristics of the overall screening process compared to standard alternative screeners are shown below.

RESULTS

Parent Questionnaire Performance on Training Data

Bootstrapped cross-validation performance metrics for the optimally parameter-tuned version of each of the variants of the parental

Table 4. Performance Comparisons Between Various Algorithms on Clinical Data

Base model	Model from this paper	AUC improvement	Mean recall improvement
2012 publication	Questionnaire	0.07, [−0.03, 0.17]	0.1, [0.02, 0.17]
M-CHAT	Questionnaire	0.01, [−0.11, 0.12]	0.06, [−0.04, 0.17]
CBCL	Questionnaire	0.06, [−0.04, 0.17]	0.11, [0.03, 0.2]
2012 publication	Questionnaire & video	0.16, [0.07, 0.25]	0.12, [0.04, 0.2]
M-CHAT	Questionnaire & video	0.08, [−0.03, 0.19]	0.1, [−0.01, 0.21]
CBCL	Questionnaire & video	0.15, [0.04, 0.26]	0.14, [0.04, 0.24]
2012 publication	Questionnaire + inconclusive	0.16, [0.02, 0.28]	0.09, [−0.02, 0.2]
M-CHAT	Questionnaire + inconclusive	−0.01, [−0.39, 0.31]	0.08, [−0.18, 0.29]
CBCL	Questionnaire + inconclusive	0.15, [0.01, 0.29]	0.11, [−0.02, 0.24]
2012 publication	Questionnaire & video + inconclusive	0.21, [0.1, 0.32]	0.19, [0.1, 0.28]
M-CHAT	Questionnaire & video + inconclusive	0.09, [−0.05, 0.23]	0.15, [0.04, 0.27]
CBCL	Questionnaire & video + inconclusive	0.2, [0.09, 0.32]	0.2, [0.09, 0.31]
Questionnaire	Questionnaire & video	0.09, [0.02, 0.15]	0.03, [−0.04, 0.09]
Questionnaire	Questionnaire + inconclusive	0.09, [−0.01, 0.17]	−0.0, [−0.09, 0.08]
Questionnaire	Questionnaire & video + inconclusive	0.14, [0.06, 0.23]	0.09, [0.01, 0.17]
Q. and video	Questionnaire & video + inconclusive	0.06, [0.01, 0.11]	0.06, [0.0, 0.13]

Each row evaluates the improvement of one of the algorithms from this paper over a "Base model" algorithm for the AUC metric, and for the average between the autism and the non-autism recalls at a response threshold point that achieves approximately 80% sensitivity. Negative values would represent a worsening of performance for a given algorithm compared to the base model. Both average values of the improvements and [5%, 95%] confidence intervals are reported. Algorithms that are labeled "inconclusive" allow up to 25% of the most difficult samples to be discarded from the metric evaluation. Note that the M-CHAT instrument is intended for use on younger children. Therefore, older children were excluded when preforming comparisons to M-CHAT in this table.

questionnaire are reported in the top of Table 3. The results for baseline variant are reported as a range rather than a single value, because the unreliability of generic feature selection leads to different sets of features selected from run to run, with varying performance results.

Parents of children included in the clinical study answered short, age-appropriate questions chosen using the robust feature selection method discussed above. The clinical performance metrics for each of the classification variants that build upon that feature selection scheme are shown in the bottom of Table 3. The difference in performance between the training and validation datasets is driven by the differences that are emphasized in Table 2. See below and the results of Table 4 for a discussion of the statistical significance of these results.

ROC curves in Figure 1 show how our parent questionnaire classification approach outperforms some of the established screening tools like MCHAT and CBCL on the clinical sample. Since clinical centers are usually interested in screening tools with a high sensitivity, we have drawn shaded regions between 70% and 90% sensitivity to aid the eye.

Combination Screening Performance on Clinical Data

ROC curves in Figure 2 show how combining the questionnaire and video classifiers into a single assessment further boosted performance on the clinical study sample. When up to 25% of the most challenging cases are allowed to be determined, inconclusive the performance on the remaining cases is shown in Figure 3. Note that the ROC curves in these figures for M-CHAT contain only younger children (mostly under four years of age) due to the fact that this instrument is not intended for older children. A same-sample comparison between M-CHAT and the ML screeners can be seen in the age binned figures (Figures 4 and 5).

Results for Young Children

Young children are of particular interest given the desire to identify autism as early as possible. Results restricted to only children less than four years old are shown in Figures 4 and 5.

Statistical Significance

For the training data, sample sizes are large enough that statistical limitations are minimal. However, results reported for the clinical data have significant statistical limitations. In this section we compare the performance of the screening algorithms on the clinical data that we have discussed in this paper: (1) the questionnaire-based algorithm of,[13] (2) M-CHAT, (3) CBCL, (4) the questionnaire-based algorithm of this paper, and (5) the combined questionnaire plus video algorithm of this paper. Direct comparisons in performance between many of these algorithms are reported along with statistical significances in Table 4.

DISCUSSION

We have introduced a novel machine learning algorithm based on a parental questionnaire and another based on short home videos recorded by parents and scored by a minimally trained analyst. We have discussed pitfalls such as data sparsity, and the mixed ordinal and categorical nature of the questions in our training data. We have also identified several important confounding factors that arise from differences between the training and application settings of the algorithms. We have shown novel feature encoding, feature selection, and feature aggregation techniques to address these challenges, and have quantified their benefits. We have shown the benefits of allowing some subjects with lower certainty output from the algorithms to be classified as inconclusive. We have also shown the benefits of combining the results of the two algorithms into a single determination.

By specializing the machine learning models on a dichotomy of age groups, we found that the screener for younger children capitalized on non-verbal behavioral features such as eye contact, gestures, and facial expressions, while the screener for older children focused more on verbal communication and interactions with other children. For more details please refer to the Supplementary Material.

The methods and resulting improvements shown in this paper are expected to translate well into other clinical science applications

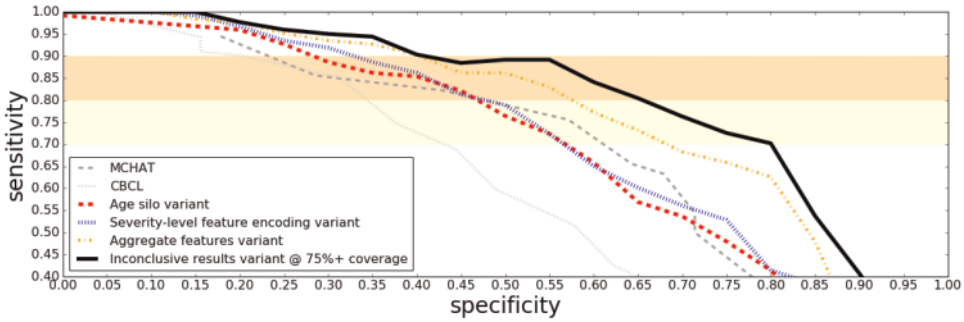

Figure 1. ROC curves on the clinical sample for various questionnaire based autism screening techniques, ordered from the least to most sophisticated. Note that unlike Figures 2 through 3 and 4, 168 children are included in this sample (six children did not have videos available).

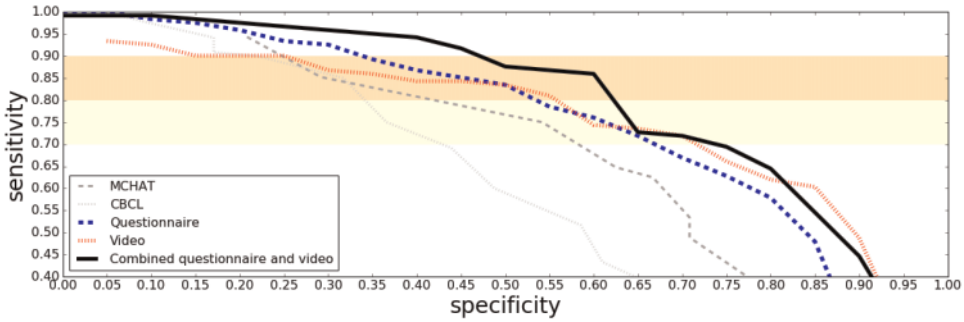

Figure 2. ROC curves on the clinical sample for the questionnaire and the video based algorithms, separately and in combination. The established screening tools MCHAT and CBCL are included as baselines.

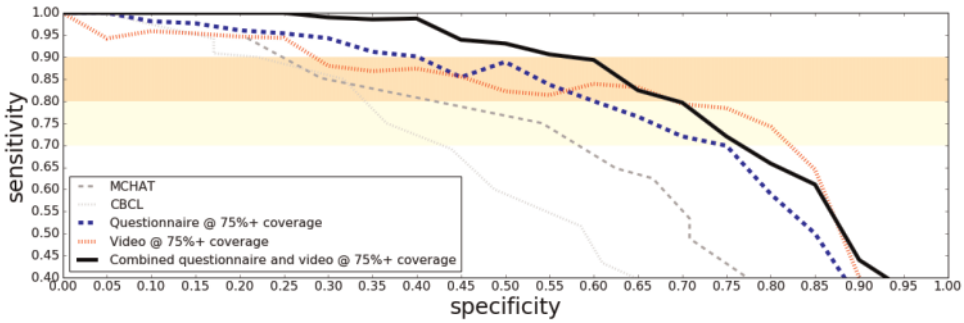

Figure 3. ROC curves on the clinical sample for the questionnaire and the video based algorithms, separately and in combination. Inconclusive determination is allowed for up to 25% of the cases. The established screening tools MCHAT and CBCL are included as baselines.

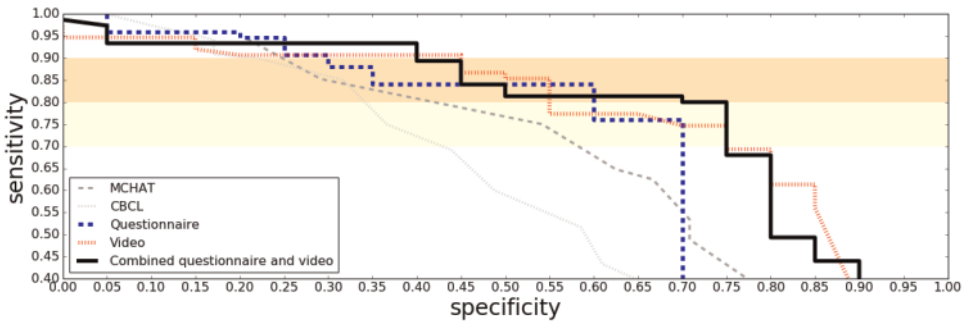

Figure 4. ROC curves on the clinical results for children under four years of age, for the questionnaire and the video based algorithms, as well as the combination. Comparisons with the established (nonmachine learning) screening tools MCHAT and CBCL are also shown.

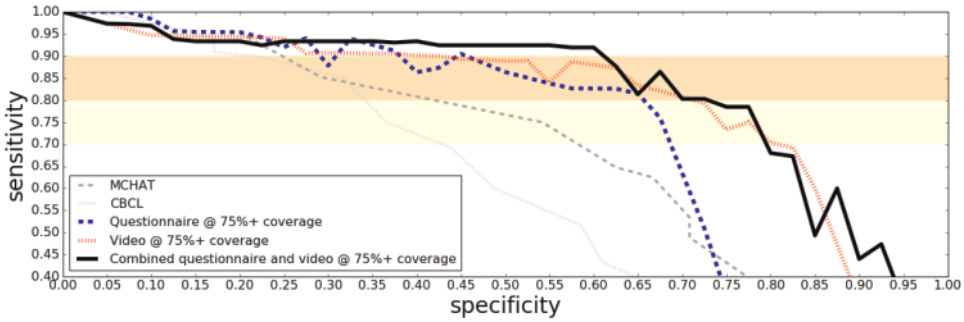

Figure 5. ROC curves on the clinical results for children under four years of age, for the questionnaire and the video based algorithms, as well as the combination, restricted to the children who were not determined to have an inconclusive outcome (tuned to have at most 25% allowed to be inconclusive). Comparisons with the established (nonmachine learning) screening tools MCHAT and CBCL are also shown.

including screening for cognitive conditions such as dementia for the elderly and physical conditions such as concussions in adults. Further, we expect that these methods would apply well to any other survey based domain in which the application context is different from the training context.

Significant further improvements may be possible. Initial studies have identified probable improvements to the machine learning methodology as well as improved methods for handling the biases between the training data and application settings. A new clinical trial with larger sample sizes is underway that will make it possible to validate new improvements resulting from these studies as well as to improve confidence in the high performance of our algorithms.

CONCLUSION

Machine learning can play a very important role in improving the effectiveness of behavioral health screeners. We have achieved a significant improvement over established screening tools for autism in children as demonstrated in a multi-center clinical trial. We have also shown some important pitfalls when applying machine learning in this domain, and quantified the benefit of applying proper solutions to address them.

FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

COMPETING INTERESTS

All authors are affiliated with Cognoa Inc. in an employment and/or advisory capacity.

CONTRIBUTORS

All listed authors contributed to the study design as well as the drafting and revisions of the paper. All authors approve of the final version of the paper to be published and agree to be accountable for all aspects of the work.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Durkin MS, Maenner MJ, Meaney FJ. Socioeconomic inequality in the prevalence of autism spectrum disorder: evidence from a U.S. cross-sectional study. *PLoS One* 2010; 5 (7): e11551.
2. Christensen DL, Baio J, Braun KV, *et al*. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2012. *MMWR Surveill Summ* 2016; 65 (3): 1–23.
3. Zwaigenbaum L, Bryson S, Lord C, *et al*. Clinical assessment and management of toddlers with suspected autism spectrum disorder: insights from studies of high-risk infants. *Pediatrics* 2009; 123 (5): 1383–91.
4. BernierMao RA, Yen J. Diagnosing autism spectrum disorders in primary care. *Practitioner* 2011; 255 (1745): 27–30.
5. Achenbach TM, Rescorla LA. *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families. 2001.
6. Lord C, Rutter M, Le Couteur A. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* 1994; 24 (5): 659–85.
7. Lord C, Rutter M, Goode S, *et al*. Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J Autism Dev Disord* 1989; 19 (2): 185–212.
8. Lord C, Petkova E, Hus V, *et al*. A multisite study of the clinical diagnosis of different autism spectrum disorders. *Arch Gen Psychiatry* 2012; 69 (3): 306–13.
9. Wall DP, Dally R, Luyster R, *et al*. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS One* 2012; 7 (8): e43855.
10. Duda M, Kosmicki JA, Wall DP, *et al*. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Transl Psychiatry* 2014; 4 (8): e424.
11. DudaDaniels MJ, Wall DP. Clinical evaluation of a novel and mobile autism risk assessment. *J Autism Dev Disord* 2016; 46 (6): 1953–1961.
12. Fusaro VA, Daniels J, Duda M, *et al*. The potential of accelerating early detection of autism through content analysis of youtube videos. *PLoS One* 2014; 16;9 (4): e93533.
13. Cognoa, Inc. Palo Alto: CA. https://www.cognoa.com/.