

## ORIGINAL ARTICLE

## Crowdsourced validation of a machine-learning classification system for autism and ADHD

M Duda<sup>1,2</sup>, N Haber<sup>1,2</sup>, J Daniels<sup>1,2</sup> and DP Wall<sup>1,2</sup>

Autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD) together affect >10% of the children in the United States, but considerable behavioral overlaps between the two disorders can often complicate differential diagnosis. Currently, there is no screening test designed to differentiate between the two disorders, and with waiting times from initial suspicion to diagnosis upwards of a year, methods to quickly and accurately assess risk for these and other developmental disorders are desperately needed. In a previous study, we found that four machine-learning algorithms were able to accurately (area under the curve (AUC) > 0.96) distinguish ASD from ADHD using only a small subset of items from the Social Responsiveness Scale (SRS). Here, we expand upon our prior work by including a novel crowdsourced data set of responses to our predefined top 15 SRS-derived questions from parents of children with ASD ( $n=248$ ) or ADHD ( $n=174$ ) to improve our model's capability to generalize to new, 'real-world' data. By mixing these novel survey data with our initial archival sample ( $n=3417$ ) and performing repeated cross-validation with subsampling, we created a classification algorithm that performs with  $AUC=0.89 \pm 0.01$  using only 15 questions.

*Translational Psychiatry* (2017) 7, e1133; doi:10.1038/tp.2017.86; published online 16 May 2017

## INTRODUCTION

Autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD) are among the most common diagnoses in children today, estimated to affect 1.5% and 9.5% of the pediatric population in the United States, respectively.<sup>1,2</sup> Although they are distinctly defined conditions, ASD and ADHD have considerable overlaps in their behavioral signatures, including impaired social interactions and increased impulsivity.<sup>3</sup> These symptomatic similarities can often complicate differential diagnosis for clinicians.

Developmental and neuropsychiatric disorders including ASD and ADHD are diagnosed through extensive behavioral examination by a team of clinicians. These diagnostic appointments, usually only accessible after referral from a general pediatrician, are rigorous and can last several hours. The time- and personnel-intensive nature of these appointments set the maximum capacity of specialized clinics far below the increasing demand for diagnostic evaluations. This bottleneck contributes to delays of months to even years from initial parental suspicion to clinical diagnosis,<sup>4</sup> despite many reports of parental concerns in children as young as 16 months,<sup>5,6</sup> and stable clinical identification of ASD as early as the age of 2.<sup>7–9</sup> These compounding roadblocks result in average ages of diagnosis of 4.5 years for ASD and 7 years for ADHD.<sup>1,10</sup>

Currently, there is no screening tool or diagnostic instrument to directly differentiate between ASD and ADHD. Previously, we applied machine learning to score sheets of gold-standard autism diagnostic instruments such as the Autism Diagnostic Observation Scale<sup>11</sup> (ADOS) and Autism Diagnostic Interview-Revised<sup>12</sup> (ADI-R) to develop screening tools that could accurately distinguish ASD from non-ASD with a reduction in questions of up to 92%.<sup>13–17</sup> To build on these methods and to fill the gap left by the lack of a

dedicated ASD-ADHD screener, we applied machine-learning to score sheets of the Social Responsiveness Scale<sup>18</sup> (SRS) to determine (1) whether the SRS could be used to distinguish ASD from ADHD and (2) whether this distinction could be made using a small subset of the questionnaire's 65 items. In that experiment, we demonstrated that only five items were necessary to discriminate between the ASD and ADHD classes with >96% accuracy.<sup>19</sup> In the present work, we employ improved machine learning methods to train a model on the entirety of our archival sample, and test the performance of this classifier on a novel, independently collected validation set.

## MATERIALS AND METHODS

## Data collection

We obtained Stanford institutional review board approval (eProtocol 35161) to collect survey data from parents of children aged 2–17 years. We utilized crowdsourcing recruitment methodology,<sup>20–24</sup> allowing our outreach to extend across the United States. Mainstream social media networking platforms, such as Facebook, Twitter, general press releases and Yahoo Groups, were targeted to inform the community of the study via recruitment flyers that clearly expressed our goals and a link to the study website. In addition, we engaged with developmental delay and parent support groups as well as relevant startup companies (for example, Cognoa, Palo Alto, CA, USA) to send direct invitation links for participation to parents who had subscribed to their email lists.

The survey was administered online via RedCap;<sup>25</sup> all responses were anonymous. We electronically consented the parents and collected the child's general demographic information (gender, age, race, ethnicity and annual household income) as well as how they learned about the study and whether the child had any existing clinical diagnosis of a developmental or neuropsychiatric disorder. We also asked the parents to answer 15 questions about their child's daily behavior identified from

<sup>1</sup>Division of Systems Medicine, Department of Pediatrics, Stanford University, Stanford, CA, USA and <sup>2</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. Correspondence: Dr DP Wall, Pediatrics and Biomedical Data Science, Stanford University, 1265 Welch Road x141, Stanford, CA 94305, USA. E-mail: dpwall@stanford.edu

Received 3 September 2016; revised 1 February 2017; accepted 16 March 2017

machine learning analysis of the SRS. These behavioral questions corresponded to the top-ranked items identified in our initial analysis.<sup>19</sup>

Survey responses were collected from 26 January to 12 March 2016. During this time frame we had 2506 total participants, 1693 of which submitted complete responses. Of the completed responses, 1643 participants responded to our optional crowdsourcing question, inquiring how they learned of the study. In all, 38.2% ( $n=628$ ) are from email listservs of organizations, 21.4% ( $n=351$ ) are from parent support groups, 18.7% ( $n=307$ ) are from Facebook, 3.4% ( $n=56$ ) are from friend referrals, 1.5% ( $n=24$ ) are from other social media outlets (such as forums), 1.2% ( $n=20$ ) are from news articles, 0.4% ( $n=6$ ) are from Twitter and 15.3% ( $n=251$ ) are from other platforms not specified above.

As our selection criteria were not limited by diagnosis, we selected the subset of responses from parents of children with only ASD ( $n=248$ ) or only ADHD ( $n=174$ ) to serve as our survey sample. For this survey data set, diagnoses of ASD or ADHD were provided as parent report. The average age of the ASD survey sample was  $8.2 \pm 4.3$  years and the average age of the ADHD survey sample was  $10.4 \pm 3.6$  years. The survey ASD sample was 80.6% male ( $n=200$ ) and the survey ADHD sample was 65.5% male ( $n=114$ ).

The data set from our initial analysis comprised 2925 complete SRS score sheets compiled from the Boston Autism Consortium, the Autism Genetic Resource Exchange and the Simons Simplex Collection data repositories. In the archival data set, diagnoses of ASD were physician-confirmed and diagnoses of ADHD were reported as part of an extensive family medical history. For each ASD ( $n=2775$ ) and ADHD ( $n=150$ ) subject, we subset the 15 items that corresponded to our survey questions and refer to this data set here as our archival sample. The average age of our archival ASD sample was  $9.2 \pm 3.9$  years and the average age of the ADHD survey sample was  $11.3 \pm 3.2$  years. The archival ASD sample was 83.9% male ( $n=2328$ ) and the archival ADHD sample was 62% male ( $n=93$ ). It is important to note that the majority of ADHD subjects in the archival sample were siblings of ASD subjects. This is a consequence of the data collection protocols of the studies included in the databases listed above.

## Machine learning

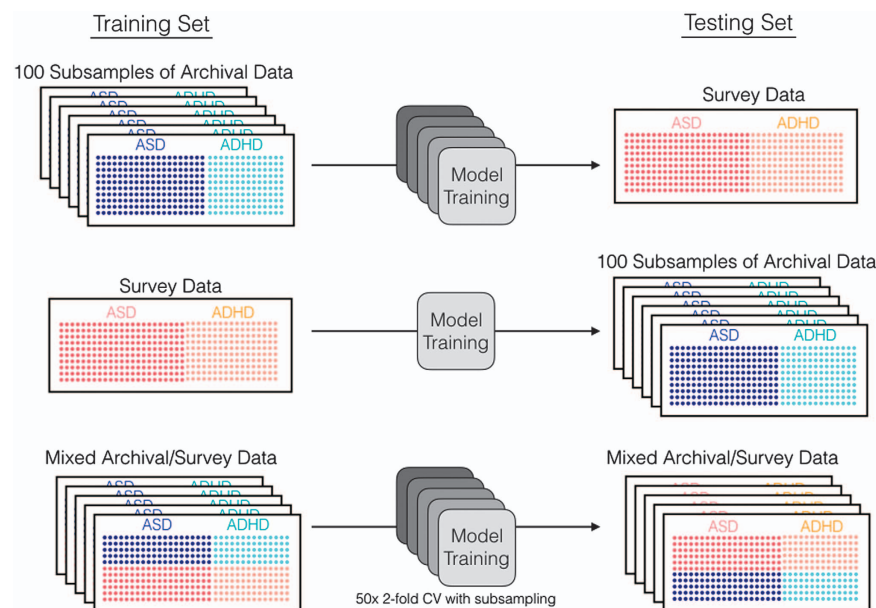
The aim of this study was to select the algorithm that best distinguishes between ASD and ADHD, given a set of items derived from the SRS and to test the performance of this classifier on a novel data set. All machine learning training and testing was carried out in Python using the package Scikit Learn.<sup>26</sup> We selected the best-performing algorithms from our initial

analysis to test in this validation experiment; support vector classification (SVC), Logistic Regression with  $l_1$  Regularization (Lasso), Logistic Regression with  $l_2$  Regularization (Ridge) (Referred to as 'Logistic Regression' in our initial publication) and linear discriminant analysis (LDA). We elected to also include Elastic Net (ENet), a logistic regression model that incorporates both  $l_1$  and  $l_2$  regularization, as both the Lasso and Ridge models performed well in our previous work.

We performed three unique machine learning experiments to test how well our classifiers can generalize to new data. First, we trained each of our five algorithms on the archival sample and tested on the survey sample. Owing to the imbalance of the ASD ( $n=2775$ ) and ADHD ( $n=150$ ) classes in our archival sample, we performed random subsampling on the ASD training set to achieve the same ratio of ASD to ADHD (1.25:1) as the survey sample. The 100 resulting training sets consisted of 214 ASD and 150 ADHD archival data points. We fit all five models to each subsampled training set, using the Grid Search function to optimize the parameters for each model. The Grid Search function accepts arrays of possible parameter values for each algorithm as input and performs an internal cross-validation (CV) on the training data set to identify the best combination of parameter values. By default, stratified three-fold CV is used to determine the optimal combination of parameter values. The parameter values that achieve the highest accuracy on the held out fold are chosen, and the model is fit to the entire training set using those optimal parameter values. Parameter value arrays tested here were as follows:  $C=(0.01, 0.1, 1, 10)$  for Lasso and Ridge;  $C=(0.01, 0.1, 1, 10)$ ,  $\text{kernel}=(\text{linear}, \text{rbf})$  for SVC;  $\alpha=(0.05, 0.1, 0.15, \dots, 1)$ ,  $l_1 \text{ ratio}=(0.05, 0.1, 0.15, \dots, 1)$  for ENet;  $\text{solver}=(\text{lsqr}, \text{eigen})$ ,  $\text{shrinkage}=(\text{none}, \text{auto})$  for LDA. After training and parameter tuning were complete, we tested each of the five unique models on our entire survey sample and recorded the receiver operating characteristic-area under the curve (ROC-AUC). To estimate the average performance of each model on the novel testing data, we repeated this training and testing process 100 times, always subsampling from the archival data for the training set and testing on the entire validation set.

Next, we used the survey sample as our training data set and tested on the archival sample. Each of our five algorithms was trained using the previously described Grid Search parameter tuning method on the entirety of the survey sample. To estimate the performance of these five survey-trained algorithms, we tested each classifier on 100 random subsamples of the archival data set, using the 1.25:1 ASD to ADHD ratio above, and recorded the ROC-AUC for each trial.

Finally, we combined both the archival and survey samples into a single mixed data set. In order to both train and test our classifiers on this mixed



**Figure 1.** Representation of our machine-learning pipeline. In Trial 1, each of the five models was trained on 100 random subsamples of the archival data and tested on the same survey sample. In Trial 2, each model was trained once on the entire survey sample and tested on 100 random subsamples of the archival data. In Trial 3, each algorithm was trained and tested on 50 rounds of twofold CV with subsampling. CV, cross-validation.

sample we performed 50 rounds of two-fold CV with subsampling. We chose to implement this two-fold CV structure to both preserve the approximate ratio of training to testing data from the first two trials and to estimate the performance of our algorithms as an average over 100 independent trials, as was done in the first two trials. This allowed for direct comparisons between the results of each of the three machine learning experiments in this study. During each fold of the CV, both the training and testing folds were subsampled with the previous ratio of 1.25:1 ASD to ADHD, resulting in 231 ASD and 162 ADHD randomly shuffled archival/survey data points per fold. ROC–AUC values were recorded for each of the five algorithms over the 100 CV trials. Experimental designs for each of our three trials are summarized in Figure 1.

## RESULTS

We tested the classification performance of five machine learning algorithms to distinguish between ASD and ADHD using three different training and testing data samples. Figure 2 compares the ROC–AUC for each training set/algorithm combination over 100 testing trials. From this boxplot, we see that the performance of each algorithm is highly variable depending on the data set used for training or testing.

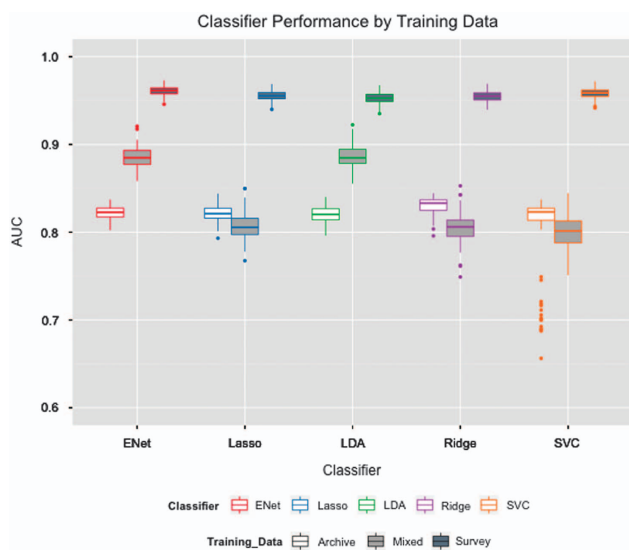
To investigate how variance between and within the archival and survey data sets contributed to the variance in classifier performance, we plotted the distributions of responses in the ASD and ADHD classes for each data set (Figure 3). For the 15 questions analyzed, this boxplot illustrates the distribution of responses ('Always' (4), 'Often' (3), 'Sometimes' (2) and 'Never' (1))

for subjects from either the ASD or ADHD class for both the archival (red) and survey (blue) samples. For each box, both the median (line) and mean (diamond) of that subset is shown. From this plot, we see that the distribution of responses from the ASD class is comparable between the archival and survey data for almost all of the questions. However, it is clear that this pattern is not present in the ADHD sample, as the means/medians from the archival and survey samples do not match for any of the 15 response distributions. Table 1 lists the Kolmogorov–Smirnov test *P*-values for answer distributions between archive and survey data sets for all 15 questions.

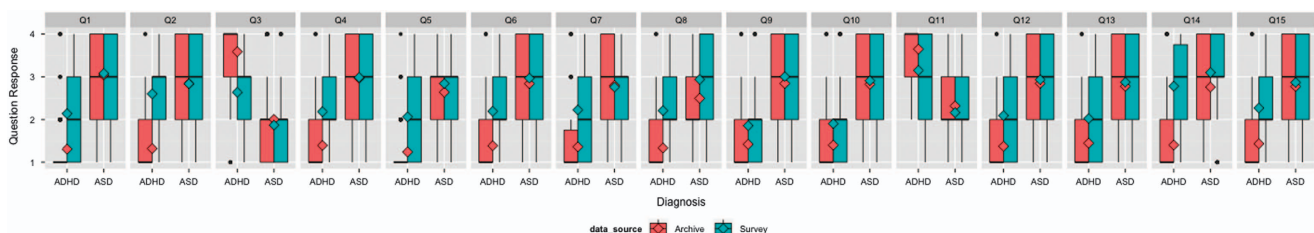
To further explore how the differences in answer distributions affect the predictive power of our algorithms, we performed two final model training and testing experiments—one solely on archival data and another solely on survey data. Both were carried out using 50 repetitions of two-fold CV, and each fold of the CV in the archival experiment was subsampled to retain the 1.25:1 ASD to ADHD ratio used throughout. Paired with the results from our prior experiments, we were able to compare the performance of each of our five classifiers across five different combinations of training and testing data (Figure 4).

By viewing the data in this way, we see that all five models performed best ( $AUC > 0.90$ ) when the archival data set was used for training. This indicated that separation of the archival data into ASD and ADHD was fairly straightforward, which is supported by the answer distributions in Figure 3. As survey data were included in the testing set, the overall performance of the model generally decreased, and for the three simpler models (Lasso, Ridge and SVC) the survey-trained algorithm tested on survey data had the lowest AUC of all training/testing combinations. However, it was evident that the ENet and LDA classifiers generalized better to survey data ( $AUC = 0.89 \pm 0.01$ ), significantly outperforming the other three models in both the mixed-trained/mixed-tested and survey-trained/survey-tested experiments ( $P < 2.2 \times 10^{-16}$ ). Figure 5 shows the ROC and precision–recall curves, as well as prediction score distributions for both the ENet (A) and LDA (B) classifiers generated from a five-fold CV with subsampling of the mixed data set. Although both models had very similar classification performance, the distribution of prediction scores differed largely between them. The prediction scores from the LDA model were concentrated toward either extreme, while we saw a more widespread distribution of prediction scores calculated from the ENet model. This similarity between these generative and discriminative models suggested that there is some truth to the generative assumptions in the LDA model, that, because of this, the model was more confident in certain regions and that these might represent reasonable scores.

To understand which of our 15 features were most useful for classification, we also plotted the feature weights for both ENet (A) and LDA (B) over the five-fold CV with subsampling on the mixed data set (Figure 6). Here, we saw a very consistent pattern in feature weights between the two models despite using differing methods for coefficient calculation, further supporting the



**Figure 2.** Classification performance boxplot of five algorithms trained on archival, survey and mixed data samples over 100 validation trials.



**Figure 3.** Distribution of responses to each of the 15 questions for both the archival ( $n = 2925$ ) and survey ( $n = 422$ ) sets. The coded response scale corresponds to answers of 'Always' (4), 'Often' (3), 'Sometimes' (2) and 'Never' (1) for each question. For each box, both the median (line) and the mean (diamond) of that subset is shown.

4 convergence between the generative and discriminative approaches and the potential that some features may indeed be more informative than others for the distinction of ASD and ADHD (shown on plot in red) in this data sample.

## DISCUSSION

The current process of diagnosing developmental and neuropsychiatric disorders such as ASD or ADHD is time-intensive, subjective and complicated by overlaps in behavioral symptoms. Furthermore, the bottlenecks created by the referral system and the extensive waitlists at specialized diagnostic clinics correspond

to time delays from initial suspicion to diagnosis of a year or longer. Considering the high incidence of both ASD and ADHD coupled with the similarities in symptomatology between the two disorders, there is a clear need for screening tools that can quickly and accurately estimate risk for either condition. In our previous work, we demonstrated that machine learning could be used to distinguish between ASD and ADHD with high accuracy using only the responses to five parent-directed questions.<sup>19</sup> However, that analysis was limited by the availability of ADHD data in our archival data repository. Here, we utilized a novel data set of 422 electronically collected responses to our 15 top-ranked questions, termed the survey sample, to both increase the number of ADHD cases in our analysis and to estimate each algorithm's ability to generalize to new, 'real-world' data.

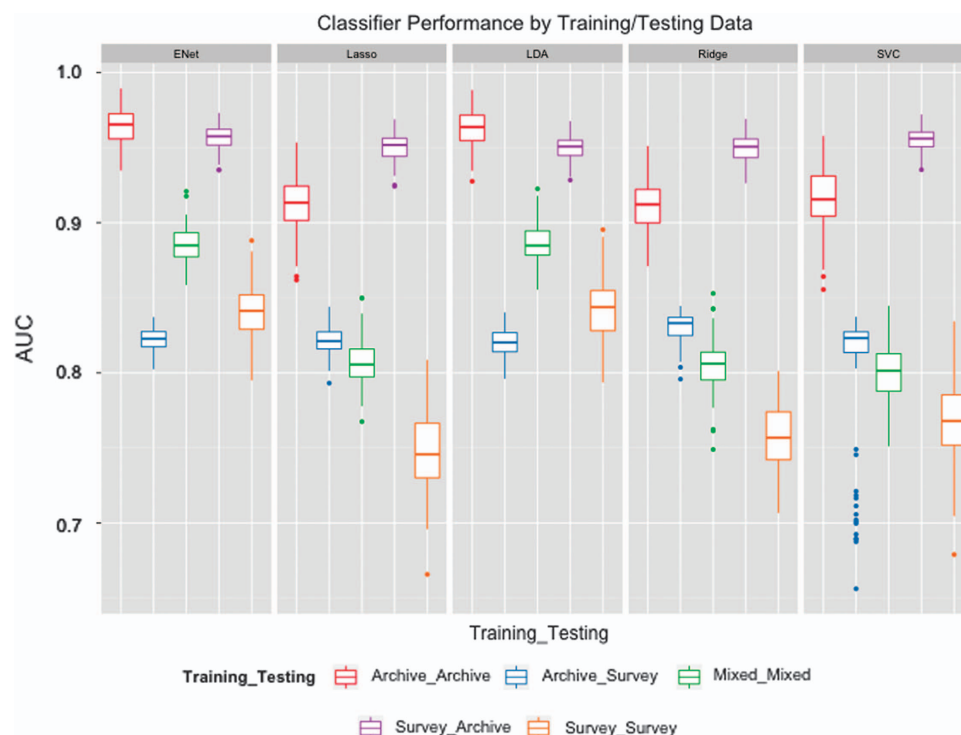
There was an inherent difference in gender distribution and age between the ASD and ADHD classes in both the survey and archival data sets. We computed the Pearson correlation between age and sum total of the coded responses to our 15 classifier questions and determined that there was no significant correlation between age and question response ( $R=0.054$ ). To test the effect on gender on the question sum total, we performed two Kolmogorov–Smirnov tests, one comparing the distribution of totals between male and female subjects in the entire ASD population (both archival and survey) and one comparing the same in the entire ADHD population. The  $P$ -values for these tests were 0.2763 and 0.6844, respectively, indicating that the distributions of score totals between male and female subjects in each diagnostic class were not significantly different.

A significant difference in distributions was evident between the archival and survey data sets, particularly in the ADHD class. In the archival sample, the responses for ADHD subjects were more uniform and on average less severe than the ADHD responses in the survey sample. There are many factors that could possibly account for the differences in answer distributions. The archival ADHD data consisted solely of siblings of children with autism, resulting in a biased and highly specific sample. Conversely, the

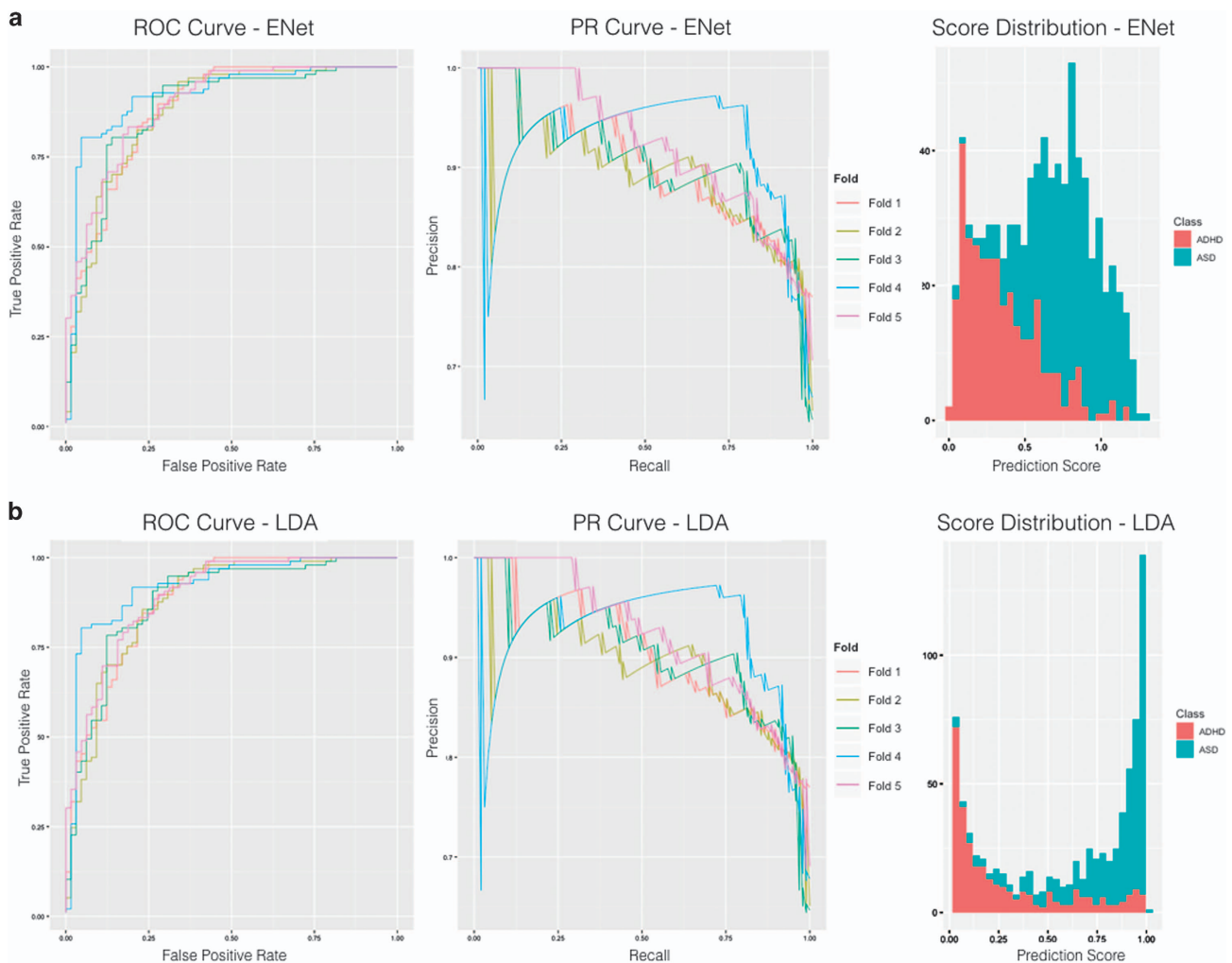
**Table 1.** Kolmogorov–Smirnov test  $P$ -values for answer distributions between archive and survey data sets for all 15 questions

| Item | $ASD_{archive} \sim ASD_{survey}$ | $ADHD_{archive} \sim ADHD_{survey}$ | $Archive \sim survey$ |
|------|-----------------------------------|-------------------------------------|-----------------------|
| Q1   | 0.9616                            | $2.20E-16^a$                        | $1.10E-06^a$          |
| Q2   | 0.3817                            | $2.20E-16^a$                        | $0.003604^a$          |
| Q3   | $0.002261^a$                      | $2.20E-16^a$                        | $0.001386^a$          |
| Q4   | 0.5188                            | $2.20E-16^a$                        | $4.87E-06^a$          |
| Q5   | $0.002592^a$                      | $5.55E-15^a$                        | 0.195                 |
| Q6   | 0.2602                            | $2.20E-16^a$                        | $0.005366^a$          |
| Q7   | 0.09608                           | $4.89E-15^a$                        | $0.0006825^a$         |
| Q8   | $1.85E-10^a$                      | $2.20E-16^a$                        | $0.001935^a$          |
| Q9   | 0.0722                            | $2.17E-06^a$                        | $5.36E-05^a$          |
| Q10  | 0.1032                            | $8.61E-07^a$                        | $0.0007416^a$         |
| Q11  | 0.08979                           | $4.13E-06^a$                        | $0.04507^a$           |
| Q12  | 0.1523                            | $1.41E-10^a$                        | $0.000634^a$          |
| Q13  | 0.5095                            | $2.98E-09^a$                        | $0.0198^a$            |
| Q14  | $6.17E-05^a$                      | $2.20E-16^a$                        | $5.24E-05^a$          |
| Q15  | 0.426                             | $5.55E-16^a$                        | 0.05345               |

Abbreviations: ADHD, attention deficit hyperactivity disorder; ASD, autism spectrum disorder. <sup>a</sup>Indicates significance.



**Figure 4.** Performance of the five machine-learning models on five different training and testing set combinations across 100 trials.



**Figure 5.** ROC and PR curves and prediction score distributions for ENet (a) and LDA (b) classifiers using fivefold CV with subsampling on the mixed data set. CV, cross-validation; ENet, Elastic Net; LDA, linear discriminant analysis; PR, precision–recall; ROC, receiver operating characteristic.

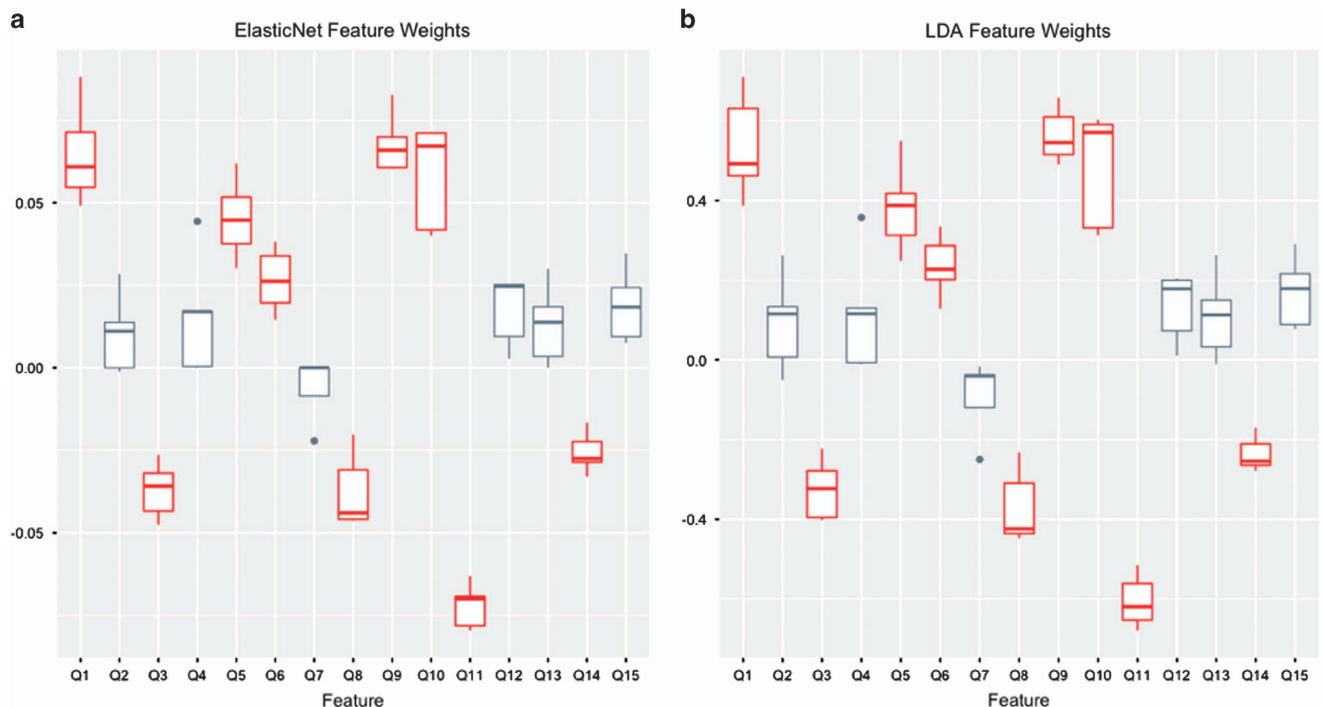
only selection criterion for our survey sample was that the child be between the ages of 2 and 17, making these data completely independent of family structure and arguably more representative of the ‘real-world’ population, possibly explaining the higher variance found in this sample. In addition, the archival responses used in this analysis came from the official SRS, which was likely administered in-person with a clinical professional present, and the survey responses were collected only for the 15-question subset, administered electronically. These administrative differences could possibly contribute to the different distributions; however, the crowdsourced data collection strategy implemented here tests the actual use of our classification system as a mobile, and therefore more accessible, tool for detection of autism and ADHD. By creating a classifier that can generalize to highly variable data, we take important steps toward building confidence in its accuracy within a general pediatric population.

The results of our first experiment, which used archival data to train the five models and predict survey data, were lower than our previously reported results ( $AUC = 0.82 \pm 0.02$  in the present study versus  $AUC = 0.96$  in the original experiment). This drop in performance was expected, especially considering the significant difference in answer distributions between the two cohorts. Nevertheless, a classification accuracy of 82% has promise for clinical use, especially considering that no other screening tool to directly distinguish autism and ADHD is currently available.

However, we elected to utilize these novel ‘real-world’ data not only as a validation set, but also to improve our classification system as a whole by including some survey samples in our training set. By combining survey and archival data for training, we create a classification algorithm that can generalize well to unseen data ( $AUC = 0.89 \pm 0.01$ ), even when those data have more natural variability like the kind seen in our survey sample.

In our machine learning experiments, the ENet and LDA models exhibited the best classification performance. This was not surprising, considering that these models are built to handle complex correlations between variables. LDA<sup>27</sup> explicitly takes into account correlation between inputs in its generative criterion while ENet<sup>28</sup> was intended to solve correlation issues with Lasso. When groups of variables are highly correlated, Lasso tends to select only one of the variables to have a nonzero coefficient, with the selection being unstable to noise. Also, the performance of Lasso is often lower than Ridge when there are many samples and when the variables are highly correlated.

These results support the potential of creating a quick, accurate and widely accessible method for differentiating risks between ASD and ADHD. Our success in crowdsourcing over 2500 responses to our survey in just 6 weeks indicates that mobile administration of such a tool would be possible, and would be well received by primary caregivers of children at risk in the general pediatric population. Furthermore, the simplicity of our



**Figure 6.** Feature weights of ENet (a) and LDA (b) models over fivefold CV with subsampling on the mixed data set. Important features (those with largest absolute mean feature weight) are indicated in red, and are consistent between the two models. CV, cross-validation; ENet, Elastic Net; LDA, linear discriminant analysis.

algorithm would allow for real-time calculation of risk scores and instantaneous feedback to caregivers. The combination of this classifier with others (e.g. those designed for distinction of just autism from non-autism) could result in an ensemble classification system that can return a risk report, not only between ASD and ADHD, but a multitude of common developmental delays. Such a mobile screening platform could provide useful information to care providers regardless of geographic location or socioeconomic status. Moreover, the data captured by such a mobile system could prove to be a useful supplement to the standard clinical encounter, providing clinical practitioners a detailed assessment of the patient before they even walk in for their appointment.

Continued evaluation of this classifier is needed to assess its ability to generalize to new data and correctly predict risk for either ASD or ADHD. As new data become available through crowdsourcing or other methods, our model can be adapted to include new data points and further improve the generalizability of the classifier. Our validated algorithm represents a novel approach aimed at differentiating autism from ADHD, and shows promise for use as a mobile assessment of risk between the two disorders.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGMENTS

The authors would like to acknowledge the families that participated in or provided data for this study, as well as the Coulter, Beckman, and Hartwell Foundations for partial support of the project.

## REFERENCES

- 1 Developmental Disabilities Monitoring Network Surveillance Year Principal I. 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**(Suppl 2): 1–21.

- 2 Visser SN, Danielson ML, Bitsko RH, Holbrook JR, Kogan MD, Ghandour RM et al. Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003–2011. *J Am Acad Child Adolesc Psychiatry* 2014; **53**: 34–46 e32.
- 3 Berenguer-Fornier C, Miranda-Casas A, Pastor-Cerezuela G, Rosello-Miranda R. [Comorbidity of autism spectrum disorder and attention deficit with hyperactivity. A review study]. *Rev Neurol* 2015; **60**(Suppl 1): S37–S43.
- 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 Veness C, Prior M, Eadie P, Bavin E, Reilly S. Predicting autism diagnosis by 7 years of age using parent report of infant social communication skills. *J Paediatr Child Health* 2014; **50**: 693–700.
- 6 Kozlowski AM, Matson JL, Horowitz M, Worley JA, Neal D. Parents' first concerns of their child's development in toddlers with autism spectrum disorders. *Dev Neuropsychol* 2011; **14**: 72–78.
- 7 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.
- 8 Worley JA, Matson JL, Mahan S, Kozlowski AM, Neal D. Stability of symptoms of autism spectrum disorders in toddlers: an examination using the Baby and Infant Screen for Children with aUtism Traits-Part 1 (BISCUIT). *Dev Neuropsychol* 2011; **14**: 36–40.
- 9 Kleinman JM, Ventola PE, Pandey J, Verbalis AD, Barton M, Hodgson S et al. Diagnostic stability in very young children with autism spectrum disorders. *J Autism Dev Disord* 2008; **38**: 606–615.
- 10 Pastor P, Reuben C, Duran C, Hawkins L. Association between diagnosed ADHD and selected characteristics among children aged 4–17 years: United States, 2011–2013. *NCHS Data Brief* 2015; 201.
- 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 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**: 659–685.
- 13 Wall DP, Dally R, Luyster R, Jung JY, Deluca TF. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS ONE* 2012; **7**: e43855.
- 14 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.
- 15 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**: e93533.

- 16 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.
- 17 Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Transl Psychiatry* 2015; **5**: e514.
- 18 Constantino JN, Davis SA, Todd RD, Schindler MK, Gross MM, Brophy SL *et al*. Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. *J Autism Dev Disord* 2003; **33**: 427–433.
- 19 Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. *Transl Psychiatry* 2016; **6**: e732.
- 20 David MM, Babineau BA, Wall DP. Can we accelerate autism discoveries through crowdsourcing? *Res Autism Spect Disord* 2016; **32**: 80–83.
- 21 Hong H, Gilbert E, Abowd GD, Arriaga RI. In-group questions and out-group answers: crowdsourcing daily living advice for individuals with autism. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*; 18–23 April 2015; Seoul, Republic of Korea. ACM: New York, NY, USA, 2015.
- 22 Jeong JW, Morris MR, Teevan J, Liebling DJ. A Crowd-Powered Socially Embedded Search Engine. ICWSM, 2013.
- 23 Krippendorff K. Reliability in content analysis. *Hum Commun Res* 2004; **30**: 411–433.
- 24 Swan M. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. *J Med Internet Res* 2012; **14**: e46.
- 25 Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009; **42**: 377–381.
- 26 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; **12**: 2825–2830.
- 27 Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugenetic* 1936; **7**: 179–188.
- 28 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Roy Stat Soc B* 2005; **67**: 301–320.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

© The Author(s) 2017