Machine Learning Techniques to Detect Autism Spectrum Disorder and Other Correlating Conditions.

Arjun Pesaru, Hiranmai Devarasetty, Navya Sri Alugubilli and Sai Nikhil Kunapareddy

01. SUMMARY

The project titled "Machine Learning Techniques to Detect Autism Spectrum Disorder and Other Correlating Conditions" aims to pioneer a predictive model for detecting Autism Spectrum Disorder (ASD) through the application of machine learning techniques. ASD, a complex neurodevelopmental disorder, manifests across a spectrum of symptoms influenced by genetic and environmental factors, often coinciding with conditions such as ADHD, Dyslexia, or Dyspraxia. Our major aim is to understand the influence of these widely different factors and incorporate them into our model.

To fulfill this objective, the project leverages comprehensive datasets encompassing surveys to forecast ASD and potential comorbidities. Initially, it utilizes the autism spectrum screening for children dataset, offering insights into the demographic characteristics and behavioral patterns of individuals on the spectrum. Each survey question serves as an independent attribute, facilitating a nuanced understanding of its role in accurate disorder prediction.

The workflow encompasses meticulous data preprocessing, including the handling of missing values and categorical variable encoding, followed by the training of a logistic regression model. Cross-validation techniques are employed to ascertain optimal model parameters, and the model's efficacy is assessed through the precision-recall curve.

This project aspires to advance diagnostic capabilities for ASD and related conditions, shedding light on the potential of machine learning methodologies in not only detecting ASD but also identifying congruent conditions in both pediatric and adult populations.

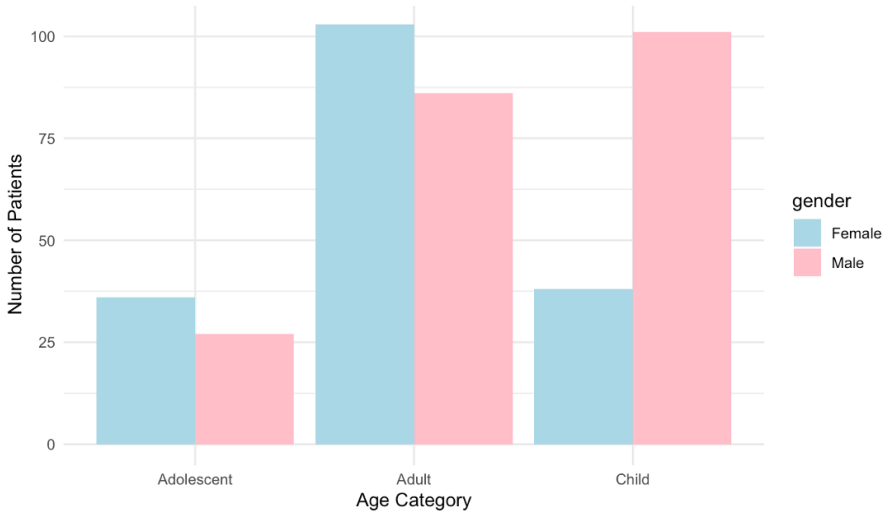

Figure 01: A bar graph explaining the distribution of the respondent's age in the dataset.

02. METHODS

The following is a step-by-step procedural explanation of the methods utilized in framing, exploring, and solving the problem statement.

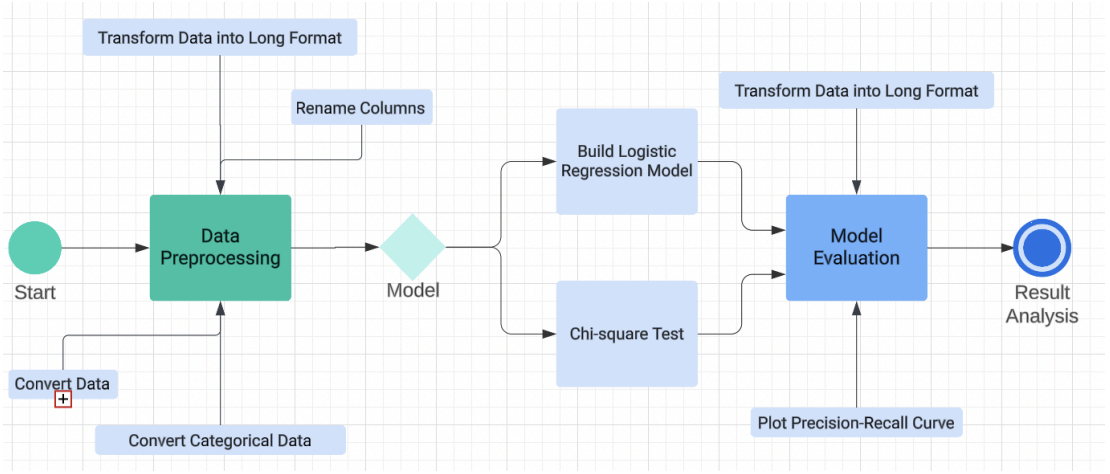

Figure 02: A visual representation of the project's workflow.

2.1 Data preprocessing

Essential libraries such as 'dplyr', 'ggplot2', 'tidyverse' and 'caret' were imported into rstudio to load, pre-process and build a valid machine learning model. The dataset (a .csv file) was loaded using read.csv() function and underwent preprocessing, including column selection, renaming, filtering and conversion into categorical data wherever necessary. Data was transformed into a long format for visualization purposes using pivor_longer() and mutate() functions..

An exploratory data analysis was conducted using ggplot2, to gain insights into the distribution of various demographic information and predictor variables related to the diagnosis of autism spectrum disorder.

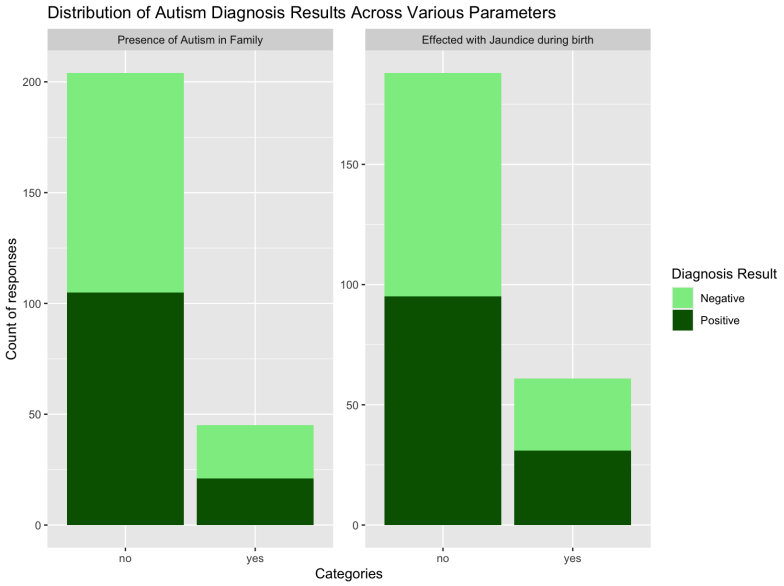

Figure 03: Results of the exploratory data analysis.

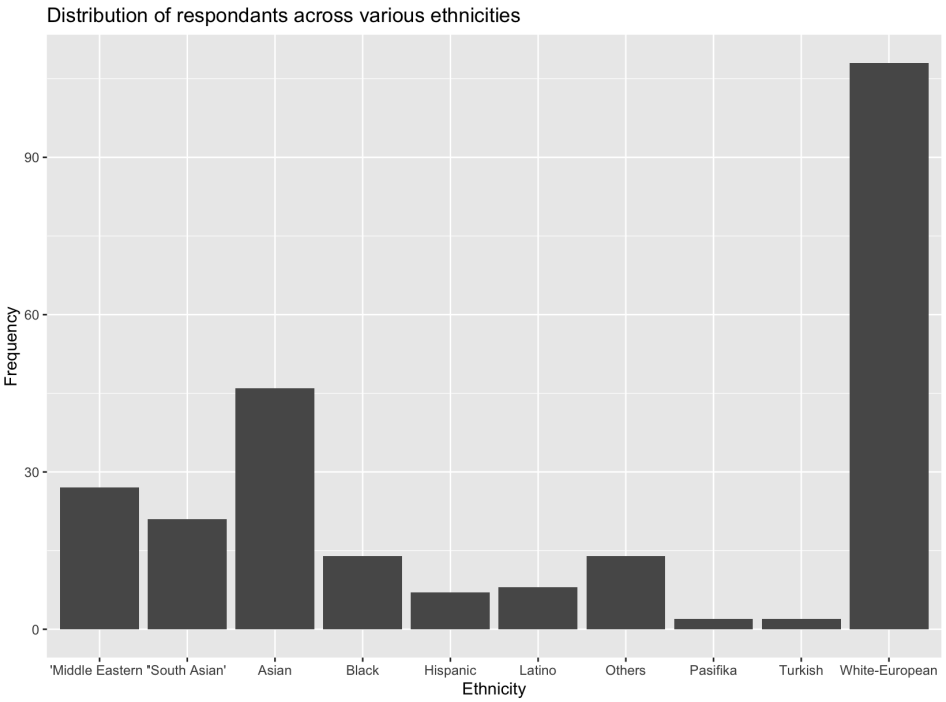

Figure 04: Distribution of respondents ethnicity.

2.2 Model building

The dataset was split equally into training and testing sets to facilitate model evaluation. This was achieved using the createDataPartition() function. Chi-square tests were performed to assess the relationship between the diagnosis of autism spectrum disorder and other attributes included in the analysis. A logistic regression model was constructed using glm() function by selecting appropriate predictor variables (based on the values of chi-squared test) as predictors and the diagnosis of autism spectrum disorder as the response variable.

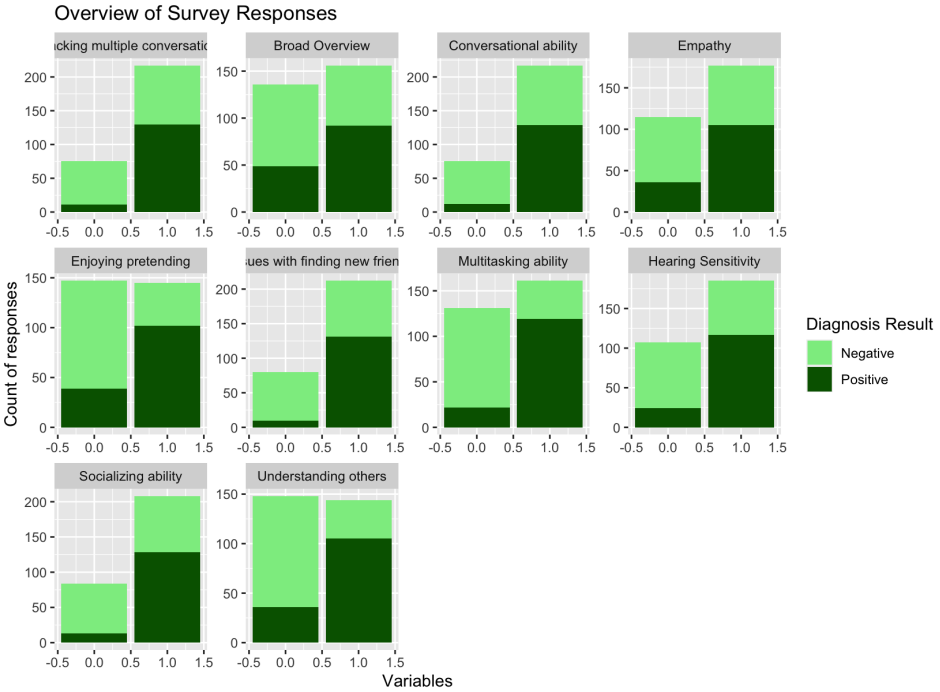

Figure 05: A bar plot offering insights into the distribution of predictor variables in 'Autism in Children' dataset.

2.3 Model evaluation

The constructed model was initially evaluated through the summary() function and later parameters like sensitivity, specificity and precision-recall curve are used to validate and compare various models among other datasets.

Given below is a chi-squared test result of data available from 'Autism in Children' dataset signifying that the gender of the respondent, effect of jaundice during birth, history of autism in family and ethnicity has no statistical significance to our target variable.

Parameter	P-value	X squared value
Gender	0.937	0.006
Diagnosed with Jaundice during birth	1.000	0.000
History of ASD in family	0.996	1.746 e-05
Ethnicity	0.013	22.35

Table 01: Chi-squared test metrics of attributes related to demographics, family and patient history.

03. RESULTS

Given below is an overview of the three different models built using the children, adolescents, and adults dataset.

Model	Sensitivity	Specificity
Children	0.87	0.73
Adolescent	0.87	0.71
Adult	0.83	0.84

Table 02: Overview of model evaluation metrics.

Following are the individual chi-squared test results of all the considered parameters along with their model summaries.

Parameter	Children		Adolescent		Adult	
	P-value	X^2	P-value	X^2	P-value	X^2
Hearing Sensitivity	1.26 e-07	27.92	0.213	1.547	2.22 e-05	18.00
Broad Overview	0.002	9.379	0.256	1.285	0.000	12.08
Tracking multiple conversations	2.54 e-06	22.13	0.000	12.23	5.03 e-05	16.43
Multitasking ability	**3.43 e-13**	**52.94**	**4.45 e-05**	**16.66**	2.71 e-08	30.89
Conversational ability	0.000	12.18	0.011	6.316	8.63 e-10	37.61
Socializing ability	0.000	14.14	0.004	8.225	**2.10 e-14**	**58.42**
Empathy	4.59 e-05	16.60	0.004	7.927	5.50 e-07	25.07
Enjoying pretending	5.22 e-06	20.75	0.003	8.476	0.000	12.76
Understanding others	1.59 e-09	36.41	0.001	10.14	**6.12 e-12**	**47.28**
Issues with finding new friends	7.86 e-06	19.97	0.000	14.20	0.000	14.85

Table 03: Overview of feature selection metrics.

Following are the receiver operating characteristic curves of all the three models.

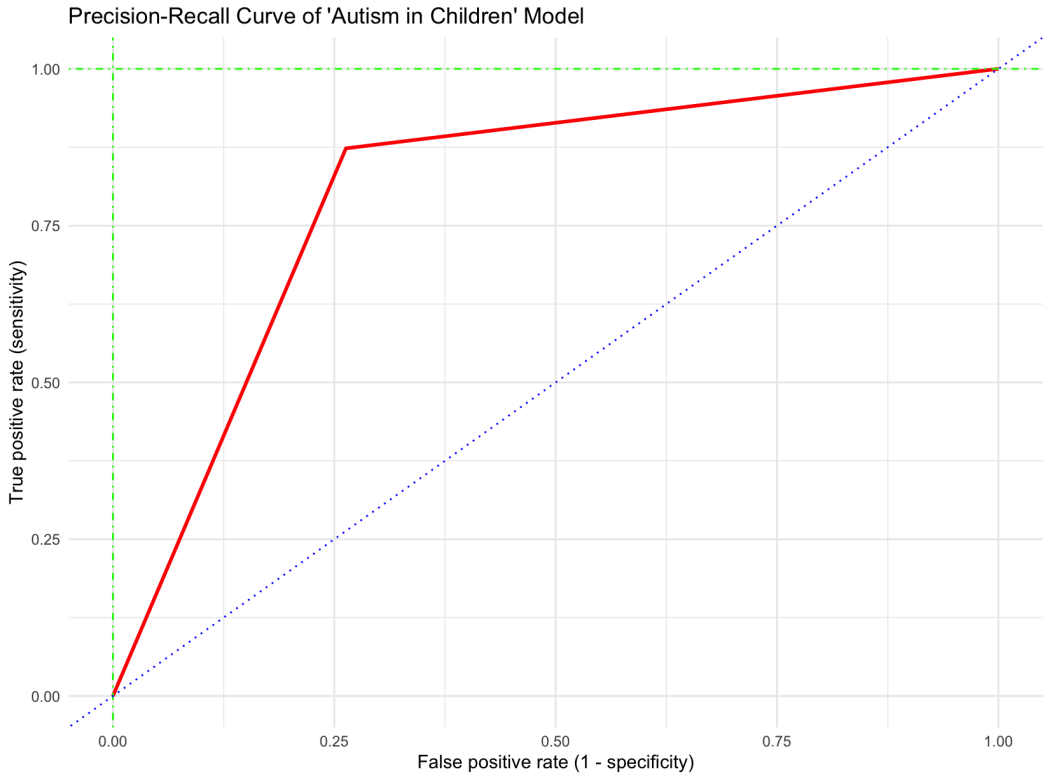

Figure 06: ROC curve of 'Autism in Children' model.

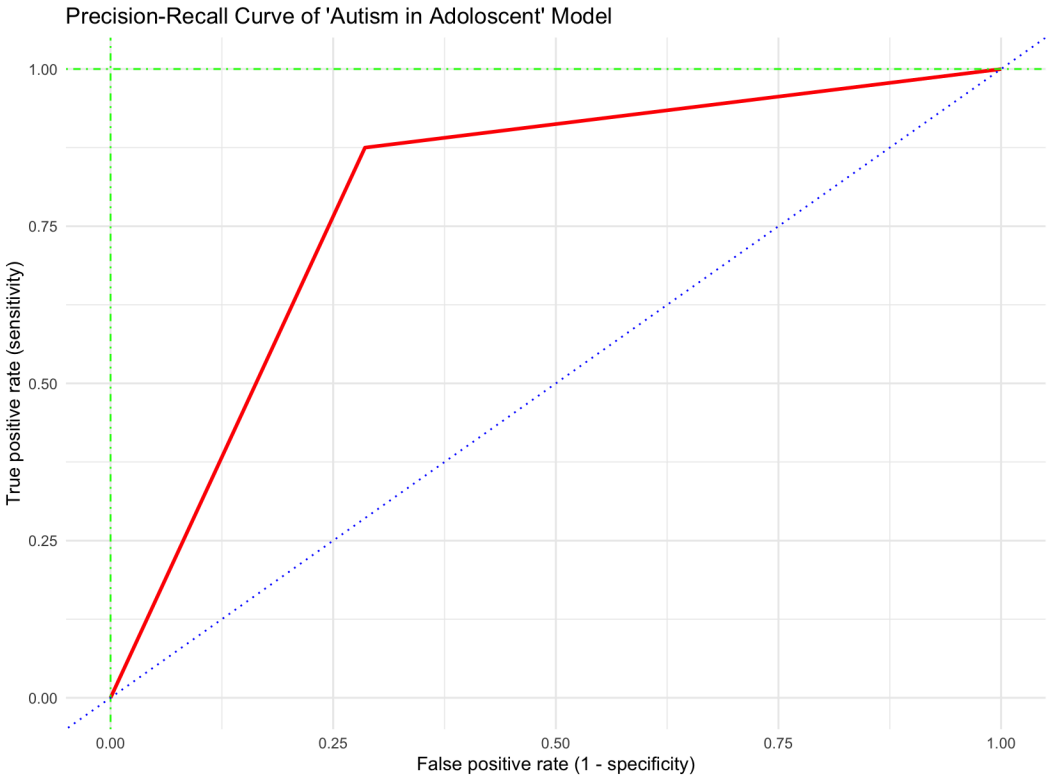

Figure 07: ROC curve of 'Autism in Adolescent' model.

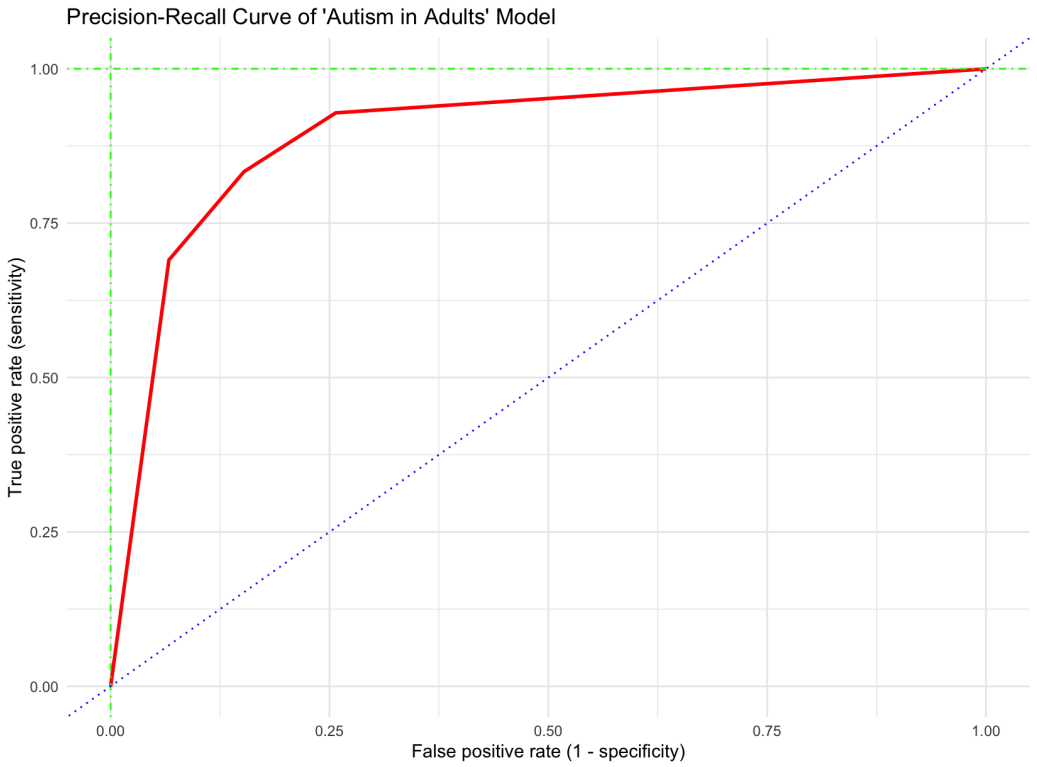

Figure 08: ROC curve of 'Autism in Adults' model.

04. DISCUSSION

In this study, we aimed to develop a predictive model for an autism diagnosis that spans children, adolescents, and adults, leveraging behavioral factors common across age groups. The analysis encompassed the integration of datasets spanning various age groups, aiming to identify predictors of autism diagnosis that remain consistent across different developmental stages.

The logistic regression models developed in our study demonstrated promising performance in predicting autism diagnosis across diverse age groups. Evaluation of the model's performance revealed satisfactory sensitivity, specificity and precision-recall curve across age groups.

The predictor 'multitasking ability' has been the sole influential predictor for the datasets related to children and adolescent ages. On the other hand, 'socializing ability' and 'understanding others' have been the most influential parameters for the adult ages. These results give an insight into the behavioral traits that an examiner has to look for while examining an individual. They can also give an insight into other correlating conditions that a respondent might be suffering with apart from autism. For example 'multitasking ability' has been the top predictor for children and adolescent datasets. This can suggest that (although it needs to be statistically proven) children and adolescents who suffer from autism might also suffer with ADHD (Attention-deficit/hyperactivity disorder).

Having said that, it is important to deploy more powerful predictive models like neural networks to be able to handle large numbers of parameters which we weren't able to do due to the limitations of the dataset. In future work, several aspects of our project could be improved to enhance its impact and applicability. Firstly, expanding the dataset to include larger and more diverse samples from different geographic regions can improve the generalizability of the findings and the model's external validity. Additionally, incorporating longitudinal data to track developmental trajectories and changes in predictive factors over time can provide deeper insights into the dynamic nature of autism symptoms and inform personalized interventions. Moreover, exploring the integration of biomarkers, genetic data, and neuroimaging findings into the predictive model can enhance its accuracy and facilitate customized diagnostics and treatment planning.

05. STATEMENT OF CONTRIBUTIONS

Arjun Pesaru

- Worked on the project presentation, ensuring clarity and effectiveness in communicating key concepts.
- Created visualizations to enhance understanding and presentation of data insights.
- Worked on project proposal and literature survey.

Navya Sri Alugubilli

- Participated in research and documentation tasks, including ML model building for the Adolescent Dataset.
- Contributed to interpreting the results and contextualizing findings within existing literature**.**
- Worked on project proposal and literature survey.

Hiranmai Devarasetty

- Conducted a literature survey to provide foundational knowledge for the project.
- Collaborated on ML model building for the Adult dataset and contributed to interpreting the results.
- Played a role in contextualizing findings within the existing literature.

Sai Nikhil Kunapareddy

- Contributed to ML model building for the Child dataset and participated in interpreting the results.
- Assisted in contextualizing findings within the existing literature.
- Edited the report for clarity, coherence, and adherence to formatting guidelines.

06. REFERENCES

[01] Md Delowar Hossain, Muhammad Ashad Kabir, Adnan Anwar and Md Zahidul Islam, “Detecting Autism Spectrum Disorder using Machine Learning Techniques”.
[02] Autistic Spectrum Disorder Screening Data for Children DOI: 10.24432/C5659W.
[03] https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults
[04] https://aspiedent.com/blog/relationship-between-autism-adhd-dyspraxia-dsylexia/
[05] http://chadd.org/about-adhd/co-occuring-conditions/
[06] https://nda.nih.gov/edit_collection.html?id=2776