JOURNAL OF UNDERGRADUATE RESEARCH

Streamlining Autism Diagnostics: A Machine Learning Approach to Simplifying the ADOS-2 for Accessible ASD Assessment

Aniruth Venkedesh[1], Erin D. Bigler[2] , & Brandon Zielinski[1]

[1]*College of Medicine, University of Florida*
[2] *Brigham Young University, Utah*

Mentor: Dr. Brandon Zielinski, Department of Pediatric Neurology, University of Florida

Abstract

The Autism Diagnostic Observation Schedule-2 (ADOS-2) is the current gold standard assessment tool supporting the diagnosis of Autism Spectrum Disorder (ASD). However, the administration of the ADOS-2 is resource-intensive, requiring specialized training and significant financial investment, leading to delays in diagnosis and limited availability. As a result, many children face significant obstacles in obtaining a timely ASD diagnosis, crucial for accessing early intervention services. This study analyzed a dataset of ADOS-2 assessments taken across 108 children utilizing a simple machine learning approach in order to simplify the assessment down to the most predictive features. Utilizing a recursive feature elimination (RFE) with multivariate linear regression, six key features from the ADOS-2 assessment were identified as playing a critical role in predicting the overall ASD severity score: echolalia, quality of response, stereotyped words, conversation, eye contact, and quality of rapport. The optimized model also demonstrated a robust predictive capability, achieving a Pearson correlation coefficient greater than 0.9 after bootstrapping the algorithm to run 1,000 times. These findings suggest that a streamlined approach to ASD diagnosis may greatly enhance diagnostic efficiency and improve accessibility for early intervention. Future work should focus on validating these findings in a clinical setting, with a larger dataset, and integrating them with other measures of ASD in children.

Keywords: Autism Spectrum Disorder (ASD), ADOS-2 assessment, predictive modeling

Introduction

Autism Spectrum Disorder (ASD) is an early developmental disorder characterized by restricted and repetitive behaviors as well as challenges in communication and social behaviors (American Psychiatric Association, 2013). Research has shown that early diagnosis is critical for accessing intervention that may significantly improve prognosis and developmental outcomes (Dawson et al., 2010; Daniels & Mandell, 2014). Modern diagnosis relies primarily on the Autism Diagnostic Observation Schedule Second Edition (ADOS-2), a comprehensive

assessment of social and communicative behaviors that is considered the gold standard for ASD diagnosis across the world (Lord et al., 2012; McCrimmon & Rostad, 2013). Nevertheless, timely and accurate autism assessments continue to be a challenge for many families, with studies such as Zuckerman et al. (2015) reporting an average delay of 2.7 years between initial parental concern and formal diagnosis due to factors such as high cost of assessment and limited provider availability. These findings are further supported by Daniels and Mandell (2014), which highlights disparities in timely ASD diagnosis due to socioeconomic status, geographic location, and community resources. With the CDC estimating that 1 in every 36 children in the country may have autism, it is now more critical than ever to introduce metrics that address accessibility issues so that more children may receive the treatment they require (Centers for Disease Control and Prevention, 2024).

The growing integration of machine learning and computer models in healthcare presents an opportunity to approach these challenges with a more data-driven approach. Similar approaches have been proven effective in simplifying the screening process for various mental health assessments. For instance, the PHQ-9 and GAD-7 were derived from longer instruments like the Hamilton Depression Rating Scale (HAM-D) using statistical item reduction, enabling quick, reliable screening in primary care (Kroenke et al., 2001; Spitzer et al., 2006). In the context of autism research, simple machine learning models, such as a classifier model or clustering algorithm, have been shown to analyze large datasets and provide key insights into patterns underlying ASD diagnosis (Bone et al., 2016). This study investigates diagnosis under the ADOS-2 by utilizing machine learning algorithms to determine the key predictive features in ASD. If successful, this approach could reduce diagnostic complexity and provide the first stepping stone for a more simplified ASD diagnostic metric that addresses key accessibility issues.

Methods

Dataset and Participants

The study analyzed a de-identified clinical dataset of ADOS-2 assessments collected from 108 children aged 4 to 13 who underwent autism evaluations at the University of Utah. The dataset was originally collected under IRB approval for clinical and research use, and all analyses for the current study were conducted in accordance with applicable data use and ethical guidelines. Out of the 108 children, 58 were diagnosed with ASD and 50 were non-autistic

controls. The data included scores for various behavioral and conversational features measured by trained ADOS-2 administrators. Due to the age and verbal fluency of the participants, the dataset primarily included Module 3 of the ADOS-2, which is designed for verbally fluent children under the age of 16 and assesses a range of social, communication, and behavioral metrics (McCrimmon & Rostad, 2013). The data was also preprocessed to handle missing entries, ensuring that only complete and valid assessments were included in the analysis.

Feature Selection and Model Development

Feature selection was conducted using recursive feature elimination (RFE), a machine learning algorithm designed to identify the most predictive features in a regression model by recursively removing the least important feature (Guyon et al., 2002). RFE begins by training a model on the entire set of features and ranking them based on their contribution to the model's overall predictive accuracy, typically using measures like coefficient weights or feature importances, depending on the model type (Chen & Guestrin, 2016; Kuhn & Johnson, 2013). With each iteration, the model is retrained by removing the least important feature. This process continues until a predetermined number of features remains, or until removing additional features significantly reduces the model's predictive accuracy.

RFE has been successfully used in biomedical and clinical research for reducing dimensionality while preserving model integrity, such as in gene selection for cancer classification (Guyon et al., 2002), Parkinson's detection (Lamba et al., 2022), and in streamlining neuroimaging feature sets for autism diagnosis (Deshpande et al., 2013). Its recursive nature and flexibility across model types make it a powerful and interpretable method for feature selection in healthcare AI applications (Kuhn & Johnson, 2013), particularly where overfitting due to high-dimensional data is a concern.

Preliminary analysis was conducted on the dataset to determine the best type of regression model to use with the RFE algorithm. Initially, an XGBoost regressor was selected for its ability to handle non-linear correlations and interactions between features. However, further analysis revealed a consistent pattern of overfitting given the small dataset, prompting the use of a simpler regression model. This led to the testing of a multivariate linear regression model and a logarithmic regression model. The performance of both models was compared using standard evaluation metrics such as the Pearson correlation coefficient and mean squared error (MSE).

In both metrics, the multivariate linear regression model significantly outperformed the logarithmic regression model, demonstrating a much stronger predictive capability ($r > 0.9$). Given these findings, the multivariate linear regression was selected as the final model for the RFE algorithm, ensuring both predictive accuracy and interpretability.

Validation and Optimal Feature Determination

With a linear regression selected as the base regression model, the data was split randomly into a training and testing set following an 80/20 split. The RFE was set to iterate through the features, recursively removing the least important one until it reached a predetermined number of retained features. To validate the robustness of the selected features, a bootstrapping approach was utilized to run the RFE process 1000 times with different randomized subsamples of the dataset. This procedure helped ensure that the identified features remained consistent over a variety of random subsamples of the dataset.

To determine the optimal number of features for diagnosing ASD, the RFE was applied with varying numbers of retained features, and the Pearson correlation coefficient was recorded for each case and compared as the number of features increased. This approach allowed for an empirical evaluation of how model performance changed as more features were included, helping to identify the point at which additional features no longer provided significant predictive benefits.

Results

Feature Selection and Correlation Growth

The results revealed a logarithmic growth pattern, in which the correlation increased rapidly at first but plateaued after a certain number of features (**Figure 1**).

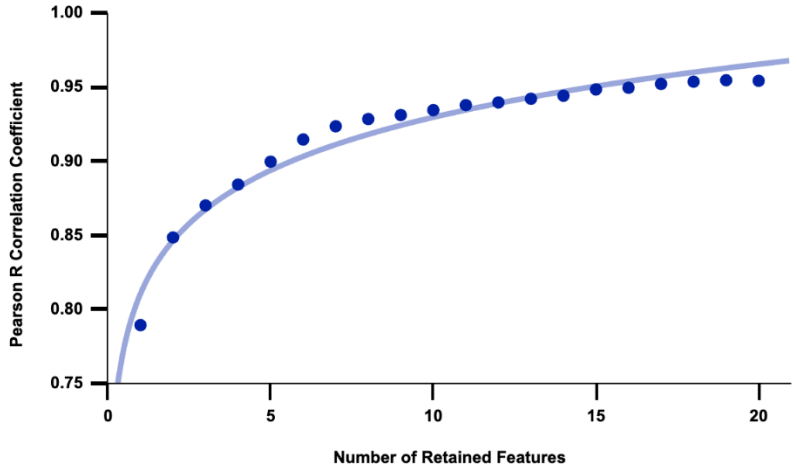

Figure 1. A graphical representation of the relationship between the number of retained ADOS-2 features and the Pearson correlation coefficient of the regression model; a logarithmic regression line is fitted to the data to depict the growth pattern.

The highest correlation coefficient was observed when the maximum number of features were included in the analysis. The greatest increase in correlation occurred between features 1 to 5, after which the gains became progressively smaller. Notably, models with six selected features achieved a Pearson correlation of 0.9141, capturing the majority of the predictive power observed in models with a greater number of features. Notably, six features also mark the start of the plateau phase, where additional features contribute only marginal improvements to the model's performance. This plateau was identified by modelling the correlation values with a logarithmic regression ($r = 0.981$) and analyzing the rate of change by taking the derivative—a technique similar to the "elbow method" described in previous research (Guyon et al., 2002; Satopaa et al., 2011). Beyond six features, the rate of change dropped below 0.01, reinforcing the decision to treat this as the optimal saturation point.

The rate of change in the correlation coefficient between five and six features was still relatively high; however, beyond six features, the increase in correlation dropped to less than 0.01 per additional feature.

Key Predictive Features

To better understand the relative importance of each feature, a feature importance analysis was conducted across all 1000 iterations of the RFE process. Each time the model was trained, feature importance was ranked, and average rankings were calculated for features that consistently demonstrated the highest predictive capability. **Table 1** presents the six most

important features identified by the model with their corresponding average rankings. Lower average values indicate greater importance in predicting ASD severity scores across bootstrapped runs, with quality of rapport, eye contact, and echolalia emerging as the top three contributors.

Table 1. Most Important Features and Average Rankings

Average Ranking	Feature	Description
1.15	Quality Rapport	The level of engagement and emotional connection established between the child and the examiner.
1.68	Eye Contact	The consistency and appropriateness of gaze during social interactions.
1.75	Echolalia	The immediate repetition of words or phrases, often without a clear communicative intent.
1.89	Conversation	The quality of engagement and reciprocal verbal communication between the child and examiner.
2.38	Quality Response	The appropriateness and coherence of a child's responses in a structured interaction as determined by the examiner.
3.66	Stereotypes Words	The presence of repetitive or scripted language in speech.

Note. Descriptions adapted from the *Autism Diagnostic Observation Schedule, Second Edition* (ADOS-2; Lord et al., 2012) and relevant clinical literature (e.g., McCrimmon & Rostad, 2013; Bone et al., 2016).

Out of the six features identified, eye contact and quality of rapport contributed the most to predictive accuracy, followed closely by echolalia and conversation (**Figure 2)**. Quality of response and stereotyped words were slightly less influential but still played a significant role in the model's performance.

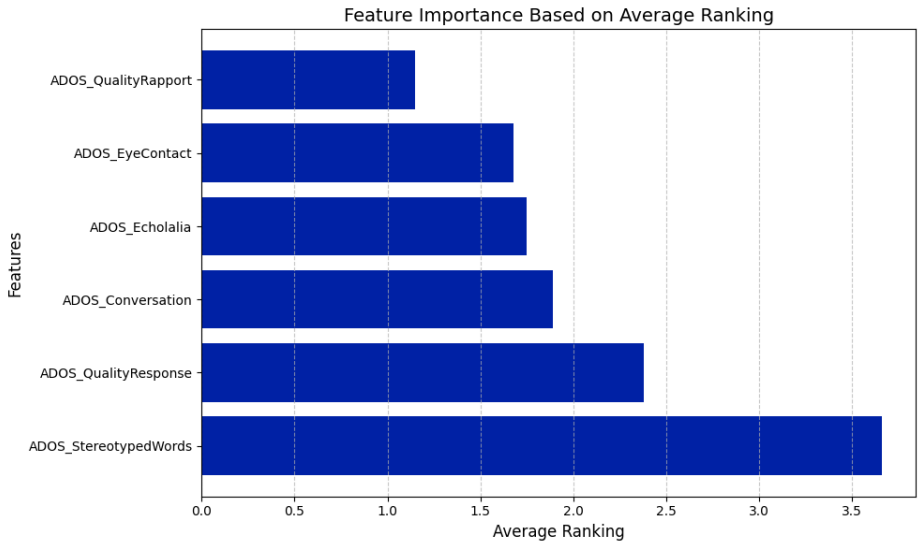

Figure 2. A bar graph illustrating feature importance of rankings provides a clear visualization of how each variable influenced the final prediction.

Model Performance and Predictive Accuracy

Moreover, the multivariate regression model, trained under the six most predictive features, also demonstrated high predictive accuracy when it came to correlating the features with total ASD severity. The average Pearson correlation coefficient over the 1000 samples was >0.9, demonstrating a strong positive relationship between the six features identified and total ASD severity score, further reinforcing their predictive capabilities.

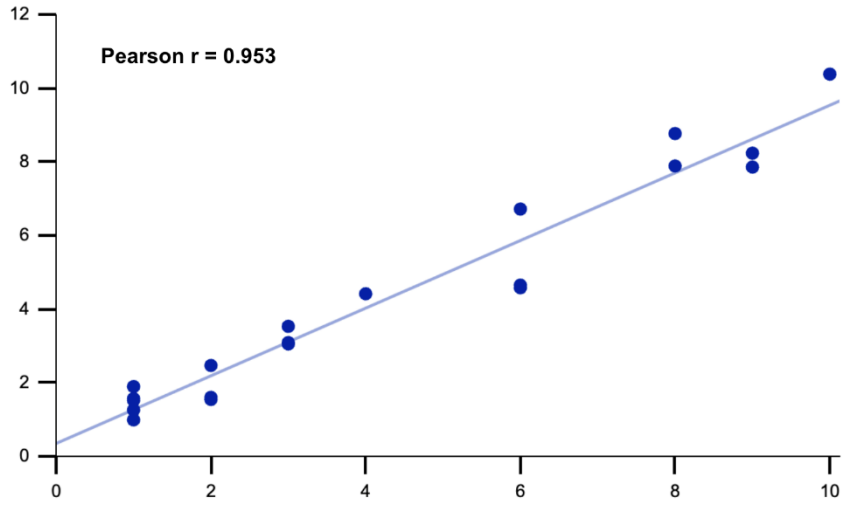

Figure 3. A scatter plot comparing predicted and actual ASD severity scores for the final sample in the bootstrap, demonstrating high predictive capabilities.

Discussion and Conclusion

These results suggest that a streamlined subset of ADOS-2 features can be used to accurately predict ASD severity scores with minimal loss of predictive power. With a distinct plateau and logarithmic growth pattern as the number of features increased, the findings indicate that the majority of predictive strength in the ADOS-2 can be captured by a core set of features, while adding additional features contributes only to marginal improvements. Accordingly, this distinct growth pattern highlights the potential for simplifying the ADOS-2 assessment without sacrificing its effectiveness in a clinical setting.

An interesting point of significance in the data is when the number of retained features in the RFE was set to six, as this represented the smallest feature set that maintained statistical significance ($r>0.90$) while preserving high predictive accuracy. These six features, defined above, align closely with the commonly observed behavioral characteristics in ASD and highlight areas with potential for future research (Lord et al., 2012; Gotham et al., 2007). For instance, features such as echolalia and stereotyped words capture repetitive and restricted language patterns, which are widely considered core traits of ASD (Eigsti et al., 2011). The inclusion of conversation and quality of response in the reduced feature set further emphasizes the importance of social reciprocity in determining ASD severity. These features suggest that conversational ability plays a significant role in distinguishing between ASD severity levels, reinforcing the importance of assessing naturalistic social exchanges in diagnostic evaluations. Notably, research in depression has shown that natural language processing (NLP) can effectively detect clinically relevant linguistic patterns in social media use (Resnik et al., 2015). Accordingly, future research in ASD may benefit from a similar approach, integrating NLP to analyze conversational behavior and offer more efficient diagnosis.

Moreover, the inclusion of quality of rapport as the top-ranking features underscores the importance of nonverbal communication in assessing ASD severity. While the language-based indicators outlined above offer valuable information about receptive abilities, this feature reflects the importance of analyzing the subtle, relational aspects of social communication that are often disrupted in individuals with ASD (Mundy, 2016; Jones & Klin, 2013). Quality of rapport, which emerged as the most predictive feature in almost every iteration of the model, captures the overall ease of communication and connection established between the examiner and the child (Lord et al., 2012). Its consistent high ranking suggests that the relational “feel” of the

interaction, though somewhat subjective in nature, holds a significant diagnostic value. Consequently, experienced assessors heavily influence ultimate diagnosis . The relative importance of this factor reinforces the notion that ASD severity is not solely based on observable behaviors in isolation, but also by the quality and depth of social interactions as a whole (Ozonoff et al., 2005). As such, further research in this area should focus on ways to objectively quantify/measure these features with the potential use of technologies such as facial expression analysis or real-time computer vision (Hashemi et al., 2014; Rudovic et al, 2018).

Notably, eye contact emerging as the second most predictive feature in the ADOS-2 is consistent with recent research and a rising interest in using eye tracking to diagnose ASD. For instance, Jones et al. (2023) showed that eye tracking technology had a sensitivity of 78% when identifying clinically diagnosed children with a certain autism diagnosis. Research has shown the cross-cultural validity of these findings, with Al-Shaban et al. (2023) demonstrating that Arabic-language adaptation of similar eye-tracking technology resulted in comparable diagnostic accuracy, supporting the global applicability.

Additionally, the results also raise important considerations for how these findings can be integrated into clinical practice. With the six features alone demonstrating a statistically significant correlation ($r > 0.95$), a shorter, more focused version of the ADOS-2 may be feasible in certain settings where time, cost, or access to trained examiners is an important consideration. The findings suggest that clinicians may be able to narrow down the focus to more diagnostically prevalent features without sacrificing diagnostic fidelity. Importantly, there is high potential for this to be used as an early screener for guidance on whether or not a full ADOS-2 assessment should be pursued in community-based or low-resource environments, where delays in initial assessments can prevent children from receiving timely early intervention (Zuckerman et al., 2015). Such a screener could even be administered remotely, following machine learning models that are able to use questionnaires and home videos to identify ASD with relatively high accuracy (Abbas et al., 2018). Further studies should investigate the clinical validity of the streamlined feature set by analyzing real-world prospective data in comparison to the full ADOS-2 protocol.

Acknowledgements

This work was supported by the University of Florida AI Scholars Program, which provided resources and guidance for making this possible. Special thanks to Dr. Brandon Zielinski for his mentorship and expertise throughout the development of this project. Appreciation is also extended to Dr. Molly Prigge and Dr. Brandon Zielinski for their foundational contributions to data collection, which enabled this research.

References

Abbas, H., Garberson, F., Glover, E., & Wall, D. P. (2018). Machine learning approach for early detection of autism by combining questionnaire and home video screening. *Journal of the American Medical Informatics Association*, *25*(8), 1000–1007. https://doi.org/10.1093/jamia/ocy039

Al-Shaban, F. A., Ghazal, I., I. Richard Thompson, Klingemier, E. W., Aldosari, M., Hawraa Al-Shammari, Fatema Al-Faraj, Saba El-Hag, Tolefat, M., Ali, M., Nasir, B., & Frazier, T. W. (2023). Development and validation of an Arabic language eye-tracking paradigm for the early screening and diagnosis of autism spectrum disorders in Qatar. *Autism Research*, *16*(12), 2291–2301. https://doi.org/10.1002/aur.3046

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.

Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *57*(8), 927–937. https://doi.org/10.1111/jcpp.12559

Centers for Disease Control and Prevention. (2025, April 15). *Data and Statistics on Autism Spectrum Disorder*. Autism Spectrum Disorder (ASD); CDC. https://www.cdc.gov/autism/data-research/index.html

Chen, T., & Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, *1*(1), 785–794. https://doi.org/10.1145/2939672.2939785

Daniels, A. M., & Mandell, D. S. (2014). Explaining differences in age at autism spectrum disorder diagnosis: A critical review. *Autism*, *18*(5), 583–597. https://doi.org/10.1177/1362361313480277

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., Donaldson, A., & Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. *Pediatrics*, *125*(1), e17-23.

Deshpande, G., Libero, L. E., Sreenivasan, K. R., Deshpande, H. D., & Kana, R. K. (2013). Identification of neural connectivity signatures of autism using machine learning. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00670

Eigsti, I.-M., de Marchena, A. B., Schuh, J. M., & Kelley, E. (2011). Language acquisition in autism spectrum disorders: A developmental review. *Research in Autism Spectrum Disorders*, *5*(2), 681–691. https://doi.org/10.1016/j.rasd.2010.09.001

Gotham, K., Risi, S., Pickles, A., & Lord, C. (2006). The Autism Diagnostic Observation Schedule: Revised Algorithms for Improved Diagnostic Validity. *Journal of Autism and Developmental Disorders*, *37*(4), 613–627. https://doi.org/10.1007/s10803-006-0280-1

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, *46*(1/3), 389–422. https://doi.org/10.1023/a:1012487302797

Hashemi, J., Tepper, M., Vallin Spina, T., Esler, A., Morellas, V., Papanikolopoulos, N., Egger, H., Dawson, G., & Sapiro, G. (2014). Computer Vision Tools for Low-Cost and Noninvasive Measurement of Autism-Related Behaviors in Infants. *Autism Research and Treatment*, *2014*, 1–12. https://doi.org/10.1155/2014/935686

Jones, W., Klaiman, C., Richardson, S., Aoki, C., Smith, C., Minjarez, M., Bernier, R., Pedapati, E., Bishop, S., Ence, W., Wainer, A., Moriuchi, J., Tay, S.-W., & Klin, A. (2023). Eye-Tracking–Based Measurement of Social Visual Engagement Compared With Expert Clinical Diagnosis of Autism. *JAMA*, *330*(9), 854–865. https://doi.org/10.1001/jama.2023.13295

Jones, W., & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, *504*(7480), 427–431. https://doi.org/10.1038/nature12715

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Lamba, R., Gulati, T., & Jain, A. (2022). A Hybrid Feature Selection Approach for Parkinson's Detection Based on Mutual Information Gain and Recursive Feature Elimination. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-021-06544-0

Lord, C., Rutter, M., Dilavore, P. C., & Risi, S. (2012). *Autism diagnostic observation schedule, second edition (ADOS-2)*. Wps.

McCrimmon, A., & Rostad, K. (2013). Test Review: Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part II): Toddler Module. *Journal of Psychoeducational Assessment*, *32*(1), 88–92. https://doi.org/10.1177/0734282913490916

Mundy, P. C. (2016). *Autism and Joint Attention*. Guilford Publications.

Ozonoff, S., Goodlin-Jones, B. L., & Solomon, M. (2005). Evidence-based assessment of autism spectrum disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *34*(3), 523–540. https://doi.org/10.1207/s15374424jccp3403_8

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., & Boyd-Graber, J. (2015). Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. https://doi.org/10.3115/v1/w15-1212

Rudovic, O., Lee, J., Dai, M., Schuller, B., & Picard, R. W. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, *3*(19), eaao6760. https://doi.org/10.1126/scirobotics.aao6760

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011, June 1). *Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior*. IEEE Xplore. https://doi.org/10.1109/ICDCSW.2011.20

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Zuckerman, K. E., Lindly, O. J., & Sinche, B. K. (2015). Parental Concerns, Provider Response, and Timeliness of Autism Spectrum Disorder Diagnosis. *The Journal of Pediatrics*, *166*(6), 1431-1439.e1. https://doi.org/10.1016/j.jpeds.2015.03.007