Written Report on Factors that Help Diagnose Autism

The Repos - Jeffrey Bohrer, Alexandra Green, Anna Zhang, Kevin Lee

2025-03-20

I. Introduction

Autism Spectrum Disorder (ASD) remains a highly prevalent condition despite modern strides made in medical technology. It is reported that nearly 2.2% of adults are affected by ASD, and growing awareness has led to an uptick in diagnoses, particularly in adults who went undiagnosed early in life (Hirota 2023). However, ASD screening tests for all age groups currently contain significant inaccuracies. For example, the most widely used toddler screening test, CHAT-R/FAs, was recently found to produce false negatives in 25% of cases. In contrast, the most commonly used adult autism screening test – the Autism-Spectrum Quotient (AQ) – was found to have limited predictive value in certain populations (Aishworiya 2023; Curnow 2023). Therefore, it has become critical to identify stronger predictors or explore underlying relationships to have more accurate tests and models to predict ASD in adults.

In this study, we will focus on identifying the features that most greatly affect the probability of being encouraged to pursue a diagnosis within this particular questionaire, as created by Prof. Fadi Thabtau of the Manukau Institute of Technology. The data was sourced from users of his app, ASDTests, which screens its users for potential indicators of autism using a ten-question survey. Admittedly then, we are working with the population of individuals or those with relationship to individuals seeking screening for ASD, which could prevent generalization to greater populations. The data set being used will contain nine demographic characteristics – ranging from ethnicity to presence of neonatal jaundice – along with the answers of each individual to the ten behavioral questions of this survey.

Because ASD is difficult to identify and can significantly impair an individual's quality of life, understanding the relationship between demographics, certain behaviors, and their association with autism could encourage individuals to seek diagnosis and gain self-understanding. These adults who receive a positive diagnosis can then access the necessary resources for support. Accordingly, our research question is: what aspects of this questionnaire are most closely associated with being encouraged to pursue an autism diagnosis?

For these purposes, both univariate and bivariate data analyses will be pursued as a means of assessing relevant avenues for further investigation and model fitting. We relabeled the features representing questions in the data set with their actual wording for interpretability, in addition to representing "Yes" as a 1 and "No" as a 0 for their values. The exploratory data analysis of the demographic data and questions can be found in the following pages.

i. Univariate EDA

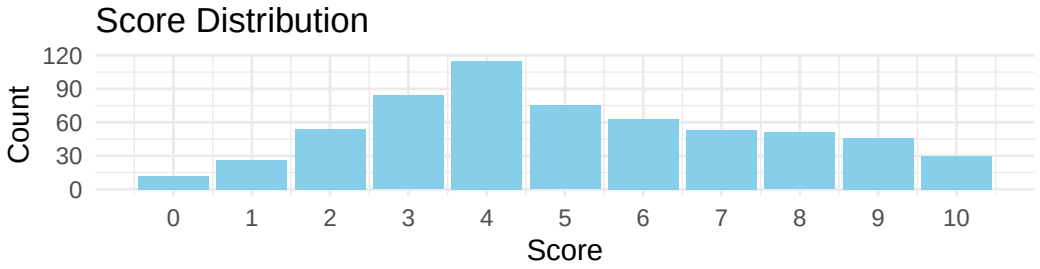


Figure 1: Distribution of total scores across dataset

The variable we are most interested in – the final score of users – ranges from 0 to 10, for each of the 10 behavioral questions. The mean score is 5.077, and the median is 5, both of which are relatively high considering that scores above 6 warrant further diagnostic evaluation. However, because suspecting a diagnosis is a reason for taking the test to begin with, these statistics are reasonable reflections of the test-taking population. We also observe that roughly 30% of subjects are encouraged to seek a diagnosis due to a score higher than 6. We also observe the IQR to be 4 points, as most subjects have a score between 3 and 7 points inclusive. In the context of the data, such a spread is reasonable, and no outliers exist which is understandable given the limited range of scores.

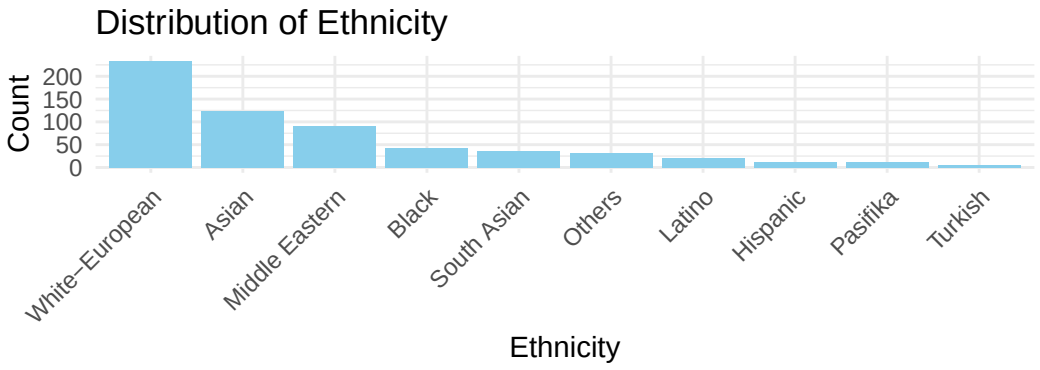


Figure 2: Distribution of ethnicities across dataset

We observe here that the most common ethnicity was White-European, with over 200 observations in our data set. Secondly and thirdly, we observe the Asian and Middle Eastern populations to be above 100 and slightly below 100 observations, respectively. All other ethnicities have fewer than 50 observations in our data set, suggesting it may be more difficult to conclude about these populations. We can now check whether the proportion of individuals encouraged to seek a diagnosis varies by ethnicity:

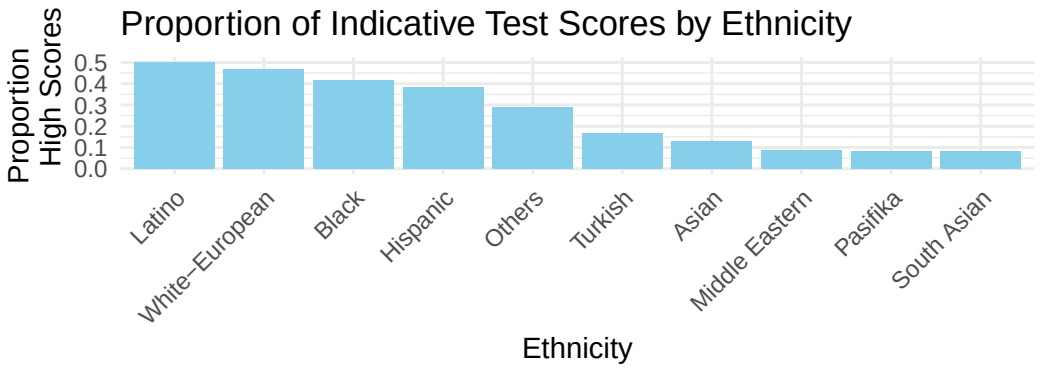


Figure 3: Proportion of indicative score across ethnicities

As we can see here, the population with the highest proportion of test takers receiving a high test score was the Latino population, with close to half achieving a score higher than a 6. Since this is a comparably uncommon population in our data set, the question remains of whether this could be representative of the greater Latino population or merely due to the smaller sample size in our data set. Additionally, the White-European, Black, and Hispanic populations also demonstrate a greater proportion of high test scores. The variation in these proportions according to ethnicity also warrants further exploration into how the distribution of test scores differs by ethnicity. Specifically, the questions remain of whether this means the rates of under- or over-diagnosis vary for different ethnic groups, and if so, what alterations in this questionnaire would be necessary to ameliorate these errors.

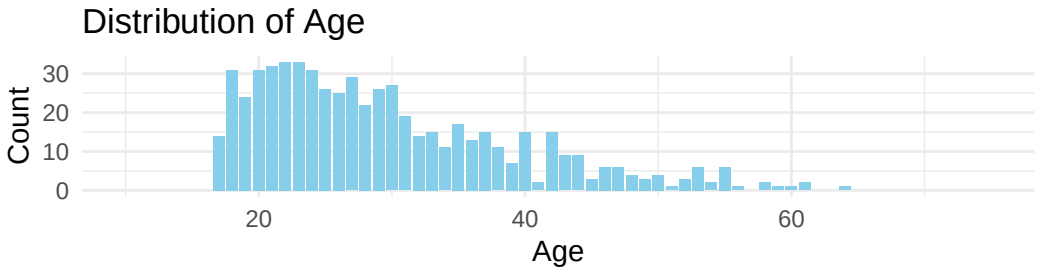


Figure 4: Age distribution across dataset

Generally, we observe the distribution of ages to possess a strong right-skewness, with a mean age of 30.22 and a median age of 27. There is a clear peak in the age distribution roughly around the early to mid-20s. The ages range from a minimum of 17 to a maximum of 383 – a false observation that should be filtered from the data. The IQR is 13 years, which is a fairly small spread given the range of ages, as we observe that the majority of test-takers are under 30 years old.

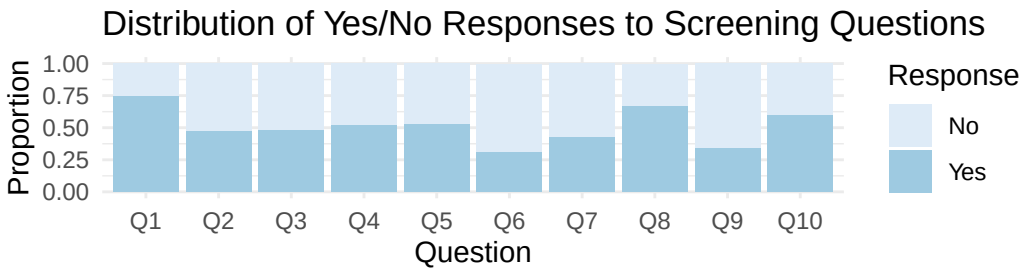


Figure 5: Distribution of Y/N responses for Q1-10. Details of what each question represents can be found in the data dictionary and in appendix.

The two questions with the highest proportion of "Yes" responses were Q1 and Q8, suggesting that these behaviors are common among respondents. On the other hand, Q6 and Q9 had noticeably lower "Yes" responses, potentially indicating difficulties in such areas. Most questions, however, had roughly even proportions between "Yes" and "No" responses.

ii. Bivariate EDA

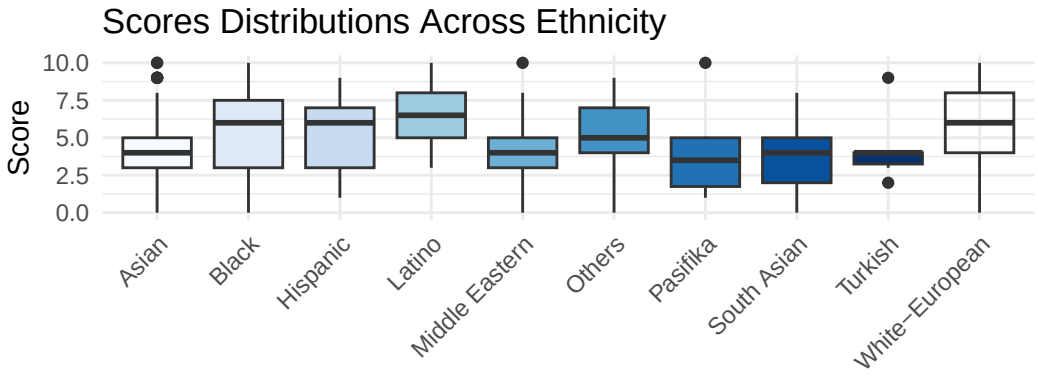


Figure 6: Score distribution across ethnicities

Here we present several graphs displaying bivariate relationships. With our first set of box plots, we find further evidence for our previous suspicion that test score distributions differ by

ethnicity. Although most ethnicities have a median within the range of roughly 3 to 6, some ethnicities, like the White-European, Black, and Hispanic populations, demonstrate a greater spread through their larger IQRs, while the Turkish, Middle Eastern, and Asian populations are much more concentrated around their medians. However, most ethnicities appear to have values almost entirely across the range of 0 to 10 in their test scores.

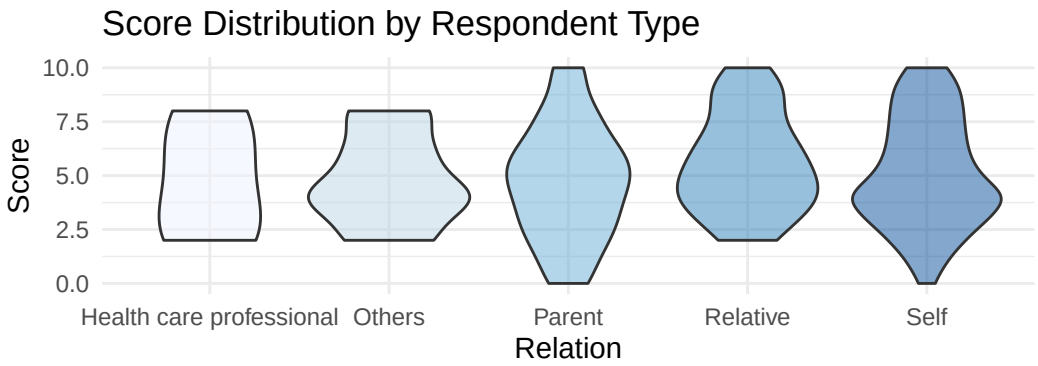


Figure 7: Distribution of scores by respondent type

Interestingly, we can also observe how the relationship between the test taker and the subject of the questions may lead to different distributions of test scores. In the case when it is filled out by a relative or health care professional for example, the observed test score is 2 or greater, while in the case of the test taker being the subject or the parent of the subject, some still received a score of 0. This could reflect how personal biases or relationships affect truthfulness during the test.

This initial exploration leads us to investigate particularly how our demographic data may impact the odds of being encouraged to seek a formal diagnosis, as observed by a high score on this screening test. Though we cannot definitively answer whether it is due to social or cultural perceptions of autism within and surrunding these subgroups or true differences in rates of its presence, it is worth identifying whether over- or under-diagnosis for certain populations is possible.

II. Methodology

i. Choosing Predictors

Firstly, a drop-in-deviance test between a logistic null model without predictor variables and only an intercept, and a logistic model with a single predictor was systematically conducted across ethnicity, gender, presence of neonatal jaundice, and relationship to the subject as a means of assessing which predictors provide a statistically significant improvement in model fit against this null condition. Additionally, we also tested the interaction effects of these demographics; we note that we did not further explore potential interactions with jaundice as that is a neonatal condition and we find it unlikely to have any relationship with the other

variables. Additionally, we did not test ethnicity and relation's interaction effect due to the high number of levels for both (10 and 5, respectively), so distribution of observations across these subgroups limits our sample sizes. These tests were done to gain information on the relationship between demographic data and test scores. The hypothesis for this test can be observed below, where $\beta_{predictor}$ represents the coefficient for a single associated predictor variable (also noted for the case of a single-level variable):

$$H_0 : \beta_{predictor} = 0$$

$$H_a : \beta_{predictor} \neq 0$$

The formulas for the models compared are the following:

$$\text{Null: } \log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \beta_0$$

$$\text{Alternative: } \log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \beta_0 + \beta_{predictor} x_{predictor}$$

Similarly, for our interaction effects, where x_1, x_2 are the predictors for which we test the interaction effect, again using the example of single-level variables:

$$H_0 : \beta_{12} = 0$$
$$H_a : \beta_{12} \neq 0$$

$$\text{Null: } \log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$\text{Alternative: } \log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12}(x_1 * x_2)$$

The following table summarizes these results.

Table 1: Non-obvious predictors drop-in-deviance test results

	Deviance	G statistic (respect to null)	p-value
Null	739.4	NA	NA
Ethnicity	645.3	94.1	0
Relation	738.194	1.206	0.877
Age	737.863	1.537	0.215
Gender	734.94	4.46	0.035
Jaundice	730.073	9.327	0.002
GenderRelation	732.723	0.975	0.914
GenderEthnicity	637.969	4.86	0.846

	Deviance	G statistic (respect to null)	p-value
GenderAge	729.134	4.587	0.032
RelationAge	726.285	10.283	0.036
EthnicityAge	639.044	6.217	0.718

The results show that the demographics that are statistically significant for the model fit are ethnicity, gender, and, quite surprisingly, neonatal jaundice, when compared to a null model, as the p-values are either significantly less than 0.05 (as in ethnicity) or slightly below it (gender and jaundice). Additionally, we acknowledge potential interactions between gender and age and relationship to the subject and age – but as relationship and age on their own both appear largely insignificant, we choose to neglect these interaction effects in our modeling to avoid unnecessary complexity.

We now check our model conditions in all of these cases. We note that in using our statistically significant predictors later, they are all categorical and thus, do not need to assess linearity between log-odds and our predictors for the logistic regression model. We also note that presumably all of the observations of this test are independent due to no obvious spatial or temporal correlations. However, as mentioned before, the model condition at risk is randomness, as our population is primarily those who already suspect autism. This does limit the generalizability of our model to the general population, especially as taking the screening test is voluntary. However, observing the population of those seeking screening tests, we will assume randomness, and again use this primarily as an investigation of the nature of this questionnaire and associations between demographics and scores.

We further investigate the correlation between answers to questions below, as a means of assessing if the test has any pitfalls regarding multicollinearity:

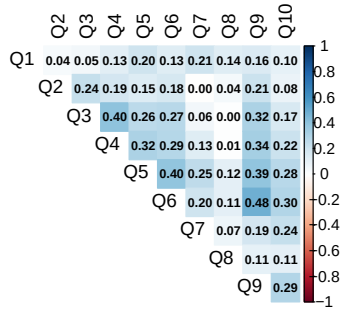


Figure 8: Correlation between AQ10 questions

By computing a correlation heat map among the ten behavior questions to check multicollinearity, we note that a few questions have strong correlations with one another. This is rather unexpected since we would expect many behaviors to be grouped together and more correlated

(e.g., social behaviors). However, the highest correlation is between questions 6 and 9, which are the questions about multitasking and liking to collect information, respectively – this is also a rather surprising pairing.

Considering this reality, principal component analysis was employed to determine the question clusters that explain the highest variance to gain further understanding of the nature of the questions and their variety in the screening tets (Szczęsna, 2022; finnstats, 2021).

Figure 9: Principal component analysis Q1-10 as a means of choosing questions that best explain the variance in the response

We can observe that Q3, Q5, and Q6 have high correlation to Q4, Q6, and Q9, respectively. Beyond these three, the questions that also observe notable loading for principal components (i.e. the parameters that explain the most variance) are Q10, Q8, Q4, Q2, and Q1 (i.e finding it hard to figure out others intentions, ease of going back to work when interrupted, focus on whole picture, difficulty to understand characters in stories, and picking up on small sounds). In general, it seems that some questions form categories that are highly correlated to explain the different subsets of behaviors in autism. However, since the answer to each question might directly be added to the final result if the answer is TRUE, we decide to focus on examining the predictive power of the demographic data on the likelihood of autism.

ii. Fitting the Model

As a means of identifying potential correlations between demographic data and the odds of a high test score, two models will be fitted and compared in terms of a drop-in-deviance test: (a) a null model fitted only to an intercept, and (b) a model that accounts for demographic data. Subsequently, the models fitted will be of the forms:

$$\text{Null: } \log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \beta_0$$

$$\text{Alternative: } \log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \beta_0 + \beta_{ethnicity}X_{ethnicity} + \beta_{gender}X_{gender} + \beta_{jaundice}X_{jaundice}$$

And the hypotheses for the test will be the following:

$$H_0 : \beta_{ethnicity} = \beta_{gender} = \beta_{jaundice} = 0$$

$$H_a : \beta_{ethnicity}, \beta_{gender}, \beta_{jaundice} \neq 0, \text{for at least one} \beta_j$$

This is to assess whether demographic predictors significantly improve model fit and our predictions of the odds of having a high score on the test, and accordingly, being encouraged to pursue a diagnosis. Additionally, AIC and BIC were included for further comparison.

Table 2: Alternative and null model comparison in terms of deviance, AIC, and BIC

Model	Residual_Deviance	Df	AIC	BIC
Null Model	739.40	608	741.40	745.81
Alternative Model	638.44	597	662.44	715.39

Based on the output, we determine that the alternative model (i.e. the one with demographic information) performs significantly better exhibiting lower deviance, AIC, but also surprisingly BIC, which considers its additional complexity. This is quite surprising also considering that ethnicity is a variable with 10 different levels, and we would expect a large penalty for this complexity.

Table 3: Drop-in-deviance test result for alternative and null models

Test	G_stat	df	p_value
Drop-in-deviance (Chi-sq)	100.956	11	0

The drop-in-deviance tests corroborates that these results are statistically significant and that we can reject the null hypothesis since the p-value is less than 0.05. Hence, the alternative model will be further assessed as a means of identifying how demographic data affects the odds of a positive screening test.

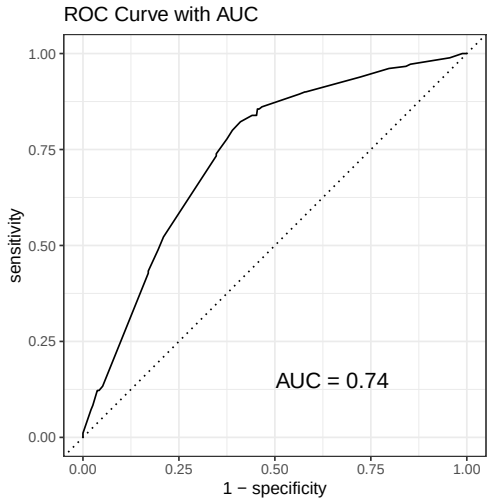
III. Results

Table 4: Output for alternative model

term	estimate	std.error	statistic	p.value
(Intercept)	-0.081	0.166	-0.488	0.625
ethnicityAsian	-1.702	0.300	-5.665	0.000
ethnicityBlack	-0.160	0.339	-0.471	0.638
ethnicityHispanic	-0.195	0.593	-0.329	0.742
ethnicityLatino	0.130	0.472	0.275	0.783
ethnicityMiddle Eastern	-2.184	0.394	-5.540	0.000
ethnicityOthers	-0.703	0.420	-1.676	0.094
ethnicityPasifika	-2.240	1.054	-2.125	0.034
ethnicitySouth Asian	-2.196	0.619	-3.546	0.000
ethnicityTurkish	-1.341	1.106	-1.212	0.225
genderm	-0.291	0.195	-1.489	0.137
jundiceyes	0.640	0.305	2.099	0.036

Among demographic predictors, subjects who identify as Asian, Middle Eastern, or South Asian appear significantly less likely to be categorized as high probability for ASD with odd ratios (ORs) of approximately 0.182, 0.113, and 0.111 respectively, compared to the baseline category of White-European. We also observe that the Pasifika population has a notable odds ratio, but was once again a rather small group of less than 30 individuals and may prove difficult to generalize. This leads us to wonder if these subgroups are more prone to under-diagnosis, or others to over-diagnosis.

In individuals with neonatal jaundice, we also observe that we expect the odds to multiply by a factor of 1.90 compared to those without, holding all else constant – indicating a positive association between this condition and receiving a positive screening test. In practice, this could suggest that individuals with neonatal jaundice should be screened for ASD at a higher rate than their counterparts, as they may require a diagnosis more often; it seems unlikely for these individuals to be experiencing over-diagnosis as the effects of neonatal jaundice are not perceptible later in life and are unlikely to impact others' perceptions of you.



The area under the curve is 0.74, which means that the model, although not a terrible fit, is not particularly remarkable. However, for a model fitted entirely to demographic data which we would normally expect to be independent of one's diagnosis, this could be considered a notable result.

IV. Discussion + Conclusion

Our model uses demographic data such as ethnicity, gender, and neonatal jaundice to predict the likelihood of a positive screening test for ASD, for this given questionnaire examining a variety of behaviors.

As previously noted, the randomness of our sample is in question, which limits the conclusions of our model and its significance. However, this assessment could serve as an encouragement to pursue further research on the subject of the relationship between particular ethnic subgroups and their rates of over- or under-diagnosis, and how neonatal jaundice impacts risk of developing ASD.

Despite a low to moderate AUC value of 0.74, the classification power of the model is notable given the circumstances of its predictors. This indicates that further investigation into how the AQ10 may over- or under-diagnose certain population demographics could still be worth pursuing. This is especially true given that the odds of a positive test appear notably higher for individuals that identify as White, Black, and Hispanic – which could either indicate over-diagnosis of these groups or under-diagnosis of others. Presence of neonatal jaundice could also potentially serve as an additional reason to screen and test individuals. Thus, broader comparisons on the impact of these demographics would be insightful.

V. References

- Aishworiya, R., Kim, V., MA, Stewart, S., Hagerman, R., & Feldman, H. M. (2023). Meta-analysis of the Modified Checklist for Autism in Toddlers, Revised/Follow-up for Screening. *PEDIATRICS*, *151*(6). https://doi.org/10.1542/peds.2022-059393
- Curnow, E., Utley, I., Rutherford, M., Johnston, L., & Maciver, D. (2023). Diagnostic assessment of autism in adults – current considerations in neurodevelopmentally informed professional learning with reference to ADOS-2. *Frontiers in Psychiatry*, *14*. https://doi.org/10.3389/fpsyt.2023.1258204
- Hirota, T., & King, B. H. (2023). Autism spectrum Disorder. *JAMA*, *329*(2), 157. https://doi.org/10.1001/jama.2022.23661
- Marin, Z. (2021, April 26). GLM fit: Algorithm did not converge – How to fix it. Statology. https://www.statology.org/glm-fit-algorithm-did-not-converge/
- finnstats. (2021, May 14). Principal component analysis (PCA) in R. https://www.r-bloggers.com/2021/05/principal-component-analysis-pca-in-r/
- Szczęsna, K. (2022). PCA in R. RPubs. https://rpubs.com/KarolinaSzczesna/862710

VI. Additional Materials

In Figure 5, the distribution of Y/N responses for Q1-10 is illustrated. Here are the details of the 10 behavioral questions (also shown in the data dictionary).

Q1, noticing small sounds;

Q2, finding it difficult to work out character intentions;

Q3, finding it easy to read between lines;

Q4, big picture-oriented;

Q5, can tell if someone listening to me is bored;

Q6, can multitask;

Q7, can tell feelings from faces;

Q8, can go back to work when interrupted;

Q9, enjoy collecting info on categories;

Q10, find it difficult to work out people's intentions.

One thing to note is that while we renamed the ten variables to more descriptive names in our dataset for clarity, we decided to use the Q1–Q10 labels in the visualization for simplicity and cleaner presentation.

In the methodology section, we first performed five drop-in-deviance tests for each of the demographic predictor. In table 1, a summary of the results of the five tests is presented. We would like to show the full output of each test here.

Table 5: Drop-in-deviance test for ethnicity

term	df.residual	residual.deviance	df	deviance	p.value
highProb ~ 1	608	739.4	NA	NA	NA
highProb ~ ethnicity	599	645.3	9	94.1	0

Table 6: Drop-in-deviance test for relation

term	df.residual	residual.deviance	df	deviance	p.value
highProb ~ 1	608	739.400	NA	NA	NA
highProb ~ relation	604	738.194	4	1.206	0.877

Table 7: Drop-in-deviance test for age

term	df.residual	residual.deviance	df	deviance	p.value
highProb ~ 1	608	739.400	NA	NA	NA
highProb ~ age	607	737.863	1	1.537	0.215

Table 8: Drop-in-deviance test for gender

term	df.residual	residual.deviance	df	deviance	p.value
highProb ~ 1	608	739.40	NA	NA	NA
highProb ~ gender	607	734.94	1	4.46	0.035

Table 9: Drop-in-deviance test for jaundice

term	df.residual	residual.deviance	df	deviance	p.value
highProb ~ 1	608	739.400	NA	NA	NA
highProb ~ jundice	607	730.073	1	9.327	0.002