An Investigation of Demographic and Behavioral Factors of Autism Screening Tests

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

2025-04-28

I. Introduction

Autism Spectrum Disorder (ASD) is a highly prevalent condition with nearly 2.2% of adults 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 found to produce false negatives in 25% of cases; 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). Thus, it is critical to identify stronger predictors and explore underlying relationships for more accurate tests 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 a questionnaire created by Prof. Fadi Thabtah 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 (Faizunnabi 2024; Thabtah 2017). The data set has over 600 observations and contains nine demographic characteristics – ranging from gender to history of neonatal jaundice – along with the binary answers to the ten behavioral questions of this survey. An individual of 383 years old was observed as a clear data entry error, so we restricted age to under 120 years old to remove implausible ages. We relabeled the features representing behavioral questions in the data set with their actual wording for interpretability (such as renaming from Q6 to i_can_multitask), in addition to representing "Yes" as a 1 and "No" as a 0 for their values.

Because ASD is difficult to identify and can significantly impair an individual's quality of life, understanding the relationship between demographics, behaviors, and their association with autism could encourage individuals to seek the diagnosis they may need, which then enables access to the necessary resources for support. Accordingly, our research question is: what characteristics of an individual are most closely associated with a high score on this screening test?

i. Univariate EDA

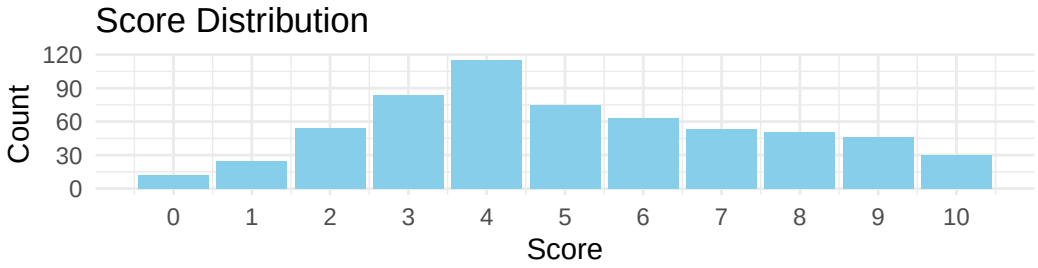


Figure 1: Distribution of total scores across dataset

The final score of individuals ranges from 0 to 10 for each of the 10 behavioral questions. The mean score is 5.084, and the median is 5, which are relatively high considering that scores above 6 warrant further diagnostic evaluation. However, as suspecting a diagnosis is a reason for taking the test to begin with, these are reasonable reflections of the test-taking population. We also observe that roughly 29.6% of subjects are encouraged to seek a diagnosis due to a score higher than 6; the IQR is 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, as no outliers exist.

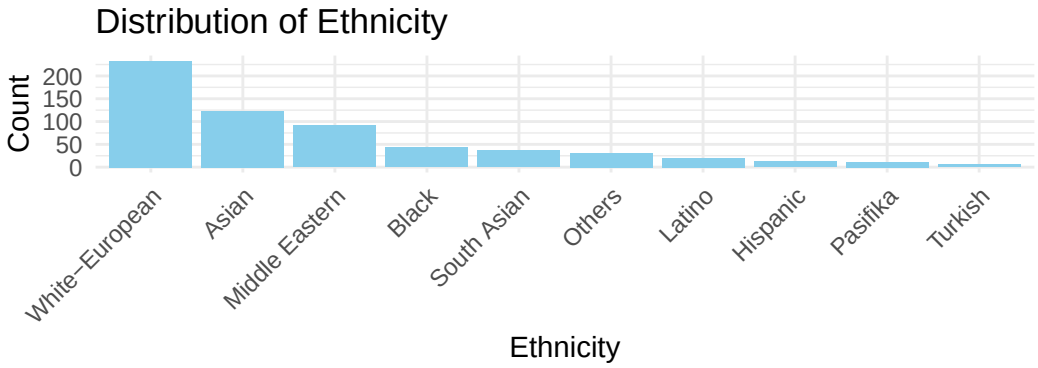


Figure 2: Distribution of ethnicities across dataset

White-Europeans comprised the most common ethnicity of our respondents, with over 200 observations in our data set, while the Asian and Middle Eastern populations have over 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 draw conclusions about those groups. We will later observe the differing score distributions as categorized by ethnicity to better understand the relationship between score and ethnicity.

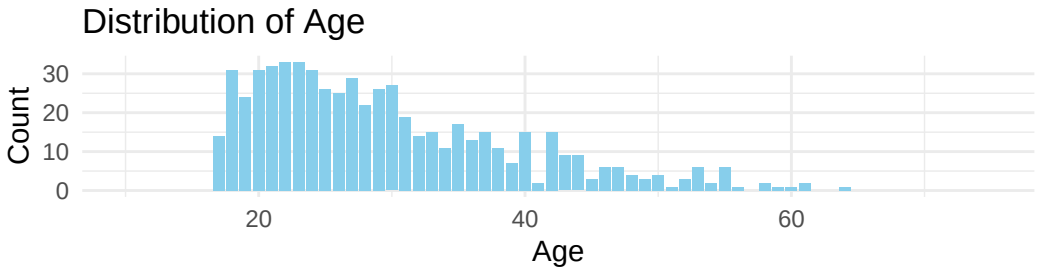


Figure 3: Age distribution across dataset

The distribution of ages possesses a strong right-skewness, with a mean age of 29.63 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 64 years old. The IQR is 13 years, which is a fairly small spread given the range of ages, as the majority of test-takers are under 30 years old.

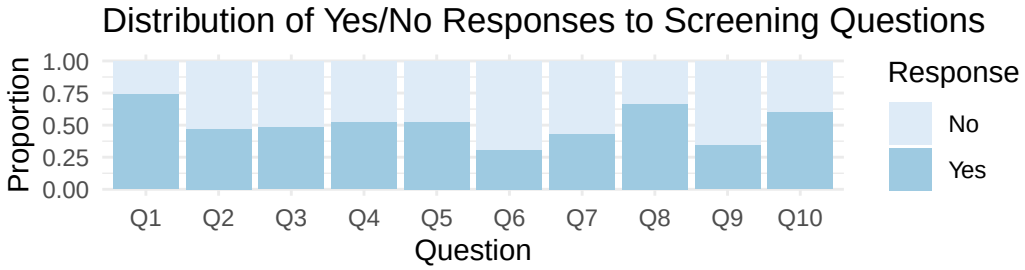


Figure 4: Distribution of Y/N responses for Q1-10 (question details found in appendix).

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

ii. Bivariate EDA

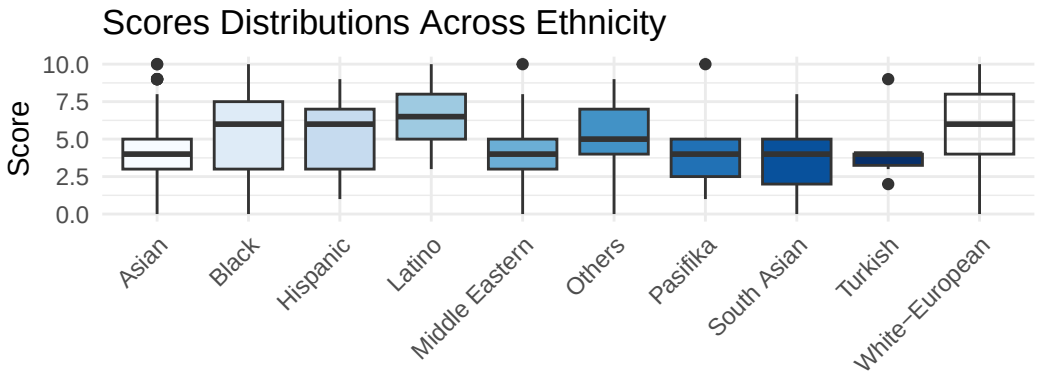


Figure 5: Score distribution across ethnicities

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.

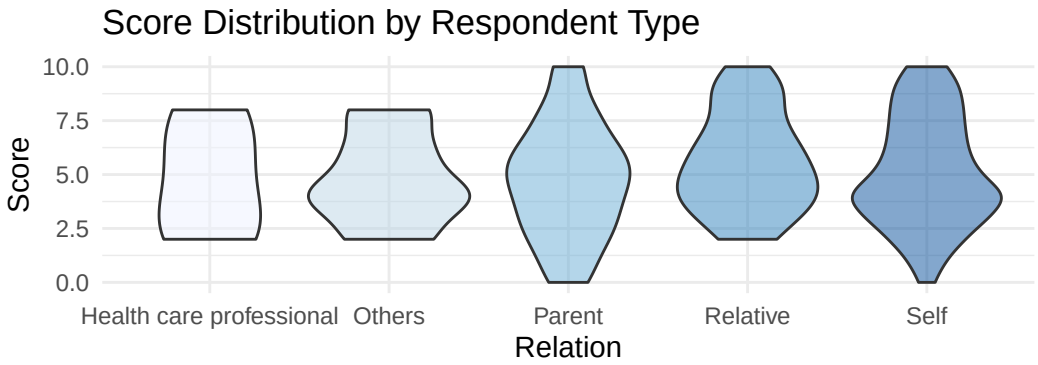


Figure 6: Distribution of scores by respondent type

Interestingly, we can observe how the relationship between the test taker and the subject appears to lead to differing distributions of test scores. 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.

We wonder if this could reflect how personal biases or relationships affect truthfulness during the test.

We choose to continue to investigate 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 surrounding these subgroups or true differences in rates of its presence, it is worth identifying whether over- or under-diagnosis for certain populations is more probable.

II. Methodology

i. Choosing Predictors

We aim to use predictors to determine the probability of obtaining a high score on the screening test. Therefore, we choose logistic regression as the model. 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 explore potential interactions with jaundice as a neonatal condition 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 the distribution of observations across these subgroups severely limits our sample sizes across coefficients. 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	738.699	NA	NA
Ethnicity	645.118	93.581	0
Relation	737.478	1.221	0.875
Age	728.342	10.357	0.001
Gender	734.117	4.582	0.032
Jaundice	729.426	9.273	0.002
GenderRelation	731.886	0.973	0.914
GenderEthnicity	637.567	5.053	0.83
GenderAge	724.316	0.031	0.86
RelationAge	716.49	9.925	0.042
EthnicityAge	638.937	5.864	0.753

The results show that the demographics that may prove statistically significant for our future model are ethnicity, gender, age, and, quite surprisingly, neonatal jaundice, when compared to a null model, as the p-values are either below 0.05. Additionally, we acknowledge potential interactions between relationship to the subject and age – but as relationship to the test subject on its own appears largely insignificant as a predictor, we choose to neglect this interaction effect in our modeling to minimize complexity. We retain the potentially statistically significant predictors for our final model.

We now check our model conditions for these tests. We note that in our choice of statistically significant predictors later, they are all categorical and thus, we 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 apparent spatial or temporal correlations. However, the model condition at risk is randomness, as our population is primarily those who already suspect autism. This does limit the generalization of our model to the overall population, especially as taking the screening test is voluntary, which leads to a clear response bias. However, pertaining to the population of individuals seeking screening tests, we will assume randomness within this group, and again use this primarily as an investigation of the nature of this questionnaire and associations between demographics and scores. We note an additional limitation that certain subgroups (e.g., the Pasifika ethnic group) possess small sample sizes with less than 30 observations. We further investigate the

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

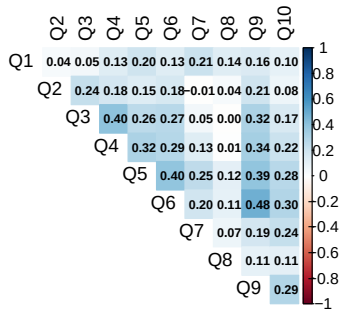


Figure 7: Correlation between AQ10 questions

Unexpectedly, only a few questions have moderate correlations with one another when typically we would expect many behaviors to be grouped together and more correlated (e.g., social behaviors). The highest correlation is between questions 6 and 9, which are the questions about multitasking and liking to collect information, respectively – a rather surprising pairing. Accordingly, principal component analysis was used to determine the question clusters that explain the highest variance to gain further understanding of the nature of the variety of questions in the screening test (Szczęsna, 2022; finnstats, 2021).

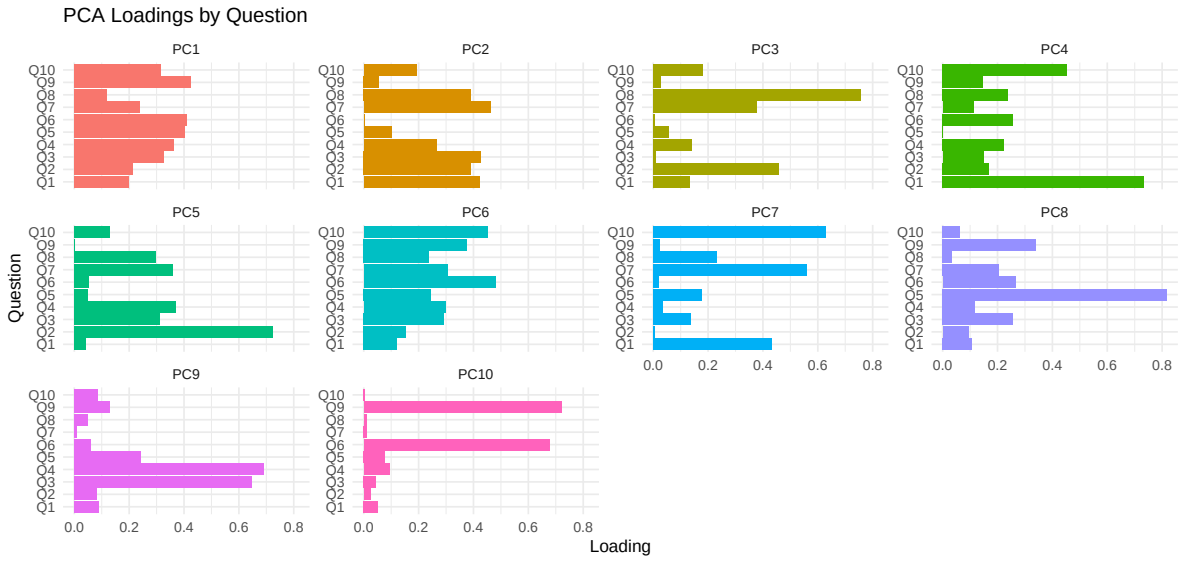


Figure 8: Principal component analysis Q1-10 as a means of choosing questions that best explain the variance in the response

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

To identify correlations between demographic data and the odds of a high test score, two models will be fitted and compared by a drop-in-deviance test: (a) a null model fitted only to an intercept, and (b) an alternative model that accounts for our chosen demographic predictors.

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

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

And the hypotheses for the test will be the following:

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

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

This is to assess whether the chosen demographic predictors significantly improve model fit and our predictions of the odds of obtaining a high score on the test, and consequently, being encouraged to pursue an ASD diagnosis. We include AIC and BIC for further comparison. We note that we reasonably assume no multicollinearity between demographic information.

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

Model	Residual_Deviance	Df	AIC	BIC
Null Model	738.70	607	740.70	745.11
Alternative Model	638.12	595	664.12	721.46

Based on the output, we observe that the alternative model (i.e. the one with demographic information) performs better exhibiting lower deviance, AIC, but also surprisingly BIC, which more harshly penalizes its higher complexity.

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

Test	G_stat	df	p_value
Drop-in-deviance (Chi-sq)	100.5752	12	0

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

III. Results

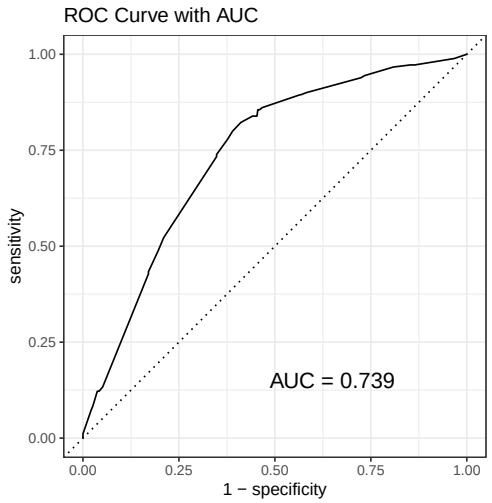
A final model was fitted and is shown below. Note that age was discarded as a predictor because it showcased a p-value much greater than 0.05 (0.74) (see additional materials b.). Below is the model that includes all previously mentioned parameters without age.

Table 4: Output for alternative model w/o age

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.080	0.166	-0.483	0.629	-0.407	0.245
ethnicityAsian	-1.702	0.300	-5.665	0.000	-2.321	-1.137
ethnicityBlack	-0.160	0.339	-0.471	0.637	-0.835	0.500
ethnicityHispanic	-0.195	0.593	-0.328	0.743	-1.429	0.948
ethnicityLatino	0.130	0.472	0.276	0.782	-0.808	1.068
ethnicityMiddle Eastern	-2.184	0.394	-5.540	0.000	-3.031	-1.466
ethnicityOthers	-0.703	0.420	-1.676	0.094	-1.573	0.091
ethnicityPasifika	-2.138	1.059	-2.018	0.044	-5.057	-0.455
ethnicitySouth Asian	-2.196	0.619	-3.546	0.000	-3.644	-1.130
ethnicityTurkish	-1.341	1.106	-1.212	0.226	-4.304	0.511
genderm	-0.293	0.195	-1.498	0.134	-0.677	0.090
jundiceyes	0.639	0.305	2.096	0.036	0.040	1.241

We note that gender and many ethnic groups possess p-values greater than 0.05. Among statistically significant 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 expected odds ratios (ORs) of approximately 0.182, 0.113, and 0.111 respectively, compared to the baseline category of White-European, holding all else constant. This leads us to wonder if these subgroups are more prone to under-diagnosis or others to over-diagnosis. The Pasifika population also has a notable odds ratio (0.118), but cannot generalize due to small sample size.

In individuals with a history of neonatal jaundice, we expect the odds of a positive screening test to multiply by a factor of 1.895 compared to those without, holding all else constant. In practice, this could suggest that individuals with neonatal jaundice should be screened for ASD at a higher rate than their counterparts, as we note it seems unlikely for these individuals to experience bias in their screening and diagnoses since the effects of neonatal jaundice are nearly imperceptible as adults. For that reason, we do not suspect over-diagnosis for this group.



The area under the curve is 0.739, which means that the model, although not a terrible fit, is likely insufficient to explain the variation in the odds of a high score on its own. However, as a model fitted entirely to demographic data that we would expect to be independent of one's screening result, this is a notable result.

IV. Discussion + Conclusion

Our model predicts the likelihood of a positive ASD screening demographica results, based on a test examining various behaviors and attitudes. Although the AUC value is modest (0.739), the model's classification performance is notable given the predictors which are not derived from the test. Nonetheless, the non-randomness of the sample and small subgroup sizes limit the conclusions. These findings, however, highlight the need for further investigation into the relationship between ethnicity, neonatal jaundice, and ASD diagnosis. Future research, with a properly random sample and large sample sizes for every subgroup, could prove beneficial to generalizing these conclusions. In particular, the notable higher odds of a positive test among White, Black, and Hispanic individuals could suggest potential for over-diagnosis and test bias, and neonatal jaundice could act as a risk factor for ASD. Learning of the former could encourage further steps in addressing biases within healthcare practices, while evidence of the latter could encourage early screening for individuals with the condition. Overall, broader comparisons and analysis on the bias of these test could provide stronger frameworks for understanding ASD diagnostic pattern in current healthcare and psychology.

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
- Faizunnabi, F. (2024). Autism Screening. https://www.kaggle.com/datasets/faizunnabi/autism-screening
- finnstats. (2021, May 14). Principal component analysis (PCA) in R. https://www.r-bloggers.com/2021/05/principal-component-analysis-pca-in-r/
- 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/
- Szczęsna, K. (2022). PCA in R. RPubs. https://rpubs.com/KarolinaSzczesna/862710
- Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com

VI. Additional Materials

a. Figure 5: Details

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.

b. Alternative Final Model With Age

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.194	0.362	-0.534	0.593	-0.904	0.519
ethnicityAsian	-1.682	0.305	-5.511	0.000	-2.310	-1.107
ethnicityBlack	-0.150	0.340	-0.442	0.659	-0.828	0.512
ethnicityHispanic	-0.191	0.593	-0.322	0.747	-1.425	0.952
ethnicityLatino	0.147	0.475	0.310	0.757	-0.796	1.090
ethnicityMiddle Eastern	-2.161	0.399	-5.409	0.000	-3.016	-1.432
ethnicityOthers	-0.689	0.422	-1.634	0.102	-1.562	0.109
ethnicityPasifika	-2.114	1.061	-1.992	0.046	-5.036	-0.425
ethnicitySouth Asian	-2.173	0.623	-3.489	0.000	-3.625	-1.098
ethnicityTurkish	-1.322	1.107	-1.194	0.233	-4.287	0.533
genderm	-0.293	0.195	-1.497	0.134	-0.677	0.090
age	0.003	0.010	0.353	0.724	-0.016	0.023
jundiceyes	0.628	0.307	2.050	0.040	0.026	1.233

A model that took into account age was previously fitted as it proved to be significant in the drop-in-deviance test with respect to the intercept. However, further inspection indicated it was not a significant predictor, as it had a p-value of 0.724. Subsequently, it was removed as a predictor.