International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2020.121927

Scopus

Research Article

IJCRR
Section: Healthcare
Sci. Journal Impact Factor: 6.1 (2018)
ICV: 90.90 (2018)

Analysis of Autistic Spectrum Disorder Screening Data for Adolescents

Tava Manggai Muthisamy[1], Manoj Jayabalan[2], Muhammad Ehsan Rana[1]

[1]School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, 57000, Malaysia; [2]Faculty of Engineering & Technology, Liverpool John Moores University, Liverpool, L33AF, UK.

ABSTRACT

Background: The increase in the prevalence of Autism Spectrum Disorder (ASD) and the complications related to effectively diagnosing the disorder has led to an urgent need for the development of easy implementing and effective screening methods.

Problem: Diagnosing ASD is often a lengthy process that requires extremely careful assessment by clinical experts for an accurate diagnosis. This results in lengthy waiting times and high cost for an ASD diagnosis. Although ASD can be a life-long disorder, early treatments and education can greatly improve a person's symptoms and ability to function. Hence, fast and effective screening methods that are easily accessible is imperative to help health professionals and inform individuals whether they require a formal clinical diagnosis for ASD.

Purpose: Thus, this study focused on analysing a dataset from an ASD screening tool, to meticulously process the raw data and derive meaningful insights from it. This study made use of a dataset that contained information about adolescent from age 12 to 16. The steps involved from the pre-processing of the data to analysing the data to deriving meaningful insights from the data are discussed in detail. Findings from the data are critically discussed and suggestions for improvement are also provided.

Conclusion: In a nutshell, this study aims to serve as a complete guide to effectively manage, from raw data to derive meaningful insights, regarding ASD screening.

Key Words: Autism spectrum disorder (ASD), Review, Adolescent, Autism Spectrum Quotient, Hetero Anamnestische Persoonlijkheidsvragenlijst (HAP), Quantitative Checklist for Autism in Toddlers (Q-CHAT).

INTRODUCTION

Autism spectrum disorder (ASD) is a complex neurodevelopment disorder that affects social functioning, communication and behaviour of an individual [1]. People with ASD often face difficulty in communicating and interacting with other people. They tend to have restricted interests and repetitive behaviours [2]. This usually affects a person's ability to function properly in school, work and other areas of life. ASD is a life-long condition that requires extensive educational, vocational and community support.

Although autism has a significant genetic component, it is primarily diagnosed through behavioural characteristics [3]. Thus, diagnosing ASD is often a lengthy process that requires extremely careful assessment by clinical experts for an accurate diagnosis. This results in lengthy waiting times and high cost for an ASD diagnosis. Although ASD can be a life-long disorder, early treatments and education can greatly improve a person's symptoms and ability to function. Hence, fast and effective screening methods that are easily accessible is imperative to help health professionals and inform individuals whether they require a formal clinical diagnosis for ASD. The features that comprise behavioural and individual characteristics are a key component of ASD screening. For effective analysis of obtained data, the data must be preprocessed (i.e. cleaned and transformed) and the right analytical techniques and tools must be applied to generate valuable insights that would help in ASD screening.

In this study, ASD Screening Dataset for Adolescent is used. The purpose of this study is to process and analyse the ASD Screening Dataset for Adolescent and derive some meaningful hypothesis. The dataset consists of information collected using a mobile app of subjects between 12 to 16 years of age. Their particulars such as age, gender, ethnicity, country of

Corresponding Author:

Manoj Jayabalan, Faculty of Engineering & Technology, Liverpool John Moores University, Liverpool, L33AF, UK.
Email: m.jayabalan@ljmu.ac.uk

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 14.07.2020 **Revised:** 16.08.2020 **Accepted:** 18.09.2020 **Published:** 06.10.2020

residence are recorded. The key information in this dataset would be their response to a set of ten questions asked to determine the autistic symptoms. These ten questions are the outcome of research by Allison et al. [4]. The authors aimed to identify 10 items from the Autism Spectrum Quotient (AQ) (Adult, Adolescent, and Child versions) and on the Quantitative Checklist for Autism in Toddlers (Q-CHAT) with good test accuracy. Their research shortlisted these 10 questions that serve as 'red flag' for frontline health professionals to aid their decision on making a referral for a full diagnostic assessment for an autism spectrum condition in children and adults. In the dataset, the answers are recorded in binary form with 1 indicating a yes and 0 indicating a no. The results of each respondent are then accumulated and those who scored above seven for the ten questions in the questionnaire are categorized as ASD. The dataset also consists of information to other questions such as, if the subject had jaundice before if the subject is known as autistic, if the respondent has used the app before and what is the relationship of the respondent to the person being tested.

Related works

Due to the growing prevalence of autism, many research works have focused on developing new methods to screen ASD most cost-effectively and quickly. There have been several clinical and non-clinical diagnosis methods developed to diagnose autism at any age. The methodologies applied to detect ASD in different ages of individuals are customized to effectively detect ASD for the particular age group. Thus, in this section, related work in ASD screening categorized by different age groups, namely Infants, Children, and Adults will be reviewed.

ASD Screening for Infants

Although autism can be diagnosed at any age, it is said to be a developmental disorder because symptoms generally appear in the first two years of life. Thus, researchers have focused their research work specifically on detecting ASD at the infant stage. A study was performed to detect early autistic features in infants that are born premature and underweight by Limperopoulos et al.[5] For this research, the authors studied 91 ex-preterm infants below 1500g during birth. Infants underwent conventional MRI studies and data concerning demographic, prenatal, intrapartum, acute postnatal, and the short-term outcome was collected for all infants. Descriptive statistics were calculated on all measures and multiple logistic regression analysis procedures were then applied to identify the final predictive model. Subsequently, follow-up assessments were performed at a mean age of 22 months. Final results showed that 26 % of ex-preterm infants had a positive result on the autism screening tool. Early autistic behaviours seem to be an under-recognized feature of very low birth weight infants. The results from this study show that early screening should be a necessity during infancy to detect symptoms of early autism. It can then be followed by clinical autism testing in those with positive screening results.

In another research by Olliac et al.[6], the authors aimed to assess the effectiveness of PREAUT grid tool to test infants less than 12 months old. The PREAUT grid evaluates the infant's ability to spontaneously engage in synchronous and joyful interactions. The authors assessed the ability of the PREAUT grid to predict ASD in low-risk individuals by screening 12,179 infants with the PREAUT grid at four (PREAUT-4) and nine (PREAUT-9) months of age. The infants who screened positive were proposed for a further clinical assessment. 100 infants were identified and proposed for further clinical assessment. However, only 45 of them received follow-up diagnosis and among those, 22 were healthy, 10 were diagnosed with ASD, seven were with intellectual disability (ID), and six had another developmental disorder. This research showed that the PREAUT grid holds the potential for early detection of ASD. The combined use of PREAUT grid with other ASD assessment tools may improve the early diagnosis of ASD and other neurodevelopmental disorders at an infant stage.

ASD Screening for Children

Most research works in this field are targeted at detecting ASD for children. One of the most common tools used for ASD screening is the Autism Diagnostic Interview-Revised (ADI-R) method. It is an exam-based assessment that consists of 93 questions that must be answered by a care provider within a focused session that often spans over 2.5 hours. In this research work by, the authors used machine learning techniques to study the complete sets of answers to the ADI-R. Dataset consisted of 891 children diagnosed with autism and 75 children who did not meet the criteria for an autism diagnosis. 15 different machine learning algorithms were tested in this study and the Alternating Decision Tree (ADTree) algorithm was found to have high sensitivity and specificity in the classification of individuals with autism. The interesting findings from this research suggest that 7 of the 93 questions contained in the ADI-R were sufficient to classify autism with 99.9% statistical accuracy. Thus, this research is significant for a fast and accurate behavioural detection of autism and greatly helps in addressing the time burden issue. This study could assist in streamlining the clinical diagnostic process overall, leading to faster screening and earlier treatment of individuals with autism.

In another interesting research by Quach [8], the author aimed to analyse genetic data on a genomic scale by creating a generalized autism diagnosis method. The author made use of Artificial Neural Network (ANN)'s capability to correctly classify large sets of input when presented with enough data. This research used a multilayer perceptron network architec-

ture with a backpropagation training method. The trained ANN network was fed with large microarray gene expression data to determine if a child is having ASD. Gene expression level data were collected from 222 children with and without ASD aged 18-36 months. Of the 222 subjects, 150 were used for training and 72 were used for independent testing. This research was able to successfully identify children with ASD at a 73 % accuracy rate. Although the accuracy rate is not sufficient to be accepted as a clinical testing method, this research holds a lot of potentials and opens up new areas for improvement.

In 2016, the authors Liu et al. proposed a new approach to use facial scanning for the screening of ASD [9]. This study analysed if patterns from facial scanning can be potentially used to identify children with ASD. Machine learning algorithm was applied to the classification of patterns. The Support Vector Machine (SVM) classification algorithm was used to analyse an eye movement dataset from a face recognition task to classify children with and without ASD. The result of the study was promising as it achieved 88.51 % of accuracy. With effective implementation, this research would highly contribute to a faster screening of ASD and earlier treatment of individuals with autism.

ASD Screening for Adults

Compared to ASD screening of children as discussed in the previous section, screening ASD for adults are often more challenging. They may not have been diagnosed at a younger age due to symptoms that were not strong enough to classify them as autistic. Thus, a few types of research have focused on studying ASD diagnosis specifically for adults. One of such research was developed by Sadeghi et al. (2017). In this study, the authors proposed an automatic screening method for recognition of ASDs from healthy controls (HCs) based on their brain functional abnormalities. Brain functional networks of 60 young adult males (29 ASDs and 31 HCs) were estimated from subjects' task-free fMRI data. The SVM classification algorithm was applied in this study, to assess the performance of the system. The results showed an accuracy of 92 %. This study suggests that besides monitoring of individual's characteristics and behaviours, the local parameters of the brain functional network can potentially be used for autism screening.

In more recent work, the authors Heijnen-Kohl et al. [11] examined if ASD symptoms in older adults above 60 years old can be detected with the Dutch informant personality questionnaire, Hetero Anamnestische Persoonlijkheidsvragenlijst (HAP). 40 adults with autism were compared with another 43 patients with a different psychiatric diagnosis. It was observed that the scales used in this questionnaire was able to effectively distinguish between individuals with or without ASD. Thus, the HAP questionnaire can be used as a screening tool for ASD symptoms in elderly people. This tool can function as a screening tool to determine if the further clinical assessment was required for an adult suspected with autism.

The review shows that some many different methodologies and tools are being actively developed and tested by researches with the main aim to make the ASD screening process faster, cost-effective and easily accessible for patients and medical practitioners. Different types of patients' data are often used in these studies. Although having huge sets of patients' data greatly help researches in developing more substantial tools; there must be equal attention given to privacy and security of these data. Especially in medical researches as this, in which personal data of patients are very often collected and used. Written consent from all the test subjects must be made compulsory and the terms and conditions stated must be strictly adhered by the researches. A framework should also be in place to secure the data during the research stage and upon completion, the data should be discarded accordingly. It was found during this review that, only a handful of literature are stressed on data privacy and data use, namely, [2,5,6,12–14].

Methods

The section describes the complete data cleaning process that is applied in this study. Each stage namely the data exploration, data pre-processing, data transformation and data analytics will be discussed in detail. Data exploration analysis was performed on the obtained dataset to understand the data and its characteristics. SAS Studio program was used to explore the characteristics of the data, size of data, completeness of the data, the correctness of the data and the possible relationship amongst data elements in the dataset. Visual explorations were used to identify the missing values, outliers and noisy data among the hundreds of data in the dataset. The identified issues such as missing values and noisy data were treated using the same SAS Studio platform. The methods applied for data pre-processing and the justifications are well described in the corresponding section. The cleaned data was then transformed into the standard qualitative data format. All the continuous data were transformed to categorical data in this stage.

The Autistic Spectrum Disorder screening data for adolescents was used in this study. This dataset contains a total of 104 observations and 21 variables. This data was collected through a questionnaire via the mobile app by asking participants the patient's particulars such as age, gender, ethnicity, country of residence if they had jaundice if they have autism and ten standard questions that are used to determine if the patient has ASD. The list and type of 21 variables in this dataset are shown in Table 1.

Table 1: Types of Data Attribute

No	Variable	Type
1	A1_Score	Nominal
2	A2_Score	Nominal
3	A3_Score	Nominal
4	A4_Score	Nominal
5	A5_Score	Nominal
6	A6_Score	Nominal
7	A7_Score	Nominal
8	A8_Score	Nominal
9	A9_Score	Nominal
10	A10_Score	Nominal
11	Age	Ratio
12	Gender	Nominal
13	Ethnicity	Nominal
14	Jaundice	Nominal
15	Autism	Nominal
16	Country_of_res	Nominal
17	Used_app_before	Nominal
18	Result	Interval
19	Age_desc	Nominal
20	Relation	Nominal
21	Class/ASD	Nominal

The initial data exploration was performed and identified that for each variable there are 104 records. This indicates that there are no missing values in the dataset. Each of the data attributes was further analysed to identify any outliers, missing values or inconsistency as discussed in subsequent sections.

A1-A10 Score: For attribute A1_Score to A10_Score, the results are represented in a binary form of 0 or 1 based on the answer for questions by the respondent. The frequency of the data for each attribute was checked to confirm if there are any sparse class. It was noticed that all the 0 and 1 answers are normally distributed, and no question score stands out.

Age: The age attribute was analysed and that participant of all ages within the adolescent age range of 12 to 16 has been screened in this test. It can be noted that the highest number of participants was 16 years old. The age distribution was further analysed no outliers are detected.

Gender: The distribution of male and female respondents was almost equal. There were four more female participants compared to male respondents. Further analysis was done on the data to determine if they were balanced gender for all ages of respondents. There was more than 40% representation from both male and female. Thus, it can be considered as normally distributed data.

Ethnicity: Initial exploratory analysis on ethnicity attribute revealed that there are some missing values in the dataset that are indicated by "?". There is a total of 6 rows that have a missing value for ethnicity. This missing value has to be cleaned for an overall analysis of this dataset.

Jaundice: For attribute known as jaundice the answers are given in yes or no form. The frequency of the answers is normal.

Austin: For attribute known as Austin, the answers are given in yes or no form. There is no apparent outlier from this result. However, the attribute names seem to have spelling errors, as the correct terms should be "autism".

Contry_of_res: The attribute contry_of_res indicates the country of residents of the respondents for this ASD screening questionnaire. The frequency of the countries under this attribute seems normal. The highest number of respondents came from the United Kingdom, followed by the United States and Argentina. Other participants came from various countries around the world.

Used_app_before: Used_app_before attribute only has 'yes' and 'no' options. Based on the analysis, there are 96% of respondents answered 'no' to this question. Although the percentage for the 'yes' and 'no' may not seem balanced, this could be logical, for someone to use the app for the first time to do the ASD Screening test. There are fewer chances for one respondent to perform this test twice. Thus, it is more likely that only new users participate in this questionnaire, resulting in a higher percentage of 'no' answers.

Result: The result is an important attribute in this dataset as it is used to determine if a respondent is classified as ASD. Result values are counted from the number "1"s that are scored in question A1 to A10. Result obtained that is equal to or higher than 7 are classified as ASD. The result shows that there is a higher percentage of respondents who can be classified as ASD as their result value is higher than 7. Out of the 104 observations, there are 5 respondents who have answered "1" to all ten questions from A1 to A10. This could be a suspicious noisy data, as sometimes human errors or random answers could lead to such a situation.

Age_desc: This attribute indicates the age range for adolescents. Two age categories are observed in this data set. As they are both within the 12 to 16 years range, they both can be accepted as valid data. However, the 12-15 years group will be transformed and standardized as 12-16 years.

Relation: The relation attribute indicates the relationship of the person filling up the questionnaire with the person being tested for ASD. It could be observed that there is one value that is not defined but shown as the character "?". There is a total of 6 rows that have this missing value for relation variable. Another interesting point to note here is that these 6

missing values are consistent with the missing values for ethnicity. It is highly possible that these particular respondents ignored this information while filling up the questionnaire. This missing value has to be cleaned for an overall analysis of this dataset.

Class/ASD: The last attribute known as Class/ASD shows the result of the ASD screening. For respondents who scored seven or more in the result column, are classified as ASD. This attribute shows values in "yes" or "no" form. This data appears normal with no apparent sparse class.

Data Pre-Processing

Missing Values

Two variables have missing values, namely ethnicity and relation. Firstly for ethnicity, as ethnicity has a high dependence on the country a person is from; thus, the missing values of ethnicity will be amputated with the mode of ethnicity for that particular country. The first missing value for ethnicity comes from a person who resides in American Samoa. The mode of ethnicity for American Samoa is first searched and then it is used to replace the missing value. The next missing value for ethnicity comes from a person who resides in Albania. The ethnicity of other respondents from Albania is first analysed to determine the mode. The mode data is used to replace the missing value. The missing value for ethnicity comes from a person who resides in Belgium. The ethnicity of other respondents from Belgium is first analysed to determine the mode. The mode data is used to replace the missing value.

There is two missing value for ethnicity that comes from people residing in Afghanistan. The ethnicity of other respondents from Afghanistan is first analysed to determine the mode. However, the analysis shows that there has not been any input for the ethnicity of those from Afghanistan. However, as Afghanistan is a middle eastern country, there are high possibilities for the respondent to be a middle easterner. Thus, "middle eastern" is chosen as the replacement for the missing value in this case.

The last missing value for ethnicity comes from a person who resides in Argentina. The ethnicity of other respondents from Argentina is first analysed to determine the mode. The mode data is used to replace the missing value. Most people from Argentina are Hispanic. Thus, "Hispanic" is chosen to impute for the missing value in this case.

Missing Values: Relation

The next variable that has missing values is the 'Relation' variable. For relation, the mode for the variable is analysed. The analysis shows that the mode for the relation variable is 'self'. This indicates that most of the respondent has taken the ASD screening test for themselves. As this dataset shows the results of adolescents, this age group is likely capable of using the mobile app and performing the screening test for themselves. Thus, the unknown relation values would be replaced with 'self' value.

Inconsistent Data: Age_Desc

In the Age_Desc variable in the dataset, inconsistent data were identified. 7 respondents have indicated their age group as 12-15 years. As this was not consistent with the rest of the dataset, this value was changed to 12-16 years to standardize the dataset. This does not leave any impact on the data as 12 to 15 years is still within the 12 to 16 years range.

Rename Variables

It was noted in that some variables have spelling errors on their name. Although this error does not have any significant impact on the dataset, however, corrected names would give a better understanding of the data. Thus, attribute names jaundice, Austin and contry_of_res was replaced.

Suspicious Noisy Data

Out of the 104 observations, 5 respondents have answered "1" to all ten questions from A1 to A10. This could be a suspicious noisy data, as sometimes human errors or random answers could lead to such a situation. Thus, further analysis was performed on these 5 respondents' data. It was noted that the respondents have answered all the other questions accordingly. Thus, it can be considered that the answer "1" to all ten questions from A1 to A10 is intentional and does not impact the overall dataset.

Data Transformation

In the data transformation step, all the continuous data will be transformed to categorical data to standardize the dataset as data mining algorithms work better in one qualitative or quantitative form rather than having mixed data. Thus, in this analysis, all the data will be transformed into a qualitative form. The continuous variables that are transformed in this dataset are A1 to A10 Score and age variables. The result variable is not transformed as it already has a categorical representation as Class/ASD variable. For A1 to A10 score, the current binary values 1 and 0 are transformed to represent 'Yes' and 'No' and are given in new A1 to A10 columns. For the age variable, two categories are created known as age group and represented with values '12-14' or '15-16'.

RESULTS & DISCUSSION

The cleaned and transformed data was then uploaded to HDFS in Apache Hadoop. The dataset was then converted to table form and used for analysis using Hive queries. A connection was also established from the Hortonworks Platform (HDP) to Tableau. The dataset was retrieved by Tableau

from HDP and was used for the data analytics step. Five hypotheses were formulated from the analysis using the visualization tool. All the hypothesis are then interpreted and are critically discussed.

Hypothesis 1: "*Only 4 out of 10 questions (A7, A8, A9 and A10) need to be answered yes, to be classified as ASD"*.

This hypothesis 1 was tested using HiveQL queries. The result showed that, when answers to question A7, A8, A9 and A10 is yes, the result is 100% positive for ASD Screening. 28 respondents have answered yes to these 4 questions, and the results showed all 28 of them to be classified as ASD. This could be due to these 4 questions are targeted on the strongest symptoms to identify ASD. This is a significant finding that can greatly help to further reduce the time taken to screen for ASD and make the process more efficient. This hypothesis can be further tested with different age groups or bigger sample groups and effectively implemented if high accuracy achieved.

Hypothesis 2: "*Jaundice is not associated with having Autism Spectrum Disorder"*.

Some researchers believe that if an individual had jaundice (most often) in an infant stage, they are more likely to develop ASD. Many research works have focused on this topic. A particular study by [15], who conducted a systematic review of relevant works to assess the relationship between Jaundice and ASD, concluded in their literature that a causal relationship between jaundice and ASD cannot be established.

Thus, this theory was tested in this dataset for adolescent and no relationship was established between jaundice and autism. Out of 63 respondents who were classified as ASD, 54 of them recorded of not being jaundice before. Only 14 % of the respondents who had jaundice before were classified as ASD. Thus, there is not enough evidence to suggest any association of jaundice with ASD. Figure 1 illustrates the Jaundice & ASD Relationship.

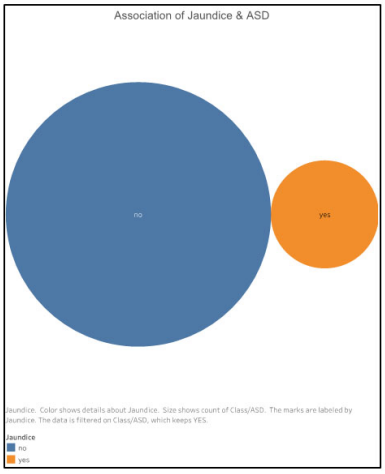

Figure 1: Association of Jaundice & ASD.

Hypothesis 3: *"The ethnicity of an individual is not associated with Autism Spectrum Disorder"*.

A growing number of studies from Europe suggest an increased frequency of autism in children of immigrant parents [16]. In contrast, other studies from North America tend to conclude that neither maternal ethnicity nor immigrant status is related to the rate of autism spectrum disorders. Finding from this dataset suggest that there is no association between the ethnicity of an individual to ASD. This dataset contained data of people of more than seven ethnicities, and none of them had a significant reaction as being more or less prone to be tested positive for ASD. It can be noted from Figure 2 that, almost all the ethnicities have an equal share of yes and no to be classified as ASD. Only the data for White-European may stand out compared to the rest. However, as White-Europeans were the highest number of participants in this questionnaire, it is likely to appear so. In percentage, only 74% % of the White Europeans tested positive as ASD, thus it cannot be considered as a significant value. Thus, it can be concluded that the ethnicity of an individual is not associated with ASD.

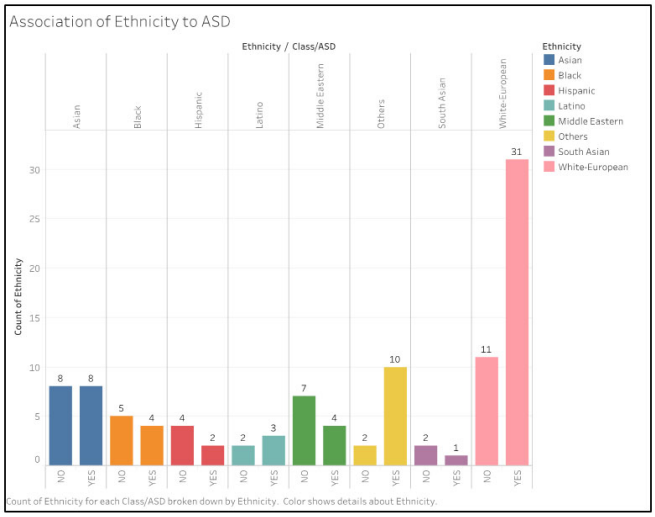

Figure 2: Association of Ethnicity to ASD.

Hypothesis 4: *"Self-awareness to screen for ASD is high among adolescents"*.

Findings from data show that self-awareness to screen for ASD is high among adolescents. About 51 respondents took the test for themselves (Figure 3). This could be due to their age at where they are more aware of behavioural differences compared to others and are curious to confirm if they have autism. Many research works are being carried out to study on teenagers dealing with Autism. In a study by The Centre on Secondary Education for Students with Autism Spectrum Disorder (CSESA), it was found that depression is more common among teenagers with ASD than teenagers without ASD. Rates of the major depressive disorder

have been reported as high as 37% in adolescents with ASD compared to about 5% of adolescents in the general population (CSESA, 2014).

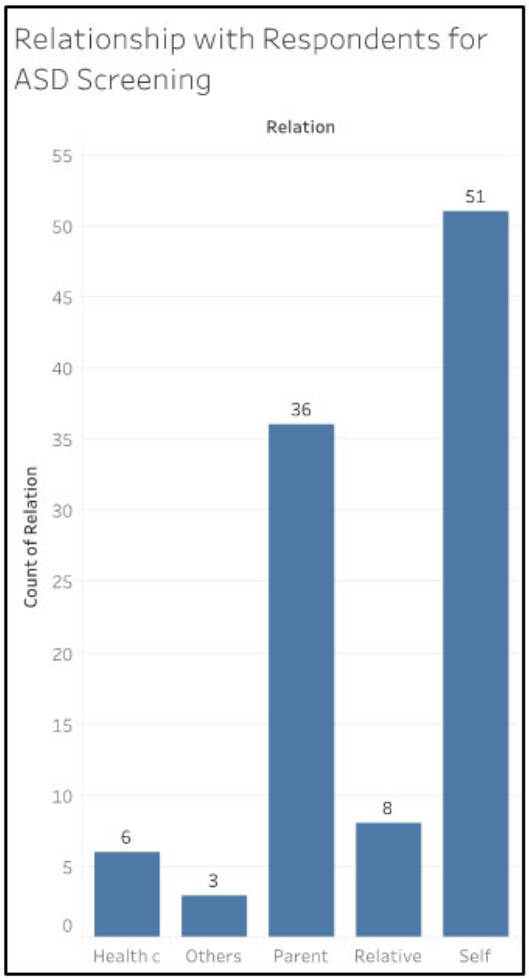

Figure 3: Relationship with Respondents for ASD Screening.

Hypothesis 5: *"Language barrier hinders the reach of the ASD Screening to certain parts of the world"*.

This dataset consists of respondents from 33 different countries. However, a closer look at the data revealed that the ASD screening test did not reach the audience in certain parts of the world. From Figure 4, it can be noted that towards the northern part of Asia and Russia, there has not been any respondent who has taken the ASD screening test completely. The countries in this part of the world are countries such as China, Japan, Mongolia, Kazakhstan, Uzbekistan, Ukraine, Russia, etc who predominantly are not well versed in English. Although the questionnaire is made available on a mobile app that is accessible by everyone, however, the language barrier could be the reason that hinders the reach of the ASD Screening tool to this part of the world. For a wider reach, the questionnaire could be translated into other languages in the future.

Figure 4: Country of Residents of Respondents.

DISCUSSION

Many interesting findings were discovered in this study. A significant finding from this study highlights that there is a high potential for ASD screening to be made in an even more time-efficient manner by reducing the number of questions asked. Four main questions have been identified from this study (i.e. A7, A8, A9 and A10), whereby if answered yes to all four of these questions, there is a high possibility of being classified as ASD. With effective implementation of this finding, it can greatly help to achieve the main aim of this research, to aid the development of easily implemented and effective screening methods. Hence, this hypothesis should be further tested in the future with different age groups or bigger sample groups to strengthen the evidence.

Besides, some common concepts and theories being researched by subject matter experts in this field were also tested using this dataset. Some researchers believe that if an individual had jaundice (most often) in an infant stage, they are more likely to develop ASD. However, the findings of this study confirmed that there is no association of jaundice with ASD. Out of the 63 respondents who were classified as ASD in this dataset, 54 of them recorded of not being jaundice before. Thus, the high percentage indicate that there is not enough evidence to suggest any association of jaundice with ASD.

Secondly, there has been a growing number of studies that suggest an increased frequency of autism in children of immigrant parents [16]. In contrast, some other studies tend to conclude that neither maternal ethnicity nor immigrant status is related to the rate of autism spectrum disorders. Finding from this dataset suggest that there is no association between the ethnicity of an individual with ASD. This dataset contained data of people of more than seven ethnicities, and none of them had a significant reaction as being more or less prone to be tested positive for ASD.

Besides that, findings from this study also displayed that self-awareness to screen for ASD is high among adolescents. This could be due to their age at where they are more aware

of behavioural differences compared to others and are curious to confirm if they have autism. Many research works are being carried out to study on teenagers dealing with Autism, and this study could be one of the stepping stone to confirm the related theories.

Final finding from this study revealed that the ASD screening test has not been effective in reaching the global audience. It could be higher due to language barriers that no responses have been received from certain parts of the world, in particular, the northern part of Asia and Russia, where the English language is not commonly used. For a wider reach, the questionnaire could be translated into other languages in the future.

These findings are valuable to further improve the ASD screening to be more efficient. Some improvements that should be looked into are, firstly, to reduce the number of questions asked in the research. Further studies should be conducted to confirm if all the ten questions (A1 to A10) is really necessary and if it could be reduced to be more effective. Secondly, other personal particulars collected such as ethnicity, if the respondent had jaundice before, or if the respondent has autism could be eliminated from the screening process as it does not add much value for the screening of ASD. Besides revamping the question set, translation of the question to different languages would also serve to be more useful for a larger audience and accurate screening of ASD.

CONCLUSION

First and foremost, this study is very important as it gives an overall understanding of autism spectrum disorder and the causes and effects of it. Besides, the related work section looks deeper into studies that have been carried out in the past for screening of autism in infant, children and adults. Many interesting methods and tools were discussed in this study that can be easily adapted and improved for use by medical practitioners and individuals who are affected by this disorder. In the method section, detailed descriptions were given to enlighten the audience on the complete data mining process. Steps involved from the pre-processing of the data to analysing the data to deriving meaningful insights from the data were discussed in detail. The different tools and technologies applied in this study such as SAS Studio Program, HDFS HIVE and Tableau were also introduced and the relevant commands, source code and queries used were also shared. Findings of the data were then critically discussed and suggestion for improvement was provided. Thus, this study serves as a complete guide to effectively manage from raw data to derive meaningful insights, regarding ASD screening.

REFERENCES

1. Towle PO, Patrick PA. Autism Spectrum Disorder Screening Instruments for Very Young Children: A Systematic Review. Autism Res Treat. 2016;2016:1-29. DOI:10.1155/2016/4624829
2. Carpenter LA, Boan AD, Wahlquist AE, et al. Screening and direct assessment methodology to determine the prevalence of autism spectrum disorders. Ann Epidemiol. 2016;26(6):395-400. DOI:10.1016/j.annepidem.2016.04.003
3. 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(4):e100-8. DOI:10.1038/tp.2012.10
4. Allison C, Auyeung B, Baron-Cohen S. Toward brief "red flags" for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 cases and 3,000 controls. J Am Acad Child Adolesc Psychiatry. 2012;51(2):202-212. DOI:10.1016/j.jaac.2011.11.003
5. Limperopoulos C, Bassan H, Sullivan NR, et al. Positive Screening for Autism in Ex-preterm Infants: Prevalence and Risk Factors. Paediatrics. 2008;121(4):758-765. DOI:10.1542/peds.2007-2158
6. Olliac B, Crespin G, Laznik MC, et al. Infant and dyadic assessment in early community-based screening for autism spectrum disorder with the PREAUT grid. PLoS One. 2017;12(12):1-23. DOI:10.1371/journal.pone.0188831
7. Wall DP, Dally R, Luyster R, Jung JY, DeLuca TF. Use of artificial intelligence to shorten the behavioural diagnosis of autism. PLoS One. 2012;7(8). DOI:10.1371/journal.pone.0043855
8. Quach KT. Application of Artificial Neural Networks in Classification of Autism Diagnosis Based on Gene Expression Signatures. 2012. DOI:10.1364/AO.56.00B222
9. Liu W, Li M, Yi L. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. Autism Res. 2016;9(8):888-898. DOI:10.1002/aur.1615
10. Sadeghi M, Khosrowabadi R, Bakouie F, Mahdavi H, Eslahchi C, Pouretemad H. Screening of autism based on task-free fMRI using the graph-theoretical approach. Psychiatry Res - Neuroimaging. 2017;263(January 2016):48-56. DOI:10.1016/j.pscychresns.2017.02.004
11. Heijnen-Kohl SM., Kok R., Wilting RMH., Rossi G, van Alphen SP. Screening of Autism Spectrum Disorders in Geriatric Psychiatry. Mouvements. 2017;72(4):27. DOI:10.1007/s10803-017-3185-2
12. Hedley D, Brewer N, Nevill R, Uljarević M, Butter E, Mulick JA. The Relationship Between Clinicians' Confidence and Accuracy, and the Influence of Child Characteristics, in the Screening of Autism Spectrum Disorder. J Autism Dev Disord. 2016;46(7):2340-2348. DOI:10.1007/s10803-016-2766-9
13. Craig F, Fanizza I, Russo L, et al. Social communication in children with autism spectrum disorder (asd): Correlation between DSM-5 and autism classification system of functioning—social communication (ACSF: SC). Autism Res. 2017;10(7):1249-1258. DOI:10.1002/aur.1772
14. Nygren G, Sandberg E, Gillstedt F, Ekeroth G, Arvidsson T, Gillberg C. A new screening programme for autism in a general population of Swedish toddlers. Res Dev Disabil. 2012;33(4):1200-1210. doi:10.1016/j.ridd.2012.02.018
15. Amin SB, Smith T, Wang H. Is neonatal jaundice associated with autism spectrum disorders: A systematic review. J Autism Dev Disord. 2011;41(11):1455-1463. doi:10.1007/s10803-010-1169-6
16. Keen D V., Reid FD, Arnone D. Autism, ethnicity and maternal immigration. Br J Psychiatry. 2010;196(4):274-281. doi:10.1192/bjp.bp.109.065490
17. Depression in Adolescents With Diabetes.Pdf.; 2014.