METIS
CONSULTING

Therapy Outreach Analytics Report

Aisha Duze
Metis Consultancy Services
11/20/2023

EXECUTIVE SUMMARY:

Puresouls Learning Centre embarked on a series of therapy outreach programs in the period 2017-2022.This report delivers insights from the data gathered over the period. Here are highlights from the report:

1) 2018 recorded the highest attendance rates with 355 entries while 2019 had the lowest with only 76 entries. 2022 had the second highest attendance with 309 entries.
2) After cleaning and sorting the data, a total of 1,269 unique attendees were recorded for the period. 9 of these attendees returned at later dates, bringing the total number of data entries to 1,278.
3) A total of 774 or 60% of the entries are male while 503 or 39.4% were female. One attendee's gender was not stated and we were not able to determine this from the name alone. As such it's labelled 'Not stated'.
4) The age range in years is 0 to 61 years old (ages determined as at date of therapy outreach).
5) 59% of attendees are in school while almost 39% are not in school. 48.5% of the attendees in school attend mainstream schools.
6) Cerebral palsy is the most mentioned diagnosis for previous and present diagnoses. The second most mentioned present diagnosis is autism while for pervious diagnosis, it is Down's syndrome.
7) Speech and motor skills dominate the "Expectations" and "Child's needs" column responses.

1. INTRODUCTION:

The nature of the data available only allows for descriptive analyses and so this report provides the outcomes of descriptive analysis on the therapy outreach data.

The data was provided in excel format by the Centre and contained data collected in outreach attendees by month and year of attendance. Two outreach programs each held in 2017, 2018, and 2022, none held in 2020 while 2019 and 2021 had one outreach each. Here is the break-down:

- 2017 - April and November
- 2018 – March and August
- 2019 – March only
- 2021 – March only
- 2022 – February and July

The analysis helped reveal some information hidden the data as well as ways to improve data collection and management at the Centre. It also points the path to improved data quality as well as better-informed decision-making using data generated internally at the Centre. Last but not least, this report presents the Centre with an opportunity to contribute to the existing body of knowledge on special needs education, particularly in Africa.

2. OBJECTIVES OF DATA ANALYSIS:

This analysis was aimed at extracting beneficial information from the therapy outreach data sets. Other objectives of the exercise include:

- Presenting a clean, organized data set.
- Delivering two detailed reports, one for internal use while the other will be for inclusion in reports, presentations, grant requests etc.
- Identifying the strengths and weaknesses of the data collection and management processes at the Centre.
- Determining the best course of action to improve data collection and management at the Centre.

3. DATA SOURCE:

The data provided was collected at a series of therapy outreach programs. It was subsequently entered into excel spreadsheets and provided as a folder containing 5 files, one file for each year that the therapy outreaches held. The data sets were presented as excel files comprised of responses to questionnaires arranged in columns. The questionnaires had been self-administered. The data showed that there were at least 3 variations of the questionnaire, each varying slightly in the number and arrangement of the questions. The table in **Fig.1** below shows the questions from each questionnaire type:

s/n	Questionnaire Type 1	Questionnaire Type 2	Questionnaire Type 3
1	Outreach Date	Outreach Date	Outreach Date
2	Name of Child	Name of Child	Name of Child
3	Date of Birth	Date of Birth	Date of Birth
4	Gender	Gender	Gender
5	Position in family	Position in family	Position in family
6	Schooling	Schooling	Schooling
7	Type of School	Type of School	Type of School
8	Child's diagnosis previously known	Child's diagnosis previously known	Child's diagnosis previously known
9	Length of Child's pregnancy (in months)	Length of Child's pregnancy (in months)	Length of Child's pregnancy (in months)
10	Mode of delivery	Mode of delivery	Mode of delivery
11	Reason for CS	Reason for CS	Reason for CS
12	Did child cry immediately after birth	Did child cry immediately after birth	Did child cry immediately after birth
13	Did child have jaundice at birth	Did child have jaundice at birth	Did child have jaundice at birth
14	What form of resuscitation was given afterwards before he/she cried?	What form of resuscitation was given afterwards before he/she cried?	What form of resuscitation was given afterwards before he/she cried?
15	Any delay in sitting, walking, etc	Any delay in sitting, walking, etc	Any delay in sitting, walking, etc
16	If yes, state	If yes, state	If yes, state
17	Any seizure	Any seizure	Any seizure
18	Does he/she have any other medical problems?	Does he/she have any other medical problems?	Does he/she have any other medical problems?
19	What are your expectations for participating in this year's programme?	What are your expectations for participating in this year's programme?	What are your expectations for participating in this year's programme?
20	Is your child on any intervention programme?	Is your child on any intervention programme?	Is your child on any intervention programme?
21	What do you attribute your child's condition to?	What is the greatest need of your child at this moment?	What do you attribute your child's condition to?
22	What is the greatest need of your child at this moment?	What does your child eat?	What is the greatest need of your child at this moment?
23	What does your child eat?	Diagnosis	What does your child eat?
24	DIAGNOSIS		Diagnosis
25			LOCATION
26			ADDRESS

Fig. 1

Limitations on the data include spelling errors, missing entries, incomplete entries, duplicate registrations on the same date, missing dates and errors in date entries, and last but not least, non-empirical responses to questions. These limitations can be attributed to the quality of the questionnaire and questions asked, the data collection method i.e., self-administration, and the data entry process.

4. ANALYTICAL METHODS:

There were several iterated steps taken to clean and analyze the data. They are as follow:

1. Data cleaning and Preprocessing:
    a. Using Excel, the data sets were scrutinized to assess form and layout.
    b. The different years were merged into one spreadsheet and again scrutinized. It was observed that due to the difference in questionnaire types, the columns did not align. This was rectified and the data merged again.
    c. The next step was to shorten the column headings which were originally entered as the questions from the questionnaire. Changes seen in **Fig.2** below:

Old Column Headings	New Column Headings
Outreach Date	Date
Name of Child	Name
Date of Birth	Date of Birth
Gender	Gender
Position in family	Birth order
Schooling	In school
Type of School	Type of school
Child's diagnosis previously known	Previous Diagnosis
Length of Child's pregnancy (in months)	Pregnancy duration
Mode of delivery	Mode of delivery
Reason for CS	Reason for CS
Did child cry immediately after birth	Cried at birth
Did child have jaundice at birth	Jaundiced at birth
What form of resuscitation was given afterwards before he/she cried?	Form of resuscitation
Any delay in sitting, walking, etc	Developmental delay
If yes, state	Type of delay
Any seizure	Any seizure
Does he/she have any other medical problems?	Other medical issues
What are your expectations for participating in this year's programme?	Expectations
Is your child on any intervention programme?	Intervention programme
What do you attribute your child's condition to?	Condition attributed to
What is the greatest need of your child at this moment?	Child's need
What does your child eat?	Food preferences
Diagnosis	Diagnosis
LOCATION	Location
ADDRESS	Address

Fig.2

d. The next step was to identify the missing data and impute missing values where possible. It involved the following:
    i. The date columns contained a lot of missing values and some cells were filled using ‘ “ ‘ signs perhaps to indicate replicated or blank entries. These blank cells were imputed with the nearest plausible date entry.
    ii. Another common anomaly was the use of ‘_____’ for blank or missing data points. All such cells were either cleared and left empty or filled with ‘Unknown’ or ‘Not stated’.
    iii. Due to perhaps errors in data entry, some dates of birth postdated the outreach dates and had to be removed altogether before final analysis.
    iv. The file was exported into Python and date formats were changed using Pandas in Python.
    v. The ages of the outreach attendees were calculated and included in a final version of the data set.
e. Following up was data validation and integrity checks.
    i. For many of the columns, the responses were harmonized for easier analysis. For example: In a column, blank cells, ‘nil’, ‘none’, ‘unknown’ could all be changed to ‘Unknown’ or ‘Not stated’ (in the case where the column cross-validates another). There were instances where complex, wordy responses were given to Yes/No questions, and in some cases, unrelated, thus invalidating the response. As much as possible, these responses were converted to binary responses.
    ii. Some responses to questions following Yes/No responses did not validate the initial response. Where possible, the responses were edited while in most cases, they were ignored at the expense of the quality of the data contained in the columns. Examples are the ‘Form of resuscitation’, ‘Type of delay’, ‘Type of school’ columns, to list a few.
    iii. Inconsistency in data entry called for harmonization in entries, spellings, font size, font case, format etc.
    iv. Exploratory data analysis was done using Excel, Python and PowerBI. Visualizations were made suing Python and PowerBI.

Version control has been maintained. Copies of the data at each iteration are available.

5. KEY FINDINGS:

The analysis is purely descriptive due to the nature of the data available. The findings have been categorized into Biometric and Therapy information:

5.1 Biometrics:

a. **Date:** Naturally, the first year 2017 recorded the highest number of repeat outreach visitors. They are:
    i. Akanji Idowu- returned in 2021

ii. Aishat Lawal- returned in 2018
iii. Iyanuoluwa Akinrinade- returned in 2021
iv. Adebisi Gbolahan- returned in 2019
v. Samuel Adewole- returned in 2022.
Others are:
vi. Oyinkansola Adeleke- attended in 2018 and 2021
vii. Abdulsalam Mubarak- attended in 2021 and 2022

There are other attendees with same names but difference in their dates of birth or gender made it impossible to confirm whether or not they were the same person. They are:

viii. Ibrahim Lawal – Different gender and date of birth, attended 20th and 21st April 2017
ix. Daniel Idowu with different dates of birth, attended in February and July 2022.

Other insights from the **Date** column include:

x. 2018 had the highest number of outreach attendees at 355. Outreaches held in March and August of that year. 98 attended in March while 257 attended in August.
xi. 2019 had the lowest attendance rate with only 76 recorded. There was also just one outreach which held in March of that year.
xii. A study of duplicates revealed data collection/entry errors where some attendees were recorded twice on the same day. Such entries were deleted.

See **Fig.3** below for the actual count of attendance by year:

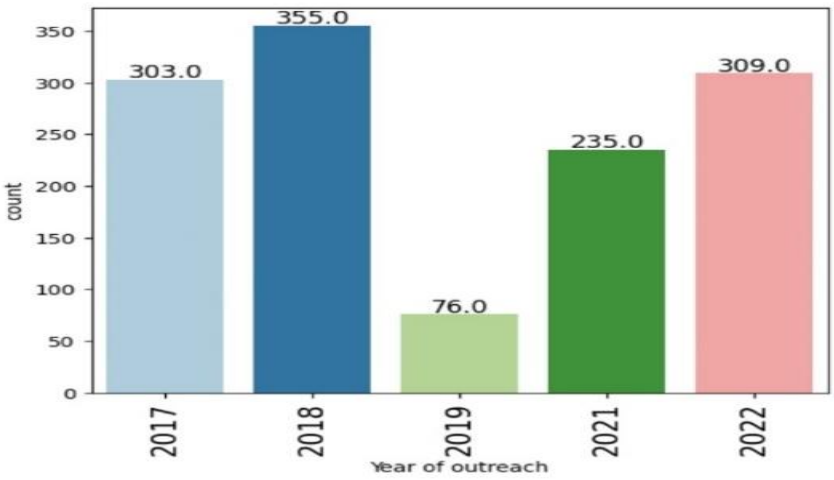
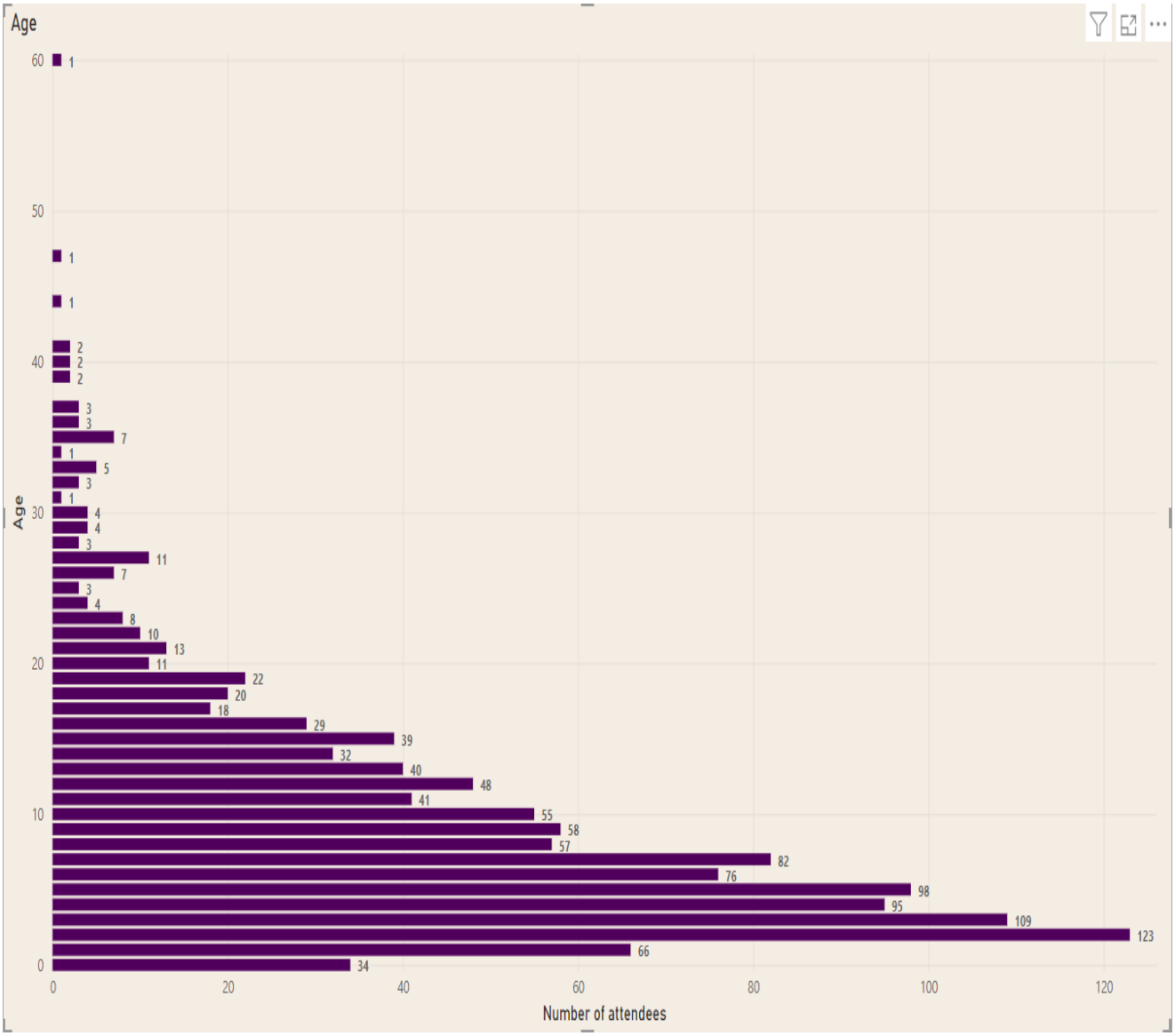


Fig.3

b. **Age:** The oldest attendee was 60 years old as at the date of outreach. That's the outlier in the data set. The age was calculated in years and so attendees less than a year old were captured as age 0. With the outlier removed, the average age of attendees is 9 years. The figure below shows the age



distribution:

Fig.4

Approximately 10% of the attendees are aged 2, 90% are aged 20 and below while a third or 67% are aged 10 and below.

c. **Position in family:** The data shows that approximately 38% of the respondents were first born children. 3 twin births were recorded, making 0.23% and one 14th position. The distribution is visualized in **Fig.5** below:

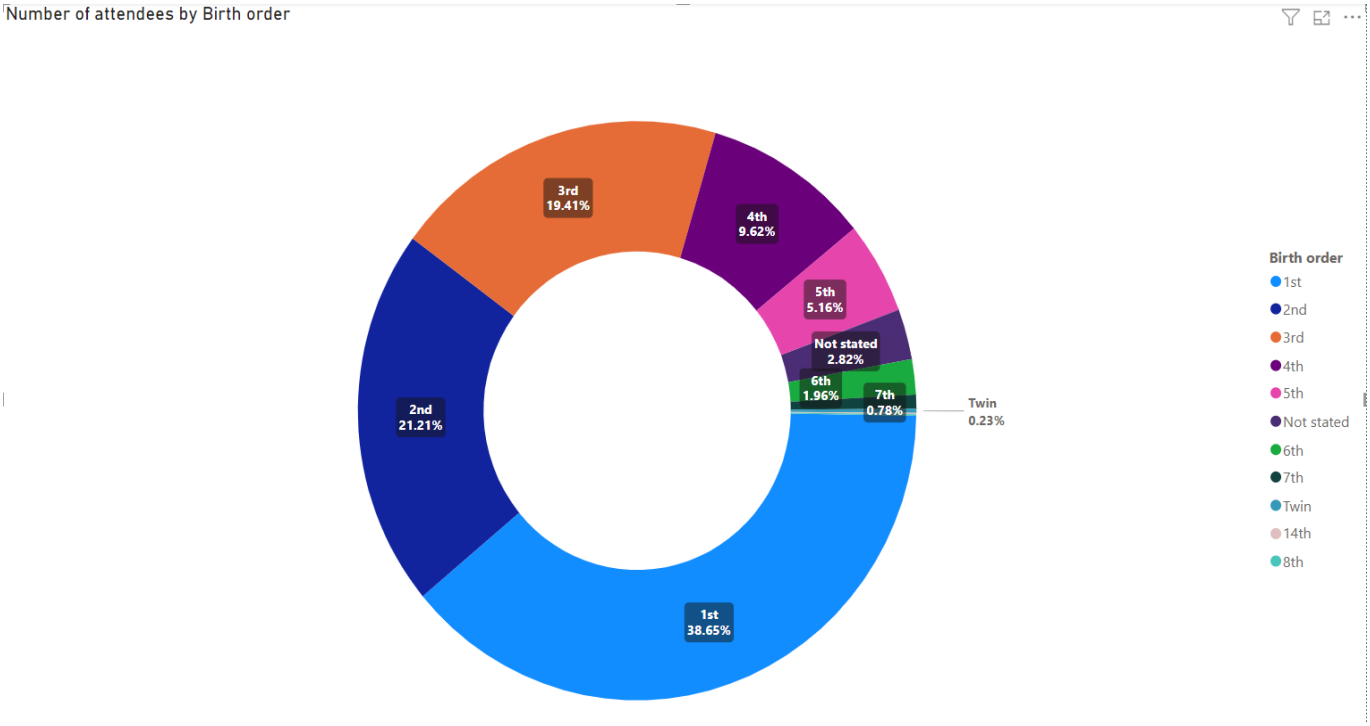


Fig.5

5.2 Therapy Data

a. **Schooling:** The data shows that 59% of the attendees were in school at the time of the outreaches. 38% were not schooling. Out of those schooling, 48% are in mainstream schools while the rest are spread across other types of schools. Unfortunately, the column had a lot of varied response that were outside the scope of the question itself. Also, the 'Type of School' column failed to validate the "In school" column. In some instances, the respondent will say the attendee is not in school and go ahead to specify a type of school or answer yes and go ahead to say the attendee is no longer in school. **Fig.6** and **Fig. 7** show the distributions of these columns.

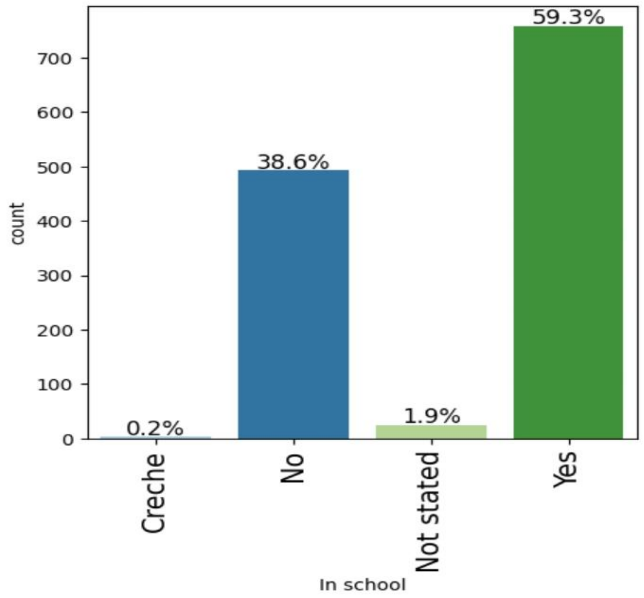

59.3%
38.6%
1.9%
0.2%
count
700
600
500
400
300
200
100
0
Creche
No
Not stated
Yes
In school

Fig.6

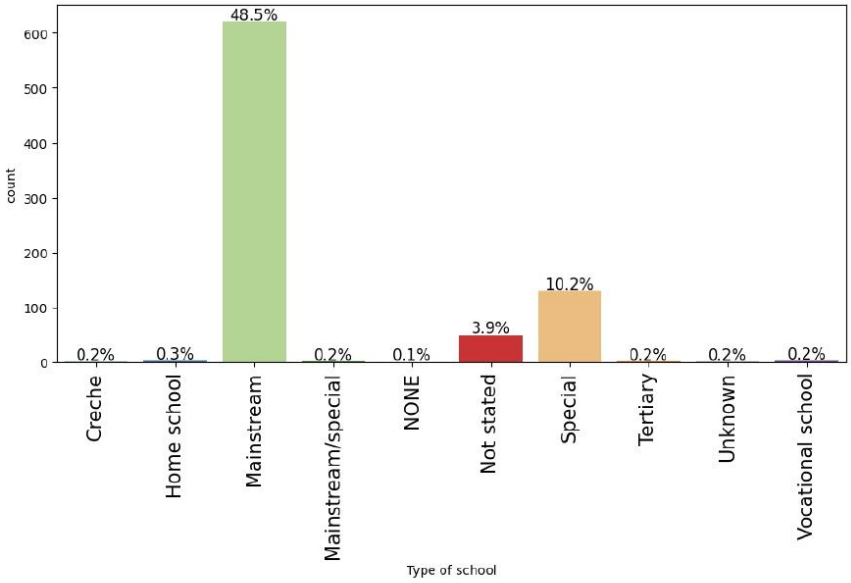

48.5%
10.2%
3.9%
0.2%
0.3%
0.2%
0.1%
0.2%
0.2%
0.2%
count
600
500
400
300
200
100
0
Creche
Home school
Mainstream
Mainstream/special
NONE
Not stated
Special
Tertiary
Unknown
Vocational school
Type of school

Fig.7

b. **Past and present diagnoses:** Here we find a lot of variation in responses, not all of which are empirical in nature. It affects the quality of the data available. Using Natural Language Processing, we were able to extract the most common diagnoses as visualized below:

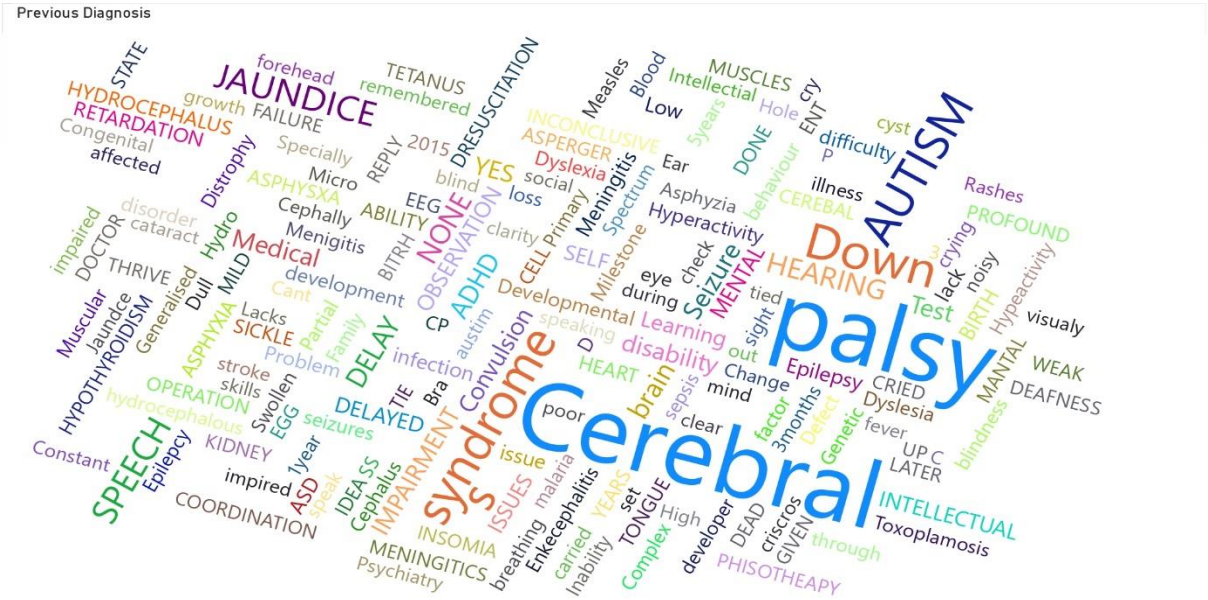


Fig.8

Fig.9

Figures 8 & 9 show cerebral palsy, autism and Downs syndrome to be the top 3 diagnoses.

c. **Pregnancy Stats:** The data on pregnancies is as follows:

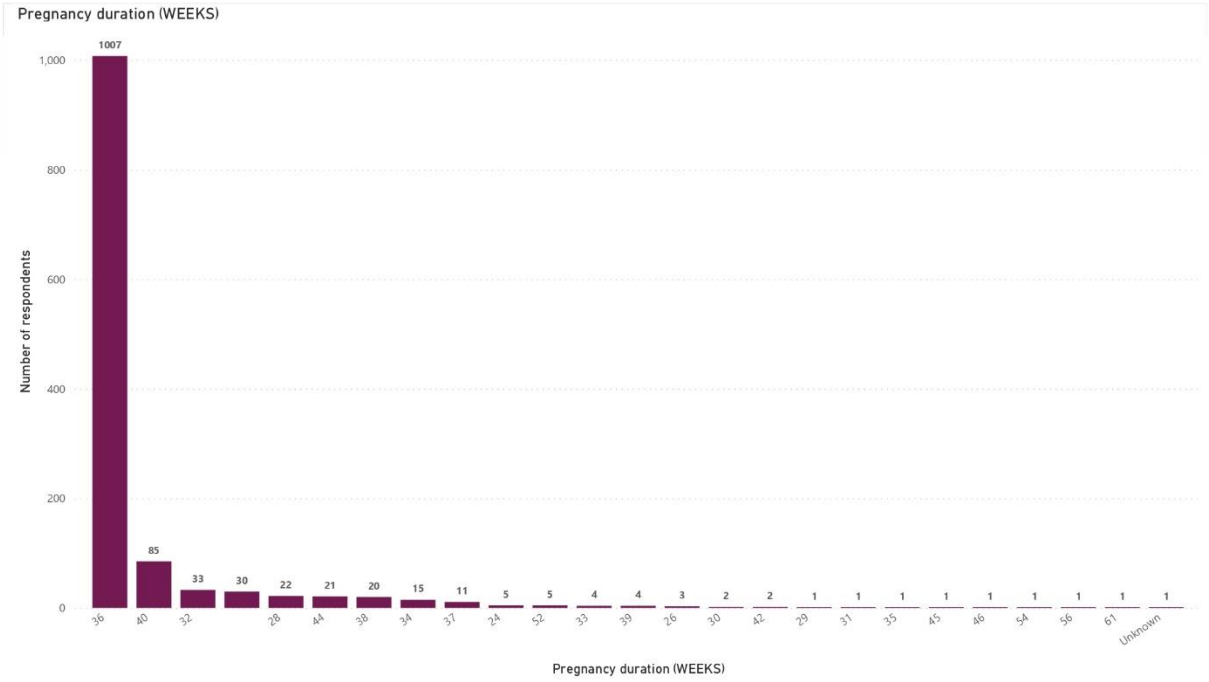


Fig.10

Approximately 80% of the respondents carried their pregnancies for at least 36 weeks. The longest pregnancy duration recorded was 61 weeks while the lowest was 24 weeks. 60% of the births at 24 weeks were Cesarean sections (CS).

d. **Birth stats:**

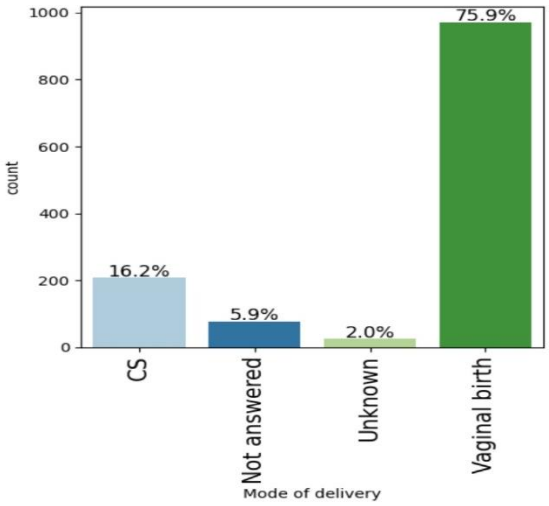


Fig.11


Reason for CS
Unknown
LABOUR
DELAYED
PELVIC
Complications
ADVICE
Recommended
PLACENTA
Delay
DOCTOR
Baby
SMALL
WATER
CERVIX
Blockage
Pregnancy
Bleeding
Prolonged
UNABLE
RECOMMENDATON
DOCTORS
BLOCKED
AHEAD
FIBROD
blood
COME
vaginal
breach
bring
VERY
TIME
withstand
tight
Due
push
induced
line
request
hours
SIZE
expected
High
Low
Long
DOWN
Ailment
head
Lost
legs
age
retardation
observed
processing
well
Premature
Leaking
birth
Milemethomy
complication
wasn't
24
TWINS
BURSTING
PELVIS
mother
woke
womb
properly
passed
Prevoius
Bloody
fell
first
FAINTED
showed
uterine
EARLY
delivery
body
medical
malposition
tired
SURVIVAL
strenght
Bridge
up
tiles
give
pelving
Dr
plancenta
INJURY
wrongly
consecutive
Excess
herself
deliver
Emergency
safety
Went
PELRIC
1st
fibriod
bleading
yes
SPINAL
positioned
previal
Cant
through
twin
LUPUS
cervise
RISK
Inabilty
child
position
forth
days
Bridged
mom
pain
week
comes
OUT
CS
flow
breathing
Swollen
convulting
issues
pressure
issue
mum
Inability
PROLONG
Fibroid
EDD
intra
big
previous
SHIFTING
Weakness
none
before
open
prevent
broke
Breech
FIBROMYALGIA
developing
dilating
preclamsia
progressing
Laeling
coma
growth
scan

Fig.12

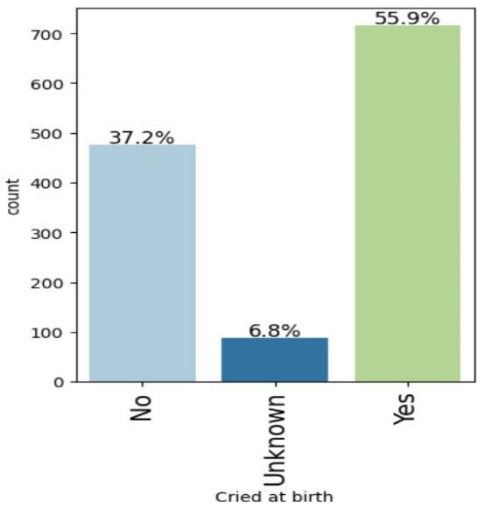

55.9%
37.2%
6.8%
count
700
600
500
400
300
200
100
0
No
Unknown
Yes
Cried at birth

Fig.13

Fig.14

Fig.15

- 75% of the total births were vaginal births and only 16% were CS. The reasons for the CS were mostly 'Unknown' but the next most common response was delayed labour.

- 56% of the respondents cried at birth and of those that required resuscitation, most were either tapped or given oxygen as a means of resuscitation. Only 21% of the respondents say the child was jaundiced at birth.

e. **Developmental delays:**

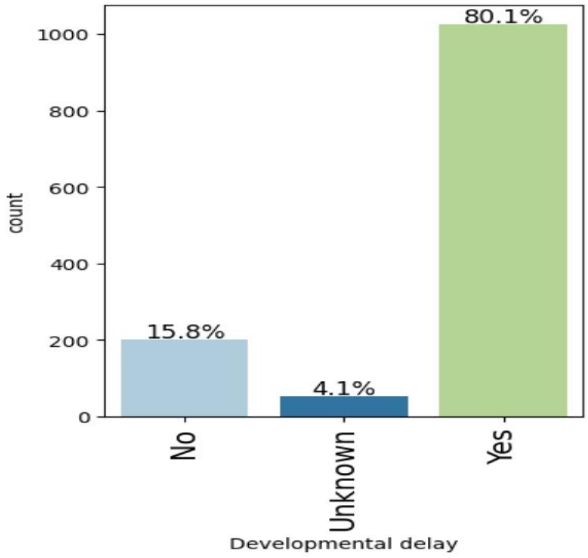


Fig.16

Fig.17

80% of the respondents had developmental delays. Speech and motor skill ranked top in the types of delays mentioned. This of course is cross-validated in the 'Child's needs' column as seen in the next visualization.

f. **Child's needs:** Speech and motor skills dominate this column's responses,

Fig.18

g. **Seizures:** A little over a third of the respondents have seizures. 12% did not respond to the question.

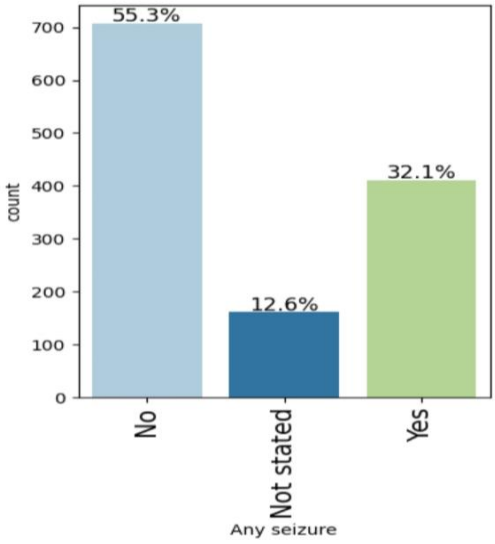


Fig.19

6. INSIGHTS FOR STAKEHOLDERS:

The following insights have been gleaned from the data:

- The wide age range of outreach attendees calls for a more focused approach to the exercise as the needs will vary across the ages. It would be more effective to split them into age groups.
- With 90% of the attendees under age 20 and 67% under age 10, it would be beneficial to focus therapy outreaches to children and young adults between the ages 2 to 20. It will help gather more useful information around the correlations and causalities around the learning and intellectual challenges.
- Most of the attendees are in mainstream schools and for many of them, the individuals with learning needs did not have major food preferences. It would be beneficial to further study these variables to find out the extent to which they are indeed due to availability or affordability. It can help further define the types of intervention required to mitigate and better manage these conditions.
- A more targeted approach makes it easier to define goals, track achievements and deliver results. It is recommended that the Centre identify it's target demographic, therapy strengths and goals and allow these factors drive subsequent outreach programs.

7. DATA QUALITY AND RELIABILITY:

The following observations were made about data quality & reliability:

1. Data Quality Issues: The data exhibited inconsistencies, inaccuracies, and missing values, impairing its reliability and usability for comprehensive analysis.
2. Questionnaire Design Challenges: The questionnaire used in the data collection process revealed shortcomings in clarity and specificity, contributing to ambiguous responses and difficulties in interpretation. There was a lot of variation in the responses resulting in a noisy data set. Variations in responses in the following columns rendered them unreliable:
    i. Other medical problems
    ii. Intervention program
    iii. Condition attributed to
    iv. Food preferences

8. RECOMMENDATIONS:

The following questions have been raised and to improve outcomes, should be answered prior the subsequent outreach sessions:

1. What is the reason for the outreach?
2. What's the purpose behind data collection?

3. What determines the period chosen for outreach?
4. What trackable data is collected for first time attendees?
5. What data will be measured for return attendees?
6. What is the difference between the question on previous diagnosis and the one on what the child's condition is attributed to?
7. How do you differentiate food preferences from availability due to financial capacity, location, parent/ guardian's choices, etc?

Recommendations for future therapy outreaches:

1. Group the attendees by age and attend to them separately.
2. Create separate forms for return attendees and first-time attendees. Have a different set of questions for return attendees. The new questionnaire should be designed to track progress, not answer the same questions as before.
3. To minimize data collection errors, there's need to train staff to be able administer the questionnaires rather than using self-administration. Also begin to consider entering the data digitally at the point of collection. This will minimize errors in data collection and entry.
4. Establish rigorous data cleaning protocols to address inconsistencies, outliers, and missing values systematically before analysis.
5. Analytical tools tend to be case sensitive. The huge variations in column entries makes it difficult to properly identify patterns. NLP can help but it's better to avoid wide variations especially for questions where multiple choice will suffice. There needs to be a separate training on Survey/Questionnaire design. in the interim, the following changes are recommended:
    - Use multiple choice questions for Gender, Schooling (and all other Yes/no questions), Type of school, Mode of delivery (make an exhaustive list), Diagnoses (also an exhaustive list, it shouldn't be open to guesses). 'Expectations' should collect responses that can be monitored and measured or evaluated and documented over time. Ideally, close-ended questions with restricted responses.
    - Restrict responses to Position in family. Include 'number of births', "number of children", columns to get more detailed information and cross validate the columns.
    - The column containing answers on form of resuscitation is polluted with wrong responses. Split the question into two. First, ask if resuscitation was given and then ask what kind.
    - Specify medical diagnosis for each question on diagnosis. If possible, collect the medical report to validate the response, especially for previous diagnosis.

- 'Mode of delivery' contains a lot of irrelevant responses. This can be mitigated by changing the question to multiple choice, restricting the response options and using the interview method of data collection.
- Split "Intervention" question into 2, a. Yes/No, b. Type of intervention

6. Ask questions to generate responses that will bear empirical inquiry.

10. CONCLUSION:

In the culmination of our data analysis, it becomes evident that the data at hand posed substantial challenges, hindering the extraction of meaningful insights. The inherent messiness and lack of coherence within the dataset limited our ability to draw robust conclusions regarding the outreach program. Despite these obstacles, the experience has provided valuable lessons and opportunities for improvement in future data collection endeavors.

11. ACKNOWLEDGMENTS:

I wish to acknowledge Mrs. Dotun Akande for her tireless support and continued cooperation as well as Mr. Gbenga for always attending to our endless enquiries.