**ACE/APP Data Audit: 03-31-2020**

**Breaking down missing data into categories**

|  |  |
| --- | --- |
| Category | Description |
| Single-entered data | Data that are not double entered (removed before going to data portal) |
| Duplicate data | Data pairs that have the same id+visit with different age or id+age with different visit numbers (removed before going to data portal) |
| Orphaned data | Data from which the participant id’s are not from tblSubject (master list of participants) |
| Non-primary data | Data that are from participants who are not marked as “primary” (i.e siblings, excluded) |
| Visit incomplete | Data from the participants which haven’t completed the visit within the time pont (scheduled to be collected/entered) |
| Uncertain | Data that are tracked by coordinators as non-missing but not available in ACCESS (need to look for physical copies) |
| Verified missing data | Data that are tracked by coordinators as missing |

*Table 1. Categories of missing data and definitions*

**Single-entered Data**

* Still waiting for double entry for medical data
* Brianna and Alexa interviewed/hired volunteers, but the process would be on hold due to the pandemic situation.
* Currently 120 participants are involved across medical data (GI History, Autoimmune, Physical Exam, and Tanner Staging)

**Orphaned Data**

* Orphaned data is defined as all data which id is not in the list of tblSubject (master list of primary participants).
* Brianna checked the list and figured out most of them where they came from.
* Further on, Soo will weekly put an updated list into a folder where Brianna can check and audit.

**Duplicate Data**

* Duplicate data is defined as either: 1) data pairs with **the same id and visit** (with different ages) OR 2) data pairs with **the same id and age** (with different visits)
* Same as orphaned data list, Soo will weekly put an updated list into a folder where Brianna can check and audit; the list has been small, and Brianna fixes the data quickly.
* Currently 4 participants (0.67%) are involved across ADOS, CDI-WG, and DAS.

**Suspected Missing Data (“Uncertain” category)**

* A major ongoing project is to account for suspected missing data, which is defined as following: The status of the data is **not** tracked as “Missing Data”, but the data is missing in ACCESS database, when the participant completed the visit where the data belongs to.
* Missing data with miscellaneous reasons
* Sameera, a Senior volunteer, took over this piece of data auditing. It is currently working in progress. More detailed update can come from Brianna.

**User Logs**

* 3 bugs report in March: Longitudinal IQ, ADIS, adding data to Einat’s query
* Average time of using the data portal is 9.2 mins.
* The results by months are shown below:

*Table 2. Number of Sessions in March, 2020 by Username for dataportal.explore*

|  |  |
| --- | --- |
| User Name | # of Sessions |
| park | 12 |
| dwyer | 9 |
| waizbard | 7 |
| heath | 6 |
| wunordahl | 6 |
| lee | 5 |
| young | 1 |

*Table 3. How Many Time Tables Queried in March, 2020?*

|  |  |
| --- | --- |
| Table Name | Freq |
| tbl\_longitudinal\_iq | 14 |
| tbl\_ados | 11 |
| tbl\_msel | 11 |
| tbl\_scan\_details | 10 |
| tbl\_dm\_status | 7 |
| tbl\_cbcl | 6 |
| tbl\_mori\_regions | 4 |
| tbl\_srs | 4 |
| tbl\_adi | 3 |
| tbl\_mori\_total | 3 |
| tbl\_ssp | 3 |
| tbl\_tcv | 3 |
| tbl\_adis | 2 |
| tbl\_demographics | 2 |
| tbl\_growth\_measurement | 2 |
| tbl\_vineland | 2 |
| tbl\_cdi\_wg | 1 |
| tbl\_cdi\_ws | 1 |
| tbl\_cshq | 1 |
| tbl\_eowpvt | 1 |
| tbl\_gort | 1 |
| tbl\_ppvt | 1 |
| tbl\_rbs | 1 |
| tbl\_ssp2 | 1 |
| tbl\_sti | 1 |

**Significant Age Outliers**

*Table 4. Sample age outliers table*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| id | visit | avg\_age | measure\_list | visit\_mean\_age | visit\_median\_age | age\_gap |
| 100604-200 | 1 | 74.09 | scq | 43.39 | 43.72 | 30.37 |
| 101370-100 | 3 | 113.93 | adi; cdi\_ws; eowpvt; ppvt; rbs; scq; srs; sti | 112.33 | 112.39 | 1.54 |
| 102268-100 | 4 | 160.95 | gort | 154.66 | 153.81 | 7.14 |

* This audit list is to double check if the visit number or date is correctly entered for the flagged participants.
* The full list is in S:/MIND/RESEARCH/APP/APP Database/7 – Data Auditing/age\_outliers\_list.csv
* Variable descriptions:
  + **Measure\_list** = list of measures within the corresponding id and visit pair is flagged as age outliers (i.e the measure\_outlier\_list variable contains age variables for the corresponding id and visit pairs.)
  + **Avg\_age** = average value of the ages in the “measure\_list” variable
  + **Visit\_mean\_age** and **visit\_median\_age** = mean and median age of the overall visit for the corresponding id and visit pair
  + **Age\_gap** = absolute value of the difference between avg\_age and visit\_median\_age

**Eligibility Assessment Missing**

* Eligibility stated by Brianna:
  + For APP: any ASD missing ADOS, Mullen, or ADI & any TD missing SCQ
  + For GAIN/NAPP/BRAIN: any participant missing ADOS, Mullen, SCQ & any ASD missing ADI.
* 25 SCQ’s, 1 ADI, and 1 MSEL total
* Full list of id’s and visits is available S:/MIND/RESEARCH/APP/APP Database/7 – Data Auditing/eligibility\_measures\_missing.csv

**Other Audit Lists generated (for Brianna)**

* List of unscorable (both whole and partial) data
* Lists to modify tracking tables