Breaking down missing data into categories

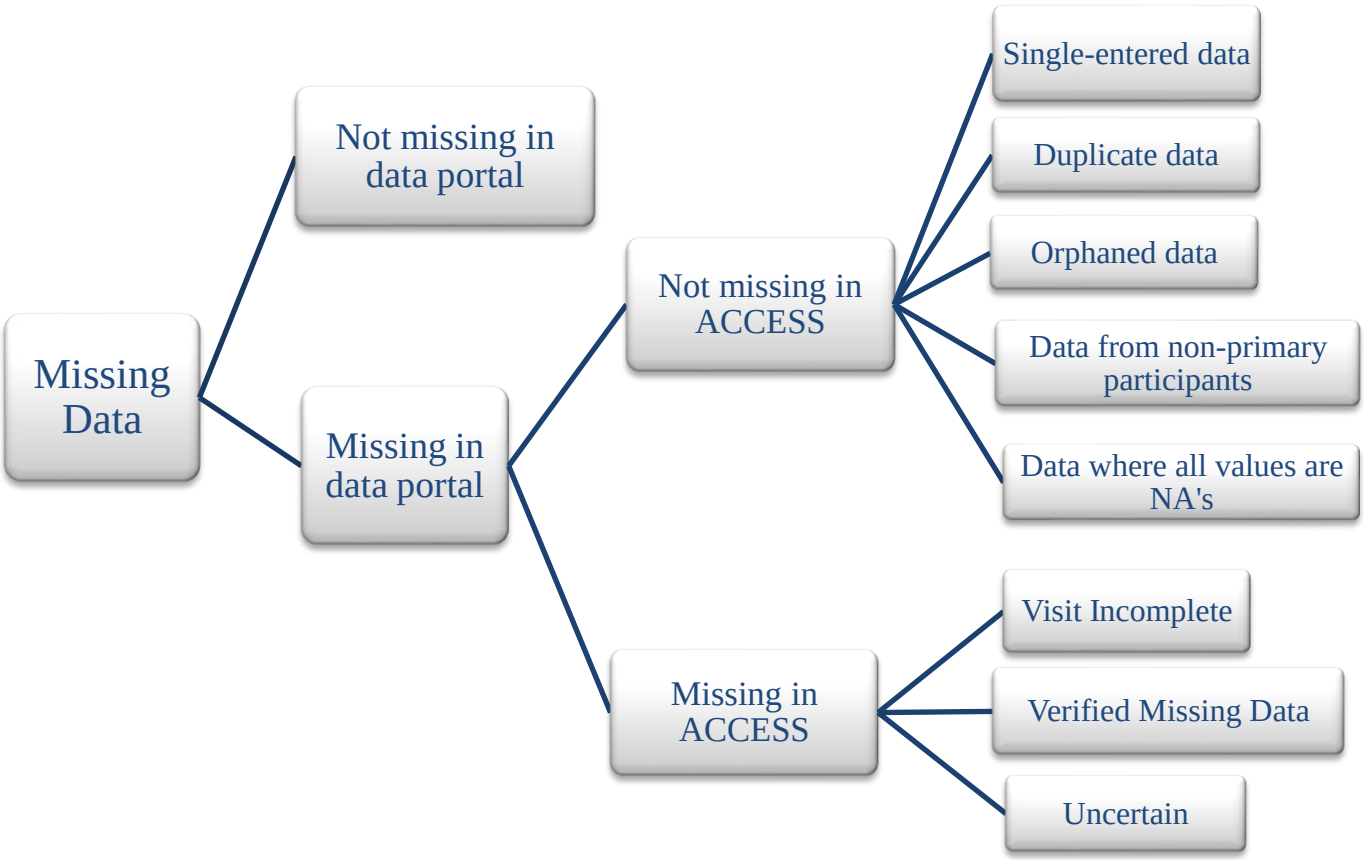


Table 1. Categories of missing data and definitions

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

Single-entered Data

Table 2. Number of single-entered data by measure

Measure	Frequency
Autoimmune (Child)	53
GI History	54
Growth Measurement	16
Physical Exam	54
Tanner Staging (Puberty)	7

- Table 2 above shows how many single-entered data in ACCESS database that are older than 3 months.
    - We discussed and decided in the last data meeting in August to include all the single-entered data (older than 3 months) in behavior measures.
    - The table below shows approximation of the behavior data that is now included in the dataportal despite a designation of only single entry.

Measure	Frequency
ADI	6
ADIS	3
ADOS	11
CBCL_6-18	2
CBQ	2
CCC	5
CDI-WG	2
CDI-WS	3
CELF	10
CSHQ	5
DAS-EY	4
DAS-SA	6
Demograph	5
DSM	4
DXCF	20
EOWPVT3	3
EOWPVT4	10
GORT	7
Harter	2
HPT	2
MASC-Adult	3

MASC-Child	3
MSEL	1
PAL	11
PPVT3	4
PPVT4	6
RBS	4
SCARED-Child	3
SCARED-Parent	3
SCQ	2
SRS-Child	1
SSP	2
SSP2	4
STI	28
TMCQ	4
VABS	16

Data Rows with NA's

Table 3. Number of data rows with all NA's by measure

Measure	Frequency	# "Validated"
CBCL	6	5
CCC	2	1
Harter	1	1
MSEL	32	26
RBS	7	7
SSP	23	7

- Table 3 above shows the number of rows in various data tables which contained empty or null values in every field within the row (i.e., an entirely blank row) in ACCESS database. Curiously, as the "# validated" column shows, the majority of these empty rows were designated as "validated" within the subject tracking database.
    - We plan to check the physical files of this data to determine whether the data exists or not.
    - If the data is missing, then we will delete the row and change the tracking status to "missing"; if actual data exists, we will re-enter the data to replace NA's.
    - Current status (out of 71): 47 validated, 17 missing, 6 ineligible, 1 invalid

Missing Key Assessment at Time 1

Table 4. Number of missing data for key assessments

Measure	Frequency	# who continued T2
ADI	2	0
ADOS	17	1 (GAIN T2)
MSEL	25	1 (Whole Visits Done)

- Table 5 above shows the number of obligatory assessments missing in **ACCESS db** where participants **1)** completed visit 1 according to **tblSubject** AND 2) at least **one month** has passed from **dates of assessment**.
    - ADI: 1 dropped, 1 invalid
    - ADOS: 7 ineligibles, 4 dropped, 6 given but not validated
    - MSEL: 13 ineligibles, 5 dropped, 4 given but not validated, 1 from CHARGE, 1 invalid, 1 validated but missing (107320-100)

Orphaned Data

Table 5. Number of orphaned data by measures

Measure	Frequency
ADOS	2
Autoimmune	4
CBCL_1-5	6
CBQ	1
CSHQ	2
Demograph	12
DxCf	32
EDQ	2
EOWPVT	1
FPS	16
Growth Measurement	8
HPT	2
MSEL	2
Physical Exam	1
PPVT	1
RBS	8
Scan Details	2
SCARED – Parent	3
SCQ	1
SRS – Child	2
SRS – Dad	2
SRS – Mom	81
SSP	3
SSP2	3
STI	1
TCV	3
TMCQ	2

- Orphaned data is defined as all data which id is not in the list of tblSubject (master list of primary participants)
- From last month report, number orphaned data slightly increased due to 112811-100 no longer a primary participant.

Duplicate Data

Table 6. Number of duplicate data pairs by measure

Measure	Frequency
EOWPVT	3
PPVT	1
Scan Details	1
Service & Treatment Inventory	1

- 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)
- Soo was able to figure out and clean up the majority of previous duplicate cases. However, a few remaining cases still need to be addressed:
    - EOWPVT: available both in EOWPVT3 & EOWPVT4 tables; which ones should we use? (Age equivalents & SS differ a bit between versions)
    - PPVT: same exact information (including dates) was entered in two different visits (1 vs 3); I believe visit 1 information is wrong.
    - STI: Similar to PPVT case
    - Scan Details: Need to delete one of the rows (entry is a bit different from each other)

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
    - Current main reason: Measures that were not collected but also were not tracked properly, data that are available in the physical copies that need to be re-entered
- Sameera, a Senior volunteer, took over this piece of data auditing. Brief report from her work according to Brianna:
    - She has been sorting out the data based on the list Soo handed to Brianna this past month.
    - Handful amount of data were found missing in ACCESS even though the copies are available; after Sameera sorts out the complete list, volunteers will re-enter the data.

Things to Remember for This Month!

- When pulling the data, please make sure to read through all the information about the participant and the visit (early columns in the data pull) before looking through the data 😊

Outlier Analysis

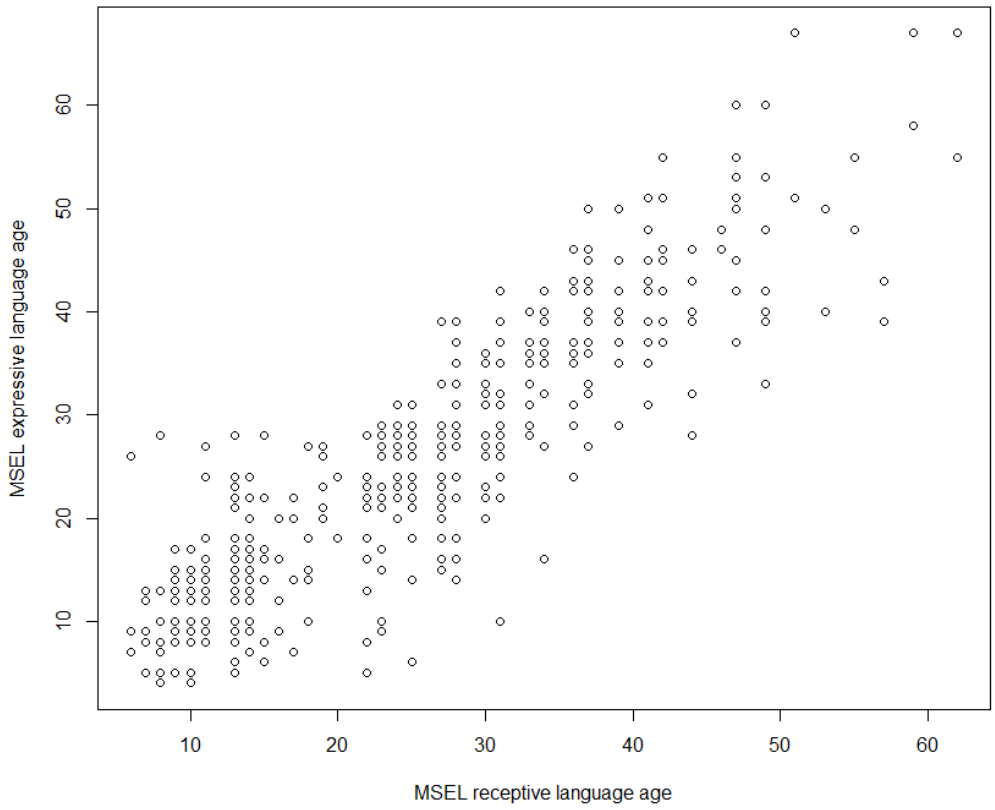
Outlier analyses are conducted on each numeric variable at each time point to detect outliers in single variable distributions using Tukey's fence method Such outliers are flagged as extreme values within the tails of the distribution for that variable at each timepoint across all subjects. Notes about which variables are flagged as outliers are made available in each table and appear in data downloads to alert analysts.

Table 4 shows data for selected tables (as an example) with at least one outlier for a given data row.

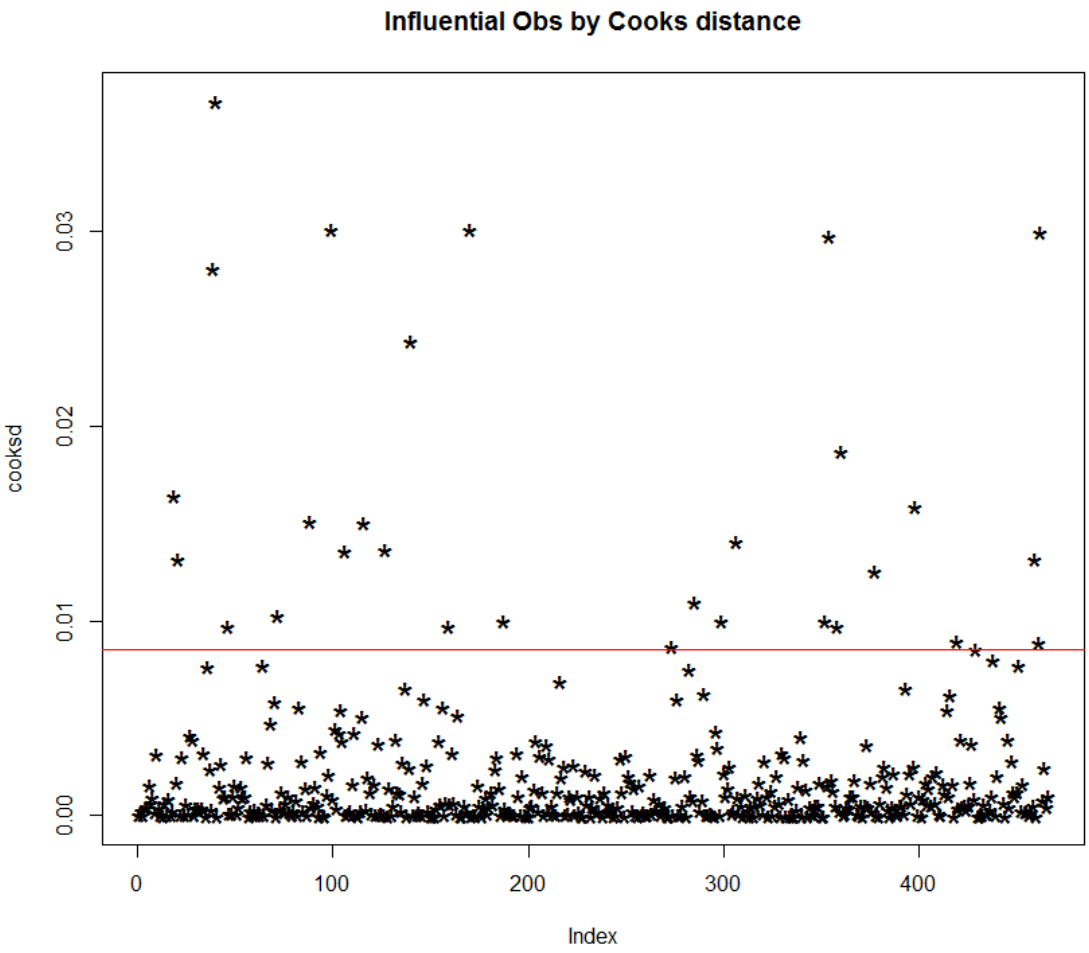
Table 4: Extreme scores

Measure	Outliers	No outliers
ADOS	0	687
ADI	13	450
CBCL	179	875
DAS	84	349
EOWPVT	19	645
PPVT	26	627
MSEL	24	580
Vineland	155	848

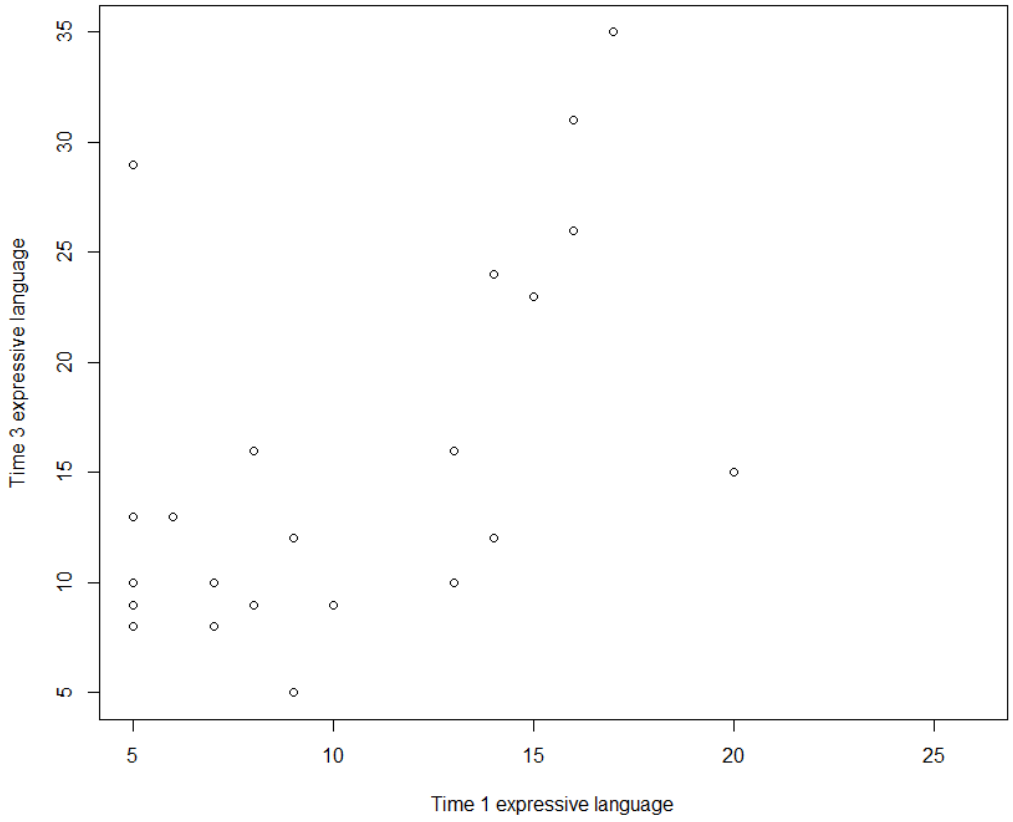
Additional outlier analysis is conducted using regression techniques looking at bivariate scatterplots of pairs of variables from similar measures. As an example, Figure 1 below shows a scatterplot of expressive language age and receptive language age from the MSEL. As expected, the correlation is high (r=.90); however, a few data points show receptive language age scores that are much lower than respective expressive language scores.



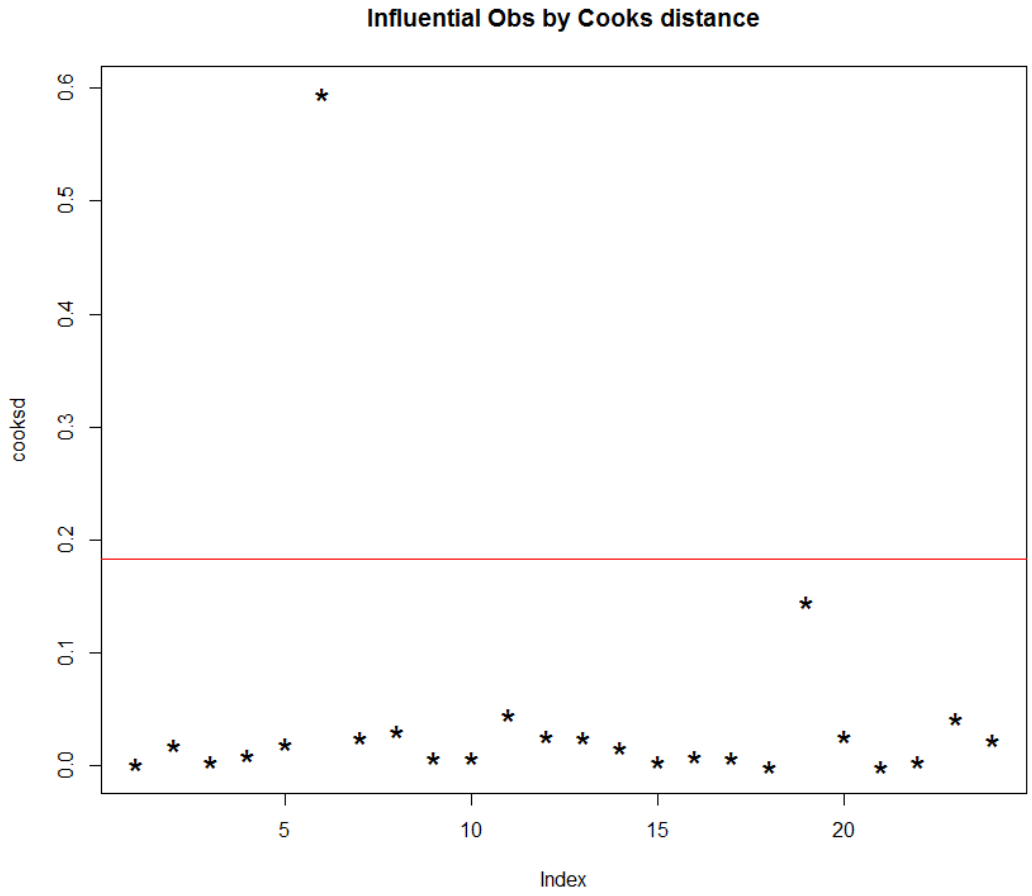
Outlier analysis is done using Cook's method and data points with a Cook's D value larger than 4 times the mean Cook's value are flagged, as shown in Figure 2 below. Cases with extreme Cook's values need to be investigated to determine whether scores are valid and can be used in analyses. This is especially true with cases having high expressive language scores but low receptive language scores. For instance, one subject has an expressive language age equivalent score of 29 months, but a receptive language age score of 6 months which suggests either an invalid MSEL protocol, or a data entry or scoring error. Such an outlier case might not normally be noticed without such higher-level outlier analysis and erroneously be used in data analysis. Such cases can at least be flagged in any downloaded data sets to alert analysts.



A similar regression approach to outlier detection is conducted on variables measured longitudinally. As shown in Figure 3 below, Time 1 expressive language age equivalents from the MSEL are regressed against Time 3 expressive language age equivalents. To the degree that the data is valid, scores should be correlated. The correlation is modest, r=.57.



Analysis of Cook's distance in Figure 4 reveals at least one point with significant discrepancy across the two time points. The most discrepant data is from a child who gained 24 months age equivalent points over the time frame; suggesting the possibility that the Time 1 assessment was invalid and needs to be investigated. Data from a second child who lost roughly 5 months of expressive language ability from T1 to T3 does not show a significant Cook's D value, but is certainly elevated beyond the rest of the sample and might also be investigated.



Regression diagnostics for outliers with bivariate scatterplots are being run for all pairs of variables either measured longitudinally or from different instruments which measure the same construct at the same timepoints.