MAPP Twin Study Data – Document Summary

Datasets

Four datasets are provided, each in both SAS and Excel formats. The vast majority of the MAPP I data is included in the 2 **primary datasets:**

- **REV1_twin_baseline:** This contains all baseline and person-level data. There is one record per person.
- **Final_twin_longitudinal:** This consists of all variables that were collected at more than one visit. There is one record per person/visit combination. The "person-level" variables cohorttype and twin_pid are included in every record of this dataset for easy reference.

The other two datasets are as follows:

- **Cmed_child_level**: CMED (concomitant medications) dataset with multiple records per person. This is as-collected data, listing all medications as reported. No data cleaning was done on this dataset.
- **famhx_child_level:** Like the CMED data, the family history dataset contains multiple records per person, and is thus provided separately from the other datasets. Participants reported any first-degree blood relatives (parents, grandparents, aunts, uncles, siblings, and children) with a history of any of a set of pre-specified chronic pain or psychiatric conditions. There was no attempt to ascertain the lack of any such conditions for any relatives. No data cleaning was done on this dataset.

A SAS format catalog is also included, which you can reference if you are working in SAS. Otherwise, refer to the spreadsheet "REV1 MAPP I SAS Format Definitions.xlsx" for a cross-reference of each format code (numeric value) and text description. The format names are listed in each data dictionary (see below) as appropriate.

Data Dictionaries

A data dictionary is included for each of the four datasets. The data dictionaries for the CMED and FAMHX child-level datasets consist of a simple listing of each of the variables included, as well as variable labels and SAS format names. The data dictionaries for the baseline and longitudinal datasets each have 3 tabs:

- Data: This is the primary data dictionary, with columns providing variable name, type, label, and format. Additional columns provide the following information:
   - o Yes/No indicators as to whether each variable is included in each of the 3 primary datasets
   - o Module: CRF name or other summary indicator for a group of variables (e.g., Neuro, COPCs)
   - o Source: Paper CRF (completed by the RC at an in-clinic visit); Survey (completed by the participant online) or Other (e.g., Lab results)
   - o Indication of whether each variable is Raw or Derived
   - o Indication of whether Imputation is used for each variable

- Derived Variables: All variables listed as Derived in the Data tab have an entry in the Derived Variables tab that includes columns for Description, Calculation, and SAS Code. In cases where the calculation and/or SAS code are complex or involve multiple variables, a reference is provided to a Word document with more detailed information.

- Imputations: All variables listed as Yes for Imputation in the Data tab have an entry in the Imputations tab that includes a column for "Imputation Notes." In cases where the imputation process is complex or involves multiple variables, a reference is provided to a Word document with more detailed information.

Additional Supporting Documentation

Module-Specific Documentation is provided in the following areas:
- o Chronic Overlapping Pain Conditions (COPCs)
- o Concomitant Meds (CMED)
- o Multiple Ability Self-Report Questionnaire (MASQ)
- o International Personality Item Pool Questionnaire (IPIP)
- o Genitourinary Pain Index Questionnaire (GUPI)
- o PROMIS CRF Calculations
- o Medical History