File Merging Instructions

The competition data consists of four separate sets of files:

- `patient_xxxx.csv` is the patient-level file for dataset `xxxx`, with `xxxx` ranging from 0001 to 3400. It contains one record per patient (identified by **id.patient**) and contains patient covariates **V1-V5** as well as the identifier for the practice to which the patient belongs (**id.practice**).
- `patient_year_xxxx.csv` is the time-varying patient outcome file for dataset `xxxx`. It contains one record per patient per year, and has the patient's monthly average medical expenditures, **Y**, for the year. Note: not all patients exist in all years. This can be due to patients aging into Medicare or dying. Years where a patient does not exist are not indicative of missing data to be imputed.
- `practice_xxxx.csv` is the time-invariant practice-level file for dataset `xxxx`. It contains one record per practice (identified by **id.practice**) and contains practice covariates **X1-X9**.
- `practice_year_xxxx.csv` is the time-varying practice file for dataset `xxxx`. It contains one record per practice per year, averages of patient-level covariates for all patients belonging to the practice for the year, **V1_C-V5_C**, as well as the total number of patients, **n.patients**, and the average monthly medical expenditures per patient, **Y**. It also contains the indicator for treatment status, **Z**, and the post period, **post**.

The star schema below shows how to merge these files together to create datasets for analysis. The bold italicized variables denote the "primary key" for each file, i.e. the uniqueness level, while the arrows indicate which variables can be used to join tables together. For example, the fact that the **id.practice** variable for the practice table is in bold italics indicates that the practice data is unique at the level of **id.practice**, while the arrow between **id.practice** in the practice table and **id.practice** in the patient table indicates that those files should be joined on **id.practice**.

- For Track 1 participants, to create the full dataset for a given replicate you will need to merge the `patient_xxxx.csv` file to the `patient_year_xxxx.csv` file on **id.patient** to add outcomes, then you will need to merge to `practice_xxxx.csv` on **id.practice** to get practice covariates, and to `practice_year_xxxx.csv` on **id.practice** and **year** to get treatment status. Importantly, note that a variable named **Y** exists in both the `practice_year_xxxx.csv` and `patient_year_xxxx.csv` files. For Track 1 analysis, participants should keep the **Y** variable only from `patient_year_xxxx.csv` and drop the **Y** variable from `practice_year_xxxx.csv`. Please note that the Track 1 datasets are split into three separate `.zip` files, each containing one third of the datasets – `track1a.zip` includes all files for datasets 0001–1200, `track1b.zip` contains 1201–2400, and `track1c.zip` contains 2401–3400.
- For Track 2 participants, to create the full dataset for a given replicate you will need only to merge the `practice_xxxx.csv` file to the `practice_year_xxxx.csv` file on **id.practice**.