TableReadIn.R

Reads in these tables:

- **Demographic**
    - Create factor for SEX, RACE, RACEWBO, HISPANIC_YN

- **Enrollment**

- **Death**

- ***Merge Demographic, Enrollment, and Death tables***
    - Calculate age @ enrollment start by (ENR_START_DATE – BIRTH_DATE)/ 365.25
    - Filter so only those 18+ years old are included

- **Dispensing (HP)**
    - Convert NDC Code to Labeler and Product part only
    - Create variables for drug type (GLP1RA, SGLT2I, Combination)
        - From NDC_Codes.R and NDC_Codes_Other.R

- **Encounter**
    - (ignore part about provider codes & primary payer type categories & provider specialty)
    - Add hospitalization indicator
        - 1 if ENC_TYPE is “EI”, “IP”, or “OS”
        - 0 otherwise

- ***Merged Diagnosis and Condition***
    - For condition data, make ONSET_DATE = REPORT_DATE if missing ONSET_DATE
    - For “diagnosis” date, used ADMIT_DATE from diagnosis
    - For “diagnosis” date, used ONSET_DATE from condition
        - Changed variable name to ADMIT_DATE
    - Row-binded diagnosis and condition tables to get “ehr_diagnosis”
    - Add outpatient (AV, OA) and inpatient (ED, EI, IP, OS) encounter indicators
    - Create variable “Condition” to specify condition based on ICD9/10 codes

- **Lab Result**
    - Uses LOINC_Codes.R for categorization
    - Create variable “HBA1C_Baseline” to be set to RESULT_NUM if LAB_LOINC is LOINC_HBA1C, RESULT_UNIT is “%”, and RESULT_MODIFIER is “EQ”
        - NA otherwise

    - Create variable “Creatinine_Baseline” to be set to RESULT_NUM if LAB_LOINC is LOINC_Creatinine, RESULT_UNIT is “mg/dL”, and RESULT_MODIFIER is “EQ”
        - NA otherwise

    - Create variable “LDL_Cholesterol_Baseline” to be set to RESULT_NUM if LAB_LOINC is LOINC_LDL_Cholesterol, RESULT_UNIT is “mg/dL”, and RESULT_MODIFIER is “EQ”
        - Else if RESULT_UNIT is “mmol/L”, variable set to RESULT_NUM*18
        - NA otherwise

    - Create variable “HDL_Cholesterol_Baseline” to be set to RESULT_NUM if LAB_LOINC is LOINC_HDL_Cholesterol, RESULT_UNIT is “mg/dL”, and RESULT_MODIFIER is “EQ”
        - NA otherwise

    - Create variable “Total_Cholesterol_Baseline” to be set to RESULT_NUM if LAB_LOINC is LOINC_Total_Cholesterol, RESULT_UNIT is “mg/dL”, and RESULT_MODIFIER is “EQ”
        - NA otherwise

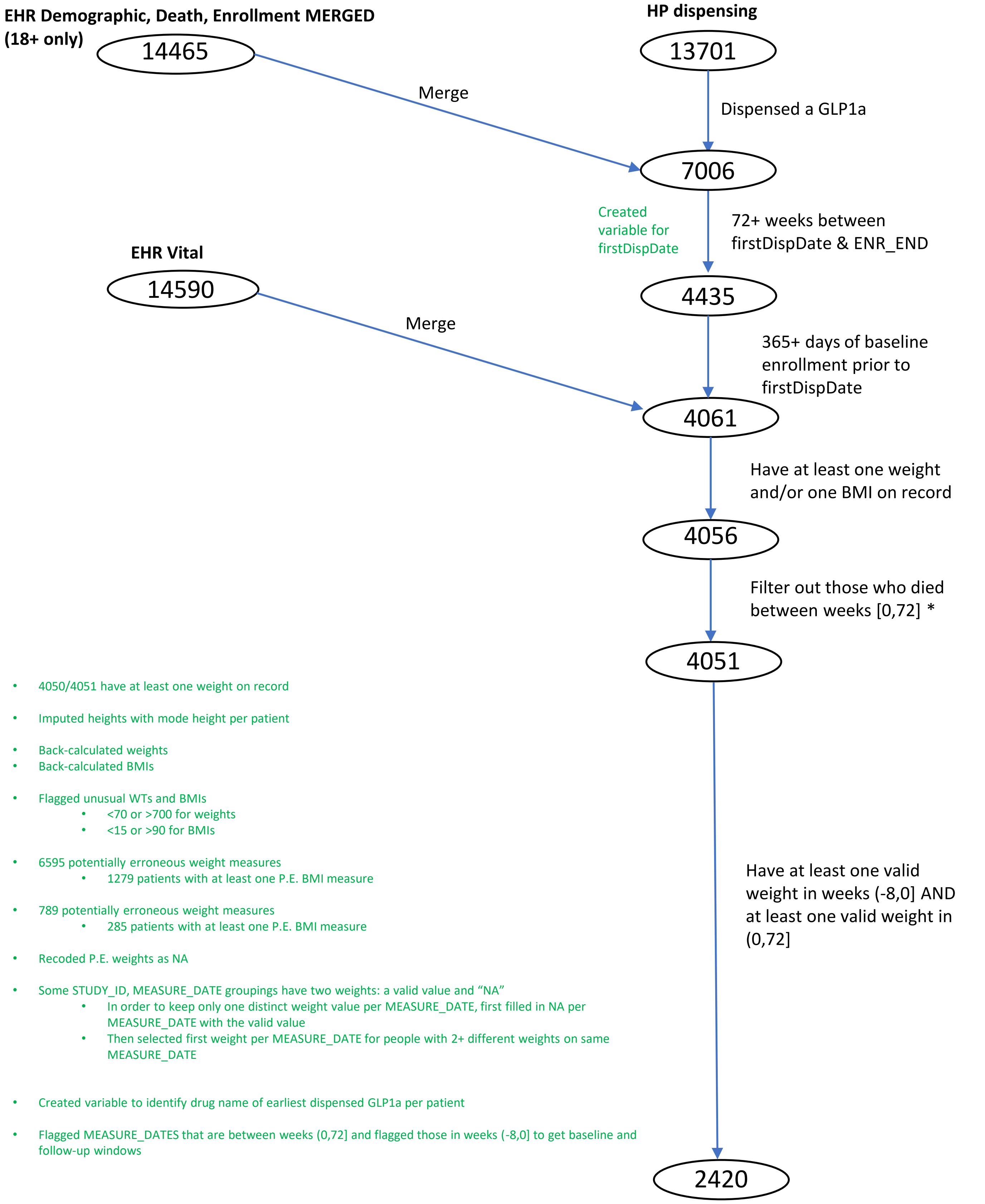
- **Vital**

- **Procedures**
    - Create indicator for bariatric procedures if PX is LAPARO_GASTRIC_BYPASS, LAPARO_GASTRIC_BANDING, LAPARO_SLEEVE_GASTRECTOMY, or MISC_GASTRIC_PROCEDURE
        - Based on Bariatric_CPT_Codes.R

    - Create variable “has_bariatric_proc” if patient has 1 or more bariatric procedures

Changed dates to date data type for all the above
Saved unmerged data frames into “ReadInDataFrames0.rda”

Inclu_Exclusion_Criteria_Filtering.RMD



EHR Demographic, Death, Enrollment MERGED (18+ only)
14465
HP dispensing
13701
Dispensed a GLP1a
Merge
7006
Created variable for firstDispDate
72+ weeks between firstDispDate & ENR_END
EHR Vital
14590
Merge
4435
365+ days of baseline enrollment prior to firstDispDate
4061
Have at least one weight and/or one BMI on record
4056
Filter out those who died between weeks [0,72] *
4051
Have at least one valid weight in weeks (-8,0] AND at least one valid weight in (0,72]
2420

- 4050/4051 have at least one weight on record
- Imputed heights with mode height per patient
- Back-calculated weights
- Back-calculated BMIs
- Flagged unusual WTs and BMIs
    - <70 or >700 for weights
    - <15 or >90 for BMIs
- 6595 potentially erroneous weight measures
    - 1279 patients with at least one P.E. BMI measure
- 789 potentially erroneous weight measures
    - 285 patients with at least one P.E. BMI measure
- Recoded P.E. weights as NA
- Some STUDY_ID, MEASURE_DATE groupings have two weights: a valid value and "NA"
    - In order to keep only one distinct weight value per MEASURE_DATE, first filled in NA per MEASURE_DATE with the valid value
    - Then selected first weight per MEASURE_DATE for people with 2+ different weights on same MEASURE_DATE
- Created variable to identify drug name of earliest dispensed GLP1a per patient
- Flagged MEASURE_DATES that are between weeks (0,72] and flagged those in weeks (-8,0] to get baseline and follow-up windows

- 2458 have a baseline weight
- 2460 have a baseline BMI
- 1009/2420 have 2+ weights in baseline window
- 1022/2420 have 2+ BMIs in baseline window
- Chose a random weight for baseline weight (for those with 2+ baseline weights)
- Chose a random weight for baseline BMI (for those with 2+ baseline BMIs)
- Created variable for baseline weight and baseline BMI

Saved disp_enr_vital11 as final merged df **up to this point**.
NOTE: disp_enr_vital11 contain 4051 patients. The only thing separating 4051 from 2420 cohort is that the 2420 cohort have both a baseline and a follow-up weight, and the remaining 4051 – 2420=1631 do not. Though these 1631 will not be in our final cohort, we are keeping them in this analytic data frame so that MNAR mixed models can be performed later to find factors related to missing data.

* Their ENR_END_DATE is after their DEATH_DATE, but we will consider their DEATH_DATE and their new ENR_END_DATE, which effectively disqualifies them based on inclusion criteria of 72+ weeks of continuous enrollment.

Inclu_Exclusion_Criteria_Filtering2.RMD

Load in “ReadInDataframes0.RDA” from TableReadInr.R and “disp_enr_vital11.RDA”

Overarching goal of this RMD is to **merge disp_enr_vital11 with the diagnosis, encounter, procedures, and lab result tables** and **refining these variables to be fit for a Table One with baseline** conditions, lab results, etc.

- Set disp_enr_vital11 to “df” and select relevant variables (STUDY_ID, firstDispDate, first_Drug_Name, SEX, RACE_WBO, HISPANIC_YN, AGE, baseline_WT, baseline_BMI, has_BLN_and_FU)

DIAGNOSIS

- Merge in diagnosis flags with variables STUDY_ID, ADMIT_DATE, Condition
- Pregnancy
    - 78/2420 found to have “pregnant” on record
        - 7 are “male”
    - 71/2420 “diagnosed” pregnant before firstDispDate
    - 12/2420 “diagnosed” pregnant after firstDispDate
    - 1/2420 “diagnosed” pregnant ON firstDispDate
    - 7/2420 “diagnosed” pregnant between weeks [0,72]
    - 8/2420 “diagnosed” pregnant between months [-9,0]
    - NOTE: the same STUDY_ID may have multiple ADMIT_DATE entries for the same pregnancy
    - Eliminate those who are diagnosed pregnant in weeks [0,72] and/or in months [-9,0] relative to firstDispDate
        - 13 eliminated

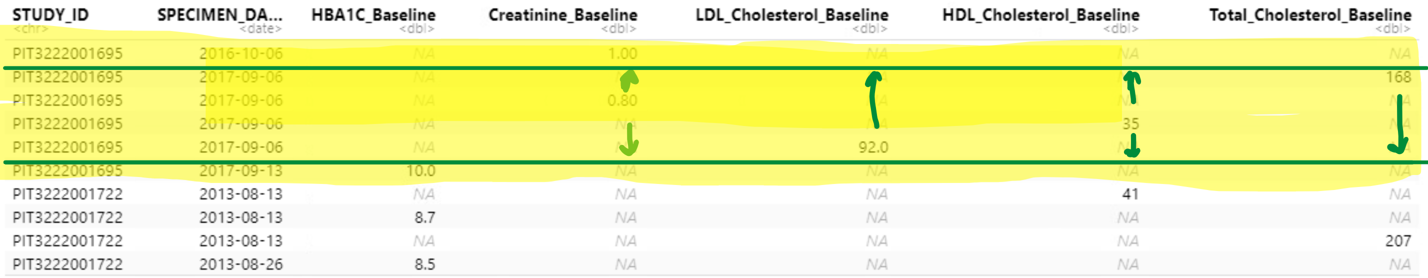
2420

2407

- Require ADMIT_DATE ≤ 365 days prior to firstDispDate as window for all conditions to show up in Table One (since Table One reflects baseline)
    - 2314/2407 have at least one valid* condition in which its ADMIT_DATE is in [-365,0] days
    - Created indicator variable “is_BLN_Condition” to mark whether the ADMIT_DATE is in [-365,0] days
        - “NA” Conditions are still included in “1” if their ADMIT_DATE is in the baseline range

LAB RESULTS

- Merge in lab result flags with variables STUDY_ID, SPECIMEN_DATE, HBA1C_Baseline, Creatinine_Baseline, LDL_Cholesterol_Baseline, HDL_Cholesterol_Baseline, Total_Cholesterol_Baseline
- Filter out SPECIMEN_DATEs with no lab results
    - i.e. include only the rows with at least one baseline lab result
- Require SPECIMEN_DATE ≤ 365 days prior to firstDispDate as window for lab results to show up in Table One (since Table One reflects baseline)
    - 2096/2407 have at least one lab result in which its SPECIMEN_DATE is in [-365,0] days
    - 2783 /4038 have at least one lab result in which its SPECIMEN_DATE is in [-365,0] days
    - Created indicator variable “is_BLN_LabResult” to mark whether the SPECIMEN_DATE is in [-365,0] days
- Choose most recent baseline lab result per category per person with multiple baseline lab



STUDY_ID <chr>	SPECIMEN_DA... <date>	HBA1C_Baseline <dbl>	Creatinine_Baseline <dbl>	LDL_Cholesterol_Baseline <dbl>	HDL_Cholesterol_Baseline <dbl>	Total_Cholesterol_Baseline <dbl>
PIT3222001695	2016-10-06	NA	1.00	NA	NA	NA
PIT3222001695	2017-09-06	NA	NA	NA	NA	168
PIT3222001695	2017-09-06	NA	0.80	NA	NA	NA
PIT3222001695	2017-09-06	NA	NA	NA	35	NA
PIT3222001695	2017-09-06	NA	NA	92.0	NA	NA
PIT3222001695	2017-09-13	10.0	NA	NA	NA	NA
PIT3222001722	2013-08-13	NA	NA	NA	41	NA
PIT3222001722	2013-08-13	8.7	NA	NA	NA	NA
PIT3222001722	2013-08-13	NA	NA	NA	NA	207
PIT3222001722	2013-08-26	8.5	NA	NA	NA	NA

- Similar to how we populated NA WT and BMI values with fill(var, .direction = “downup”), we will **group by STUDY_ID and SPECIMEN_DATE** and then fill the NA values for each baseline type if there is an available value in one of the other rows
    - This allows us to then condense each SPECIMEN_DATE to one row instead of 4+

* “valid” denotes the condition being one that we categorized for this study based on the codes in ICD9_10_Codes.R. If a condition shows as “NA”, it means that it is a condition that is not in this list

- Choose most recent baseline lab result per category per person with multiple baseline lab results *cont.*
    - Create temp which includes only baseline lab results (in the [-365,0] window)
        - Group by STUDY_ID and arrange by descending SPECIMEN_DATE so that most recent SPECIMEN_DATE per patient is on slice 1
        - Fill NA lab result values with fill(HBA1C_Baseline, .direction = “up”) when grouped by STUDY_ID
            - If the value in the first slice (row of the most recent SPECIMEN_DATE) is valid, it will not be populated by the below value
            - But if the value in the first slice is NA and the value in the second slice is valid, the value in the second slice will populate itself in the first slice
            - This way, the original first slice values (from most recent SPECIMEN_DATE) still get “priority”
            - New “first slice” of STUDY_ID & SPECIMEN_DATE groupings will include original lab results where valid AND filled in lab results from the second most recent valid lab results
                - Regardless, all the lab results here were still collected within the baseline window
            - Store these “first slices” into a df so that each patient has their own row with baseline lab results
            - Merge this df with the main merged df

Encounter

- EHR_encounter2 from selecting STUDY_ID, Hospitalization (boolean), ADMIT_DATE from EHR_encounter
- Create variable for number of total hospitalizations between [-365,0] days of firstDispDate
    - Get distinct STUDY_ID & firstDispDate groupings from main merged df
    - Left join this with ehr_encounter2
    - Filter so that only ADMIT_DATES in [-365,0] are included
    - New totalHospitalizations variable is sum of hospitalization booleans per patient
    - Join with main merged df
    - If totalHospitalizations variable = NA, set it = 0 since it means there was no ADMIT_DATEs in [-365,0] for any condition, including hospitalizations

Procedures

- EHR_procedures2 from selecting STUDY_ID, PX_DATE, is_bariatric_proc (boolean) from EHR_procedures
- Already have indicator for whether a PX_DATE coded for a bariatric proc
- Now create indicator variable for whether it’s a baseline bariatric proc
- Based on above variable, create indicator variable for whether a patient has at least one baseline bariatric proc

    - 35/4038 have had a baseline bariatric procedure
    - 28/2407 have had a baseline bariatric procedure

Saved main merged df into df8 in “ReadInDataframes1.RDA”

Table_One_1.RMD

For all the following tables, patients who both **have BLN & FU AND those who don't** are included (n = 4038)

Conditions

- Created separate factor variable for each condition (e.g. "Diabetes.f"
- Made df filtered to include only baseline conditions (PX_DATE in BLN)
    - Necessary for Table One
    - Made indicator "Diabetes_BLN.f" of whether patient has positive record of each condition being diagnosed in BLN
        - "Yes" if sum of non-NA values in Diabetes.f column is 1+
- Made another df filtered to include only outside-of-baseline conditions (PX_DATE not in BLN)
    - Made indicator "Diabetes_out.f" of whether patient has positive record of each condition being diagnosed outside of BLN
        - "Yes" if sum of non_NA values in Diabetes.f column is 1+

Lab Results

- Separate dataset for just lab results, filtering so that only baseline lab results are included

Total Hospitalizations

- Separate dataset for just total hospitalization, totalHosp_BLN variable already calculates number of hospitalizations in baseline

Bariatric Procedures

- Separate dataset for just bariatric procedures , has_BLN_BariProc already indicates whether one has a baseline bariatric procedure

- Merged the above tables

10/3/2021: made the 3 table ones
Next just compile in an email