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

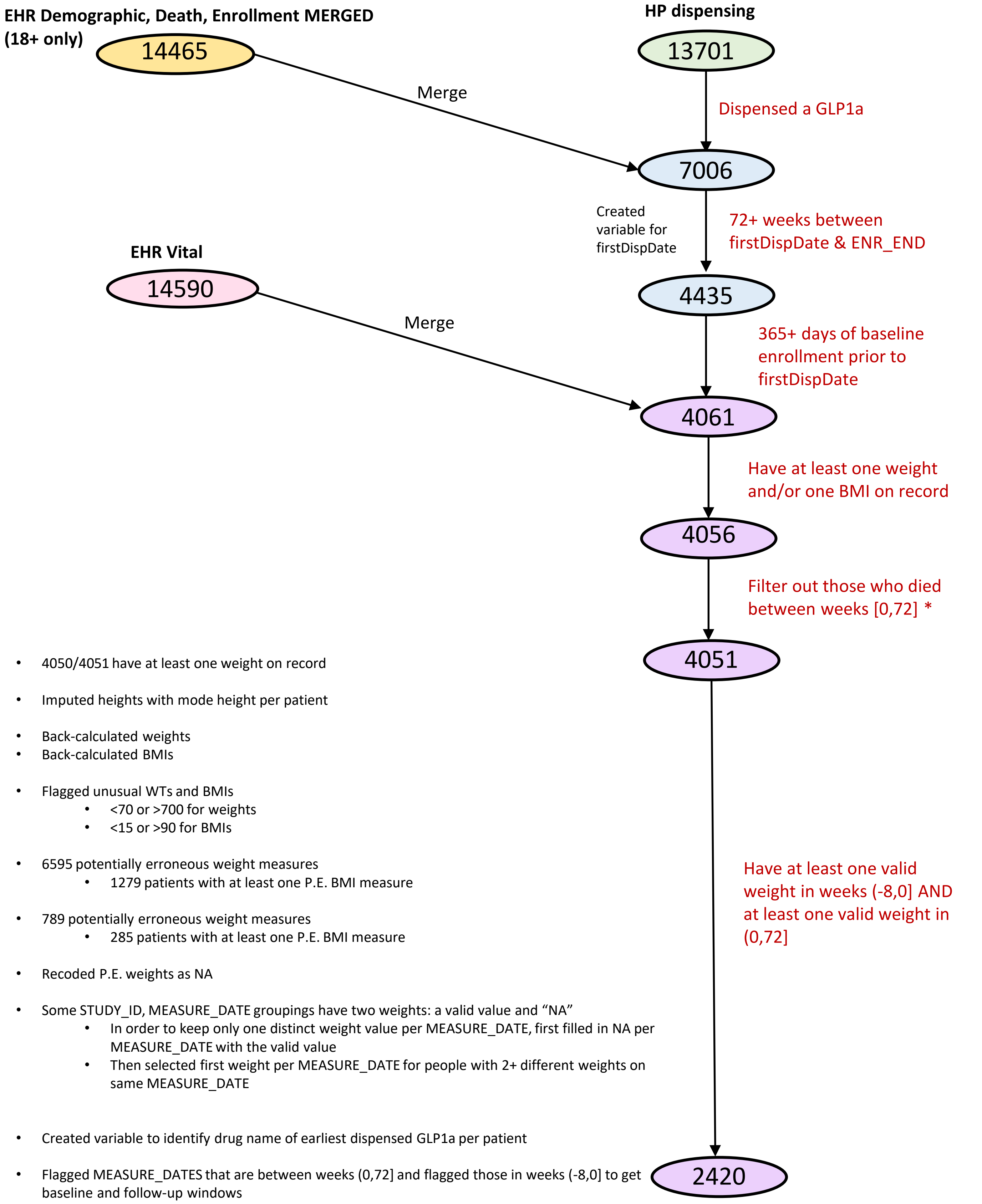
- **Vital**

- **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 (Filtering actions are in **red**)



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

4435

Merge

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

- 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

Have at least one valid weight in weeks (-8,0] AND at least one valid weight in (0,72]

2420

- 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

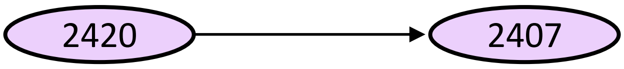
Load in "ReadInDataframes0.RDA" from TableReadInr.R and "disp_enr_vital11.RDA"

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

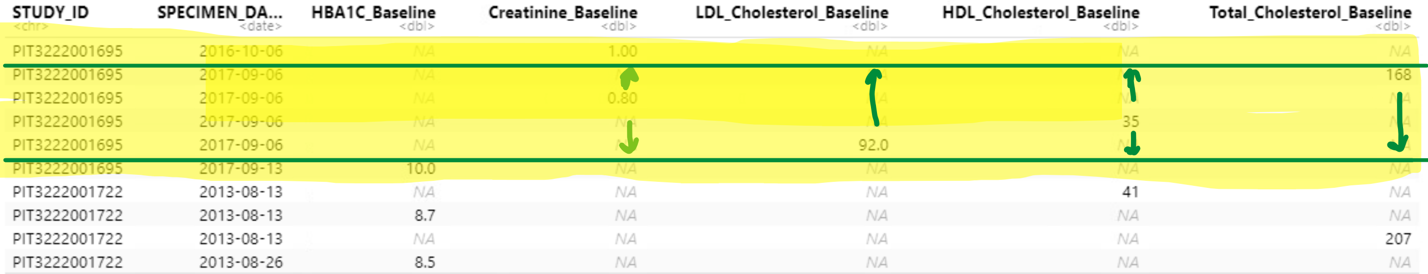
- Merge in diagnosis flags with variables STUDY_ID, ADMIT_DATE, Condition
- Pregnancy
    - 78/2420 found to have "pregnant" on record
        - 7 are "male"
    - 71/2420 "diagnosed" pregnant before firstDispDate
    - 12/2420 "diagnosed" pregnant after firstDispDate
    - 1/2420 "diagnosed" pregnant 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



- 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

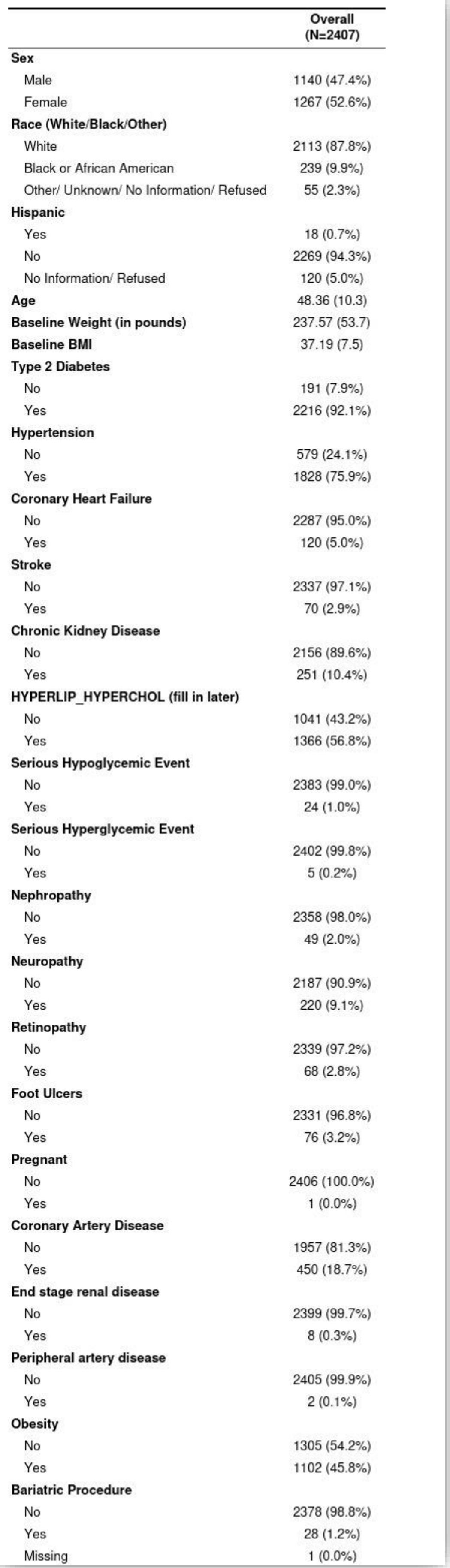
- Separate dataset for just total hospitalization, totalHosp_BLN variable already calculates number of hospitalizations in baseline

Bariatric Procedures

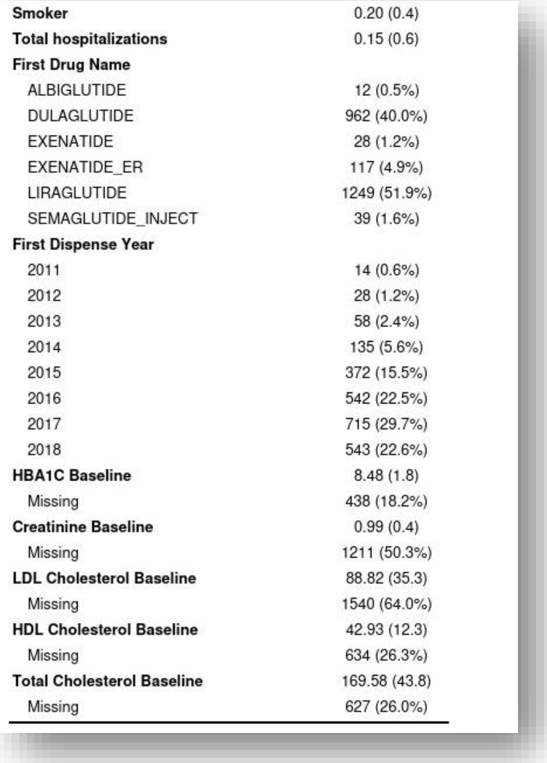
- Separate dataset for just bariatric procedures , has_BLN_BariProc already indicates whether one has a baseline bariatric procedure

- Merged the above tables
- Now prepared to create table ones

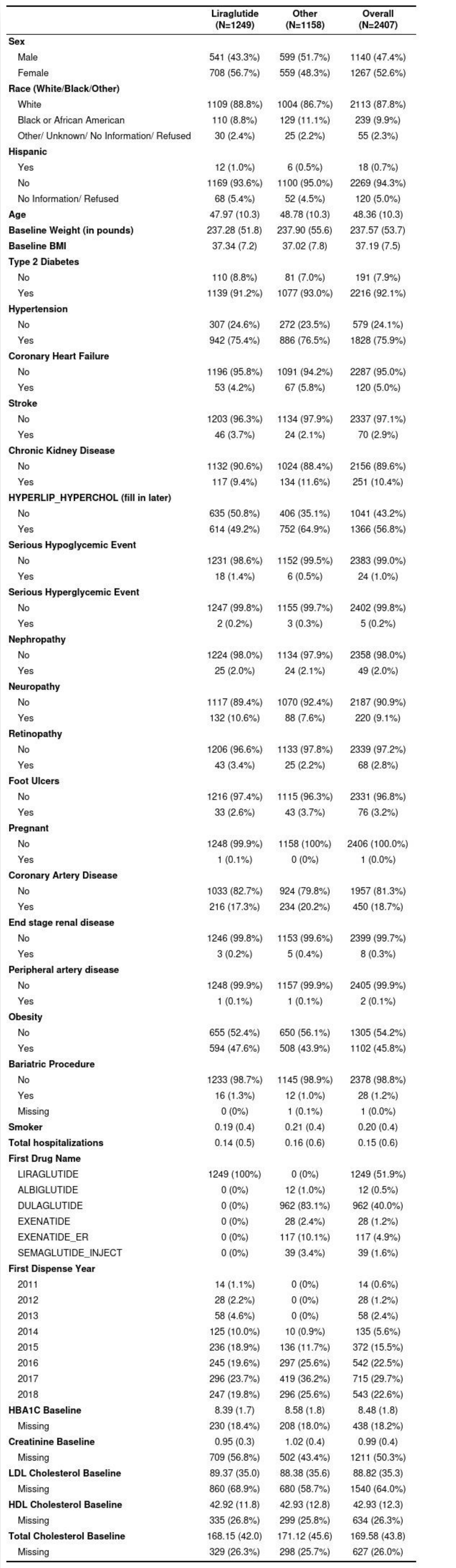
- **Table Ones:**
  1. With the 2407
  2. With the 2407, split by liraglutide or not
  3. With the 4038, split by having baseline & follow-up vs. without



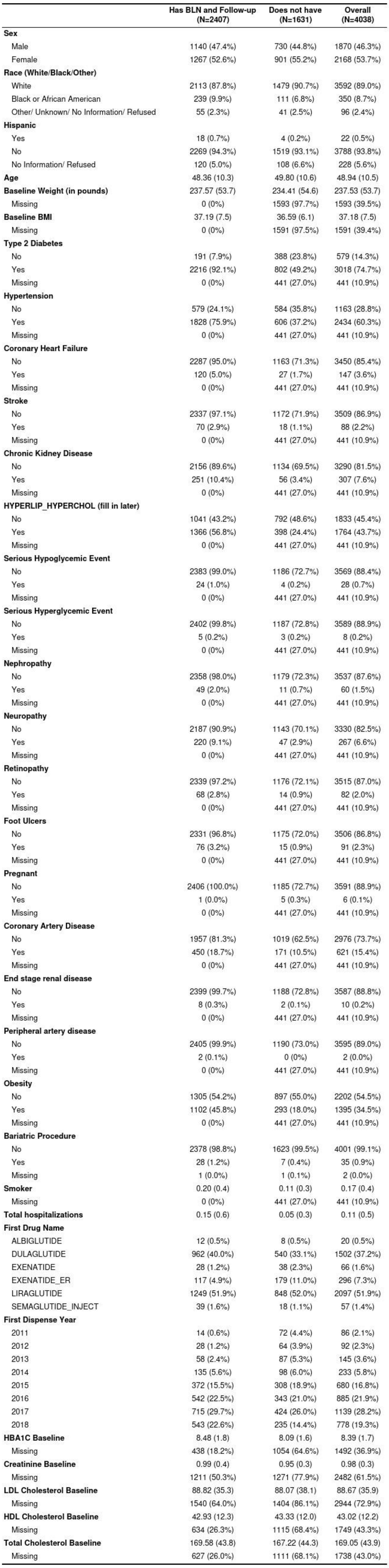
	Overall (N=2407)
Sex	
Male	1140 (47.4%)
Female	1267 (52.6%)
Race (White/Black/Other)	
White	2113 (87.8%)
Black or African American	239 (9.9%)
Other/ Unknown/ No Information/ Refused	55 (2.3%)
Hispanic	
Yes	18 (0.7%)
No	2269 (94.3%)
No Information/ Refused	120 (5.0%)
Age	48.36 (10.3)
Baseline Weight (in pounds)	237.57 (53.7)
Baseline BMI	37.19 (7.5)
Type 2 Diabetes	
No	191 (7.9%)
Yes	2216 (92.1%)
Hypertension	
No	579 (24.1%)
Yes	1828 (75.9%)
Coronary Heart Failure	
No	2287 (95.0%)
Yes	120 (5.0%)
Stroke	
No	2337 (97.1%)
Yes	70 (2.9%)
Chronic Kidney Disease	
No	2156 (89.6%)
Yes	251 (10.4%)
HYPERLIP_HYPERCHOL (fill in later)	
No	1041 (43.2%)
Yes	1366 (56.8%)
Serious Hypoglycemic Event	
No	2383 (99.0%)
Yes	24 (1.0%)
Serious Hyperglycemic Event	
No	2402 (99.8%)
Yes	5 (0.2%)
Nephropathy	
No	2358 (98.0%)
Yes	49 (2.0%)
Neuropathy	
No	2187 (90.9%)
Yes	220 (9.1%)
Retinopathy	
No	2339 (97.2%)
Yes	68 (2.8%)
Foot Ulcers	
No	2331 (96.8%)
Yes	76 (3.2%)
Pregnant	
No	2406 (100.0%)
Yes	1 (0.0%)
Coronary Artery Disease	
No	1957 (81.3%)
Yes	450 (18.7%)
End stage renal disease	
No	2399 (99.7%)
Yes	8 (0.3%)
Peripheral artery disease	
No	2405 (99.9%)
Yes	2 (0.1%)
Obesity	
No	1305 (54.2%)
Yes	1102 (45.8%)
Bariatric Procedure	
No	2378 (98.8%)
Yes	28 (1.2%)
Missing	1 (0.0%)



Smoker	0.20 (0.4)
Total hospitalizations	0.15 (0.6)
First Drug Name	
ALBIGLUTIDE	12 (0.5%)
DULAGLUTIDE	962 (40.0%)
EXENATIDE	28 (1.2%)
EXENATIDE_ER	117 (4.9%)
LIRAGLUTIDE	1249 (51.9%)
SEMAGLUTIDE_INJECT	39 (1.6%)
First Dispense Year	
2011	14 (0.6%)
2012	28 (1.2%)
2013	58 (2.4%)
2014	135 (5.6%)
2015	372 (15.5%)
2016	542 (22.5%)
2017	715 (29.7%)
2018	543 (22.6%)
HBA1C Baseline	8.48 (1.8)
Missing	438 (18.2%)
Creatinine Baseline	0.99 (0.4)
Missing	1211 (50.3%)
LDL Cholesterol Baseline	88.82 (35.3)
Missing	1540 (64.0%)
HDL Cholesterol Baseline	42.93 (12.3)
Missing	634 (26.3%)
Total Cholesterol Baseline	169.58 (43.8)
Missing	627 (26.0%)



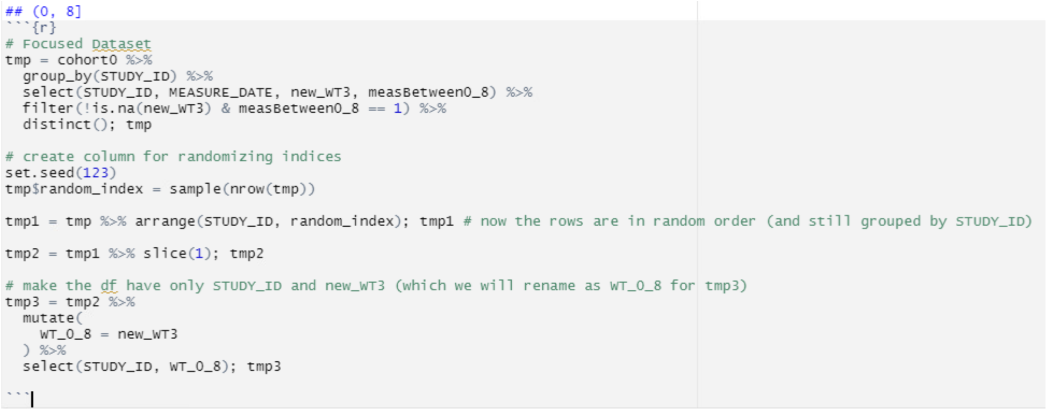
	Liraglutide (N=1249)	Other (N=1158)	Overall (N=2407)
Sex			
Male	541 (43.3%)	599 (51.7%)	1140 (47.4%)
Female	708 (56.7%)	559 (48.3%)	1267 (52.6%)
Race (White/Black/Other)			
White	1109 (88.8%)	1004 (86.7%)	2113 (87.8%)
Black or African American	110 (8.8%)	129 (11.1%)	239 (9.9%)
Other/ Unknown/ No Information/ Refused	30 (2.4%)	25 (2.2%)	55 (2.3%)
Hispanic			
Yes	12 (1.0%)	6 (0.5%)	18 (0.7%)
No	1169 (93.6%)	1100 (95.0%)	2269 (94.3%)
No Information/ Refused	68 (5.4%)	52 (4.5%)	120 (5.0%)
Age	47.97 (10.3)	48.78 (10.3)	48.36 (10.3)
Baseline Weight (in pounds)	237.28 (51.8)	237.90 (55.6)	237.57 (53.7)
Baseline BMI	37.34 (7.2)	37.02 (7.8)	37.19 (7.5)
Type 2 Diabetes			
No	110 (8.8%)	81 (7.0%)	191 (7.9%)
Yes	1139 (91.2%)	1077 (93.0%)	2216 (92.1%)
Hypertension			
No	307 (24.6%)	272 (23.5%)	579 (24.1%)
Yes	942 (75.4%)	886 (76.5%)	1828 (75.9%)
Coronary Heart Failure			
No	1196 (95.8%)	1091 (94.2%)	2287 (95.0%)
Yes	53 (4.2%)	67 (5.8%)	120 (5.0%)
Stroke			
No	1203 (96.3%)	1134 (97.9%)	2337 (97.1%)
Yes	46 (3.7%)	24 (2.1%)	70 (2.9%)
Chronic Kidney Disease			
No	1132 (90.6%)	1024 (88.4%)	2156 (89.6%)
Yes	117 (9.4%)	134 (11.6%)	251 (10.4%)
HYPERLIP_HYPERCHOL (fill in later)			
No	635 (50.8%)	406 (35.1%)	1041 (43.2%)
Yes	614 (49.2%)	752 (64.9%)	1366 (56.8%)
Serious Hypoglycemic Event			
No	1231 (98.6%)	1152 (99.5%)	2383 (99.0%)
Yes	18 (1.4%)	6 (0.5%)	24 (1.0%)
Serious Hyperglycemic Event			
No	1247 (99.8%)	1155 (99.7%)	2402 (99.8%)
Yes	2 (0.2%)	3 (0.3%)	5 (0.2%)
Nephropathy			
No	1224 (98.0%)	1134 (97.9%)	2358 (98.0%)
Yes	25 (2.0%)	24 (2.1%)	49 (2.0%)
Neuropathy			
No	1117 (89.4%)	1070 (92.4%)	2187 (90.9%)
Yes	132 (10.6%)	88 (7.6%)	220 (9.1%)
Retinopathy			
No	1206 (96.6%)	1133 (97.8%)	2339 (97.2%)
Yes	43 (3.4%)	25 (2.2%)	68 (2.8%)
Foot Ulcers			
No	1216 (97.4%)	1115 (96.3%)	2331 (96.8%)
Yes	33 (2.6%)	43 (3.7%)	76 (3.2%)
Pregnant			
No	1248 (99.9%)	1158 (100%)	2406 (100.0%)
Yes	1 (0.1%)	0 (0%)	1 (0.0%)
Coronary Artery Disease			
No	1033 (82.7%)	924 (79.8%)	1957 (81.3%)
Yes	216 (17.3%)	234 (20.2%)	450 (18.7%)
End stage renal disease			
No	1246 (99.8%)	1153 (99.6%)	2399 (99.7%)
Yes	3 (0.2%)	5 (0.4%)	8 (0.3%)
Peripheral artery disease			
No	1248 (99.9%)	1157 (99.9%)	2405 (99.9%)
Yes	1 (0.1%)	1 (0.1%)	2 (0.1%)
Obesity			
No	655 (52.4%)	650 (56.1%)	1305 (54.2%)
Yes	594 (47.6%)	508 (43.9%)	1102 (45.8%)
Bariatric Procedure			
No	1233 (98.7%)	1145 (98.9%)	2378 (98.8%)
Yes	16 (1.3%)	12 (1.0%)	28 (1.2%)
Missing	0 (0%)	1 (0.1%)	1 (0.0%)
Smoker	0.19 (0.4)	0.21 (0.4)	0.20 (0.4)
Total hospitalizations	0.14 (0.5)	0.16 (0.6)	0.15 (0.6)
First Drug Name			
LIRAGLUTIDE	1249 (100%)	0 (0%)	1249 (51.9%)
ALBIGLUTIDE	0 (0%)	12 (1.0%)	12 (0.5%)
DULAGLUTIDE	0 (0%)	962 (83.1%)	962 (40.0%)
EXENATIDE	0 (0%)	28 (2.4%)	28 (1.2%)
EXENATIDE_ER	0 (0%)	117 (10.1%)	117 (4.9%)
SEMAGLUTIDE_INJECT	0 (0%)	39 (3.4%)	39 (1.6%)
First Dispense Year			
2011	14 (1.1%)	0 (0%)	14 (0.6%)
2012	28 (2.2%)	0 (0%)	28 (1.2%)
2013	58 (4.6%)	0 (0%)	58 (2.4%)
2014	125 (10.0%)	10 (0.9%)	135 (5.6%)
2015	236 (18.9%)	136 (11.7%)	372 (15.5%)
2016	245 (19.6%)	297 (25.6%)	542 (22.5%)
2017	296 (23.7%)	419 (36.2%)	715 (29.7%)
2018	247 (19.8%)	296 (25.6%)	543 (22.6%)
HBA1C Baseline	8.39 (1.7)	8.58 (1.8)	8.48 (1.8)
Missing	230 (18.4%)	208 (18.0%)	438 (18.2%)
Creatinine Baseline	0.95 (0.3)	1.02 (0.4)	0.99 (0.4)
Missing	709 (56.8%)	502 (43.4%)	1211 (50.3%)
LDL Cholesterol Baseline	89.37 (35.0)	88.38 (35.6)	88.82 (35.3)
Missing	860 (68.9%)	680 (58.7%)	1540 (64.0%)
HDL Cholesterol Baseline	42.92 (11.8)	42.93 (12.8)	42.93 (12.3)
Missing	335 (26.8%)	299 (25.8%)	634 (26.3%)
Total Cholesterol Baseline	168.15 (42.0)	171.12 (45.6)	169.58 (43.8)
Missing	329 (26.3%)	298 (25.7%)	627 (26.0%)



	Has BLN and Follow-up (N=2407)	Does not have (N=1631)	Overall (N=4038)
Sex			
Male	1140 (47.4%)	730 (44.8%)	1870 (46.3%)
Female	1267 (52.6%)	901 (55.2%)	2168 (53.7%)
Race (White/Black/Other)			
White	2113 (87.8%)	1479 (90.7%)	3592 (89.0%)
Black or African American	239 (9.9%)	111 (6.8%)	350 (8.7%)
Other/ Unknown/ No Information/ Refused	55 (2.3%)	41 (2.5%)	96 (2.4%)
Hispanic			
Yes	18 (0.7%)	4 (0.2%)	22 (0.5%)
No	2269 (94.3%)	1519 (93.1%)	3788 (93.8%)
No Information/ Refused	120 (5.0%)	108 (6.6%)	228 (5.6%)
Age	48.36 (10.3)	49.80 (10.6)	48.94 (10.5)
Baseline Weight (in pounds)	237.57 (53.7)	234.41 (54.6)	237.53 (53.7)
Missing	0 (0%)	1593 (97.7%)	1593 (39.5%)
Baseline BMI	37.19 (7.5)	36.59 (6.1)	37.18 (7.5)
Missing	0 (0%)	1591 (97.5%)	1591 (39.4%)
Type 2 Diabetes			
No	191 (7.9%)	388 (23.8%)	579 (14.3%)
Yes	2216 (92.1%)	802 (49.2%)	3018 (74.7%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Hypertension			
No	579 (24.1%)	584 (35.8%)	1163 (28.8%)
Yes	1828 (75.9%)	606 (37.2%)	2434 (60.3%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Coronary Heart Failure			
No	2287 (95.0%)	1163 (71.3%)	3450 (85.4%)
Yes	120 (5.0%)	27 (1.7%)	147 (3.6%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Stroke			
No	2337 (97.1%)	1172 (71.9%)	3509 (86.9%)
Yes	70 (2.9%)	18 (1.1%)	88 (2.2%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Chronic Kidney Disease			
No	2156 (89.6%)	1134 (69.5%)	3290 (81.5%)
Yes	251 (10.4%)	56 (3.4%)	307 (7.6%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
HYPERLIP_HYPERCHOL (fill in later)			
No	1041 (43.2%)	792 (48.6%)	1833 (45.4%)
Yes	1366 (56.8%)	398 (24.4%)	1764 (43.7%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Serious Hypoglycemic Event			
No	2383 (99.0%)	1186 (72.7%)	3569 (88.4%)
Yes	24 (1.0%)	4 (0.2%)	28 (0.7%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Serious Hyperglycemic Event			
No	2402 (99.8%)	1187 (72.8%)	3589 (88.9%)
Yes	5 (0.2%)	3 (0.2%)	8 (0.2%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Nephropathy			
No	2358 (98.0%)	1179 (72.3%)	3537 (87.6%)
Yes	49 (2.0%)	11 (0.7%)	60 (1.5%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Neuropathy			
No	2187 (90.9%)	1143 (70.1%)	3330 (82.5%)
Yes	220 (9.1%)	47 (2.9%)	267 (6.6%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Retinopathy			
No	2339 (97.2%)	1176 (72.1%)	3515 (87.0%)
Yes	68 (2.8%)	14 (0.9%)	82 (2.0%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Foot Ulcers			
No	2331 (96.8%)	1175 (72.0%)	3506 (86.8%)
Yes	76 (3.2%)	15 (0.9%)	91 (2.3%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Pregnant			
No	2406 (100.0%)	1185 (72.7%)	3591 (88.9%)
Yes	1 (0.0%)	5 (0.3%)	6 (0.1%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Coronary Artery Disease			
No	1957 (81.3%)	1019 (62.5%)	2976 (73.7%)
Yes	450 (18.7%)	171 (10.5%)	621 (15.4%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
End stage renal disease			
No	2399 (99.7%)	1188 (72.8%)	3587 (88.8%)
Yes	8 (0.3%)	2 (0.1%)	10 (0.2%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Peripheral artery disease			
No	2405 (99.9%)	1190 (73.0%)	3595 (89.0%)
Yes	2 (0.1%)	0 (0%)	2 (0.0%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Obesity			
No	1305 (54.2%)	897 (55.0%)	2202 (54.5%)
Yes	1102 (45.8%)	293 (18.0%)	1395 (34.5%)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Bariatric Procedure			
No	2378 (98.8%)	1623 (99.5%)	4001 (99.1%)
Yes	28 (1.2%)	7 (0.4%)	35 (0.9%)
Missing	1 (0.0%)	1 (0.1%)	2 (0.0%)
Smoker	0.20 (0.4)	0.11 (0.3)	0.17 (0.4)
Missing	0 (0%)	441 (27.0%)	441 (10.9%)
Total hospitalizations	0.15 (0.6)	0.05 (0.3)	0.11 (0.5)
First Drug Name			
ALBIGLUTIDE	12 (0.5%)	8 (0.5%)	20 (0.5%)
DULAGLUTIDE	962 (40.0%)	540 (33.1%)	1502 (37.2%)
EXENATIDE	28 (1.2%)	38 (2.3%)	66 (1.6%)
EXENATIDE_ER	117 (4.9%)	179 (11.0%)	296 (7.3%)
LIRAGLUTIDE	1249 (51.9%)	848 (52.0%)	2097 (51.9%)
SEMAGLUTIDE_INJECT	39 (1.6%)	18 (1.1%)	57 (1.4%)
First Dispense Year			
2011	14 (0.6%)	72 (4.4%)	86 (2.1%)
2012	28 (1.2%)	64 (3.9%)	92 (2.3%)
2013	58 (2.4%)	87 (5.3%)	145 (3.6%)
2014	135 (5.6%)	98 (6.0%)	233 (5.8%)
2015	372 (15.5%)	308 (18.9%)	680 (16.8%)
2016	542 (22.5%)	343 (21.0%)	885 (21.9%)
2017	715 (29.7%)	424 (26.0%)	1139 (28.2%)
2018	543 (22.6%)	235 (14.4%)	778 (19.3%)
HBA1C Baseline	8.48 (1.8)	8.09 (1.6)	8.39 (1.7)
Missing	438 (18.2%)	1054 (64.6%)	1492 (36.9%)
Creatinine Baseline	0.99 (0.4)	0.95 (0.3)	0.98 (0.3)
Missing	1211 (50.3%)	1271 (77.9%)	2482 (61.5%)
LDL Cholesterol Baseline	88.82 (35.3)	88.07 (38.1)	88.67 (35.9)
Missing	1540 (64.0%)	1404 (86.1%)	2944 (72.9%)
HDL Cholesterol Baseline	42.93 (12.3)	43.33 (12.0)	43.02 (12.2)
Missing	634 (26.3%)	1115 (68.4%)	1749 (43.3%)
Total Cholesterol Baseline	169.58 (43.8)	167.22 (44.3)	169.05 (43.9)
Missing	627 (26.0%)	1111 (68.1%)	1738 (43.0%)

Bucket_Time_Windows_1.RMD

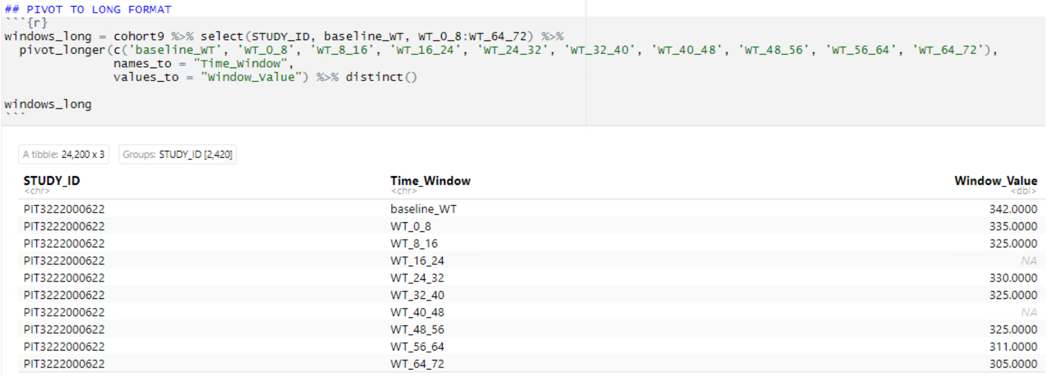
- Load in disp_enr_vital11.rda
- Created indicator for each window, mark whether each row (MEASURE_DATE) falls in the window
- Filtered to only include 2407 cohort
- Chose a random weight for windows that have 2+ valid weights per patient
    - Ran the following for each window:



~~~
## (0, 8]
```{r}
# Focused Dataset
tmp = cohort0 %>%
  group_by(STUDY_ID) %>%
  select(STUDY_ID, MEASURE_DATE, new_WT3, measBetween0_8) %>%
  filter(!is.na(new_WT3) & measBetween0_8 == 1) %>%
  distinct(); tmp

# create column for randomizing indices
set.seed(123)
tmp$random_index = sample(nrow(tmp))

tmp1 = tmp %>% arrange(STUDY_ID, random_index); tmp1 # now the rows are in random order (and still grouped by STUDY_ID)

tmp2 = tmp1 %>% slice(1); tmp2

# make the df have only STUDY_ID and new_WT3 (which we will rename as WT_0_8 for tmp3)
tmp3 = tmp2 %>%
  mutate(
    WT_0_8 = new_WT3
  ) %>%
  select(STUDY_ID, WT_0_8); tmp3

```
~~~

- Above code would generate new variable named “WT_StartNumber_EndNumber” containing a weight for that window, one-to-one per patient (got rid of 2+ weight problem)
- Successively merge this temp df with the main one to accumulate the weight variables

- Pivoted to long format:



~~~
## PIVOT TO LONG FORMAT
```{r}
windows_long = cohort9 %>% select(STUDY_ID, baseline_WT, WT_0_8:WT_64_72) %>%
  pivot_longer(c('baseline_WT', 'WT_0_8', 'WT_8_16', 'WT_16_24', 'WT_24_32', 'WT_32_40', 'WT_40_48', 'WT_48_56', 'WT_56_64', 'WT_64_72'),
               names_to = "Time_Window",
               values_to = "Window_Value") %>% distinct()

windows_long
```
~~~

A tibble: 24,200 x 3 Groups: STUDY_ID [2,420]

STUDY_ID <chr>	Time_Window <chr>	Window_Value <dbl>
PIT3222000622	baseline_WT	342.0000
PIT3222000622	WT_0_8	335.0000
PIT3222000622	WT_8_16	325.0000
PIT3222000622	WT_16_24	NA
PIT3222000622	WT_24_32	330.0000
PIT3222000622	WT_32_40	325.0000
PIT3222000622	WT_40_48	NA
PIT3222000622	WT_48_56	325.0000
PIT3222000622	WT_56_64	311.0000
PIT3222000622	WT_64_72	305.0000