In large part, the following describes what happens in /proj/DaltonLab/projects/p0013/progs/04\_Harmonizing\_CCF\_and\_MH\_data.Rmd

Essentially, that .Rmd file creates views that combine the analogous CCF and MH data sets into single tables (e.g., combines MH’s and CCF’s encounters tables into a single table).

For each table, a “union” or “union all” of both CCF’s and MH’s data was performed. Notes on what happens on each side of the union can be found in the institution-specific notes.

There is usually a “from\_ccf” flag to indicate where each row came from. Additional table-specific notes are below.

Demographics section

* NEOCARE\_COHORT\_DEMOGRAPHICS\_V
* Columns:
  + study\_id (integer)
  + birth\_date (date)
  + female (byteint)
  + race (varchar 38)
  + hispanic\_or\_latino (byteint)
  + death\_date (date)
  + icd10cod (varchar 5)
  + cv\_death (byteint)
  + all\_death\_codes (varchar 160)
  + ccf\_birth\_date (date)
  + mh\_birth\_date (date)
  + bd\_discrep\_in\_days (integer)
  + ccf\_female (byteint)
  + mh\_female (byteint)
  + ccf\_race (varchar 38)
  + mh\_race (varchar 38)
  + ccf\_hispanic\_or\_latino (byteint)
  + mh\_hispanic\_or\_latino (byteint)
  + ccf\_death\_date (date)
  + mh\_death\_date (date)
* CCF’s data preferred to MH’s data except for the following cases:
  + When a patient’s latest 1999-2017 encounter was at MH (and not tied with CCF), MH’s data value for sex was used.
  + When a patient’s MH race value was “Multiracial”, that value was used
  + When a patient’s MH hispanic\_or\_latino value was 1, that value was used.
* MH’s death data was in a separate table, not included in the demographics table like CCF’s data. Therefore, MH\_Deaths was included in the creation of this view along with MH\_COHORT\_DEMOGS and CCF\_COHORT\_DEMOGRAPHICS\_V
* A flag was created indicating when one institution marked a patient as male and the other, female.
* A column was created indicating the difference in days between the institution’s birth dates

Encounters section

* NEOCARE\_COHORT\_ENCOUNTERS\_V
* Columns:
  + study\_id (integer)
  + CONTACT\_DATE (date)
  + BMI (decimal (9, 2))
  + height\_cm (number (40, 0))
  + weight\_kg (float)
  + sbp (float)
  + dbp (float)
  + from\_ccf (byteint)
  + MHSpecialty (varchar 30)
* Notes on implausible vitals values are at the top of this section
* MH’s “specialty” column was kept as “MHSpecialty”. CCF’s specialty data is more complicated so was simply left out. It can be merged in from QHS\_OUTCOMES\_V.Encounters as needed via QHS\_UMLS\_V.Relationships (see commented-out code in query)
* Encounters included if between 1999-2017 and if occurred when patient was 18 or older

Code chunk “procs”

* DL\_NEOCARE.NEOCARE\_COHORT\_PROCEDURES\_V
* Columns:
  + study\_id (integer)
  + proc\_date (timestamp 3)
  + ConceptID (character 8)
  + cpt (varchar 100)
  + hcpcs (varchar 100)
  + CCS (smallint)
  + icd9 (varchar 5)
  + icd10 (character 7)
  + from\_ccf (byteint)
* Procedures identified by CCS and CPT are the only reliably unified fields. ICD codes (found in MH rows) do not map to CCS codes 232-244. Some CCF rows only have ConceptID.

Code chunk “geoids”

* Creates DL\_NEOCARE.NEOCARE\_COHORT\_GEOIDS, prepared from UNIONed CCF and MH GEOID tables
  1. CCF and MH GEOID data were concatenated, which created instances of identical rows.
  2. Rows were removed if they were unique but shared a study\_id-start\_date combination with rows that were not unique (i.e., if one institution had multiple GEOIDs associated with a specific patient at a specific start\_date, and the other institution had a matching GEOID for that patient at that start\_date, then the matching GEOID is kept and the non-matching one is removed).
  3. Duplicate rows were removed.
  4. An iterative process in which rows were removed was applied to the table over and over again until no more rows met the criteria for removal. The table was sorted by study\_id followed by start\_date, and then any given row (call it Row X) was removed if these three conditions were met:
     + Row X’s combination of study\_id and start\_date was not unique in the table (i.e., at least one other row had the same study\_id and start\_date as Row X).
     + Considering only rows with the same study\_id as Row X, there existed at least one start\_date earlier than Row X’s start\_date, and the largest of these start\_dates occurred in only one row, which we will call Row Y (i.e., the patient’s next earliest start\_date has a single GEOID associated with it).
     + Row X and Row Y have the same GEOID.
  5. An iterative process in which rows were removed was applied to the table over and over again until no more rows met the criteria for removal. The table was sorted by study\_id followed by start\_date, and then any given row (call it Row J) was removed if these three conditions are met:
     + Row J’s combination of study\_id and start\_date is not unique in the table (i.e., at least one other row had the same study\_id and start\_date as Row J).
     + Considering only rows with the same study\_id as Row J, there existed at least one start\_date earlier than Row J’s start\_date, and the largest of these start\_dates occurred in only one row, which we will call Row K (i.e., the patient’s next earliest start\_date has a single GEOID associated with it).
     + Row J and Row K have the same GEOID.
* Columns:
  + study\_id (integer)
  + geoid\_date (date)
  + geoid (varchar 15)
  + homeless (byteint)
* It is important to note that there remains contradictory data in NEOCARE\_COHORT\_GEOIDS in order to prevent data loss: there are many instances of a patient allegedly living in two different places on the same day. The user will have to account for this, using some process to decide which GEOID to use for any given day.
* Homelessness subsection – creates DL\_NEOCARE.NEOCARE\_COHORT\_HOMELESSNESS\_V by combining the corresponding institution-specific homelessness views

Diagnoses section

* DL\_NEOCARE.NEOCARE\_COHORT\_DIAGNOSES\_V
* Columns:
  + study\_id (integer)
  + diag\_date (timestamp 3)
  + ConceptID (character 8)
  + icd9 (varchar 100)
  + icd10 (varchar 100)
  + from\_ccf (byteint)
* Other diagnoses views
  + Both of these are indexed on study\_id and diag\_date
  + Elixhauser view
    - DL\_NEOCARE.NEOCARE\_COHORT\_ELIX\_V
    - Columns:
      * study\_id (integer)
      * diag\_date (timestamp 3)
      * elix (varchar 12)
      * from\_ccf (integer)
      * from\_mh (integer)
  + CCS view
    - DL\_NEOCARE.NEOCARE\_COHORT\_CCS\_DIAGNOSES\_V
    - Columns:
      * study\_id (integer)
      * diag\_date (timestamp 3)
      * ccs (integer)
      * from\_ccf (integer)
      * from\_mh (integer)

Labs section

* straightforwardly UNIONed views individualized to lab results of interest (see labs sections of institution-specific documentation files)
* Columns:
  + study­\_id (integer)
  + lab\_datetime (timestamp 3)
  + column named for the lab result of interest
  + from\_ccf (byteint)
* Work subsection was an analysis in order to determine how to properly create merged lab result views

Medications section

* DL\_NEOCARE.NEOCARE\_COHORT\_MEDICATIONS\_V
* Columns:
  + study\_id (integer)
  + ConceptID (character 8)
  + Strength (varchar 1156)
    - In CCF rows this is QHS\_OUTCOMES.Medications.dosage concatenated with QHS\_OUTCOMES.Medications.doseunitdescription separated by a single space
  + order\_date (timestamp 3)
  + start\_date (timestamp 3)
  + end\_date (timestamp 3)
  + from\_ccf (byteint)
* Meds identified by ConceptID
* MH STRENGTH column combined with CCF’s dosage column concatenated with dosageunitdescription
* Meds of Interest section contains the code chunk that loads the fruits of the labor of /proj/DaltonLab/projects/neocare/p0013/progs/med\_work.Rmd
  + In order to find all relevant ConceptIDs for a medication of interest, its children are collected using QHS\_UMLS\_V.Relationships where ExpandedRelationship = “isa”.
  + Doug Einstadter manually selected the child concepts that can be appropriately deemed a subset or member of the medication of interest. Each of these child concepts was fed to the MedicationConceptsTesting stored procedure (found in the “Medication Concepts” section of /proj/DaltonLab/projects/neocare/p0013/progs/stored\_procedures.Rmd) in order to grab all appropriate ConceptIDs
  + This was done for antihypertensives, antidiabetes medication, and antidepressants

ASCVD Events section

* NEOCARE\_COHORT\_ASCVD\_EVENTS\_V
* Columns:
  + study\_id (integer)
  + event\_date (date)
  + event\_type (varchar 8)
  + from\_ccf (byteint)
* The leading subsection creates a table of diagnosis ConceptIDs identifying those that are ASCVD events. This uses the Child Concepts stored procedure, running it on C0038454 (“Cerebrovascular accident”) and C0027051 (“Myocardial Infarction”)
* This is a UNION of cardiovascular deaths (taken from NEOCARE\_COHORT\_DEMOGRAPHICS\_V), heart attack and diagnoses (taken from NEOCARE\_COHORT\_DIAGNOSES\_V), and coronary artery bypass graft (CABG) and percutaneous transluminal coronary angioplasty (PTCA) procedures (taken from NEOCARE\_COHORT\_PROCEDURES\_V)

Financial Class section

* Creates DL\_NEOCARE.NEOCARE\_COHORT\_FINANCIAL\_CLASS\_V
* Columns:
  + study\_id (integer)
  + CONTACT\_DATE (date)
  + financial\_class (varchar 36)
  + from\_ccf (integer)
  + from\_mh (integer)
* Selects one financial class per study\_id / CONTACT\_DATE combination. Prefers classes found in both institutions, followed by the following order of precedence:
  + Military
  + Medicaid
  + Medicare
  + Private
  + Research/Worker’s Comp/Other/Unknown
  + Uninsured/Self-pay

Smoking Status section

* Creates DL\_NEOCARE.NEOCARE\_COHORT\_SMOKING\_V
* Columns:
  + study\_id (integer)
  + CONTACT\_DATE (date)
  + smoking\_status (varchar 7)
  + quit\_date (date)
  + packs\_per\_day (decimal (6, 4))
  + years\_smoked (decimal (6, 3))
* Creates one row per study\_id/CONTACT\_DATE combination, selecting the maximum smoking\_status, quit\_date, packs\_per\_day, and years\_smoked for each combination
* Once 2 or more smoking statuses of “current” or “former” are accumulated, subsequent non-“current” smoking statuses are coerced to “former”.
* If quit\_date is nonmissing and smoking\_status is not “current”, smoking\_status is coerced to “former”