In large part, the following describes what happens in /proj/DaltonLab/projects/p0013/progs/02\_CCF\_cohort\_views.Rmd

Essentially, this document creates views that lightly clean the raw tables of CCF data and pare them down to only include patients in the NEOCARE cohort (i.e., they are inner-joined with DL\_NEOCARE.STUDY\_ID\_KEY).

Code chunk “demogs”

* Creates DL\_NEOCARE.CCF\_COHORT\_DEMOGRAPHICS\_V on Teradata.
* DL\_NEOCARE.STUDY\_ID\_KEY inner joined with DL\_NEOCARE.CCF\_DEMOGRAPHICS on patient ID.
* Cleaning of the variables RACE and ETHNICITY:
  + Step 1
    - If RACE was missing and ETHNICITY = “Caucasian,” it was filled with “White”.
    - If RACE was missing and ETHNICITY = “African American”, it was filled with “Black/African American”.
    - If RACE was “Black”, it was changed to “Black/African American”.
    - If RACE = “Hispanic/Latino” and ETHNICITY = “Not Hispanic”, ETHNICITY was changed to missing.
    - If RACE = “Hispanic/Latino” and ETHNICITY was either missing or equal to something other than “Not Hispanic”, ETHNICITY was changed to “Hispanic or Latino”.
  + Step 2
    - If RACE = “Hispanic/Latino”, it was changed to missing.
    - If RACE = “Asian/Pacific Islander” or “Native Hawaiian/Other Pacific Islander”, it was changed to “Native Hawaiian/Other Pacific Islander”.
    - If ETHNICITY = “Caucasian” or “African American” it was changed to missing.

Code chunk “encs”

* Creates DL\_NEOCARE.CCF\_COHORT\_ENCOUNTERS\_V on Teradata.
* DL\_NEOCARE.STUDY\_ID\_KEY inner joined with DL\_NEOCARE.CCF\_ENCOUNTERS on patient ID.
* CONTACT\_DATE was changed to missing if its year was not in the range 1900-2017, inclusive.

Code chunk “diags”.

* Creates DL\_NEOCARE.CCF\_COHORT\_DIAGNOSES\_V
* DL\_NEOCARE.STUDY\_ID\_KEY inner joined with DL\_NEOCARE.CCF\_DIAGNOSES on patient ID.
* Both dx\_date and contact\_date were changed to missing if their years were not in the range 1900-2017, inclusive.

Code chunk “meds”.

* Creates DL\_NEOCARE.CCF\_COHORT\_MEDICATIONS\_V
* DL\_NEOCARE.STUDY\_ID\_KEY inner joined with DL\_NEOCARE.CCF\_MEDICATIONS.
* Each of ORDERING\_DATE, START\_DATE, and END\_DATE were changed to missing if their years were not in the range 1900-2017, inclusive.

Code chunk “labs”.

* Creates DL\_NEOCARE.CCF\_COHORT\_LABS\_V
* DL\_NEOCARE.STUDY\_ID\_KEY inner joined with DL\_NEOCARE.CCF\_LABS on patient ID.
* specimn\_taken\_time [sic] was changed to missing if its year was not in the range 1900-2017, inclusive.

Code chunk “deaths”.

* Creates DL\_NEOCARE.CCF\_COHORT\_DEATHS\_V
* QHS\_OUTCOMES\_V.Patients full-outer-joined with DL\_NEOCARE.ODI\_CauseOfDeath on patient ID, which was then inner-joined with DL\_NEOCARE.STUDY\_ID\_KEY on patient ID.
* DL\_NEOCARE.ODI\_CauseOfDeath had eleven patients who had two death dates listed. Alex Milinovich was consulted, and he said to take the death date that had a nonmissing associated value of EntityAxisCodes. For the others, he specified the correct row in each case.
* Handling of death dates:
  + The QHS\_OUTCOMES\_V.Patients table contained three death date columns:
    - OhioDeathIndexDate
    - SocialSecurityDeathDate
    - DateOfDeath
  + The DL\_NEOCARE.ODI\_CauseOfDeath table contained its own OhioDeathIndexDate column.
  + A column named death\_date\_calc was added, containing the first non-missing value of the above four (cleaned) columns in the order that they were mentioned above.
  + All the above columns were kept in the final result, though the QHS\_OUTCOMES\_V.Patients’s OhioDeathIndexDate column was renamed to pat\_OhioDeathIndexDate and DL\_NEOCARE.ODI\_CauseOfDeath’s OhioDeathIndexDate column was renamed to cod\_OhioDeathIndexDate.
* Records were excluded if death\_date\_calc was earlier than the patient’s earliest encounter in the CONTACT\_DATE column in DL\_NEOCARE.CCF\_COHORT\_ENCOUNTERS\_V