In large part, the following describes what happens in /proj/DaltonLab/projects/neocare/p0013/progs/03\_MH\_cohort\_data\_uploading\_and\_management.Rmd

These files create the MH data files for transfer to CCF. The files are created at MH, transferred to CCF, and then uploaded to Teradata. These .rds files are all located in:

/proj/neocare/2018/rawdata/mh/to\_ccf

In each of the sections below, the first step after loading the data into R was to remove all records associated with the study\_id 2094273, since it was created in error.

All empty character strings were coerced to missing before being written as flat files to upload to Teradata. The file format .csv was used unless longer character fields were present (usually free text containing quotation marks and/or commas), in which case the file format .tsv was used. When .tsv was used, each character field was first checked for the presence of tab characters—none were ever found. Thus, .tsv files, when written, were written without any quoting. All flat files were encoded as UTF-8, and missing data was written as an empty character string (i.e., “”). They were all written into the folder:

/proj/neocare/2018/rawdata/mh/to\_teradata

All flat files were uploaded into Teradata via Teradata Studio using “Data Transfer”. In order that character string data would be read in successfully, correctly, and efficiently, a few less obvious options had to be specified:

* File Encoding was specified as UTF-8.
* Concerning character fields:
  + The maximum number of characters had to be explicitly specified in many cases. These ranges were calculated in R on the data sets that were fed into the flat-file-writing functions.
  + The data type CHAR was specified for the few cases in which each value in a character column had the same number of characters; otherwise, the data type VARCHAR was specified.
  + For longer (usually free text) character fields, Teradata Studio required that “Unicode” was specified so that certain symbols could be correctly read.

Some further MH data management also occurs after uploading. Where this is the case, it is described below.

Demographics section

* File: mh\_demogs\_with\_dates.rds
* A duplicate row for the study\_id 93009 had been created in error, so the second one was removed.
* A patient who was not at least 18 years old during the study period of 1999-2017 was excluded
* Written as: mh\_demogs.csv
* Uploaded into Teradata as: DL\_NEOCARE.MH\_COHORT\_DEMOGRAPHICS.
* Columns:
  + study\_id (integer)
  + SEX (varchar 6)
  + BIRTHDATE (date)
  + RACE (varchar 34)
  + ETHNICITY (varchar 12)

Diagnoses section

* Files:
  + mh\_diags9\_with\_dates.rds
  + mh\_diags10\_with\_dates.rds
* Written as:
  + mh\_diags\_icd9.csv
  + mh\_diags\_icd10.csv
* Uploaded into Teradata as:
  + DL\_NEOCARE.MH\_DIAGS\_ICD9
  + DL\_NEOCARE.MH\_DIAGS\_ICD10.
* View DL\_NEOCARE.MH\_COHORT\_DIAGNOSES\_V creating by UNIONing the above tables, while also left-joining ICD-to-ConceptID key tables created from QHS\_UMLS\_V.Atoms
  + Columns:
    - study\_id (integer)
    - diag\_date (date)
      * coalesced first\_date and last\_date
    - ConceptID (character 8)
    - icd9 (varchar 7)
    - icd10 (varchar 8)
    - from\_ccf (byteint)
* Elixhauser subsection
  + Creates DL\_NEOCARE.MH\_COHORT\_ELIX\_V
  + Columns:
    - study\_id (integer)
    - diag\_date (date)
      * coalesced first\_date and last\_date
    - elix (varchar 12)
  + Inner joins DL\_NEOCARE.MH\_DIAGS\_ICD9 and DL\_NEOCARE.MH\_DIAGS\_ICD10 to the icd-to-elixhauser key tables created in /proj/DaltonLab/projects/neocare/p0013/progs/icd\_elix\_key\_table\_creation.R, and UNIONs them
* CCS subsection
  + Creates DL\_NEOCARE.MH\_COHORT\_CCS\_DIAGNOSES\_V
  + Columns:
    - study\_id (integer)
    - diag\_date (date)
      * coalesced first\_date and last\_date
    - ccs (integer)
  + Inner joins DL\_NEOCARE.MH\_DIAGS\_ICD9 and DL\_NEOCARE.MH\_DIAGS\_ICD10 to the icd-to-ccs key tables created in /proj/DaltonLab/projects/neocare/p0013/progs/icd\_ccs\_key\_table\_creation.R, and UNIONs them

Section “Encounters”

* File: mh\_encs\_with\_dates.rds
* Written as: mh\_encs.csv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_ENCS
* MH\_COHORT\_ENCOUNTERS\_V created
  + Columns:
    - study\_id (integer)
    - CONTACT\_DATE (date)
    - ENC\_TYPE (varchar 17)
    - DEPARTMENT\_NAME (varchar 24)
    - SPECIALTY (varchar 30)
    - HEIGHT\_CM (decimal (19, 16))
    - WEIGHT\_KG (decimal (21, 18))
    - TEMPERATURE (decimal (18, 15))
    - PULSE (smallint)
    - RESPIRATIONS (smallint)
    - BP\_SYSTOLIC (smallint)
    - BP\_DIASTOLIC (smallint)
  + filtered out encounters occurring outside 1999-2017 and occurring when patient was under 18

Section “Labs”

* File: mh\_labs\_with\_dates.rds
* Modifications
  + PROC\_ID was coerced from character to integer.
* Written as: mh\_labs.tsv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_LABS
* Unit Analysis subsection analyzes distributions of the values associated with different combinations of components and units (see also labs section of corresponding CCF file and NEOCARE harmonization documentation file)
* Results subsection
  + Creates views containing the values of select lab results.
  + Based on work also completed in 02\_CCF\_cohort\_data.Rmd
  + Columns:
    - study\_id (integer)
    - SPECIMN\_TAKEN\_TIME (timestamp)
    - Column named after the result type
* In order to obtain new labs, manually review mh\_lab\_component\_unit\_summary.csv for the Component/Unit combinations that seem to contain the desired data. This is created in the subsection “Lab Component/Unit distribution and counts”. It is in descending order of counts so that the user may choose stop when entering the decimal dust. Then, create a query that grabs labs with the desired Component/Unit combinations, converting lab values as necessary to ensure unit conformity. See 02\_CCF\_cohort\_data.Rmd for the similar process on the CCF side. The queries in the “Results” subsection may serve as a template. Make sure to always exclude rows with a lab value of 9999999.

Section “Meds”

* File: mh\_meds\_with\_dates.rds
* Written as: mh\_meds.tsv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_MEDS
* Creates DL\_NEOCARE.MH\_COHORT\_MEDICATIONS\_V
  + MH\_MEDS left joined with DL\_NEOCARE.MetroDataMapping (created by Alex Milinovich and Doug Einstadter) in order to provide UMLS ConceptIDs
  + Columns:
    - study\_id (integer)
    - ConceptID (character 8)
    - ORDERING\_DATE (date)
    - PHARM\_CLASS (varchar 33)
    - PHARM\_SUBCLASS (varchar 57)
    - MEDICATION\_NAME (varchar 72)
    - GENERIC\_NAME (varchar 60)
    - STRENGTH (varchar 26)
    - SIG (varchar 450)
    - QUANTITY (varchar 255)
    - START\_DATE (date)
    - END\_DATE (date)
    - REFILLS (varchar 68)
  + Includes meds with an order date, start date, or end date in 1999-2017

Section “Procs”

* Files:
  + mh\_procs1.rds, mh\_procs2.rds, and neo\_cpt.sas7bdat
  + These files contained a mix of ICD9CM, ICD10PCS, and CPT codes. Key tables to CCS were merged with each.
* Written as: mh\_procs2.tsv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_COHORT\_PROCS
  + Columns:
    - study\_id (integer)
    - proc\_date (date)
    - proc\_name (varchar 160)
    - cpt (varchar 14)
    - icd9 (varchar 5)
    - icd10 (char 7)
    - ccs (smallint)

Section “GEOIDs”

* File: mh\_geoid\_with\_dates.rds
* Written as: mh\_geoids.csv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_GEOIDS\_RAW
* DL\_NEOCARE.MH\_COHORT\_GEOIDS, a subset of MH\_COHORT\_GEOIDS\_RAW that was prepared according to the iterative processes described in the corresponding CCF GEOID documentation:
  + Columns:
    - study\_id (integer)
    - geoid\_date (date)
    - geoid (varchar 12)
  + Patients who are homeless are assumed to live in Cuyahoga county (39035)
* Homelessness subsection creates the view DL\_NEOCARE.MH\_COHORT\_HOMELESSNESS\_V, which reports the dates when each patient was reported as homeless
  + Columns:
    - study\_id (integer)
    - homelessness\_date (date)
  + Only includes dates in 1999-2017

Deaths section

* MH death data was uploaded months later than the above data and was not handled in 03\_MH\_cohort\_data\_uploading\_and\_management.Rmd. It was not processed at CCF in R, but rather was processed and written as a .csv file at MH before being sent to CCF. See mh\_death\_data\_prep.docx.
* File mh\_deaths.csv received from MH into CCF into the same folder as the .rds files (see above).
* Copied into the /to\_teradata/ folder (see above).
* Uploaded to Teradata as: DL\_NEOCARE.MH\_COHORT\_DEATHS.
  + Columns:
    - study\_id (integer)
    - death\_date (date)
    - death\_year (smallint)
    - place\_of\_death (byteint)
      * this would need to be decoded using the Ohio death index data dictionary
    - acme\_uc (varchar 4)
      * this is icd10 cause of death
    - rac (varchar 73)
      * “record axis codes”
      * list of cause of death codes
    - c113 (varchar 3)
    - lc113 (varchar 2)
    - tract (varchar 11)
      * place where decedent was living when they died

Financial Class section

* File: /proj/neocare/2018/rawdata/mh/to\_ccf/2020-05-01/enc\_with\_fc.sas7bdat
* Written as mh\_financial\_class.csv
* Uploaded to Teradata as MH\_FINANCIAL\_CLASS
* Creates DL\_NEOCARE.MH\_COHORT\_FINANCIAL\_CLASS\_V
  + Columns:
    - study\_id (integer)
    - CONTACT\_DATE (date)
    - financial\_class (varchar 23)
  + includes rows in MH\_FINANCIAL\_CLASS dated between 1999-2017

Smoking Status section

* File: /proj/neocare/2018/rawdata/mh/to\_ccf/2020-06-03/neocare\_smoke\_status.zip (neocare\_smoke\_status.sas7bdat is the file)
* Written as mh\_smoke.tsv
* Uploaded to Teradata as MH\_SMOKE
* Creates DL\_NEOCARE.MH\_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))
    - start\_date (date)
  + Some study\_id/contact\_date combinations had multiple rows. The maximum quit\_date, packs\_per\_day, years\_smoked, start\_date, and smoking status (unknown < never < passive < former < current) was taken for each study\_id/contact\_date combination
  + For any given study\_id, once two or more instances of “current” or “smoker” are accumulated, all non-“current” smoking statuses will be coerced to “former”
  + If quit\_date is not null and if smoking\_status is not “current”, smoking\_status is coerced to “former”
  + Other manual fixes were performed (e.g., changed “1+” to 1, “0,25” to 0.25)