In large part, the following describes what happens in /proj/DaltonLab/projects/p0013/progs/03\_clean\_and\_upload\_mh\_data\_with\_dates.Rmd

Essentially, this document edits and uploads to Teradata the patient data that were assembled at MH (see \_\_\_) and sent to CCF as .rds files. 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 on 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.

Section/Code chunk “demogs”

* 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.
* Written as: mh\_demogs.csv
* Uploaded into Teradata as: DL\_NEOCARE.MH\_COHORT\_DEMOGRAPHICS.

Section “Diags”

* 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\_COHORT\_DIAGS\_ICD9
  + DL\_NEOCARE.MH\_COHORT\_DIAGS\_ICD10.

Section “Encounters”

* File: mh\_encs\_with\_dates.rds
* Written as: mh\_encs.csv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_COHORT\_ENCS

Section “Labs”

* File: mh\_labs\_with\_dates.rds
* Modifications
  + PROC\_ID was coerced from character to integer.
  + A dash was inserted between the penultimate and last character in DEFAULT\_LNC\_ID so that it would be a correctly formatted LOINC code.
* Written as: mh\_labs.csv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_COHORT\_LABS

Section “Meds”

* File: mh\_meds\_with\_dates.rds
* Written as: mh\_meds.tsv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_COHORT\_MEDS

Section “Procs”

* File: mh\_procs\_with\_dates.rds
* Written as: mh\_procs.tsv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_COHORT\_PROCS

Section “GEOIDs”

* File: mh\_geoid\_with\_dates.rds
* Written as: mh\_geoids.csv
* Uploaded to Teradata as: DL\_NEOCARE.MH\_COHORT\_GEOIDS

Concerning MH death data

* MH death data was uploaded months later than the above data and was not handled in 03\_clean\_and\_upload\_mh\_data\_with\_dates.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 \_\_\_.
* 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.