In large part, the following describes what happens in /proj/DaltonLab/projects/p0013/progs/01\_deduplication.Rmd

1. CCF data setup
   1. A subset of CCF’s encounters table (in which each row represents a unique patient encounter) is migrated into a Teradata database. Encounters are included in this subset if they occur during the years 1999-2017, are listed as completed, and are in specified categories (see flowchart in /figures/ subfolder). This Teradata table is called CCF\_ENCOUNTERS. **See ?query?**
   2. A subset of CCF’s demographics database (in which each row represents a unique patient) is migrated into a Teradata database. Patients are included in this subset if they have an encounter in CCF\_ENCOUNTERS. **See ?query?**
   3. A subset of CCF\_DEMOGRAPHICS is extracted. Patients are included if they have encounters in CCF\_ENCOUNTERS on at least 2 different dates on which their truncated age in years was between 18 and 115, inclusive. Since an inordinate number of patient birth dates are listed as 1 January 1900, patients with this birthday were excluded under the assumption that most of these birthdays are incorrect. All non-English characters in patients’ first and last names are coerced to their English equivalent (e.g., “é” is coerced to “e”) and characters are coerced to uppercase. Furthermore, last4SSNs that were allegedly equal to “0000” are coerced to missing values, since the Social Security Administration does not assign SSNs ending in “0000.” **See the code chunk “read\_in\_ccf\_cohort”.**
2. Stage 1 deduplication: CCF and MH patients are deduplicated based on common Care Everywhere IDs (CEIDs).
   1. List of all CEIDs on file at MH that begin with “CCF” (henceforth “CCFCEIDs”) are sent to CCF (henceforth called the “MH CCFCEID list”). This included the CCFCEIDs of patients not in the cohort, but a binary variable identifying those in the cohort enabled this list to be filtered to include only patients in the MH cohort. **See the code chunk “ccf\_ceid”**.
   2. The CCF cohort table, which includes CCFCEIDs when available, is filtered to only include patients who have a CCFCEID that is in the MH CCFCEID list. These are the **Stage 1 matches**, and they are set aside as successful deduplicates. **See the code chunk “stage1\_matches”**.
   3. Patients in the CCF cohort table who did not have a CCFCEID that matched one in the MH CCFCEID list are set aside in another table, henceforth called the “Stage 2 CCF cohort table.” **See the code chunk “stage1\_ccf\_nonmatches”**.
   4. Elements of the MH CCFCEID list that did not match any CCFCEIDs in the CCF cohort table are extracted and sent back to MH, henceforth called the “Stage 1 MH non-match list.” **See the code chunk “stage1\_mh\_nonmatches”**.
3. Stage 2 deduplication: CCF and MH patients who were not deduplicated in Stage 1 are deduplicated based on first name, last name, last four digits of Social Security Number (henceforth “last4SSN”), and birth year.
   1. MH sends to CCF the internal MH patient ID, MH MRN, first name, last name, last4SSN, and birth year of patients who meet the MH study criteria and who were not deduplicated in Stage 1; henceforth, this will be called the “Stage 2 MH cohort table.” All non-English characters in patients’ first and last names are coerced to their English equivalent (e.g., “é” is coerced to “e”) and characters are coerced to uppercase. Furthermore, last4SSNs that were allegedly equal to “0000” are coerced to missing values, since the Social Security Administration does not assign SSNs ending in “0000.” **See the code chunk “receive\_phi\_of\_MH\_patients\_for\_stage\_2”.**
   2. Inner join performed on the Stage 2 CCF cohort table and the Stage 2 MH cohort table based upon total matches across Anglicized first name, Anglicized last name, last4SSN, and birth year. These are the **Stage 2 matches**. **See the code chunk “stage2”.**
      1. There were no missing values among first name or last name in either table, and there were no missing birth years in the Stage 2 CCF cohort table.
      2. However, both tables had patients with missing last4SSNs; such patients could still be deduplicated, provided that the last4SSN was missing on both ends.
      3. Per standard inner join behavior, if a row on either side matched multiple rows on the other side, all possible combinations were produced in the result; therefore, the first combination was kept and all others were removed.
   3. All rows in the Stage 2 CCF cohort table that did not have a match in the Stage 2 MH cohort table were extracted, henceforth called the “CCF non-match table.” All rows in the Stage 2 MH cohort table that did not have a match in the Stage 2 CCF cohort table were extracted, henceforth called the “MH non-match table.” **See the code chunk “consolidate\_nonmatches”.**
4. Assembling the master study data set
   1. The Stage 1 matches, Stage 2 matches, CCF non-match table, and MH non-match table were concatenated, the rows were randomly rearranged, and a study\_id was assigned to each row based on its position in the randomized table. **See the code chunk “create\_study\_data”.**
   2. This study\_id key was written as a .tsv file and uploaded to Teradata as STUDY\_ID\_KEY. **See the code chunk “write\_study\_data”.**
5. Sending the study\_ids back to MH.
   1. Since the creation of the master study data set was completed at CCF servers, an MH-specific study\_id key was sent back to MH so that their cohort patients could be correctly identified with their newly created study\_ids. **See the code chunks “stage1\_study\_ids”, “stage2\_study\_ids”, and “write\_mh\_study\_id”.** The key took the form of two tables:
      1. A table relating the newly created study\_ids to the CCFCEIDs of the Stage 1 matches (which are MH patients whose PHI were not sent to CCF).
      2. A table relating the newly created study\_ids to the Stage 2 MH cohort table.
   2. The study\_id was added to MH data, Patient IDs and MRNs were removed, and these data were sent to CCF. **See MH\_cohort\_creation.docx in /proj/DaltonLab/projects/p0013/documentation.**