

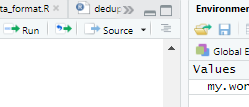
Deduplication Version 4

Cervantes, Andrew

Algorithm Protocol

For any Questions about the Protocol, code, or files please contact [jihong.yao@doh.nj.gov](mailto:jihong.yao@doh.nj.gov) or [Stella.Tsai@doh.nj.gov](mailto:Stella.Tsai@doh.nj.gov) with your questions. If they are unable to help or solve the issue at hand, You can email me at [ajcervantes99@gmail.com](mailto:ajcervantes99@gmail.com) and I will try to get back to you in a timely manner.

* Downloading R and RStudio (**Both links need to be used**)
  + If you do not already have R or RStudio, they will need to be downloaded
    - R: <https://cran.r-project.org/mirrors.html>, select somewhere in the U.S.
    - RStudio: <https://www.rstudio.com/products/rstudio/download/>
* Getting Files(If you are **NOT** using a specialized program (**Hep C**, **Hep B**, **Lyme**))
  + Go to X:\DISEASE INFORMATION\HEPATITIS C\DATA & STATS (BY YEAR)\Dedup\Summer 2019
  + Copy the “Deduplication” folder and place it where you find it convenient for your disease group to access it
  + Open RStudio
  + Select “File” and then “open file”
  + Go to the “Deduplication” folder you just moved and open it
  + Double click on the file labeled “Dedup\_general. R”
* If you **ARE** working with one of the Specialized programs (**Hep C**, **Hep B**, **Lyme**), then follow these steps instead:
  + Open RStudio
  + Select “File” and then “open file”
  + Go to X:\DISEASE INFORMATION\HEPATITIS C\DATA & STATS (BY YEAR)\Dedup\Summer 2019
  + Double click the file of your disease
  + Double click the file “Dedup\_\*\*disease name\*\*”
* Getting Data
  + Use CDRSS to download a report for your disease
    - Download with the following criteria (check these boxes):
      * Disease
      * Date for report
      * Illness onset date
      * Case status
      * Report status
      * First name
      * Last name
      * Date of birth
      * 7 addresses (address, county, and state **ONLY**)
      * Primary Phone number
      * Gender
      * Race
      * Ethnicity
    - **IMPORTANT**: Excel file of the download should have 35 columns
    - Exclude all cases with the report status “Merged”
  + Open the .cdrss file in Excel
    - Under the options, select delimited then next
    - Then select other and put the ‘^’ symbol
    - Then click next, and then finished
  + Formatting
    - Click on the date of birth column
      * With whole column selected, right click, and select “format cells”
      * First click the category “date”, then the category “custom”
      * Format the date to be “mm/dd/yyyy” and click “okay”
    - Sorting data (Recommended for improved speed)
      * Use “ctrl-A” to select all data
      * Go to the “Data” tab in excel and select “Sort”
      * Under “Sort By” select “DOB”, and sort from oldest to newest
  + Save new Data in the “Deduplication” folder as a .csv file
* Running the code
  + Open RStudio and the file that you opened earlier should be in the top left window
    - Some general things to know:
      * In R, both “=” and “<-“ mean equals
      * Any line with a ‘#’ at the front is a comment
      * The “Run” button can be found in the top right of the loft left box



* + First, highlight all code in **part 1** and run only if needed
    - Run the install packages code **only** if you are using R or the code for the first time. Running that code more than once may cause issues
    - This may take a few minutes
  + Once it finishes, look at the code in **Part 2**
    - If you moved the location of the folder you are using, edit the file path in the line “my.working.directory <- “”
      * **NOTE**: If you copy the file path and paste it, you will need to change all forward slashes (‘\’) to backslashes (‘/’)
    - Highlight all code in **Part 2** and hit run
  + Getting and editing the other file
    - In RStudio, at the bottom right of the GUI is the directory window
    - Under files, you will see all files that are in the working directory that you set above. Click the file named “dedup\_data\_format.R” or “dedup\_data\_\*\*disease\_name\*\*.R” and click on that file tab (top left)
    - On line 2, where it says read.csv (a filename, …), change the filename with the name of the .csv file you just saved
      * Make sure to use “” and to include “.csv” at the end
    - Save the file and go back to “Dedup\_general.R”
  + Look at the code in **Part 3**
    - Find the “Cutoff” variable and edit it if needed
      * For example, if you want values above or equal to 0.9, set cutoff to be 0.9
      * Recommended values are either 0.95, 0.9, or 0.85 (would not recommend setting the value lower then 0.85)
    - Highlight all the code in **Part 3** and hit Run
  + If no errors occur, run the code in **Part 4.** This is the actual program running, and will take some time
  + When the code has finished, there will be one txt file in your “Deduplication” folder
    - Code will be finished when there are some black numbers at the bottom of the console tab (Bottom left)
    - “OutputP.txt” is your cases to merge the people of, and the percent from the algorithm
    - Outputs will consist of the percent value and then the two case IDs, one belonging to each of the two people to be merged.
    - **IMPORTANT**: expect the code to take a few hours to complete depending on the size of the data (55,000 case IDs take about 2 hours)
    - Program will run when computer is asleep, but will stop if it is shut off or restarted
    - For what to do with the output, continue to “Algorithm Output”

Algorithm Output

Algorithm is designed to give a percentage that two given people should be merged. From testing, some common percent values **usually** correlate to certain merge situations, such as:

* 1: Means same name and same DOB and same address or phone number
* 0.97: Means same DOB, a nickname/full name combination and same address or phone number
  + Ex. “Rob” and “Robert”
* 0.96: Means same DOB and a slight difference in name or an extra middle initial and same address or phone number
  + Ex. “Jerry R Jones” and “Jerry Jones”
* 0.95: Means DOB is the same and one of the names are one letter off while the other name is exact and same address/phone number
  + Ex. “Bruce Springsteen” and “Bruce Springstean”
* 0.94: Means that the DOB and address or phone number are the same, but the first and last names are flipped for one person (wildcard search)
  + Ex. “Jon Smith” and “Smith Jon”
* 0.93: Means same name and address or phone number, while the DOBs are one number off
  + Ex. “01/01/1967” and “01/01/1968”
* 0.92: Same name and DOB, and one of the addresses is a jail/correctional facility
  + Ex. One home address and one address is “396 Straight Street, Paterson”
* 0.91: Means DOB is off by more than one digit but name and address or phone number match
  + Ex. “01/01/1968” and “01/01/1972” or “01/01/1955” and “01/01/1985”
* 0.9: Means DOB and name are the same, but not address or phone number

**Excel Files to Update**

Files that contain working databases used for the code. Update each of them to your benefit\*\*\* as follows:

* Jail\_addresses.csv: This files contains the street adrress of jails and correctional facilities. If your disease sees a particular address be quite common, add it to this list
* Nicknames\_list.csv: This file contains a list of nicknames and their corresponding full names. If you think of a nickname/fullname combo, add it to this list so the program can catch it for you
* Exclusions.csv: For any IDs that come up that you do not want to match, add one of the IDs to each column of the Excel document in order to prevent this match from appearing the next time you run the program

\*\*\*For any updates to the Excel Files, You can either input them yourself, or email [jihong.yao@doh.nj.gov](mailto:jihong.yao@doh.nj.gov) with what you would like to have changed

Person Merge

Names are the same without

‘-‘ or a space

NO

Address or Phone number the same

Any same names when first/last name are switched

Date of birth Is the same

Same first name (married)

Date of birth is the same

Name has “Jr” or “Sr” or a common nickname (ex. Jimmy -> James)

YES

Do Not merge People

More Research Required

Merge People

NO

NO

NO

NO

NO

YES

YES

YES

YES

YES

YES

NO

Is the date of birth Jan 1

NO

YES

Date of birth Similar (Same year, 1 digit off)

YES

NO

Tips for Merging

* Don’t merge two people unless 100% confident
* There are exceptions for every path above
* The number of lab tests can be telling for which name / DOB is correct
* Addresses and Phone numbers change often
* Use Wildcard and Soundex Search for more options
* Searches can be used to see past addresses and phone numbers
* Wildcard search with the first name in the last name spot can be used to find people who have changed their last name
* Pipl search and MyLife sometimes have date of birth records to help figure out which date of birth is the correct one
* The addresses may not match for each primary but may match when looking at the addresses in each case file