# Code to randomize birthdates, deathdates and service dates for CIBMTR Data Back to Center datasets

# version: 2.0

# date: August 22, 2013

# author: Paul K. Courtney

#

# The two source datasets were downloaded from the CIBMTR portal as Excel XLSX files,

# which were then saved as CSV files for import into R.

require(lubridate)

require(plyr)

# First, get the data read into dataframes; version 2 of the csv's indicates that the date columns

# have all been formatted using Excel in the form MM/DD/YYYY since I found that R interpreted a

# two year date of "56" as "2056".

PreTED = read.csv("PreTED2.csv")

PrePostTED = read.csv("PostTED2.csv")

# Read in the date column names for Pre and Post

preDates <- read.csv("PreDates.csv")

postDates <- read.csv("PostDates.csv")

# Then we need to fix this so it's not a dataframe but a vector:

preDates <- as.character(preDates$Pre)

postDates <- as.character(postDates$Post)

# Take away the bdate and deathdate and call them service dates:

preDatesService <- c(preDates[1:4], preDates[7:9])

postDatesService <- c(postDates[1:4], postDates[7:46])

# There are 6,777 observations in the PreTED dataset and 26,548 observations in the PostTED dataset.

# Need to find the unique crids for all observations in both datasets

mergedTED <- merge(PreTED, PostTED, by=c("crid","bdate"), all=FALSE)

uniqueCrids <- unique(mergedTED$crid)

# To change the format of date fields,

# newDates <- as.Date(PreTED$[datefield], format = "%m/%d/%y")

# but we need to replace the old with the new value in the data frame, so

# PreTED$[datefield] <- as.Date(PreTED$[datefield], format = "%m/%d/%y");

# this needs to be done over all of the date columns.

maxPre <- length(preDates)

for(i in 1:maxPre){

PreTED[[preDates[i]]] <- as.Date(PreTED[[preDates[i]]], format = "%m/%d/%Y")

}

maxPost <- length(postDates)

for(i in 1:maxPost){

PostTED[[postDates[i]]] <- as.Date(PostTED[[postDates[i]]], format = "%m/%d/%Y")

}

# We will want to associate the date offsets with the crid. The offsets will be randomly generated

# and will be different for bdate, deathdate and all the service dates will have the same offset.

# The seeds for the random number generator were obtained from the web site

# http://numbergenerator.org/random-6-digit-number-generator

# Now the offsets:

set.seed(296586)

rBirthDateOffset <- round(rnorm(6102,0,35)) # Approximately a range of +/- 4 months

set.seed(916952)

rDeathDateOffset <- round(rnorm(6102,0,35)) # Approximately a range of +/- 4 months

set.seed(653103)

rServiceDateOffset <- round(rnorm(6102,0,45)) # Approximately a range of +/- 6 months

# Create a data.frame with the unique crids and all of the new date offsets; change the

# name of the first column to just "crid" to make the merge easier

cridDateOffsets <- data.frame(uniqueCrids, rBirthDateOffset,rDeathDateOffset,rServiceDateOffset)

names(cridDateOffsets)[names(cridDateOffsets) == "uniqueCrids"] <- "crid"

# Now merge this with the PreTED and the PostTED data.frames into new data.frames.

# The "by" will ensure that each crid will be assigned the same offsets for multiple rows.

PreTEDOffset <- merge(PreTED, cridDateOffsets, by = "crid", all = TRUE)

PostTEDOffset <- merge(PostTED, cridDateOffsets, by = "crid", all = TRUE)

# Adjust the birth date

PreTEDOffset$bdate <- PreTEDOffset$bdate + days(PreTEDOffset$rBirthDateOffset)

PostTEDOffset$bdate <- PostTEDOffset$bdate + days(PostTEDOffset$rBirthDateOffset)

# Adjust the death date

PreTEDOffset$deathdate <- PreTEDOffset$deathdate + days(PreTEDOffset$rDeathDateOffset)

PostTEDOffset$deathdate <- PostTEDOffset$deathdate + days(PostTEDOffset$rDeathDateOffset)

# Now adjust the service dates

maxPreService <- length(preDatesService)

for(i in 1:maxPreService){

PreTEDOffset[[preDatesService[i]]] <- PreTEDOffset[[preDatesService[i]]] + days(PreTEDOffset$rServiceDateOffset)

}

maxPostService <- length(postDatesService)

for(i in 1:maxPostService){

PostTEDOffset[[postDatesService[i]]] <- PostTEDOffset[[postDatesService[i]]] + days(PostTEDOffset$rServiceDateOffset)

}

# Drop the date offset columns in the data.frames so that the original dates

# cannot be reconstructed:

PreTEDOffset$rBirthDateOffset <- NULL

PreTEDOffset$rDeathDateOffset <- NULL

PreTEDOffset$rServiceDateOffset <- NULL

PostTEDOffset$rBirthDateOffset <- NULL

PostTEDOffset$rDeathDateOffset <- NULL

PostTEDOffset$rServiceDateOffset <- NULL

# And finally, write the data.frames out as CSV files for use by vendors

write.csv(PreTEDOffset, file="PreTEDVendor.csv", row.names = FALSE, fileEncoding = "UTF-8")

write.csv(PostTEDOffset, file="PostTEDVendor.csv", row.names = FALSE, fileEncoding = "UTF-8")