eid_02_analysis.R

rsoren

Wed Aug 02 20:00:56 2017

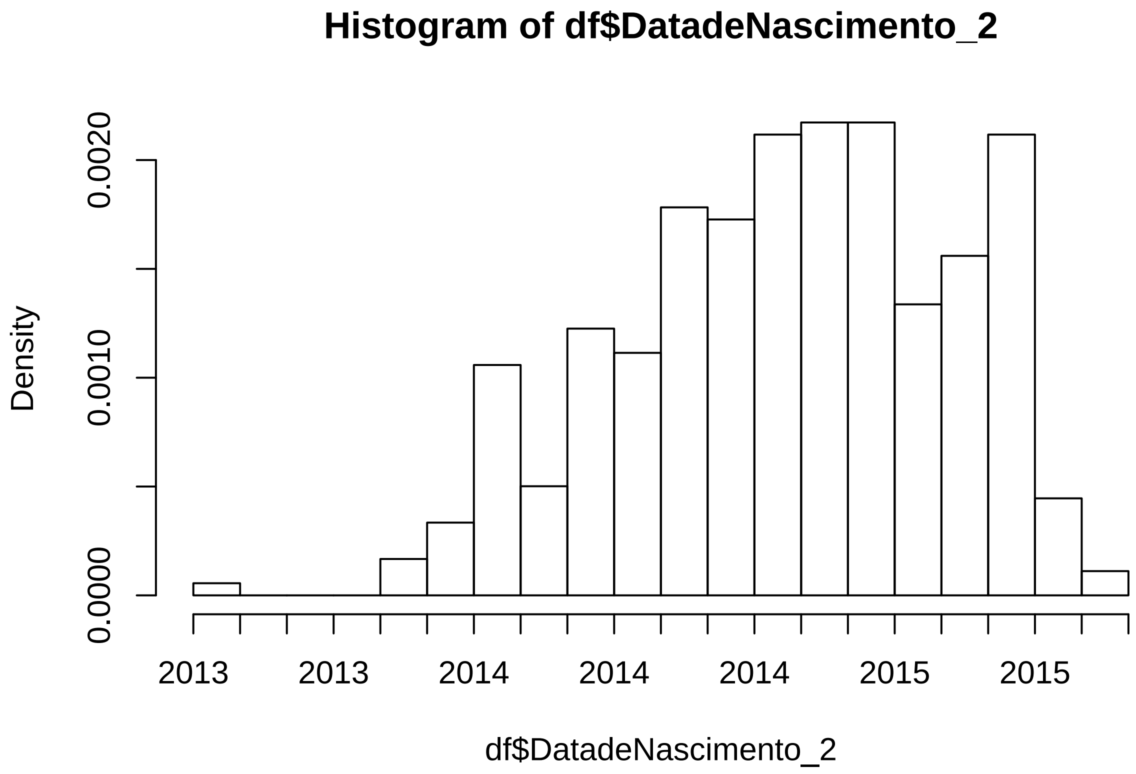
```
#
# eid_02_analysis.R
#
# Reed Sorensen
# August 2017
#

library(dplyr)
library(gmodels)
library(descr)

rm(list = ls())

dir <- "C:/Users/rsoren/Documents/prog/projects/201706_moz_research/"
df <- readRDS(paste0(dir, "_intermediate_files/DB_EID_v3.RDS"))


# distribution of dates of birth
hist(df$DatadeNascimento_2, breaks = 30)
```

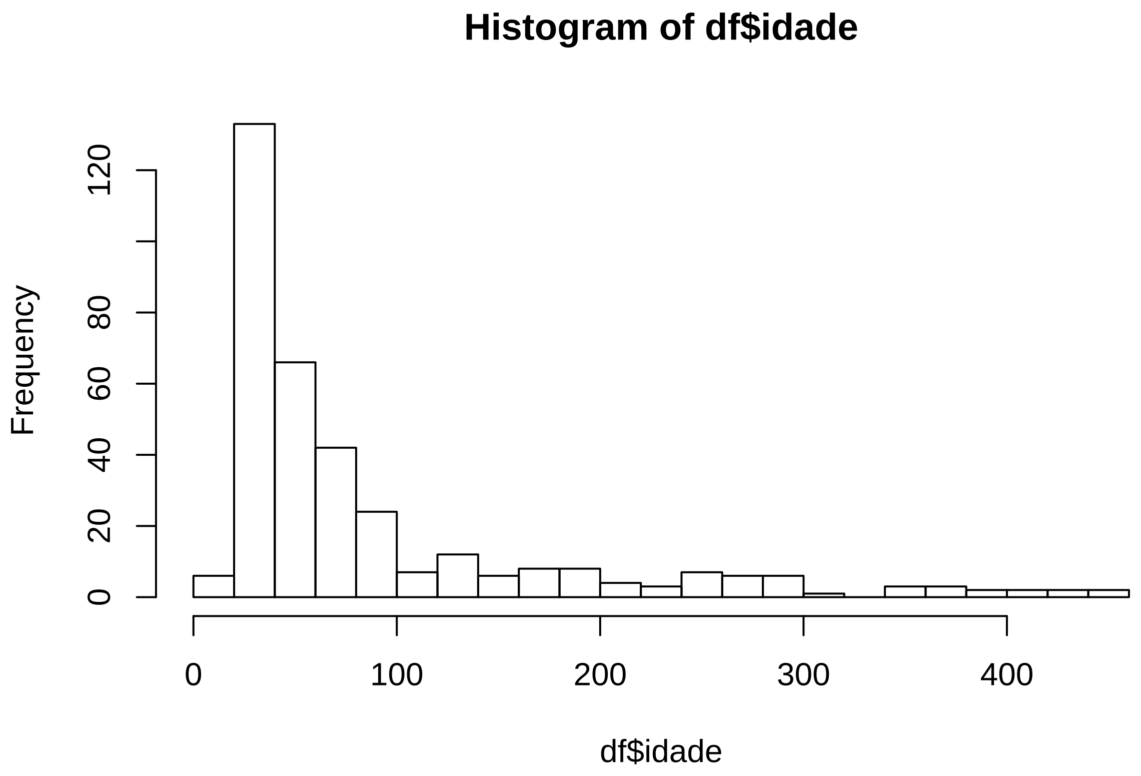


```
# Describe the sociodemographic profile of children attending EID services and tested for HIV in Maputo

# demographic variables:
# -- sex: Sexo
freq(df$Sexo, plot=F)
```

```
## df$Sexo
##       Frequency  Percent Valid Percent
## 0           188  45.8537         45.97
## 1           221  53.9024         54.03
## NA's          1   0.2439
## Total       410 100.0000        100.00
```

```
# -- age: Calculate as date of collection (colheitaUS1) minus date of birth (Data de Nacimento)
df$idade <- as.numeric(df$colheitaUS1_2 - df$DatadeNascimento_2)
hist(df$idade, breaks = 30)
```



```
# -- Above and below 6 weeks


# 49 entries didn't have a date of birth
with(df, crosstab(is.na(colheitaUS1_2), is.na(DatadeNascimento_2), plot=F))
```

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |-------------------------|
##
## ==============================================
##                          is.na(DatadeNascimento_2)
## is.na(colheitaUS1_2)    FALSE   TRUE   Total
## ----------------------------------------------
## FALSE                     353     49     402
## ----------------------------------------------
## TRUE                        6      2       8
## ----------------------------------------------
## Total                     359     51     410
## ==============================================
```

```
# Check why some colletion dates before the date of birth
# NOTE: fixed this; don't need to switch day and month from Esmeralda's file after all
#
# tmp1 <- subset(df, idade < 0,
#   select = c("DatadeNascimento", "colheitaUS1", "DatadeNascimento_2", "colheitaUS1_2"))
#


# -- provenance (health facility): Proveniencia
freq(df$Proveniencia, plot=F)
```

```
## df$Proveniencia
##                      Frequency  Percent
## CS 1  DE MAIO               24   5.8537
## CS 1 DE JUNHO               46  11.2195
## CS ALBASINE                 29   7.0732
## CS ALTO MAE                 13   3.1707
## CS BAGAMOYO                 24   5.8537
## CS CATEMBE                   4   0.9756
## CS CHAMANCULO               25   6.0976
## CS HULENE                    6   1.4634
## CS INCASSANE                 4   0.9756
## CS INHAGOIA                 18   4.3902
## CS JOSE MACAMO              31   7.5610
## CS MAGOANINE                 8   1.9512
## CS MAGOANINE TENDAS          9   2.1951
## CS MALHANGALENE             21   5.1220
## CS MAVALANE                 20   4.8780
## CS MAXAQUENE                 4   0.9756
## CS PESCADORES                5   1.2195
## CS POLANA CANICO            48  11.7073
## CS POLANA CIMENTO            8   1.9512
## CS PORTO                     7   1.7073
## CS ROMAO                    10   2.4390
## CS XIPAMANINE               30   7.3171
## CS ZIMPETO                  15   3.6585
## HMM                          1   0.2439
## Total                      410 100.0000
```

```
# Determine the proportion of mothers with known HIV status at delivery moment;
# No data for this at the moment
```

```r
# Determine the proportion of exposed children who took the first test;
# No data for this at the moment


# Determine the proportion of children tested for HIV using PCR DNA at 4 to 6 weeks at the
# EID services in Maputo city;
# Among all children, what proportion had a PCR test within 6 weeks
# -- Define no test as whether 'colheitaUS1' is missing, or if collection >6 weeks after birth
df2 <- df %>%
  mutate(
    test_after_6wk = ifelse(idade > 42, 1, 0),
    test_none = ifelse(is.na(colheitaUS1_2), 1, 0),
    test_within_6wk = ifelse(test_after_6wk | test_none, 1, 0) )

# with(df2, freq(test_after_6wk, plot=F))
# with(df2, freq(test_none, plot=F))
with(df2, freq(test_within_6wk, plot=F))
```

```
## test_within_6wk
##       Frequency Percent Valid Percent
## 0           153   37.32         42.38
## 1           208   50.73         57.62
## NA's         49   11.95
## Total       410  100.00        100.00
```

```r
# Determine the proportion of children with positive HIV who had a confirmatory test;
# Among children who took the first PCR test and were positive
#   (If ResultadoLab1 is 1, or if missing use ResultadoUS1),
# how many had a second test (how many not missing, ResultadoLAB2 | ResultadoUS2)


df3 <- df2 %>%
  mutate(
    first_test_result = ifelse(is.na(ResultadoLAB1), ResultadoUS1, ResultadoLAB1),
    first_test_positive = ifelse(first_test_result == 1, 1, 0),
    second_test_result = ifelse(is.na(ResultadoLAB2), ResultadoUS2, ResultadoLAB2),
    second_test_positive = ifelse(second_test_result == 1, 1, 0),
    had_second_test = ifelse(is.na(second_test_result), 0, 1) )

# how many missing first lab result
with(df3, table(is.na(ResultadoLAB1)) )
```

```
##
## FALSE  TRUE
##   281   129
```

```r
# among those with positive first test, how many took second test (133)
with(subset(df3, first_test_positive == 1),
  freq(had_second_test, plot=F)
)
```

```
## had_second_test
##       Frequency Percent
## 0           230   63.36
## 1           133   36.64
## Total       363  100.00
```

```r
# Determine the proportion of children with discordant results who had a third confirmatory test;

df4 <- df3 %>%
  mutate(
    discordant_posneg = ifelse(first_test_result == 1 & second_test_result == 0, 1, 0),
    discordant_negpos = ifelse(first_test_result == 0 & second_test_result == 1, 1, 0),
    is_discordant = ifelse(discordant_posneg | discordant_negpos, 1, 0),
    third_test_result = ifelse(is.na(ResultadoLAB3), ResultadoUS3, ResultadoLAB3),
    had_third_test = ifelse(is.na(third_test_result), 0, 1)
  )

table(df4$first_test_result, df4$second_test_result, exclude = NULL)
```

```
##
##           0   1   2 <NA>
##   0       0  17   0    3
##   1       6 127   0  230
##   <NA>    0  21   1    5
```

```r
with(df4, freq(is_discordant, plot=F))
```

```
## is_discordant
##       Frequency Percent Valid Percent
## 0           128   31.22         84.77
## 1            23    5.61         15.23
## NA's        259   63.17
## Total       410  100.00        100.00
```

```r
with(subset(df4, is_discordant == 1), freq(had_third_test, plot=F))
```

```
## had_third_test
##       Frequency Percent
## 0            21  91.304
## 1             2   8.696
## Total        23 100.000
```

```r
# Determine the proportion of children with the appropriate management
#   through the age of 18 months (as per the algorithm) who had a positive result;
#
# -- Make the time window smaller, try 9 months and 3 months
# -- Decide which threshold to use based on the remaining sample size
#
# Definition of appropriate management:
# 1. Positive PCR DNA test at 4-6 weeks
# 2. ART initiation [variable for this?] # DatadeiniciodoTARV
# 3a. Pos. confirmatory test --> Lifelong ART
# 3b. Pos. confirmatory test --> Neg. --> Pos. --> Lifelong ART
# 3c. Pos. confirmatory test --> Neg. --> Neg. --> Referral to clinician
#


# --Variables of interest:
# test_within_6wk
# first_test_positive
# had_second_test
# second_test_positive
# is_discordant
# had_third_test
# third_test_positive

# see "testing_cascade_logic.docs" for a visual representation
#   of the Boolean logic

df5 <- df4 %>%
  # remove children with negative first test, then no 2nd or 3rd test
  # -- must be an error, because there's no way to diagnose positive case
  filter(
    !(first_test_positive == 0 & had_second_test == 0 & had_third_test == 0) ) %>%
  mutate(
    c0 = as.integer(NA),
    # not compliant if didn't take test within 6 weeks
    c1 = ifelse(test_within_6wk == 0, 0, c0),

    # compliant if first test was negative, then took second or third tests
    c2 = ifelse(is.na(c1) & first_test_positive == 0 &
        (had_second_test | had_third_test), 1, c1),

    # not compliant if first test was negative, then didn't take second or third test
    c3 = ifelse(is.na(c2) & first_test_positive == 0 &
        !(had_second_test | had_third_test), 1, c2),

    # among people who took test within 6 weeks and had positive first test,
    #   not compliant if didn't start TARV
    c4 = ifelse(is.na(c3) & is.na(DatadeiniciodoTARV_2), 0, c3),

    # among people who took test within 6 weeks, had positive first test, and
    #   started TARV, not compliant if didn't have second test
    c5 = ifelse(is.na(c4) & had_second_test == 0, 0, c4),

    # among people on ART who took second test,
    # compliant if the second test is positive (end of algorithm)
    c6 = ifelse(is.na(c5) & second_test_positive == 1, 1, c5),

    # among people on ART who took second test and it was negative,
    # compliant if they had a third test
    c7 = ifelse(is.na(c6) & had_third_test == 1, 1, c6),

    # among people on ART who took second test and it was negative,
    # non-compliant if they didn't take a third test
    c8 = ifelse(is.na(c7) & had_third_test == 0, 1, c7)

  )

# lapply(paste0("c", 1:7), function(x) print(table(df5[, x])))

# overall frequency of following protocol
freq(df5$c8, plot=F)
```

```
## df5$c8
##       Frequency Percent
## 0           348   86.14
## 1            56   13.86
## Total       404  100.00
```

```
# split by provenance (health facility)
df6 <- df5 %>%
  group_by(Proveniencia) %>%
  summarize(
    count = sum(c8),
    total = n()) %>%
  mutate(proportion = count / total)




# Identify the factors relating to the non-compliance with the
#   algorithm for EID for PCR DNA HIV first positive test.
# -- Logistic regression

# What are the factors we want to use to predict non-compliance?
# -- Waiting on data about maternal age and site of delivery
# -- Child age at enrollment in CCR (same thing as age at first collection)
# -- Time it takes for the laboratory to get a result back to the health facility


# Among people who took test within 6 weeks and
#   had positive first test [designated as is.na(c3)],
# is age at child's first test associated with with whether they started TARV?

dat_fit1 <- df5 %>%
  filter(is.na(c3)) %>%
  mutate(started_tarv = ifelse(is.na(DatadeiniciodoTARV_2), 0, 1) )

fit1 <- glm(
  formula = started_tarv ~ idade,
  data = dat_fit1,
  family = binomial(link = "logit")
)

summary(fit1)
```

```
##
## Call:
## glm(formula = started_tarv ~ idade, family = binomial(link = "logit"),
##     data = dat_fit1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0294  -1.0115  -0.9227   1.3441   1.5687
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.301847   0.242501  -1.245    0.213
## idade       -0.001322   0.001486  -0.890    0.374
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 256.97  on 192  degrees of freedom
## Residual deviance: 256.16  on 191  degrees of freedom
##   (49 observations deleted due to missingness)
## AIC: 260.16
##
## Number of Fisher Scoring iterations: 4
```

```
# Among people who took test within 6 weeks and
#   had positive first test and started TARV [designated as is.na(c4)],
# is age at child's first test associated with with whether they had second test?

dat_fit2 <- subset(df5, is.na(c4))

fit2 <- glm(
  formula = had_second_test ~ idade,
  data = dat_fit2,
  family = binomial(link = "logit")
)

summary(fit2)
```

```
##
## Call:
## glm(formula = had_second_test ~ idade, family = binomial(link = "logit"),
##     data = dat_fit2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3908  -1.2926   0.9669   1.0062   1.5241
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.744887   0.383766   1.941   0.0523 .
## idade       -0.005225   0.002576  -2.028   0.0426 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 102.369  on 73  degrees of freedom
## Residual deviance:  97.563  on 72  degrees of freedom
##   (17 observations deleted due to missingness)
## AIC: 101.56
##
## Number of Fisher Scoring iterations: 4
```

```
# Measures of non-compliance
# [what are the variables for these?]
# -- Whether or not mother comes to take the result
# -- Missed appointment
```

```r
# Hypothesis: Child age at enrollment influences
#   whether the mother comes to take the result


# Among people with discordant results for the
# first two tests, what factors influence
# whether they got a third test?
# -- Not enough data to answer this (n=23, with 2 people getting third test)

# hypothesis: taking more time for the result to return to the health facility
# --> non-compliance [what kind; what variable?]


# Motive de nao TARV
# 1 = abandono
# 2 = nao encaminhado para o TARV
# 3 = transferido
# 4 = recusa
# 5 = sem proceso TARV
# 6 = obito
# 7 = endereco / contacto falso
# 8 = nao inicio
# 9 = sem resultado
# 10 = sem ficha de no proceso TARV
# 11 = dados nao conferem (not the same)

# Report as: This number of children don't start TARV
#            If not start TARV, frequency giving each reason

# FALSE = did not start TARV
# TRUE = started TARV
table(!is.na(df5$DatadeiniciodoTARV_2))
```

```
##
## FALSE  TRUE
##   257   147
```

```r
# among people who didn't start TARV, why not?
with(subset(df5, is.na(DatadeiniciodoTARV_2)),
  table(MotivodenaoTARV)
)
```

```
## MotivodenaoTARV
##  1  2  3  4  5  6  7  8  9 10
##  3 43 21  2  1 13  7 18 31  3
```

```r
# save as R Markdown document
# library("rmarkdown")
# setwd(paste0(dir, "EID_HIV/"))
#
# rmarkdown::render(
#   input = "eid_02_analysis.R",
#   output_format = "pdf_document"
# )
#
```

```
# setwd(dir)
```