eid_02_analysis.R

rsoren

Thu Aug 17 16:39:11 2017

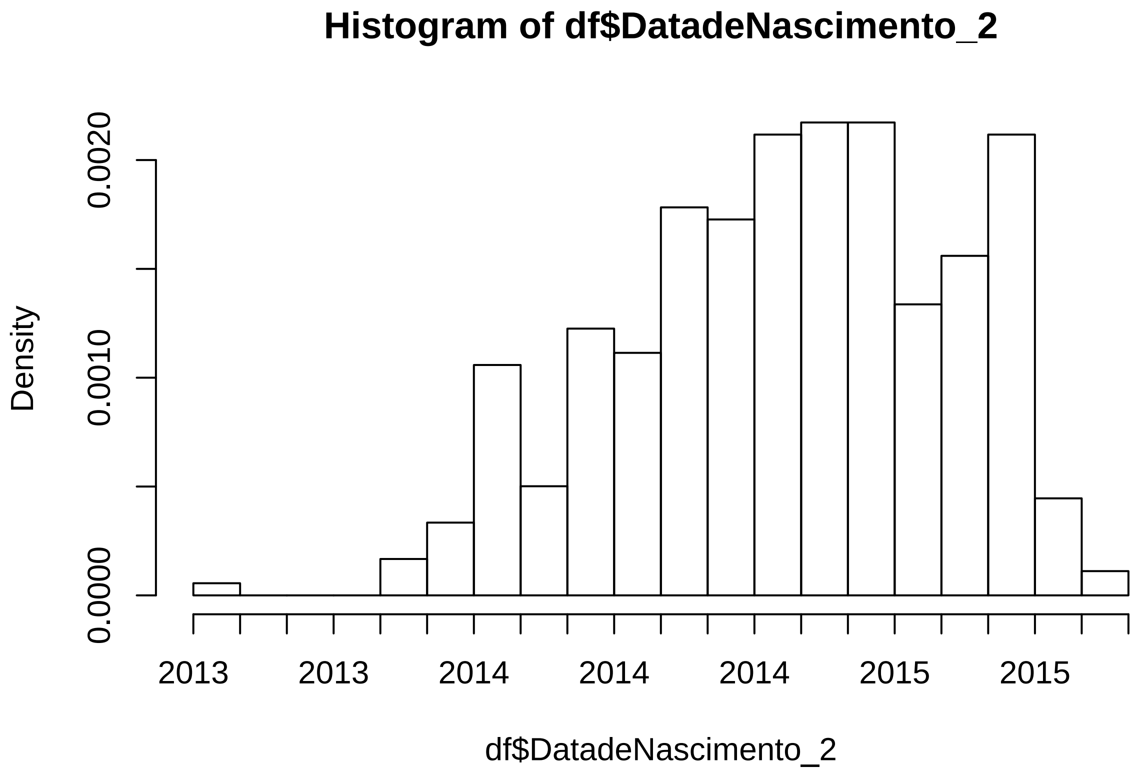
```
#
# eid_02_analysis.R
#
# Reed Sorensen
# August 2017
#

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)
```

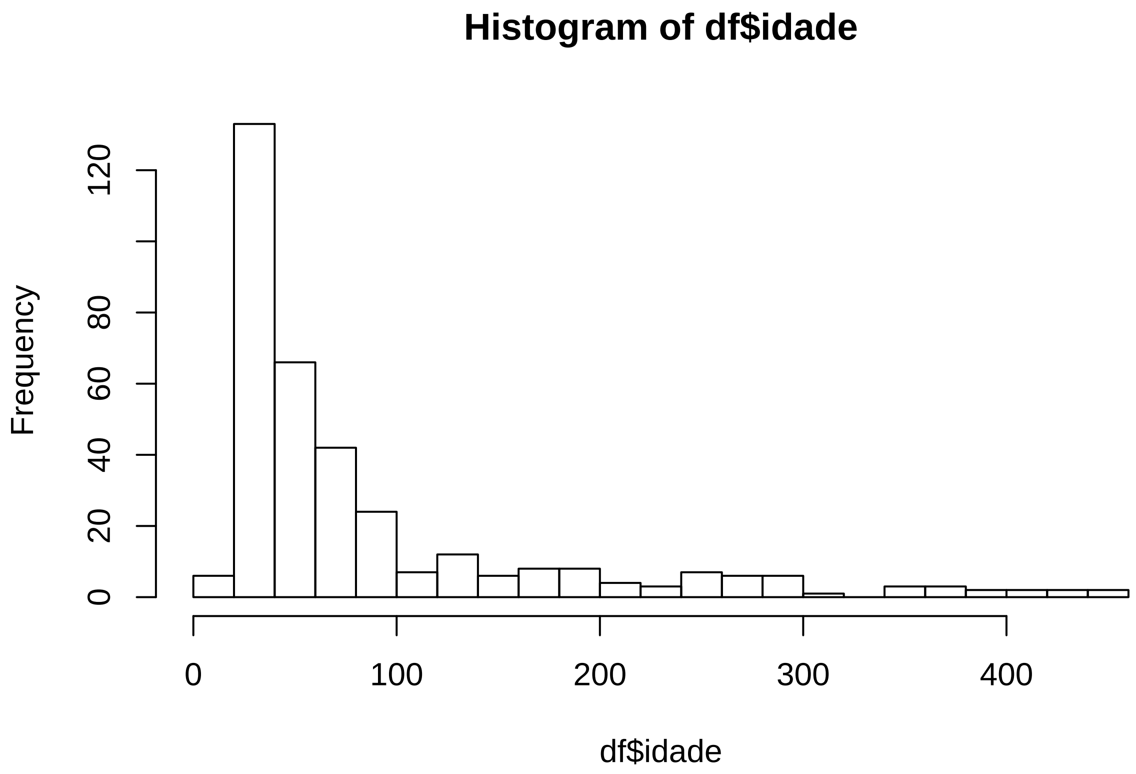


```
# Describe the sociodemographic profile of children attending EID services and tested for HIV in Maputo

# 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

# distribution of age at collection, by age group
# 0-6 weeks; 7-12 weeks; 13-18 weeks; 19-24 weeks; 25-30 weeks; 31-36 weeks; 37 +
# Note: I changed the cutoffs slightly, so the first number increases by 6 each time

df$idade_semana <- df$idade / 7
df$idade_semana_grupo <- cut(df$idade_semana, breaks = c(0,6,12,18,24,30,36,999), right = FALSE)
freq(df$idade_semana_grupo, plot=F)
```

```
## df$idade_semana_grupo
##          Frequency Percent Valid Percent
## [0,6)          147  35.854        41.643
## [6,12)         102  24.878        28.895
## [12,18)         35   8.537         9.915
## [18,24)         16   3.902         4.533
## [24,30)         12   2.927         3.399
## [30,36)         11   2.683         3.116
## [36,999)        30   7.317         8.499
## NA's            57  13.902
## Total          410 100.000       100.000
```

```
# 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


# 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
```

```
# 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
```

```
# 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
```

```
# 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
```

```
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
```

```
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
```

```
# Determine the proportion of children with the appropriate management
#   through the age of 18 months (as per the algorithm) who had a positive result;
#
# 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

df4_tmp <- df4 %>%
  mutate(exclude_if_false = first_test_positive == 0 & had_second_test == 0 & had_third_test == 0) %>%
  select(Proveniencia, ResultadoLAB1, ResultadoUS1, first_test_positive,
    had_second_test, had_third_test, exclude_if_false)

with(df4_tmp, table(exclude_if_false, exclude = ""))
```

```
## exclude_if_false
## FALSE  TRUE  <NA>
##   404     2     4
```

```
write.csv(df4_tmp, paste0(dir, "EID_HIV/check_exclusions.csv", row.names = FALSE))

  # filter(first_test_positive == 0 & had_second_test == 0 & had_third_test == 0)

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) ) %>%
  mutate(
    started_tarv = ifelse(is.na(DatadeiniciodoTARV_2), 0, 1),
    ProvUS_factor = factor(ProvUS),
    turnaround_time_1 = as.numeric(processamento1_2 - colheitaUS1_2),
    turnaround_time_2 = as.numeric(processamento2_2 - colheitaUS2_2),
    tt_30 = turnaround_time_1 >= 30,
    tt_30_2 = turnaround_time_2 >= 30
  )


# write.csv(df5, paste0(dir, "EID_HIV/eid_hiv_processed_data.csv"), row.names = FALSE)
```

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


# get results for selected variables by proveniencia

results_by_prov <- df5 %>%
  group_by(Proveniencia) %>%
  dplyr::summarize(
    appropriate_management = sum(c8, na.rm=T),
    started_tarv = sum(started_tarv, na.rm=T),
    median_age_days = median(idade, na.rm=T),
    first_test_positive = sum(first_test_result == 1, na.rm=T),
    first_test_negative = sum(first_test_result == 0, na.rm=T),
    had_second_test = sum(had_second_test, na.rm=T),
    second_test_positive = sum(second_test_result == 1, na.rm=T),
    second_test_negative = sum(second_test_result == 0, na.rm=T),
    had_third_test = sum(had_third_test, na.rm=T),
    third_test_positive = sum(third_test_result == 1, na.rm=T),
    third_test_negative = sum(third_test_result == 0, na.rm=T),
    facility_receive_1 = sum(!is.na(DataderecepcaonaCCR1), na.rm=T),
    facility_receive_2 = sum(!is.na(DataderecepcaonaCCR2), na.rm=T),
    facility_receive_3 = sum(!is.na(DataderecepcaonaCCR3), na.rm=T),
    # not sure if these 'mother_receive' variables measure the same thing,
    #   but there's no alternative
    # Note that only 1 and 3 have "oucuidador" in the variable name
    mother_receive_1 = sum(!is.na(Datadeentregaamaeoucuidador1_2), na.rm=T),
    mother_receive_2 = sum(!is.na(Datadeentregaamae2_2), na.rm=T),
    mother_receive_3 = sum(!is.na(Datadeentregaamaeoucuidador3_2), na.rm=T),
    total = n() )


pct_names <- names(results_by_prov)[!names(results_by_prov) %in% c("Proveniencia", "total")]

for (nm in pct_names) {
  results_by_prov[, paste0("pct_", nm)] <-
    results_by_prov[, nm] / results_by_prov$total
}

write.csv(results_by_prov, paste0(dir, "EID_HIV/results_by_prov.csv"))

#####
# Bivariate analysis, comparing how many children had appropriate management, versus:
# - how many children had test within 6 weeks
crosstab(df5$c8, df5$test_within_6wk, plot=F, chisq = TRUE)
```

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |-------------------------|
##
## ===========================
##           df5$test_within_6wk
## df5$c8      0     1   Total
## ---------------------------
## 0         152   158     310
## ---------------------------
## 1           0    48      48
## ---------------------------
## Total     152   206     358
## ===========================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 40.90147      d.f. = 1      p = 1.6e-10
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 = 38.91914      d.f. = 1      p = 4.42e-10
##         Minimum expected frequency: 20.37989
```

```
# - how many started TARV
crosstab(df5$c8, df5$started_tarv, plot=F, chisq = TRUE)
```

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |-------------------------|
##
## ===========================
##           df5$started_tarv
## df5$c8      0     1   Total
## ---------------------------
## 0          254    94     348
## ---------------------------
## 1            3    53      56
## ---------------------------
## Total      257   147     404
## ===========================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 95.32229      d.f. = 1      p <2e-16
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 = 92.42281      d.f. = 1      p <2e-16
##         Minimum expected frequency: 20.37624
```

```
# - how many mothers comes to take the first result
crosstab(df5$c8, is.na(df5$Datadeentregaamaeoucuidador1), plot=F, chisq = TRUE)
```

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |-------------------------|
##
## ==============================
##           is.na(df5$Datadeentregaamaeoucuidador1)
## df5$c8    FALSE   TRUE   Total
## ------------------------------
## 0           152    196     348
## ------------------------------
## 1            39     17      56
## ------------------------------
## Total       191    213     404
## ==============================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 13.04674      d.f. = 1      p = 0.000304
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 = 12.02586      d.f. = 1      p = 0.000525
##         Minimum expected frequency: 26.47525
```

```
# - how many have turn-around time before 30 days, for each test
crosstab(df5$c8, df5$tt_30, plot=F, chisq = TRUE)
```

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |-------------------------|
##
## ==============================
##           df5$tt_30
## df5$c8    FALSE   TRUE   Total
## ------------------------------
## 0           107    133     240
## ------------------------------
## 1            12     11      23
## ------------------------------
## Total       119    144     263
## ==============================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 0.4881297      d.f. = 1      p = 0.485
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 = 0.2298174      d.f. = 1      p = 0.632
##         Minimum expected frequency: 10.40684
```

```
# overall frequency of following protocol ('appropriate management')
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) %>%
  dplyr::summarize(
    count = sum(c8),
    total = n()) %>%
  mutate(proportion = count / total) %>%
  as.data.frame(.)

print(df6)
```

```
##           Proveniencia count total proportion
## 1        CS 1  DE MAIO     3    24 0.12500000
## 2        CS 1 DE JUNHO     8    46 0.17391304
## 3          CS ALBASINE     3    29 0.10344828
## 4          CS ALTO MAE     2    13 0.15384615
## 5          CS BAGAMOYO     3    24 0.12500000
## 6           CS CATEMBE     0     4 0.00000000
## 7        CS CHAMANCULO     4    25 0.16000000
## 8            CS HULENE     1     6 0.16666667
## 9         CS INCASSANE     0     4 0.00000000
## 10         CS INHAGOIA     2    18 0.11111111
## 11      CS JOSE MACAMO     7    31 0.22580645
## 12        CS MAGOANINE     2     8 0.25000000
## 13 CS MAGOANINE TENDAS     2     9 0.22222222
## 14     CS MALHANGALENE     1    19 0.05263158
## 15         CS MAVALANE     1    19 0.05263158
## 16        CS MAXAQUENE     1     4 0.25000000
## 17       CS PESCADORES     0     5 0.00000000
## 18    CS POLANA CANICO     3    46 0.06521739
## 19   CS POLANA CIMENTO     0     8 0.00000000
## 20            CS PORTO     3     7 0.42857143
## 21            CS ROMAO     2    10 0.20000000
## 22       CS XIPAMANINE     6    29 0.20689655
## 23          CS ZIMPETO     1    15 0.06666667
## 24                 HMM     1     1 1.00000000
```

```
# 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


# function for getting odds ratio and CI from regression results

get_ci <- function(model, variable, exponentiate = FALSE) {

  # model <- fit1; variable <- "idade"; exponentiate = TRUE # dev variables

  txt <- paste0("Beta coefficient of '", variable, "': ")

  output <- list(
    tmp_point <- coef(model)[variable],
    tmp_confint <- confint(model)[variable,]
  )

  if (exponentiate) {
    txt <- paste0("Exponentiated '", variable, "': ")
    output <- lapply(output, exp)
  }

  output <- lapply(output, function(x) round(x, digits = 3))
  cat("\n", paste0(
    txt, output[[1]], " (95% CI: ", paste0(output[[2]], collapse = ", "), ")"
  ))
}


# 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))

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
```

```
get_ci(model = fit1, variable = "idade", exponentiate = TRUE)
```

```
## Waiting for profiling to be done...
```

```
##
##  Exponentiated 'idade': 0.999 (95% CI: 0.996, 1.002)
```

```
# check for evidence of clustering
# rho=0 means no clustering; rho=1 means complete clustering

library("Hmisc")
deff(as.logical(dat_fit1$started_tarv), cluster = dat_fit1$Proveniencia)
```

```
##            n     clusters          rho         deff
## 242.00000000  24.00000000   0.07988521   2.29202772
```

```
# random effects model that accounts for clustering
library(lme4)

dat_fit1_re <- dat_fit1 %>%
  # have to drop facilities with low numbers, otherwise this model doesn't converge
  group_by(Proveniencia) %>%
  filter(n() >= 10) %>% # keep only facilities with n >= 10; lowest number that still converges
  as.data.frame(.) %>%
  mutate(Proveniencia = droplevels(Proveniencia))

table(dat_fit1_re$Proveniencia)
```

```
## 
##    CS 1  DE MAIO    CS 1 DE JUNHO      CS ALBASINE      CS BAGAMOYO 
##               14               33               12               16 
##    CS CHAMANCULO      CS INHAGOIA   CS JOSE MACAMO      CS MAVALANE 
##               16               11               16               11 
## CS POLANA CANICO    CS XIPAMANINE 
##               31               20
```

```
fit1_re <- glmer(
  formula = started_tarv ~ idade + (1 | Proveniencia),
  data = dat_fit1_re,
  family = binomial(link = "logit")
)

summary(fit1_re)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: started_tarv ~ idade + (1 | Proveniencia)
##    Data: dat_fit1_re
## 
##      AIC      BIC   logLik deviance df.resid 
##    177.9    186.7    -85.9    171.9      135 
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -1.0879 -0.7371 -0.5285  0.9915  2.1036 
## 
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  Proveniencia (Intercept) 0.3388   0.5821  
## Number of obs: 138, groups:  Proveniencia, 10
## 
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.173660   0.372684  -0.466    0.641
## idade       -0.003561   0.002165  -1.644    0.100
## 
## Correlation of Fixed Effects:
##       (Intr)
## idade -0.688
```

```r
ranef1 <- ranef(fit1_re)[[1]][,1]


# 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))

# check for evidence of clustering in having a second test (no clustering)
deff(as.logical(dat_fit2$had_second_test), cluster = dat_fit2$Proveniencia)
```

```
##           n    clusters         rho        deff
## 91.00000000 22.00000000 -0.02300661  0.87257877
```

```r
table(dat_fit2$Proveniencia)
```

```
## 
##       CS 1  DE MAIO       CS 1 DE JUNHO         CS ALBASINE 
##                   5                  10                   4 
##         CS ALTO MAE         CS BAGAMOYO          CS CATEMBE 
##                   7                   9                   1 
##       CS CHAMANCULO           CS HULENE        CS INCASSANE 
##                   9                   1                   0 
##         CS INHAGOIA      CS JOSE MACAMO        CS MAGOANINE 
##                   2                   6                   3 
## CS MAGOANINE TENDAS     CS MALHANGALENE         CS MAVALANE 
##                   2                   2                   1 
##        CS MAXAQUENE       CS PESCADORES    CS POLANA CANICO 
##                   1                   1                   4 
##   CS POLANA CIMENTO            CS PORTO            CS ROMAO 
##                   0                   3                   3 
##       CS XIPAMANINE          CS ZIMPETO                 HMM 
##                  11                   5                   1
```

```r
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
```

```
get_ci(model = fit2, variable = "idade", exponentiate = TRUE)
```

```
## Waiting for profiling to be done...
```

```
##
##  Exponentiated 'idade': 0.995 (95% CI: 0.989, 0.999)
```

```
# Data for turnaround time analyses

dat_fit3 <- df5

# check for clustering in turnaround time (not enough to need accounting for it)
table(dat_fit3$Proveniencia)
```

```
##
##       CS 1  DE MAIO       CS 1 DE JUNHO         CS ALBASINE
##                  24                  46                  29
##         CS ALTO MAE         CS BAGAMOYO          CS CATEMBE
##                  13                  24                   4
##      CS CHAMANCULO           CS HULENE        CS INCASSANE
##                  25                   6                   4
##        CS INHAGOIA      CS JOSE MACAMO        CS MAGOANINE
##                  18                  31                   8
## CS MAGOANINE TENDAS     CS MALHANGALENE         CS MAVALANE
##                   9                  19                  19
##       CS MAXAQUENE       CS PESCADORES    CS POLANA CANICO
##                   4                   5                  46
##   CS POLANA CIMENTO            CS PORTO            CS ROMAO
##                   8                   7                  10
##      CS XIPAMANINE          CS ZIMPETO                 HMM
##                  29                  15                   1
```

```
deff(dat_fit3$turnaround_time_1, cluster = dat_fit3$Proveniencia)
```

```
##            n     clusters          rho         deff
## 263.00000000  24.00000000   0.03069713   1.47644751
```

```
deff(dat_fit3$turnaround_time_2, cluster = dat_fit3$Proveniencia)
```

```
##             n      clusters           rho          deff
## 142.000000000  22.000000000  -0.002676309   0.972106073
```

```
# Among people who had a FIRST sample taken,
# what is the effect of faster turnaround time ('processamento1_2' minus 'colheitaUS1_2')
```

```r
# check for clustering in appropriate management (rho=0.0051)
deff(as.logical(dat_fit3$c8), cluster = dat_fit3$Proveniencia)
```

```
##            n     clusters          rho         deff
## 4.040000e+02 2.400000e+01 5.069598e-03 1.127443e+00
```

```r
# outcome: appropriate management
# -- predictor: turnaround time at least 30 days for first test
fit3 <- glm(
  formula = c8 ~ tt_30,
  data = dat_fit3,
  family = binomial(link = "logit")
)

summary(fit3)
```

```
##
## Call:
## glm(formula = c8 ~ tt_30, family = binomial(link = "logit"),
##     data = dat_fit3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4611  -0.4611  -0.3987  -0.3987   2.2680
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1879     0.3044  -7.187 6.63e-13 ***
## tt_30TRUE    -0.3045     0.4372  -0.697    0.486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 156.01  on 262  degrees of freedom
## Residual deviance: 155.53  on 261  degrees of freedom
##   (141 observations deleted due to missingness)
## AIC: 159.53
##
## Number of Fisher Scoring iterations: 5
```

```r
get_ci(model = fit3, variable = "tt_30TRUE", exponentiate = TRUE)
```

```
## Waiting for profiling to be done...
```

```
##
##  Exponentiated 'tt_30TRUE': 0.737 (95% CI: 0.308, 1.747)
```

```r
# -- predictor: turnaround time (continuous) for first test

fit4 <- glm(
  formula = c8 ~ turnaround_time_1,
  data = dat_fit3,
  family = binomial(link = "logit")
)
summary(fit4)
```

```
##
## Call:
## glm(formula = c8 ~ turnaround_time_1, family = binomial(link = "logit"),
##     data = dat_fit3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4799  -0.4564  -0.4379  -0.3771   2.5271
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.071229   0.335554  -6.173 6.72e-10 ***
## turnaround_time_1 -0.006428   0.006532  -0.984    0.325
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 156.01  on 262  degrees of freedom
## Residual deviance: 154.87  on 261  degrees of freedom
##   (141 observations deleted due to missingness)
## AIC: 158.87
##
## Number of Fisher Scoring iterations: 5
```

```
get_ci(model = fit4, variable = "turnaround_time_1", exponentiate = TRUE)
```

```
## Waiting for profiling to be done...
```

```
##
##  Exponentiated 'turnaround_time_1': 0.994 (95% CI: 0.979, 1.005)
```

```
# outcome: started TARV
# -- predictor: turnaround time at least 30 days

fit5 <- glm(
  formula = started_tarv ~ tt_30,
  data = dat_fit3,
  family = binomial(link = "logit")
)

summary(fit5)
```

```
##
## Call:
## glm(formula = started_tarv ~ tt_30, family = binomial(link = "logit"),
##     data = dat_fit3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9331  -0.9331  -0.8657   1.4432   1.5252
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.6061     0.1918  -3.160  0.00158 **
## tt_30TRUE    -0.1823     0.2629  -0.693  0.48801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 333.87  on 262  degrees of freedom
## Residual deviance: 333.39  on 261  degrees of freedom
##   (141 observations deleted due to missingness)
## AIC: 337.39
##
## Number of Fisher Scoring iterations: 4
```

```
get_ci(model = fit5, variable = "tt_30TRUE", exponentiate = TRUE)
```

```
## Waiting for profiling to be done...
```

```
##
##  Exponentiated 'tt_30TRUE': 0.833 (95% CI: 0.497, 1.397)
```

```
# -- predictor: turnaround time (continuous)

fit6 <- glm(
  formula = started_tarv ~ turnaround_time_1,
  data = dat_fit3,
  family = binomial(link = "logit")
)

summary(fit6)
```

```
##
## Call:
## glm(formula = started_tarv ~ turnaround_time_1, family = binomial(link = "logit"),
##     data = dat_fit3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9845  -0.9352  -0.8452   1.4107   1.9187
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.443066   0.200251  -2.213   0.0269 *
## turnaround_time_1 -0.005832   0.003524  -1.655   0.0980 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 333.87  on 262  degrees of freedom
## Residual deviance: 330.85  on 261  degrees of freedom
##   (141 observations deleted due to missingness)
## AIC: 334.85
##
## Number of Fisher Scoring iterations: 4
```

```r
get_ci(model = fit6, variable = "turnaround_time_1", exponentiate = TRUE)
```

```
## Waiting for profiling to be done...
```

```
##
##  Exponentiated 'turnaround_time_1': 0.994 (95% CI: 0.987, 1.001)
```

```r
# outcome: appropriate management
# -- predictor: turnaround time at least 30 days for second test
fit7 <- glm(
  formula = c8 ~ tt_30_2,
  data = dat_fit3,
  family = binomial(link = "logit")
)

summary(fit7)
```

```
##
## Call:
## glm(formula = c8 ~ tt_30_2, family = binomial(link = "logit"),
##     data = dat_fit3)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.016  -1.016  -0.714   1.348   1.727
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3920     0.2589  -1.514   0.1300
## tt_30_2TRUE  -0.8447     0.3724  -2.268   0.0233 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 174.16  on 141  degrees of freedom
## Residual deviance: 168.92  on 140  degrees of freedom
##   (262 observations deleted due to missingness)
## AIC: 172.92
##
## Number of Fisher Scoring iterations: 4
```

```r
get_ci(model = fit7, variable = "tt_30_2TRUE", exponentiate = TRUE)
```

```
## Waiting for profiling to be done...
```

```
##
##  Exponentiated 'tt_30_2TRUE': 0.43 (95% CI: 0.205, 0.886)
```

```r
# -- predictor: turnaround time (continuous) for second test

fit8 <- glm(
  formula = c8 ~ turnaround_time_2,
  data = dat_fit3,
  family = binomial(link = "logit")
)
summary(fit8)
```

```
##
## Call:
## glm(formula = c8 ~ turnaround_time_2, family = binomial(link = "logit"),
##     data = dat_fit3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9281  -0.8762  -0.8418   1.4663   1.8131
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.574588   0.292714  -1.963   0.0496 *
## turnaround_time_2 -0.006375   0.005855  -1.089   0.2762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 174.16  on 141  degrees of freedom
## Residual deviance: 172.86  on 140  degrees of freedom
##   (262 observations deleted due to missingness)
## AIC: 176.86
##
## Number of Fisher Scoring iterations: 4
```

```
get_ci(model = fit8, variable = "turnaround_time_2", exponentiate = TRUE)
```

```
## Waiting for profiling to be done...
```

```
##
##  Exponentiated 'turnaround_time_2': 0.994 (95% CI: 0.981, 1.004)
```

```
# interesting that dichotomous <>30 days is significant, but
#   a continuous measure of time is not
# -- checking a 'generalized additive model' (GAM) to visualize
#    the non-linear effect of turnaround time

library(mgcv)
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:lme4':
##
##     lmList
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

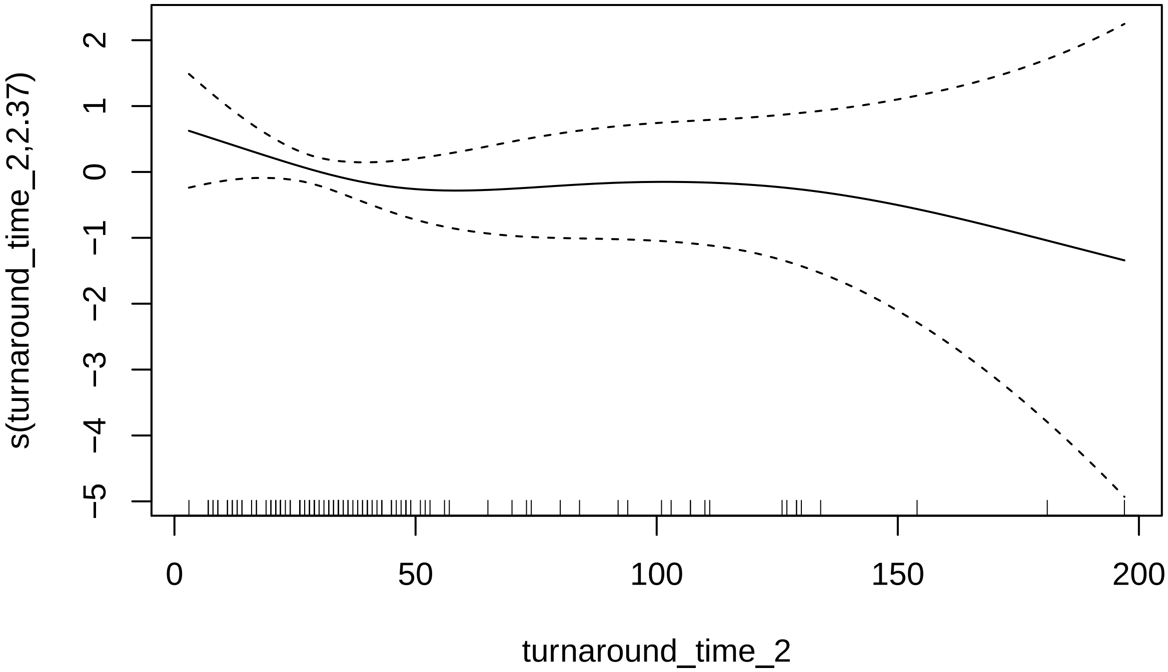
```
## This is mgcv 1.8-17. For overview type 'help("mgcv-package")'.
```

```
fit8_gam <- gam(
  formula = c8 ~ s(turnaround_time_2),
  data = dat_fit3,
  family = binomial(link = "logit")
)

summary(fit8_gam)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## c8 ~ s(turnaround_time_2)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8491     0.1851  -4.588 4.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                        edf Ref.df Chi.sq p-value
## s(turnaround_time_2) 2.37  2.996  3.046   0.394
##
## R-sq.(adj) =  0.0113   Deviance explained = 2.36%
## UBRE = 0.24493  Scale est. = 1         n = 142
```

```
plot(fit8_gam)
```



```
# 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)


# 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

with(subset(df5, started_tarv != 1), freq(MotivodenaoTARV, plot=F))
```

```
## MotivodenaoTARV
##       Frequency  Percent Valid Percent
## 1             3   1.1673        2.1127
## 2            43  16.7315       30.2817
## 3            21   8.1712       14.7887
## 4             2   0.7782        1.4085
## 5             1   0.3891        0.7042
## 6            13   5.0584        9.1549
## 7             7   2.7237        4.9296
## 8            18   7.0039       12.6761
## 9            31  12.0623       21.8310
## 10            3   1.1673        2.1127
## NA's        115  44.7471
## Total       257 100.0000      100.0000
```

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

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

```
# 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
```

```
# 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)
```