WA State HIV Testing Histories - Data Exploration and Formatting

Martina Morris and Jeanette Birnbaum

July 2, 2015

1 Data Structure

```
str(dataf)

## 'data.frame': 26134 obs. of  20 variables:
##  $ FirstVL          : num  658 19914 NA 51 9050 ...
##  $ FirstCD4cnt      : num  566 243 1406 711 858 ...
##  $ tth_ever_neg     : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ new_race         : int  2 2 1 1 1 1 3 1 1 1 ...
##  $ hst              : chr  "WA" "WA" "WA" "WA" ...
##  $ hdx_age          : int  51 25 41 34 38 33 33 41 45 19 ...
##  $ new_mode         : int  3 6 8 1 1 1 3 1 1 1 ...
##  $ TTH_lneg_DT_FLAG : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ tth_ppos_dt_flag : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ est_infect_period: int  3 3 3 3 3 3 3 3 3 3 ...
##  $ hdx_yr_qtr       : chr  "1998_3Q" "1999_3Q" "1995_2Q" "1990_" ...
##  $ HDX_DT_FLAG      : chr  "M" "M" "M" "Y" ...
##  $ adx_yr_qtr       : chr  "2003_2Q" "2000_1Q" NA NA ...
##  $ adx_DT_FLAG      : chr  "M" "M" NA NA ...
##  $ LAG_LNEG_HDX_DT  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LAG_PPOS_HDX_DT  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TTH_PREV_POS     : chr  "N" "N" "N" "N" ...
##  $ VL_DAYS          : int  181 111 NA 4032 30 3061 2618 1810 0 4461 ...
##  $ CD4_DAYS         : int  122 122 1553 3271 683 1765 30 1218 304 3195 ...
##  $ METH_USE         : chr  NA NA NA NA ...
```

2 Overview

- N = 26134

3 Raw Variable Summaries

```
##
##
##
## VARIABLE 1 : firstvl
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     623   14050   36710   87590  100000    7945
##
##      Percent missing:[1] 30.4
##
```

```
##
##
## VARIABLE 2 : firstcd4cnt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   109.0   258.0   331.7   494.0  6745.0    2830
##
##      Percent missing:[1] 10.83
##
##
##
## VARIABLE 3 : tth_ever_neg
##      var
##     1     2     3     4     5  <NA>
##  2688   605     6   505 22330     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 4 : new_race
##      var
##   White   Black    Hisp   Asian   NHoPI   AI/AN   Multi
##   18142    3631    2716     602     100     357     573
## Unknown    <NA>
##      13       0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 5 : hst
##      var
##    OO    WA  <NA>
##  5953 20180     1
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 6 : hdx_age
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    28.0    34.0    35.1    41.0    91.0     215
##
##      Percent missing:[1] 0.82
##
##
##
## VARIABLE 7 : new_mode
##      var
##           MSM           IDU       MSM/IDU      Transfus
##         16416          2132          2697           139
##          Hemo        Hetero           Ped F Pres Hetero
##           124          1994           153           437
##           NIR          <NA>
##          2042             0
##
```

```
##      Percent missing:[1] 0
##
##
##
## VARIABLE 8 : tth_lneg_dt_flag
##      var
##     1     2     3     4  <NA>
##   348  1586   684 23516     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 9 : tth_ppos_dt_flag
##      var
##     1     2     3     4  <NA>
##   818  2390   460 22466     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 10 : est_infect_period
##      var
##     1     2     3  <NA>
##  1499   913 23722     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 11 : hdx_yr_qtr
##      [1] ""
##
##      Percent missing:numeric(0)
##
##
##
## VARIABLE 12 : hdx_dt_flag
##      var
##     D     M     Y  <NA>
##  4721 18249  3021   143
##
##      Percent missing:[1] 0.55
##
##
##
## VARIABLE 13 : adx_yr_qtr
##      [1] ""
##
##      Percent missing:numeric(0)
##
##
##
## VARIABLE 14 : adx_dt_flag
##      var
```

```
##     D     M     Y  <NA>
##  2245 15551   143  8195
##
##      Percent missing:[1] 31.36
##
##
##
## VARIABLE 15 : lag_lneg_hdx_dt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   178.0   426.5   941.1  1112.0  9938.0   23516
##
##      Percent missing:[1] 89.98
##
##
##
## VARIABLE 16 : lag_ppos_hdx_dt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     0.0     4.0   318.8    15.0 10630.0   22466
##
##      Percent missing:[1] 85.96
##
##
##
## VARIABLE 17 : tth_prev_pos
##      var
##     N     Y  <NA>
## 25359   775     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 18 : vl_days
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       9     303    1514    2696   13040    7891
##
##      Percent missing:[1] 30.19
##
##
##
## VARIABLE 19 : cd4_days
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     0.0    61.0   863.9  1259.0 14550.0    2828
##
##      Percent missing:[1] 10.82
##
##
##
## VARIABLE 20 : meth_use
##      var
##      NO UNKNOWN     YES    <NA>
##     658     247     473   24756
##
##      Percent missing:[1] 94.73
```

4 Variable Transformations

4.1 Split the combined year-quarter of diagnosis and AIDS variables

```
################################################################ SPLIT COMBINED YR-QTR VARIABLE Year, quar
################################################################ quarter-year of Dx (diagnosis)
dataf$yearDx <- as.numeric(substring(dataf$hdx_yr_qtr, 0, 4))
dataf$quarterDx <- as.numeric(substring(dataf$hdx_yr_qtr, 6,
    6))
dataf$timeDx <- dataf$yearDx + (dataf$quarterDx - 1)/4
# AIDS at Dx - if missing, assumed to be false
dataf$aidsAtDx <- dataf$hdx_yr_qtr == dataf$adx_yr_qtr
dataf$aidsAtDx[is.na(dataf$aidsAtDx)] <- FALSE
# Year, quarter, and quarter-year of AIDS (diagnosis)
dataf$yearAids <- as.numeric(substring(dataf$adx_yr_qtr, 0, 4))
dataf$quarterAids <- as.numeric(substring(dataf$adx_yr_qtr, 6,
    6))
dataf$timeAids <- dataf$yearAids + (dataf$quarterAids - 1)/4
```

4.2 Now subset the data based on essentials

```
################################################################ SUBSET THE DATA - INITIAL RESTRICTIONS
if (!"year_min" %in% ls()) year_min <- 2005
if (!"year_max" %in% ls()) year_max <- 2013

# Year min and max for this run
c(year_min, year_max)

## [1] 2005 2014

# Non-sequential look
table(hst_included = dataf$hst == "WA", useNA = "ifany")

## hst_included
## FALSE  TRUE  <NA>
##  5953 20180     1

table(yearDx_included = dataf$yearDx >= year_min & dataf$yearDx <=
    year_max, useNA = "ifany")

## yearDx_included
## FALSE  TRUE  <NA>
## 19267  6724   143

table(yearDx_missing = is.na(dataf$hdx_yr_qtr))

## yearDx_missing
## FALSE  TRUE
## 25991   143

table(age_missing_and_missing_lastNeg = (is.na(dataf$hdx_age) &
    is.na(dataf$lag_lneg_hdx_dt)))

## age_missing_and_missing_lastNeg
## FALSE  TRUE
## 25919   215
```

```
# Sequential look
(hst_included <- table(hst_included = dataf$hst == "WA", useNA = "ifany"))

## hst_included
## FALSE  TRUE  <NA>
##  5953 20180     1

dataf <- subset(dataf, hst == "WA")
(yearDx_included <- table(yearDx_included = (dataf$yearDx >=
    year_min & dataf$yearDx <= year_max), useNA = "ifany"))

## yearDx_included
## FALSE  TRUE
## 14928  5252

dataf <- subset(dataf, yearDx >= year_min & yearDx <= year_max)
(age_included <- table(age_and_lastNeg_present = !(is.na(dataf$hdx_age) &
    is.na(dataf$lag_lneg_hdx_dt))))

## age_and_lastNeg_present
## TRUE
## 5252

dataf <- subset(dataf, !(is.na(hdx_age) & is.na(lag_lneg_hdx_dt)))
(Nobs1 <- nrow(dataf))

## [1] 5252
```

Excluded 20882 cases based on year and hst restrictions and missingness in age and year of diagnosis.

4.2.1 Diagnosis

Years of initial diagnosis represented:

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
##  559  542  583  539  548  558  496  512  468  447
```

Quarters of initial diagnosis represented:

```
##
##    1    2    3    4 <NA>
## 1376 1369 1270 1228    9
```

4.3 Split the combined year-quarter of diagnosis and AIDS variables

Editing For those cases when we don't know the quarter, when should the diagnosis fall? Should we evenly distribute them throughout the 4 quarters? I will do that for now:

```
################################################################ IMPUTE A QUARTER IF ONLY YEAR IS KNOWN
impute_qtr <- !is.na(dataf$yearDx) & is.na(dataf$quarterDx)
set.seed(98103)
dataf$quarterDx[impute_qtr] <- sample(4, size = sum(impute_qtr),
    replace = TRUE)
dataf$timeDx <- dataf$yearDx + (dataf$quarterDx - 1)/4
summary(dataf$timeDx, digits = 6)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 2005.00 2007.25 2009.50 2009.67 2012.00 2014.75

time_min <- min(dataf$timeDx)
time_max <- max(dataf$timeDx)

# Time min and max for this run
c(time_min, time_max)

## [1] 2005.00 2014.75
```

4.4 Tabulate and collapse race and mode of diagnosis variables

Investigating counts of race by year and mode by year:

```
table(dataf$new_race, dataf$yearDx, useNA = "ifany")

##
##           2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
##   White    340  345  342  288  319  319  281  288  251  227
##   Black    104   81  104  101   92   79   90   97   90   98
##   Hisp      75   65   90   94   87  105   76   63   79   61
##   Asian     20   23   23   28   25   26   25   31   24   38
##   NHoPI      2    5    3    0    2    1    5    7    8    5
##   AI/AN      9    6    6   12    5    9    5    5    4    6
##   Multi      9   17   15   16   18   19   14   21   12   12
##   Unknown    0    0    0    0    0    0    0    0    0    0

table(dataf$new_mode, dataf$yearDx, useNA = "ifany")

##
##                2005 2006 2007 2008 2009 2010 2011 2012
##   MSM           296  314  335  302  318  349  297  280
##   IDU            40   41   31   25   27   33   30   22
##   MSM/IDU        61   46   48   31   44   27   47   40
##   Transfus        1    0    1    1    0    0    0    0
##   Hemo            1    0    0    0    0    0    0    0
##   Hetero         69   54   54   60   39   49   21   23
##   Ped             0    2    2    2   11   10    6    3
##   F Pres Hetero  22   17   29   25   35   19   18   16
##   NIR            69   68   83   93   74   71   77  128
##
##                2013 2014
##   MSM           271  237
##   IDU            20   20
##   MSM/IDU        34   29
##   Transfus        0    0
##   Hemo            0    0
##   Hetero         20   18
##   Ped             4    4
##   F Pres Hetero  18   12
##   NIR           101  127
```

```
################################################################ COLLAPSE RACE AND MODE OF DIAGNOSIS

race_levels <- c("White", "Black", "Hisp", "Asian", "Native",
```

```
    "Multi")
mode_levels <- c("MSM", "Hetero", "Blood/Needle")
dataf <- within(dataf, {
    race <- as.character(new_race)
    race[race == "AI/AN" | race == "NHoPI"] <- "Native"
    race <- factor(race, labels = race_levels, levels = race_levels)
    mode <- as.character(new_mode)
    mode[mode == "MSM/IDU"] <- "MSM"
    mode[mode == "F Pres Hetero" | mode == "NIR"] <- "Hetero"
    mode[mode == "IDU" | mode == "Transfus" | mode == "Hemo" |
        mode == "Ped"] <- "Blood/Needle"
    mode <- factor(mode, levels = mode_levels, labels = mode_levels)
    mode2 <- factor(ifelse(mode == "MSM", "MSM", "non-MSM"))
})
```

```
table(dataf$race, dataf$yearDx, useNA = "ifany")
```

```
##
##          2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
##   White   340  345  342  288  319  319  281  288  251  227
##   Black   104   81  104  101   92   79   90   97   90   98
##   Hisp     75   65   90   94   87  105   76   63   79   61
##   Asian    20   23   23   28   25   26   25   31   24   38
##   Native   11   11    9   12    7   10   10   12   12   11
##   Multi     9   17   15   16   18   19   14   21   12   12
```

```
table(dataf$mode, dataf$yearDx, useNA = "ifany")
```

```
##
##                2005 2006 2007 2008 2009 2010 2011 2012 2013
##   MSM           357  360  383  333  362  376  344  320  305
##   Hetero        160  139  166  178  148  139  116  167  139
##   Blood/Needle   42   43   34   28   38   43   36   25   24
##
##                2014
##   MSM           266
##   Hetero        157
##   Blood/Needle   24
```

4.4.1 AIDS at diagnosis

AIDS at initial diagnosis?

```
##
## FALSE  TRUE
##  3871  1381
```

Years of AIDS diagnosis represented:

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 <NA>
##  166  208  217  257  280  235  233  201  173  162   20 3100
```

Quarters of AIDS diagnosis represented:

```
##
##    1    2    3    4 <NA>
##  547  549  532  521 3103
```

4.5 Make a flag for everHadNegTest

This variable will be coded as Yes=TRUE, No=FALSE, and Don't Know/Refused/Missing=NA

```
############################################################## CREATE everHadNegTest Define everHadNegTes
############################################################## tth_ever_neg 2015 data update: this varial
############################################################## numerically, so I have added that option
dataf <- transform(dataf, everHadNegTest = ifelse(tth_ever_neg ==
    "Y" | tth_ever_neg == 1, TRUE, ifelse(tth_ever_neg == "N" |
    tth_ever_neg == 2, FALSE, NA)))
with(dataf, table(everHadNegTest, tth_ever_neg, useNA = "always"))

##               tth_ever_neg
## everHadNegTest    1    2    3    4    5 <NA>
##          FALSE    0  552    0    0    0    0
##          TRUE  2479    0    0    0    0    0
##          <NA>     0    0    6  354 1861    0

# Now cross-check it with the lag_lneg_hdx_dt, which actually
# has the time since last negative test
(checkEver <- with(dataf, table(everHadNegTest, TID_NA = is.na(lag_lneg_hdx_dt),
    useNA = "always")))

##               TID_NA
## everHadNegTest FALSE TRUE <NA>
##          FALSE     4  548    0
##          TRUE   2403   76    0
##          <NA>     16 2205    0

# Look at actual lag_lneg_hdx_dt values by everHadNegTest
ddply(dataf, .(everHadNegTest), function(x) c(summary(x$lag_lneg_hdx_dt)))

##   everHadNegTest Min. 1st Qu. Median  Mean 3rd Qu. Max.
## 1          FALSE  112   475.0  738.0 667.0     930 1080
## 2           TRUE    0   178.5  428.0 949.3    1120 9938
## 3             NA  122   212.2  665.5 907.1    1588 2663
##   NA's
## 1  548
## 2   76
## 3 2205
```

Editing We have 4 cases with everHadNegTest=FALSE and 16 with everHadNegTest=NA but have a time since last negative test. Change their everHadNegTest flag.

```
toTRUE1 <- !dataf$everHadNegTest & !is.na(dataf$lag_lneg_hdx_dt)
toTRUE2 <- is.na(dataf$everHadNegTest) & !is.na(dataf$lag_lneg_hdx_dt)
dataf$everHadNegTest[toTRUE1] <- TRUE
dataf$everHadNegTest[toTRUE2] <- TRUE
```

More editing We have 76 cases who have everHadNegTest=TRUE but have NO time since last negative test. Change their everHadNegTest flag.

```
toFALSE <- dataf$everHadNegTest & is.na(dataf$lag_lneg_hdx_dt)
dataf$everHadNegTest[toFALSE] <- FALSE
```

```
(checkEver <- with(dataf, table(everHadNegTest, TID_NA = is.na(lag_lneg_hdx_dt),
    useNA = "always")))
```

```
##                TID_NA
## everHadNegTest FALSE TRUE <NA>
##          FALSE     0  624    0
##          TRUE   2423    0    0
##          <NA>      0 2205    0
```

Better?

4.6 Define TID, aka infPeriod

Define aidsUB=17.98 years, and lastNeg_yrs as lag_lneg_hdx_dt/365, and infPeriod as follows:

everHadNegTest	infPeriod
TRUE	min(lastNeg_yrs, aidsUB)
FALSE	min(age-16, aidsUB)
NA	NA

```
################################################################ CREATE infPeriod and then look at it

#### TEMPORARY: dataf£age=35

aidsUB <- qweibull(0.95, shape = 2.516, scale = 1/0.086)  #17.98418
dataf <- within(dataf, {
    lastNeg_yrs <- lag_lneg_hdx_dt/365
    infPeriod <- ifelse(everHadNegTest, pmin(lastNeg_yrs, aidsUB),
        ifelse(!everHadNegTest, pmin(hdx_age - 16, aidsUB), NA))
    earliestInf <- hdx_age - infPeriod
})
```

```
summary(dataf$infPeriod, digits = 3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -3.000   0.612   1.990   5.080   7.000  18.000    2205
```

```
# Number of cases who got a negative infPeriod
(neginfPeriod <- sum(dataf$infPeriod < 0, na.rm = TRUE))
```

```
## [1] 3
```

```r
# Diagnoses at or under age 16 by everHadNegTest
(a1 <- table(atunder16 = dataf$hdx_age <= 16, everHadNegTest = dataf$everHadNegTest,
    useNA = "ifany"))
```

```
##          everHadNegTest
## atunder16 FALSE TRUE <NA>
##     FALSE   619 2418 2134
##     TRUE      5    5   71
```

```r
# Diagnoses at or under age 16 by year, 2005-2013
table(atunder16count = subset(dataf, yearDx >= year_min & yearDx <=
    year_max)$hdx_age <= 16, year = subset(dataf, yearDx >= year_min &
    yearDx <= year_max)$yearDx, useNA = "ifany")
```

```
##               year
## atunder16count 2005 2006 2007 2008 2009 2010 2011 2012 2013
##          FALSE  556  539  576  533  536  546  488  500  457
##          TRUE     3    3    7    6   12   12    8   12   11
##               year
## atunder16count 2014
##          FALSE  440
##          TRUE     7
```

```r
# Now just under 16, excluding hdx_age=16 Diagnoses under age
# 16 by everHadNegTest
(a2 <- table(under16 = dataf$hdx_age < 16, everHadNegTest = dataf$everHadNegTest,
    useNA = "ifany"))
```

```
##        everHadNegTest
## under16 FALSE TRUE <NA>
##   FALSE   621 2420 2139
##   TRUE      3    3   66
```

```r
# Diagnoses under age 16 by year
table(under16count = subset(dataf, yearDx >= year_min & yearDx >=
    year_max)$hdx_age < 16, year = subset(dataf, yearDx >= year_min &
    yearDx >= year_max)$yearDx, useNA = "ifany")
```

```
##             year
## under16count 2014
##        FALSE  441
##        TRUE     6
```

```r
# Among those diagnosed at or under 16: everHadNegTest by
# mode
table(everHadNegTest = subset(dataf, hdx_age <= 16)$everHadNegTest,
    mode = subset(dataf, hdx_age <= 16)$new_mode, useNA = "ifany")
```

```
##               mode
## everHadNegTest MSM IDU MSM/IDU Transfus Hemo Hetero Ped
##          FALSE   1   0       0        0    0      1   3
##          TRUE    1   0       1        0    0      1   0
##          <NA>    2   0       0        0    0      1  40
##               mode
## everHadNegTest F Pres Hetero NIR
##          FALSE             0   0
##          TRUE              1   1
##          <NA>              0  28
```

Diagnoses younger than age 16 There are 76 cases who do not have a date of last negative test and may not fit the assumption of TID=age-16. Of those, 7 are age 16 at diagnosis and will have TID=0 using this assumption. Primary mode of transmission is Ped ('Perinatal or pediatric').

```
(young_included <- with(dataf, table(over16_or_atunder16_with_obs_infPeriod = (hdx_age >
    16 | !(hdx_age <= 16 & (!everHadNegTest | is.na(everHadNegTest)))))))

## over16_or_atunder16_with_obs_infPeriod
## FALSE  TRUE
##    76  5176

dataf <- subset(dataf, !(hdx_age <= 16 & (!everHadNegTest | is.na(everHadNegTest))))
(Nobs2 <- nrow(dataf))

## [1] 5176

summary(dataf$infPeriod, digits = 3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.617   1.990   5.100   7.000  18.000    2134
```

Excluded 76 cases due to age≤16 and no observed infPeriod data.

4.7 TID is still zero

```
# Cases who still have a zero infPeriod - treat like missing
zeroinf <- dataf$infPeriod == 0 & !is.na(dataf$infPeriod)
(table(dataf$everHadNegTest[zeroinf], useNA = "ifany"))

##
## TRUE
##   29

# Change their everHadNeg flag to NA and their infPeriod to
# NA, since TID=0 does not make sense
dataf$everHadNegTest[zeroinf] <- NA
dataf$infPeriod[zeroinf] <- NA
```

A remaining 29 cases had infPeriod=0 for some reason. Since this doesn't make much sense, we treat these as if they had missing responses for everHadNegTest and infPeriod.

4.8 Check effect of TID definition

```
# We did cap some people whose TID's were >aidsUB
(check_cap1 <- with(subset(dataf, everHadNegTest), table(original_over_aidsUB = lastNeg_yrs >
    aidsUB, infPeriod_over_aidsUB = infPeriod > aidsUB, useNA = "ifany")))

##                     infPeriod_over_aidsUB
## original_over_aidsUB FALSE
##                FALSE  2369
##                TRUE     25
```

Among those with everHadNegTest=TRUE, we capped 25 cases at aidsUB.

```
(check_cap2 <- with(subset(dataf, !everHadNegTest), table(original_over_aidsUB = lastNeg_yrs >
    aidsUB, infPeriod_over_aidsUB = infPeriod > aidsUB, useNA = "ifany")))
```

```
##                      infPeriod_over_aidsUB
## original_over_aidsUB FALSE
##                 <NA>   619
```

Among those with everHadNegTest=FALSE, no one had an original TID value.

```
(check_cap3 <- with(subset(dataf, is.na(everHadNegTest)), table(original_over_aidsUB = lastNeg_yrs >
    aidsUB, infPeriod_over_aidsUB = infPeriod > aidsUB, useNA = "ifany")))

##                      infPeriod_over_aidsUB
## original_over_aidsUB <NA>
##                FALSE   29
##                <NA>  2134
```

Among those with everHadNegTest=NA, no one had an original TID value.

5 Analysis Subset

Final subset is

- 2005 onwards
- Diagnosis made in WA state
- If missing age, must have recorded time of last negative test
- If age¡=16, must have recorded time of last negative test
- Non-missing year of diagnosis

Final look at data:

```
nrow(dataf)

## [1] 5176

if (printSummaries) {
    for (var in c("hdx_age", "timeDx", "everHadNegTest", "lastNeg_yrs",
        "infPeriod")) {
        cat("\nVARIABLE:", var, "\n")
        print(summary(dataf[, var]))
    }
}

##
## VARIABLE: hdx_age
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   28.00   36.00   37.54   46.00   83.00
##
## VARIABLE: timeDx
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2005    2007    2010    2010    2012    2015
##
## VARIABLE: everHadNegTest
##    Mode   FALSE    TRUE    NA's
## logical     619    2394    2163
##
## VARIABLE: lastNeg_yrs
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
```

```
##  0.0000  0.4904  1.1750  2.5990  3.0670 27.2300    2753
##
## VARIABLE: infPeriod
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0027  0.6438  2.0160  5.1450  7.1450 17.9800    2163
```