Format WA Data - 2017 Estimates

Jeanette Birnbaum

8/31/18

Contents

1 Raw Data Overview

1.1 Sample Size

N = 21550

1.2 Variable list

```r
str(dataf)
```

```
## 'data.frame':    21550 obs. of  21 variables:
##  $ firstvl          : num  8433 19914 35382 51 108 ...
##  $ firstcd4cnt      : num  177 243 501 636 847 ...
##  $ tth_ever_neg     : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ new_race         : Factor w/ 8 levels "White","Black",..: 2 2 1 1 1 1 3 3 1 1 ...
##  $ hst              : chr  "WA" "WA" "WA" "WA" ...
##  $ hdx_age          : int  51 25 41 34 38 33 33 41 45 19 ...
##  $ new_mode         : Factor w/ 9 levels "MSM","IDU","MSM/IDU",..: 3 6 6 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  "D" "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" ...
##  $ dx_in_king       : chr  "Y" "Y" "Y" "Y" ...
##  $ vl_days          : int  673 111 7517 4032 1396 3061 2618 1810 1607 4461 ...
##  $ cd4_days         : int  1734 122 7517 6294 4151 3857 2618 2283 2350 5356 ...
##  $ meth_use         : chr  NA NA NA NA ...
```

1.3 Variable summaries

```
## 
## 
## 
## VARIABLE 1 : firstvl
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     868   17340   38720   98900  100000    6917
## 
##      Percent missing:[1] 32.1
## 
## 
## 
## VARIABLE 2 : firstcd4cnt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   150.0   346.0   389.6   567.0  4269.0    6833
## 
##      Percent missing:[1] 31.71
## 
## 
## 
## VARIABLE 3 : tth_ever_neg
##      var
##     1     2     5  <NA>
##  3435   758 17357     0
## 
```

```
##      Percent missing:[1] 0
##
##
##
## VARIABLE 4 : new_race
##      var
##   White   Black    Hisp   Asian   NHoPI   AI/AN   Multi Unknown    <NA>
##   14492    2934    2287     585      79     292     871      10       0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 5 : hst
##      var
##    WA  <NA>
## 21550     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 6 : hdx_age
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   28.00   35.00   35.87   42.00   92.00
##
##      Percent missing:[1] NA
##
##
##
## VARIABLE 7 : new_mode
##      var
##           MSM           IDU       MSM/IDU      Transfus          Hemo
##         13526          1720          2017           122           101
##        Hetero           Ped F Pres Hetero           NIR          <NA>
##          2109           125             0          1830             0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 8 : tth_lneg_dt_flag
##      var
##     1     2     3     4  <NA>
##   518  1788  1028 18216     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 9 : tth_ppos_dt_flag
##      var
##     1     2     3     4  <NA>
##  1137  2304   485 17624     0
```

```
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 10 : est_infect_period
##      var
##     1     2     3  <NA>
##  1630  1050 18870     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>
##  8270 11011  2269     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 13 : adx_yr_qtr
##      [1] ""
##
##      Percent missing:numeric(0)
##
##
##
## VARIABLE 14 : adx_dt_flag
##      var
##    D    M    Y <NA>
## 5131 9723   58 6638
##
##      Percent missing:[1] 30.8
##
##
##
## VARIABLE 15 : lag_lneg_hdx_dt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   160.0   382.5   907.2  1054.0 11570.0   18216
##
##      Percent missing:[1] 84.53
##
##
##
```

```
## VARIABLE 16 : lag_ppos_hdx_dt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     0.0     6.0   445.8    15.0 13380.0   17624
##
##      Percent missing:[1] 81.78
##
##
##
## VARIABLE 17 : tth_prev_pos
##      var
##     N     Y  <NA>
## 20675   875     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 18 : dx_in_king
##      var
##     N     Y  <NA>
##  8246 13304     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 19 : vl_days
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       7      62    1286    2277   11870    6917
##
##      Percent missing:[1] 32.1
##
##
##
## VARIABLE 20 : cd4_days
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       8      94    1459    2629   11440    6833
##
##      Percent missing:[1] 31.71
##
##
##
## VARIABLE 21 : meth_use
##      var
##      NO UNKNOWN     YES    <NA>
##     778     215     461   20096
##
##      Percent missing:[1] 93.25
```

2 Subset based on hst=WA and year

2.1 First, split the combined year-quarter of diagnosis and AIDS variables

```
##############################################################
# SPLIT COMBINED YR-QTR VARIABLE
##############################################################
# Year, quarter, and 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
```

2.2 Subset the data based on hst=WA and year

```
##############################################################
# 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 2017
```

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

```
## hst_included
##  TRUE
## 21550
```

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

```
## yearDx_included
## FALSE  TRUE
## 14997  6553
```

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

```
## yearDx_missing
## FALSE
## 21550
```

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

```
## age_missing_and_missing_lastNeg
## FALSE
## 21550
```

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

```
## hst_included
##  TRUE
## 21550
```

```
dataf <- subset(dataf, hst=='WA')
(yearDx_included <- table(yearDx_included=(dataf$yearDx>=year_min & dataf$yearDx<=year_max), useNA='ifa
```

```
## yearDx_included
## FALSE  TRUE
## 14997  6553
```

```
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
## 6553
```

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

```
## [1] 6553
```

Excluded 14997 cases based on year and hst restrictions and missingness in age and year of diagnosis.

2.3 New sample size

New sample size is 6553

3 Year and quarter of diagnosis: cleaning it up

3.1 Years represented

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
##  554  533  579  533  546  557  492  511  456  448  461  438  445
```

3.2 Quarters represented

```
##
##    1    2    3    4 <NA>
## 1697 1671 1631 1543   11
```

3.3 Distribute unknown quarters uniformly across Q1-Q4

```r
#############################################################
# 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 2007.75 2010.75 2011.05 2014.25 2017.75
```

```r
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 2017.75
```

4 Tabulate and collapse race and mode of diagnosis variables

4.1 Race and mode by year

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

```
##
##           2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
##   White    311  311  321  281  312  317  277  284  245  225  223  205  191
##   Black     96   77   98   99   90   78   89   94   88   97   93   91  115
##   Hisp      74   66   90   96   88  106   78   64   79   64   90   74   92
##   Asian     20   21   22   27   25   24   23   29   24   39   35   35   26
##   NHoPI      2    5    1    0    3    1    5    6    5    5    3    4    3
##   AI/AN      5    6    6   11    6    8    4    5    3    6    5    9    6
##   Multi     46   47   41   19   22   23   16   29   12   12   12   20   12
##   Unknown    0    0    0    0    0    0    0    0    0    0    0    0    0
```

```r
    table(dataf$new_mode, dataf$yearDx, useNA='ifany')
```

```
##
##                 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##   MSM            293  311  332  301  316  352  297  284  268  251  276
##   IDU             38   42   32   25   27   33   30   22   20   23   36
##   MSM/IDU         65   43   50   32   43   30   47   42   33   29   22
##   Transfus         1    0    1    1    0    0    0    0    0    0    0
##   Hemo             1    0    0    0    0    0    0    0    0    0    0
##   Hetero          92   70   81   84   74   68   39   40   37   45   45
##   Ped              0    2    2    2   11   10    6    3    5    4    4
##   F Pres Hetero    0    0    0    0    0    0    0    0    0    0    0
##   NIR             64   65   81   88   75   64   73  120   93   96   78
```

```
##
##                   2016 2017
##   MSM              223  238
##   IDU               29   18
##   MSM/IDU           27   25
##   Transfus           0    0
##   Hemo               0    0
##   Hetero            62   51
##   Ped                5    6
##   F Pres Hetero      0    0
##   NIR               92  107
```

4.2 Collapse

```
#############################################################
# 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 2015 2016 2017
##   White   311  311  321  281  312  317  277  284  245  225  223  205  191
##   Black    96   77   98   99   90   78   89   94   88   97   93   91  115
##   Hisp     74   66   90   96   88  106   78   64   79   64   90   74   92
##   Asian    20   21   22   27   25   24   23   29   24   39   35   35   26
##   Native    7   11    7   11    9    9    9   11    8   11    8   13    9
##   Multi    46   47   41   19   22   23   16   29   12   12   12   20   12
```

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

```
##
##                2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
##   MSM           358  354  382  333  359  382  344  326  301  280  298  250
##   Hetero        156  135  162  172  149  132  112  160  130  141  123  154
##   Blood/Needle   40   44   35   28   38   43   36   25   25   27   40   34
##
##                2017
##   MSM           263
##   Hetero        158
##   Blood/Needle   24
```

```
table(dataf$mode2, dataf$yearDx, useNA='ifany')
```

```
##
##           2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
##   MSM      358  354  382  333  359  382  344  326  301  280  298  250  263
##   non-MSM  196  179  197  200  187  175  148  185  155  168  163  188  182
```

5 AIDS at Diagnosis

5.1 AIDS at initial diagnosis?

```
##
## FALSE  TRUE
##  4907  1646
```

5.2 Years of AIDS diagnosis represented:

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 <NA>
##  166  206  220  265  269  237  232  202  175  160  169  137  160   30 3925
```

5.3 Quarters of AIDS diagnosis represented:

```
##
##    1    2    3    4 <NA>
##  656  684  654  627 3932
```

6 Ever had a last negative test (everHadNegTest)

6.1 Coding

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

```
#############################################################
# CREATE everHadNegTest
#############################################################
# Define everHadNegTest based on tth_ever_neg
# 2015 data update: this variable was coded numerically, so I have
# added that option in.
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    5 <NA>
##          FALSE    0  738    0    0
##          TRUE  3342    0    0    0
##          <NA>     0    0 2473    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     5  733    0
##          TRUE   3229  113    0
##          <NA>     13 2460    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. NA's
## 1          FALSE  101     112    596 551.8     880  1070  733
## 2           TRUE    0     159    383 911.7    1065 11570  113
## 3             NA  122     207    569 738.5     997  2022 2460
```

6.2 Make compatible with recorded LNT dates

6.2.1 Change incorrect FALSEs

We have 5 cases with everHadNegTest=FALSE and 13 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
```

6.2.2 Change incorrect TRUEs

We have 113 cases who have everHadNegTest=TRUE but have NO time since last negative test. Change their everHadNegTest flag. Change, 9/27/17 - previously was setting to false; now, set to NA.

```
## an alternative to setting to FALSE
toNA <- dataf$everHadNegTest & is.na(dataf$lag_lneg_hdx_dt)
dataf$everHadNegTest[toNA] <- NA
```

6.2.3 Check

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

```
##                 TID_NA
## everHadNegTest FALSE TRUE <NA>
##          FALSE     0  733    0
##          TRUE   3247    0    0
##          <NA>      0 2573    0
```

7 Time since last negative test (infPeriod)

7.1 Apply age-16 assumption and summarize

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

#### TEMPORARY:
#dataf$age=35

aidsUB <- qweibull(.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   0.529   1.590   4.700   6.380  18.000    2573
```

7.2 Diagnoses younger than 16

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

```
## [1] 4
```

```
# 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   724 3240 2491
##     TRUE      9    7   82
```

```
# 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 2014 2015 2016
##          FALSE  551  528  573  527  534  545  484  499  445  441  456  434
##          TRUE    3    5    6    6   12   12    8   12   11    7    5    4
##               year
## atunder16count 2017
##          FALSE  438
##          TRUE     7
```

```
# 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   729 3243 2496
##   TRUE      4    4   77
```

```
# 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 2017
##        FALSE  440
##        TRUE     5
```

```
# 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 F Pres Hetero NIR
##          FALSE   4   1       0        0    0      1   3             0   0
##          TRUE    2   1       1        0    0      1   0             0   2
##          <NA>    2   0       0        0    0      1  49             0  30
```

There are 91 cases who do not have a date of last negative test and may not fit the assumption of TID=age-16. Of those, 10 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
##    91  6462
```

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

```
## [1] 6462
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.534   1.600   4.720   6.400  18.000    2491
```

Excluded 91 cases due to age ≤ 16 and no observed infPeriod data.

7.3 Maximum window of 18 years

```
# 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  3209
##                TRUE     38
```

Among those with everHadNegTest=TRUE, we capped 38 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>   724
```

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>
##                 <NA> 2491
```

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

8 Final analytic dataset

8.1 Reminder of data cleaning

Final subset is of size 6462 * Diagnoses included: - Year: non-missing, and 2005 onwards - Occurred in WA state - Excluded 14997 cases based on year and hst restrictions (no missingness in age and year of diagnosis in data for 2015 estimates). * Ages included: - If missing age, must have recorded time of last negative test - If age ≤ 16, must have recorded time of last negative test - Excluded 91 cases due to age ≤ 16 and no observed LNT.

8.2 Variable summaries

```
## [1] 6462
```

```
##
## VARIABLE: hdx_age
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   28.00   36.00   37.49   46.00   83.00
##
## VARIABLE: timeDx
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2005    2008    2011    2011    2014    2018
##
## VARIABLE: everHadNegTest
##    Mode   FALSE    TRUE    NA's
## logical     724    3247    2491
##
## VARIABLE: lastNeg_yrs
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.434   1.049   2.494   2.916  31.710    3215
##
## VARIABLE: infPeriod
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.5342  1.6030  4.7170  6.4010 17.9800    2491
```