Format WA Data - 2015 Estimates

Jeanette Birnbaum

10/17/2016

Contents

1 Raw Data Overview **2**

1.1 Sample Size . . . 2

1.2 Variable list . . . 2

1.3 Variable summaries . . . 2

2 Subset based on hst=WA and year **6**

2.1 First, split the combined year-quarter of diagnosis and AIDS variables . . . 6

2.2 Subset the data based on hst=WA and year . . . 6

2.3 New sample size . . . 8

3 Year and quarter of diagnosis: cleaning it up **8**

3.1 Years represented . . . 8

3.2 Quarters represented . . . 8

3.3 Distribute unknown quarters uniformly across Q1-Q4 . . . 8

4 Tabulate and collapse race and mode of diagnosis variables **8**

4.1 Race and mode by year . . . 8

4.2 Collapse . . . 9

5 AIDS at Diagnosis **10**

5.1 AIDS at initial diagnosis? . . . 10

5.2 Years of AIDS diagnosis represented: . . . 10

5.3 Quarters of AIDS diagnosis represented: . . . 10

6 Ever had a last negative test (everHadNegTest) **11**

6.1 Coding . . . 11

6.2 Make compatible with recorded LNT dates . . . 11

7 Time since last negative test (infPeriod) **12**

7.1 Apply age-16 assumption and summarize . . . 12

7.2 Diagnoses younger than 16 . . . 13

7.3 Maximum window of 18 years . . . 14

8 Final analytic dataset **15**

8.1 Reminder of data cleaning 15

8.2 Variable summaries 15

1 Raw Data Overview

1.1 Sample Size

N = 20451

1.2 Variable list

```
str(dataf)
```

```
## 'data.frame':    20451 obs. of  21 variables:
##  $ firstvl          : num  658 19914 35382 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         : Factor w/ 8 levels "White","Black",..: 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         : Factor w/ 9 levels "MSM","IDU","MSM/IDU",..: 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" ...
##  $ dx_in_king       : chr  "Y" "Y" "Y" "Y" ...
##  $ vl_days          : int  181 111 7517 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 ...
```

1.3 Variable summaries

```
##
##
##
## VARIABLE 1 : firstvl
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    1340   18700   39510  100000  100000    6652
##
##      Percent missing:[1] 32.53
##
```

```
##
##
## VARIABLE 2 : firstcd4cnt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   105.0   259.0   327.9   485.0  6745.0    2063
##
##      Percent missing:[1] 10.09
##
##
##
## VARIABLE 3 : tth_ever_neg
##      var
##     1     2     3     4     5  <NA>
##  2671   622     6   390 16762     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 4 : new_race
##      var
##   White   Black    Hisp   Asian   NHoPI   AI/AN   Multi Unknown    <NA>
##   14350    2726    2049     530      78     293     415      10       0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 5 : hst
##      var
##    WA  <NA>
## 20451     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 6 : hdx_age
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   28.00   35.00   35.77   42.00   91.00
##
##      Percent missing:[1] NA
##
##
##
## VARIABLE 7 : new_mode
##      var
##          MSM           IDU       MSM/IDU      Transfus          Hemo
##        12967          1653          1945           122           101
##       Hetero           Ped F Pres Hetero           NIR          <NA>
##         1572           107           371          1613             0
##
##      Percent missing:[1] 0
##
```

```
##
##
## VARIABLE 8 : tth_lneg_dt_flag
##      var
##     1     2     3     4  <NA>
##   391  1554   664 17842     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 9 : tth_ppos_dt_flag
##      var
##     1     2     3     4  <NA>
##   945  2173   281 17052     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 10 : est_infect_period
##      var
##     1     2     3  <NA>
##  1517   960 17974     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>
##  4231 13935  2285     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>
##  1876 12517    58  6000
##
##      Percent missing:[1] 29.34
##
##
##
## VARIABLE 15 : lag_lneg_hdx_dt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   178.0   428.0   942.4  1112.0  9938.0   17842
##
##      Percent missing:[1] 87.24
##
##
##
## VARIABLE 16 : lag_ppos_hdx_dt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     0.0     3.0   299.0    12.5 10630.0   17052
##
##      Percent missing:[1] 83.38
##
##
##
## VARIABLE 17 : tth_prev_pos
##      var
##     N     Y  <NA>
## 19876   575     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 18 : dx_in_king
##      var
##     N     Y  <NA>
##  7726 12725     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 19 : vl_days
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       5      56    1190    2047   11440    6598
##
##      Percent missing:[1] 32.26
##
##
##
## VARIABLE 20 : cd4_days
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     0.0    31.0   670.2   867.0 11440.0    2058
##
##      Percent missing:[1] 10.06
```

```
##
##
##
## VARIABLE 21 : meth_use
##      var
##      NO UNKNOWN     YES    <NA>
##     565     163     355   19368
##
##      Percent missing:[1] 94.7
```

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 2015
```

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

```
## hst_included
##  TRUE
## 20451
```

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

```
## yearDx_included
## FALSE  TRUE
## 14776  5675
```

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

```
## yearDx_missing
## FALSE
## 20451
```

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

```
## age_missing_and_missing_lastNeg
## FALSE
## 20451
```

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

```
## hst_included
##  TRUE
## 20451
```

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

```
## yearDx_included
## FALSE  TRUE
## 14776  5675
```

```r
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
## 5675
```

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

```
## [1] 5675
```

Excluded 14776 cases based on year and hst restrictions and missingness in age and year of diagnosis.

2.3 New sample size

New sample size is 5675

3 Year and quarter of diagnosis: cleaning it up

3.1 Years represented

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##  556  537  581  537  549  557  495  510  457  444  452
```

3.2 Quarters represented

```
##
##    1    2    3    4 <NA>
## 1471 1468 1389 1336   11
```

3.3 Distribute unknown quarters uniformly across Q1-Q4

```
################################################################
# 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 2007.50 2010.00 2010.12 2012.75 2015.75
```

```
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 2015.75
```

4 Tabulate and collapse race and mode of diagnosis variables

4.1 Race and mode by year

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

```
##
##           2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##   White    335  339  341  287  319  320  281  287  244  228  222
##   Black    102   80  104  100   92   79   89   95   89   96   93
##   Hisp      76   65   90   93   86  105   77   63   78   61   84
##   Asian     20   24   22   29   25   26   24   31   24   38   35
##   NHoPI      2    5    3    0    3    1    5    7    6    5    5
##   AI/AN      9    6    6   12    6    8    5    5    4    6    5
##   Multi     12   18   15   16   18   18   14   22   12   10    8
##   Unknown    0    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 2013 2014 2015
##   MSM            293  312  333  301  317  353  296  280  266  250  271
##   IDU             39   42   32   25   27   33   31   21   20   22   33
##   MSM/IDU         64   44   49   32   44   28   47   42   33   28   21
##   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          69   52   53   60   40   48   21   23   19   21   20
##   Ped              0    2    2    2   11   10    6    3    4    3    4
##   F Pres Hetero   22   17   29   25   34   19   18   15   17   12   13
##   NIR             67   68   82   91   76   66   76  126   98  108   90
```

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
##   White   335  339  341  287  319  320  281  287  244  228  222
##   Black   102   80  104  100   92   79   89   95   89   96   93
##   Hisp     76   65   90   93   86  105   77   63   78   61   84
##   Asian    20   24   22   29   25   26   24   31   24   38   35
##   Native   11   11    9   12    9    9   10   12   10   11   10
##   Multi    12   18   15   16   18   18   14   22   12   10    8
```

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

```
##
##                2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##   MSM           357  356  382  333  361  381  343  322  299  278  292
##   Hetero        158  137  164  176  150  133  115  164  134  141  123
##   Blood/Needle   41   44   35   28   38   43   37   24   24   25   37
```

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

```
##
##           2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##   MSM      357  356  382  333  361  381  343  322  299  278  292
##   non-MSM  199  181  199  204  188  176  152  188  158  166  160
```

5 AIDS at Diagnosis

5.1 AIDS at initial diagnosis?

```
##
## FALSE  TRUE
##  4215  1460
```

5.2 Years of AIDS diagnosis represented:

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 <NA>
##  166  206  216  255  279  237  234  201  175  160  167   19 3360
```

5.3 Quarters of AIDS diagnosis represented:

```
##
##    1    2    3    4 <NA>
##  573  604  571  563 3364
```

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    3    4    5 <NA>
##          FALSE    0  622    0    0    0    0
##          TRUE  2671    0    0    0    0    0
##          <NA>     0    0    6  390 1986    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  617    0
##          TRUE   2589   82    0
##          <NA>     15 2367    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   112.0    596 553.8     880 1080  617
## 2           TRUE    0   178.0    428 944.6    1113 9938   82
## 3             NA  122   210.5    366 686.8     970 2022 2367
```

6.2 Make compatible with recorded LNT dates

6.2.1 Change incorrect FALSEs

We have 5 cases with everHadNegTest=FALSE and 15 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 82 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
```

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  699    0
##          TRUE   2609    0    0
##          <NA>      0 2367    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
                  })
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -3.000   0.614   1.990   5.140   7.300  18.000    2367
```

7.2 Diagnoses younger than 16

```r
# 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   693 2604 2294
##     TRUE      6    5   73
```

```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 2014 2015
##          FALSE  553  533  575  531  537  545  487  498  447  438  447
##          TRUE     3    4    6    6   12   12    8   12   10    6    5
```

```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   696 2606 2299
##   TRUE      3    3   68
```

```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 2015
##        FALSE  448
##        TRUE     4
```

```
# 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   2   0       0        0    0      1   3             0   0
##          TRUE    1   1       1        0    0      0   0             1   1
##          <NA>    2   0       0        0    0      1  43             0  27
```

There are 79 cases who do not have a date of last negative test and may not fit the assumption of TID=age-16. Of those, 8 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
##    79  5596
```

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

```
## [1] 5596
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.623   2.000   5.150   7.360  18.000    2294
```

Excluded 79 cases due to age $\leq$ 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  2581
##                TRUE     28
```

Among those with everHadNegTest=TRUE, we capped 28 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>   693
```

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> 2294
```

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 5596 * Diagnoses included: - Year: non-missing, and 2005 onwards - Occurred in WA state - Excluded 14776 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 $\leq$ 16, must have recorded time of last negative test - Excluded 79 cases due to age $\leq$ 16 and no observed LNT.

8.2 Variable summaries

```
## [1] 5596
```

```
##
## VARIABLE: hdx_age
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   28.00   36.00   37.44   45.00   83.00
##
## VARIABLE: timeDx
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2005    2008    2010    2010    2013    2016
##
## VARIABLE: everHadNegTest
##    Mode   FALSE    TRUE    NA's
```

```
## logical     693    2609    2294
##
## VARIABLE: lastNeg_yrs
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.4877  1.1730  2.5820  3.0470 27.2300    2987
##
## VARIABLE: infPeriod
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.6226  1.9960  5.1540  7.3610 17.9800    2294
```