Format WA Data - 2018 Estimates

Jeanette Birnbaum

2019-September-05

Contents

1 Raw Data Overview

1.1 Sample Size

N = 21747

1.2 Variable list

```
str(dataf)
```

```
## 'data.frame':    21747 obs. of  24 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 ...
##  $ mi_trans_categ   : int  3 5 2 1 1 1 3 1 1 1 ...
##  $ birth_sex        : chr  "M" "M" "F" "M" ...
##  $ notes            : chr  "This csv was saved from WA_BACKCALC_DATA_201807_revised.xlsx" NA NA NA .
```

1.3 Variable summaries

```
##
##
##
## VARIABLE 1 : firstvl
##          Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##      0.0    855.5  17530.0  38810.0  99140.0 100000.0     6853
##
##      Percent missing:[1] 31.51
##
##
##
## VARIABLE 2 : firstcd4cnt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     152     349     391     568    4269    6752
##
##      Percent missing:[1] 31.05
##
##
##
## VARIABLE 3 : tth_ever_neg
##      var
```

```
##     1     2     5  <NA>
##  3600   755 17392     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 4 : new_race
##      var
##   White   Black    Hisp   Asian   NHoPI   AI/AN   Multi Unknown    <NA>
##   14542    2958    2363     588      77     288     921      10       0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 5 : hst
##      var
##    WA  <NA>
## 21747     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 6 : hdx_age
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   28.00   35.00   35.89   42.00   92.00
##
##      Percent missing:[1] NA
##
##
##
## VARIABLE 7 : new_mode
##      var
##           MSM           IDU       MSM/IDU      Transfus          Hemo
##         13629          1734          2051           122           101
##        Hetero           Ped F Pres Hetero           NIR          <NA>
##          2135           126             0          1849             0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 8 : tth_lneg_dt_flag
##      var
##     1     2     3     4  <NA>
##   653  1770  1065 18259     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 9 : tth_ppos_dt_flag
```

```
##      var
##     1     2     3     4  <NA>
##  1200  2245   516 17786     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 10 : est_infect_period
##      var
##     1     2     3  <NA>
##  1627  1050 19070     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>
##  8494 10986  2267     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>
## 5252 9690   57 6748
##
##      Percent missing:[1] 31.03
##
##
##
## VARIABLE 15 : lag_lneg_hdx_dt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   151.0   372.0   893.8  1023.0 11570.0   18259
##
##      Percent missing:[1] 83.96
```

```
##
##
##
## VARIABLE 16 : lag_ppos_hdx_dt
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     0.0     7.0   460.3    18.0 13380.0   17786
##
##      Percent missing:[1] 81.79
##
##
##
## VARIABLE 17 : tth_prev_pos
##      var
##     N     Y  <NA>
## 20826   921     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 18 : dx_in_king
##      var
##     N     Y  <NA>
##  8356 13391     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 19 : vl_days
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       7      59    1272    2222   11870    6853
##
##      Percent missing:[1] 31.51
##
##
##
## VARIABLE 20 : cd4_days
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       8      85    1441    2587   12060    6752
##
##      Percent missing:[1] 31.05
##
##
##
## VARIABLE 21 : meth_use
##      var
##      NO UNKNOWN     YES    <NA>
##     830     224     500   20193
##
##      Percent missing:[1] 92.85
##
##
##
```

```
## VARIABLE 22 : mi_trans_categ
##         Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.901   2.000  19.000
##
##      Percent missing:[1] NA
##
##
##
## VARIABLE 23 : birth_sex
##      var
##     F     M  <NA>
##  2689 19058     0
##
##      Percent missing:[1] 0
##
##
##
## VARIABLE 24 : notes
##      var
## This csv was saved from WA_BACKCALC_DATA_201807_revised.xlsx
##                                                             1
##                                                          <NA>
##                                                         21746
##
##      Percent missing:[1] 100
```

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
#################################################################
```

```r
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 2018
```

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

```
## hst_included
##  TRUE
## 21747
```

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

```
## yearDx_included
## FALSE  TRUE
## 14943  6804
```

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

```
## yearDx_missing
## FALSE
## 21747
```

```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
## 21747
```

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

```
## hst_included
##  TRUE
## 21747
```

```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
## 14943  6804
```

```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
## 6804
```

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

```
## [1] 6804
```

Excluded 14943 cases based on year and hst restrictions and missingness in age and year of diagnosis.

2.3 New sample size

New sample size is 6804

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 2018
##  552  534  579  531  544  557  491  510  459  449  459  437  441  261
```

3.2 Quarters represented

```
##
##    1    2    3    4 <NA>
## 1839 1784 1639 1531   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 2008.00 2011.00 2011.32 2014.50 2018.25
```

```
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 2018.25
```

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 2016 2017
##   White    308  310  320  279  311  315  276  283  244  224  222  203  191
##   Black     95   77   98   95   88   77   88   94   87   97   92   92  111
##   Hisp      76   66   91   97   89  109   78   64   81   64   89   75   91
##   Asian     19   21   22   26   25   23   23   29   24   38   35   35   28
##   NHoPI      2    4    1    0    3    1    5    6    5    5    3    4    3
##   AI/AN      5    6    6   11    6    7    4    5    3    6    5    9    5
##   Multi     47   50   41   23   22   25   17   29   15   15   13   19   12
##   Unknown    0    0    0    0    0    0    0    0    0    0    0    0    0
##
##           2018
##   White    121
##   Black     73
##   Hisp      38
##   Asian     14
##   NHoPI      4
##   AI/AN      2
##   Multi      9
##   Unknown    0
```

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

```
##
##                 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##   MSM            292  313  333  300  314  352  295  283  270  254  274
##   IDU             38   42   34   26   27   33   29   22   20   23   35
##   MSM/IDU         64   43   51   33   45   30   48   43   35   29   25
##   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   82   85   73   68   39   42   41   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   64   76   84   74   64   74  117   88   94   76
##
##                 2016 2017 2018
##   MSM            222  233  118
##   IDU             28   19   21
##   MSM/IDU         27   25   20
##   Transfus         0    0    0
##   Hemo             0    0    0
##   Hetero          64   49   32
##   Ped              4    6    3
##   F Pres Hetero    0    0    0
##   NIR             92  109   67
```

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   308  310  320  279  311  315  276  283  244  224  222  203  191
##   Black    95   77   98   95   88   77   88   94   87   97   92   92  111
##   Hisp     76   66   91   97   89  109   78   64   81   64   89   75   91
##   Asian    19   21   22   26   25   23   23   29   24   38   35   35   28
##   Native    7   10    7   11    9    8    9   11    8   11    8   13    8
##   Multi    47   50   41   23   22   25   17   29   15   15   13   19   12
##
##          2018
##   White   121
##   Black    73
##   Hisp     38
##   Asian    14
##   Native    6
##   Multi     9
```

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

```
##
##                2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
##   MSM           356  356  384  333  359  382  343  326  305  283  299  249
##   Hetero        156  134  158  169  147  132  113  159  129  139  121  156
##   Blood/Needle   40   44   37   29   38   43   35   25   25   27   39   32
##
##                2017 2018
##   MSM           258  138
##   Hetero        158   99
```

```
##   Blood/Needle   25   24
```

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

```
##
##           2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
##   MSM      356  356  384  333  359  382  343  326  305  283  299  249  258
##   non-MSM  196  178  195  198  185  175  148  184  154  166  160  188  183
##
##           2018
##   MSM      138
##   non-MSM  123
```

5 AIDS at Diagnosis

5.1 AIDS at initial diagnosis?

```
##
## FALSE  TRUE
##  5115  1689
```

5.2 Years of AIDS diagnosis represented:

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
##  164  205  220  264  269  238  231  202  176  159  169  136  160  105   17
## <NA>
## 4089
```

5.3 Quarters of AIDS diagnosis represented:

```
##
##    1    2    3    4 <NA>
##  690  719  663  635 4097
```

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  739    0    0
##          TRUE  3506    0    0    0
##          <NA>     0    0 2559    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  734    0
##          TRUE   3383  123    0
##          <NA>     13 2546    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  112     596    880 2400.0    1070  9344  734
## 2           TRUE    0     151    372  889.8    1020 11570  123
## 3             NA  122     207    569  738.5     997  2022 2546
```

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 123 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  734    0
##          TRUE   3401    0    0
##          <NA>      0 2669    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   0.499   1.510   4.590   6.000  18.000    2669
```

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   725 3394 2585
##     TRUE      9    7   84
```

```
# 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  549  529  573  525  532  545  483  498  448  442  454  433
##          TRUE     3    5    6    6   12   12    8   12   11    7    5    4
##                year
## atunder16count 2017 2018
##          FALSE  434  259
##          TRUE     7    2
```

```
# 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   730 3397 2590
##   TRUE      4    4   79
```

```
# 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 2018
##        FALSE  259
##        TRUE     2
```

```
# 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  50             0  31
```

There are 93 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
##    93  6711
```

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

```
## [1] 6711
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.501   1.520   4.600   6.000  18.000    2585
```

Excluded 93 cases due to age $\leq$ 16 and no observed infPeriod data.

7.3 Maximum window of 18 years

```r
# 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  3362
##                TRUE     39
```

Among those with everHadNegTest=TRUE, we capped 39 cases at aidsUB.

```r
(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>   725
```

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

```r
(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> 2585
```

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 6711 * Diagnoses included: - Year: non-missing, and 2005 onwards - Occurred in WA state - Excluded 14943 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 93 cases due to age $\leq$ 16 and no observed LNT.

8.2 Variable summaries

```
## [1] 6711
```

```
##
## 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    2008    2011    2011    2014    2018
##
## VARIABLE: everHadNegTest
##    Mode   FALSE    TRUE    NA's
## logical     725    3401    2585
##
## VARIABLE: lastNeg_yrs
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.414   1.019   2.442   2.795  31.710    3310
##
## VARIABLE: infPeriod
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.5014  1.5160  4.5980  6.0000 17.9800    2585
```

8.3 Cross-tabulation of KC diagnosis counts for SHAMP

```
# See diagnoses by year
table(dataf$yearDx)
```

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
##  549  529  574  528  532  545  484  498  448  443  454  433  435  259
```

```
# Subset to non-MSM in KC and look again
# Note that "Hetero" excludes "Blood/Needle"
nonMSMkc <- subset(dataf, dx_in_king=='Y' & mode2=='non-MSM')
table(nonMSMkc$yearDx)
```

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
##   91   78   80   99   75   62   62   80   68   84   70   72   77   72
```

```
# Create race3 which is just White/Black/Other
nonMSMkc$race3 <- ifelse(nonMSMkc$race!='White' & nonMSMkc$race!='Black', 'Other', as.character(nonMSMk
nonMSMkc$race3 <- factor(nonMSMkc$race3, levels=c('White', 'Black', 'Other'),
                         labels=c('White', 'Black', 'Other'))
table(nonMSMkc$race3, nonMSMkc$yearDx)
```

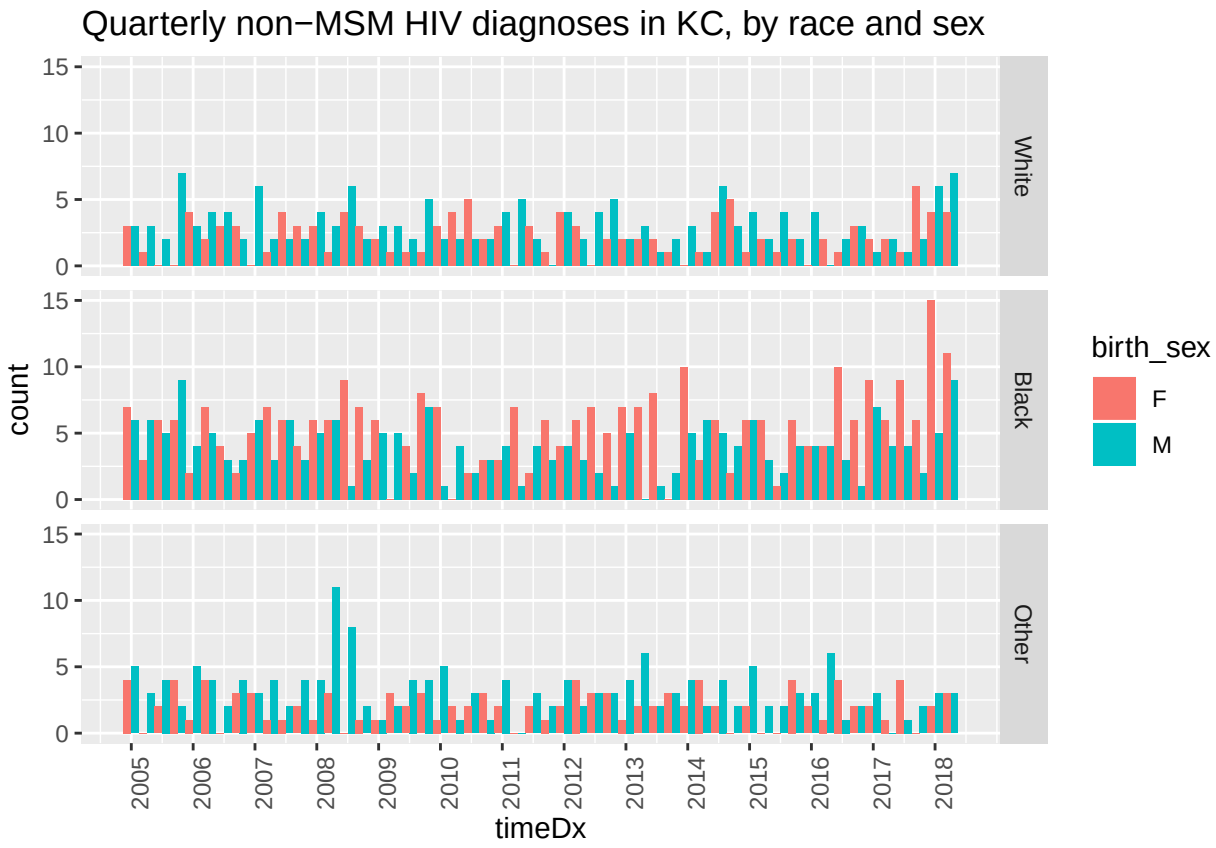
```
##
##         2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
##   White   19   25   20   26   18   22   18   24   15   23   18   15   17
##   Black   48   30   40   43   37   22   30   32   30   41   34   36   47
##   Other   24   23   20   30   20   18   14   24   23   20   18   21   13
##
```

```
##         2018
##   White   21
##   Black   40
##   Other   11
```

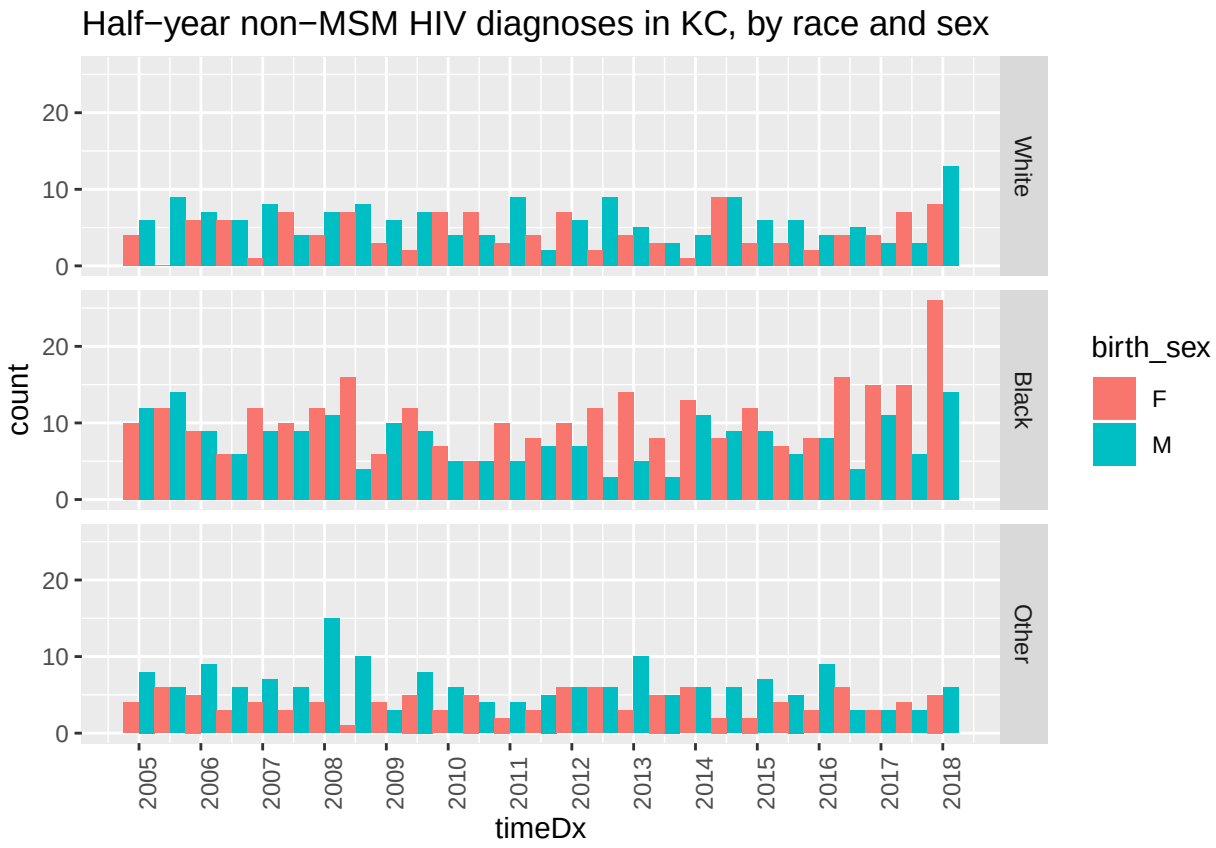
```
table(nonMSMkc$race3, nonMSMkc$timeDx)
```

```
##
##         2005 2005.25 2005.5 2005.75 2006 2006.25 2006.5 2006.75 2007
##   White    6       4      2       7    7       6      7       5    6
##   Black   13       9     11      15    6      12      7       5   11
##   Other    9       3      6       6    6       8      2       7    6
##
##         2007.25 2007.5 2007.75 2008 2008.25 2008.5 2008.75 2009 2009.25
##   White       3      6       5    7       4     10       5    5       4
##   Black      10     12       7   11      12     10      10   11       5
##   Other       5      3       6    5      14      8       3    2       5
##
##         2009.5 2009.75 2010 2010.25 2010.5 2010.75 2011 2011.25 2011.5
##   White      3       6    5       6      7       4    7       5      5
##   Black      6      15    8       4      4       6    7       8      6
##   Other      6       7    6       3      5       4    6       0      5
##
##         2011.75 2012 2012.25 2012.5 2012.75 2013 2013.25 2013.5 2013.75
##   White       1    8       5      4       7    4       5      3       3
##   Black       9    8       9      9       6   12       7      9       2
##   Other       3    6       6      6       6    5       8      4       6
##
##         2014 2014.25 2014.5 2014.75 2015 2015.25 2015.5 2015.75 2016
##   White    3       2     10       8    5       4      5       4    4
##   Black   15       9     11       6   12       9      3      10    8
##   Other    6       6      6       2    7       2      2       7    5
##
##         2016.25 2016.5 2016.75 2017 2017.25 2017.5 2017.75 2018 2018.25
##   White       2      3       6    3       4      2       8   10       11
##   Black       8     13       7   16      10     13       8   20       20
##   Other       7      5       4    5       1      5       2    5        6
```

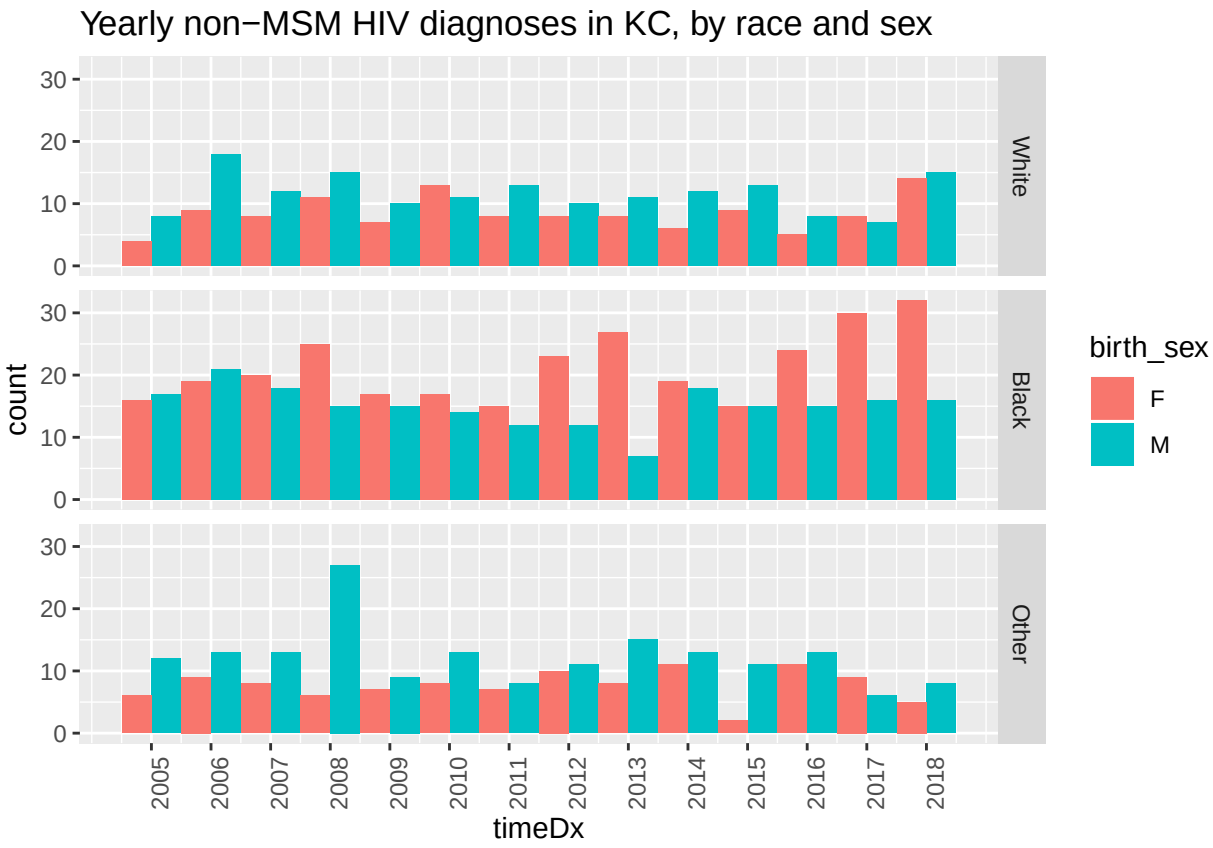
```
ggplot(nonMSMkc, aes(x=timeDx, fill=birth_sex)) +
    geom_histogram(position='dodge', binwidth=0.25) +
    facet_grid(race3~.) +
    scale_x_continuous(breaks=seq(2005,2018,by=1)) +
    theme(axis.text.x=element_text(angle=90,hjust=1)) +
ggtitle('Quarterly non-MSM HIV diagnoses in KC, by race and sex')
```




```
ggplot(nonMSMkc, aes(x=timeDx, fill=birth_sex)) +
    geom_histogram(position='dodge', binwidth=0.5) +
    facet_grid(race3~.) +
    scale_x_continuous(breaks=seq(2005,2018,by=1)) +
    theme(axis.text.x=element_text(angle=90,hjust=1)) +
ggtitle('Half-year non-MSM HIV diagnoses in KC, by race and sex')
```




```
ggplot(nonMSMkc, aes(x=timeDx, fill=birth_sex)) +
    geom_histogram(position='dodge', binwidth=1) +
    facet_grid(race3~.) +
    scale_x_continuous(breaks=seq(2005,2018,by=1)) +
    theme(axis.text.x=element_text(angle=90,hjust=1)) +
ggtitle('Yearly non-MSM HIV diagnoses in KC, by race and sex')
```


Yearly non−MSM HIV diagnoses in KC, by race and sex
count
timeDx
White
Black
Other
birth_sex
F
M
0
10
20
30
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018