WA State HIV Testing Histories - Data Exploration and Formatting

Martina Morris and Jeanette Birnbaum

June 24, 2014

1 Data Structure

```
str(dataf)

## 'data.frame': 25233 obs. of  19 variables:
##  $ FirstVL         : num  658 19914 NA 51 9050 ...
##  $ FirstCD4cnt     : num  566 243 1406 711 858 ...
##  $ tth_ever_neg    : chr  NA NA NA NA ...
##  $ 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 ...
##  $ 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 ...
##  $ TTH_lneg_DT_FLAG: chr  NA NA NA NA ...
##  $ LAG_LNEG_HDX_DT : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TTH_ppos_DT_FLAG: chr  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 = 25233

3 Raw Variable Summaries

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
```

```r
# 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

```r
################################################################ SUBSET THE DATA - INITIAL RESTRICTIONS
year_min <- 2005
year_max <- 2013

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

## hst_included
## FALSE  TRUE  <NA>
##  5478 19752     3

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

## yearDx_included
## FALSE  TRUE  <NA>
## 19052  6038   143

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

## yearDx_missing
## FALSE  TRUE
## 25090   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
## 25016   217

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

## hst_included
## FALSE  TRUE  <NA>
##  5478 19752     3

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

## yearDx_included
## FALSE  TRUE
## 14940  4812
```

```
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
## 4812

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

## [1] 4812
```

Excluded 20421 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
##  560  543  584  540  546  556  496  517  470
```

Quarters of initial diagnosis represented:

```
##
##    1    2    3    4 <NA>
## 1280 1251 1133 1139    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)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2005    2007    2009    2009    2012    2014

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

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
```

```
##   White    340  345  344  288  319  317  281  294  252
##   Black    104   82  104  103   91   79   91   98   91
##   Hisp      76   65   90   94   86  105   76   63   79
##   Asian     20   24   22   28   25   26   25   31   24
##   NHoPI      2    5    3    0    2    1    5    7    8
##   AI/AN      9    6    6   12    5    9    5    5    5
##   Multi      9   16   15   15   18   19   13   19   11
##   Unknown    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            295  312  336  300  322  347  296  281
##   IDU             40   42   32   26   26   32   30   22
##   MSM/IDU         61   45   48   31   41   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   22   23
##   Ped              0    3    2    2    9   10    6    4
##   F Pres Hetero   22   17   29   25   35   19   18   16
##   NIR             71   70   82   95   74   72   77  131
##
##                 2013
##   MSM            273
##   IDU             20
##   MSM/IDU         35
##   Transfus         0
##   Hemo             0
##   Hetero          20
##   Ped              4
##   F Pres Hetero   18
##   NIR            100
```

```
########################################################################## 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)
})
```

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

```
##
##          2005 2006 2007 2008 2009 2010 2011 2012 2013
```

```
##   White   340  345  344  288  319  317  281  294  252
##   Black   104   82  104  103   91   79   91   98   91
##   Hisp     76   65   90   94   86  105   76   63   79
##   Asian    20   24   22   28   25   26   25   31   24
##   Native   11   11    9   12    7   10   10   12   13
##   Multi     9   16   15   15   18   19   13   19   11
```

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

```
##
##                2005 2006 2007 2008 2009 2010 2011 2012 2013
##   MSM           356  357  384  331  363  374  343  321  308
##   Hetero        162  141  165  180  148  140  117  170  138
##   Blood/Needle   42   45   35   29   35   42   36   26   24
```

4.4.1 AIDS at diagnosis

AIDS at initial diagnosis?

```
##
## FALSE  TRUE
##  3531  1281
```

Years of AIDS diagnosis represented:

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 <NA>
##  166  209  215  257  278  235  232  202  167   13 2838
```

Quarters of AIDS diagnosis represented:

```
##
##    1    2    3    4 <NA>
##  507  509  486  469 2841
```

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
dataf <- transform(dataf, everHadNegTest = ifelse(tth_ever_neg ==
    "Y", TRUE, ifelse(tth_ever_neg == "N", FALSE, NA)))
with(dataf, table(everHadNegTest, tth_ever_neg, useNA = "always"))
```

```
##                tth_ever_neg
## everHadNegTest    D    N    R    Y <NA>
##          FALSE    0  511    0    0    0
##          TRUE     0    0    0 2182    0
##          <NA>   364    0    6    0 1749
```

```
# 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     2  509    0
##          TRUE   2099   83    0
##          <NA>     15 2104    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   354.0    596 596.0     838 1080
## 2           TRUE    0   181.0    431 931.4    1118 9938
## 3             NA  122   210.5    569 790.1    1274 2022
##   NA's
## 1  509
## 2   83
## 3 2104
```

Editing We have 2 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
```

More editing We have 83 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  592    0
##          TRUE   2116    0    0
##          <NA>      0 2104    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
##  -2.000   0.643   2.030   5.230   7.490  18.000    2104
```

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

## [1] 1

# 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   589 2112 2039
##     TRUE      3    4   65

# 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  557  538  578  534  536  544  488  505  460
##          TRUE     3    5    6    6   10   12    8   12   10

# 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   591 2114 2043
##   TRUE      1    2   61
```

```
# 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 2013
##        FALSE  461
##        TRUE     9
```

```
# 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   1
##          TRUE    1   0       0        0    0      1   0
##          <NA>    1   0       0        0    0      1  38
##                mode
## everHadNegTest F Pres Hetero NIR
##          FALSE             0   0
##          TRUE              1   1
##          <NA>              0  25
```

Diagnoses younger than age 16 There are 68 cases who do not have a date of last negative test and may not fit the assumption of TID=age-16. Of those, 6 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
##    68  4744
```

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

```
## [1] 4744
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.649   2.030   5.240   7.560  18.000    2039
```

Excluded 68 cases due to age≤16 and no observed infPeriod data.

4.7 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  2096
##                TRUE     20
```

Among those with everHadNegTest=TRUE, we capped 20 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>   589
```

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> 2039
```

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] 4744

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