Philadelphia Testing Histories

Jeanette Birnbaum

November 10, 2015

1 Data Basics

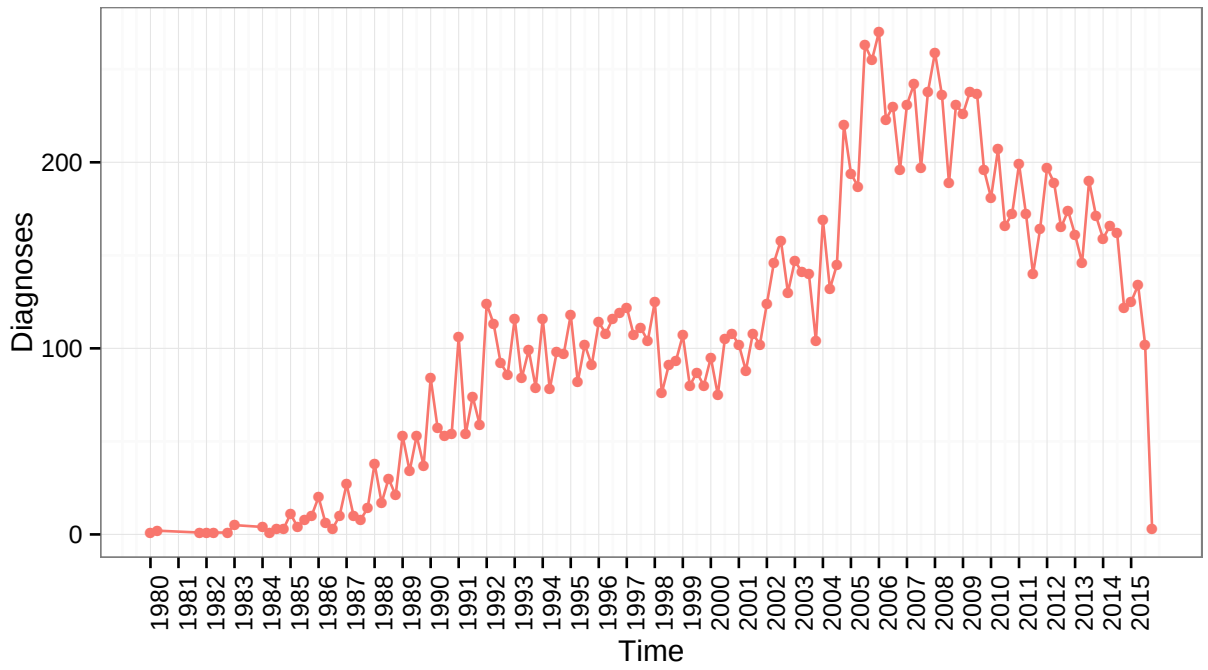
1.1 Initial dataset

```
# Size of formatted data
nrow(dataf)

## [1] 15037

# Years of data
table(dataf$yearDx)

##
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
##    3    1    3    5   11   33   39   59  106  177  248  293  415  378  389  393
## 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011
##  457  444  385  354  383  400  558  532  666  899  919  908  915  897  726  675
## 2012 2013 2014 2015
##  725  668  609  364
```

1.2 Diagnoses over time

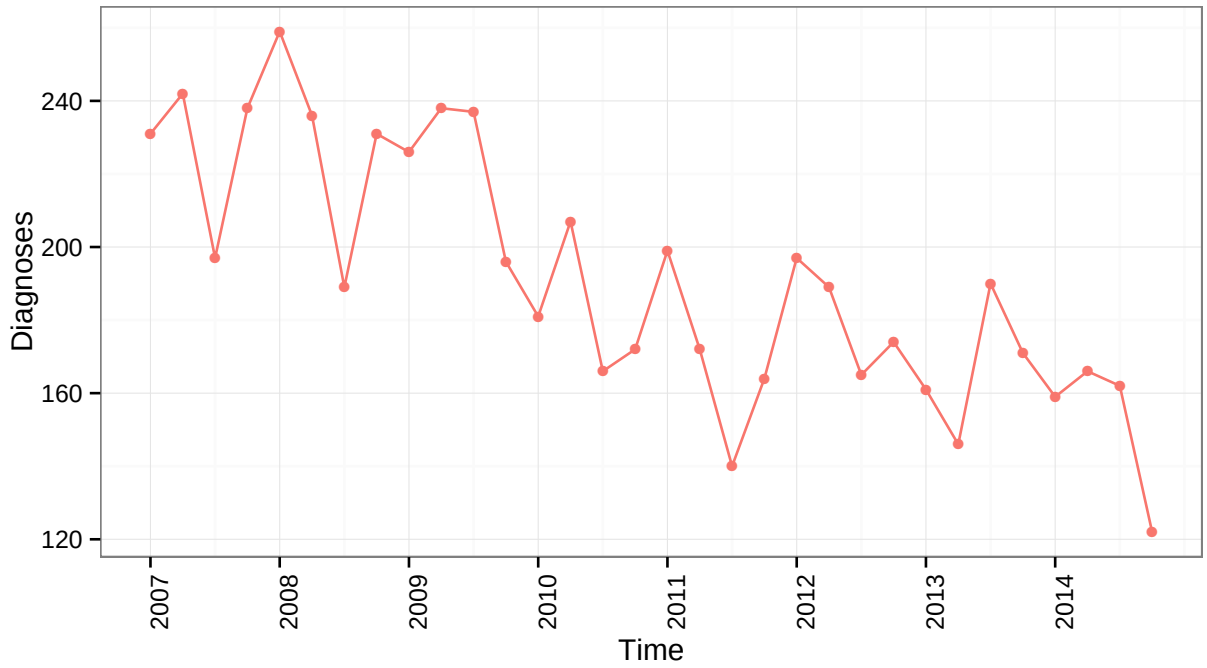
1.2.1 All time periods

```
plotDiagnoses(dataf)
```



1.2.2 Only 2007-2014

```
plotDiagnoses(subset(dataf, yearDx >= 2007 & yearDx <= 2014))
```



1.2.3 Subset decision

Let's limit to 2008-2013, assuming stable reporting by 2008 and no reporting delay affecting 2013 data.

```
dataS <- subset(dataf, yearDx >= 2008 & yearDx <= 2013)
```

2 Impact of missing date info

2.1 Identifying cases with missing date info

```
# Identify cases with some missing date info, using the 'flag' variable that the
# formatting script creates
(flags <- unique(dataS$flag))

##  [1] " Missing day"
##  [2] ""
##  [3] " Missing month"
##  [4] " everHadNegTest inconsistent with infPeriod"
##  [5] " Missing day; Missing day special case"
##  [6] " Missing month; Missing day"
##  [7] " Missing day; infPeriod > aidsUB"
##  [8] " Missing month; Missing month special case"
##  [9] " Missing day; everHadNegTest inconsistent with infPeriod"
## [10] " infPeriod > aidsUB"
## [11] " Missing month; infPeriod > aidsUB"
## [12] " Illogical last negative; everHadNegTest inconsistent with infPeriod"
## [13] " Missing day; Missing day special case; Illogical last negative; everHadNegTest inconsistent with infPeriod"

missFlags <- flags[grepl("Missing", flags)]
missCases <- dataS$flag %in% missFlags

# Tabulate those cases with some missing date info (missDate) against
```

```r
# everHadNegTest. There will be values in all cells because people can have
# missing date info in either their dx date, lneg date, or both
(missTable <- table(missDate = missCases, everHadNegTest = dataS$everHadNegTest,
    useNA = "ifany"))
```

```
##         everHadNegTest
## missDate FALSE TRUE <NA>
##    FALSE  1127  338 1671
##    TRUE    127 1264   79
```

```r
# Marginal values for everHadNegTest
(missColSum <- colSums(missTable))
```

```
## FALSE  TRUE  <NA>
##  1254  1602  1750
```

```r
# Marginal values for missDate, missing date info
(missRowSum <- rowSums(missTable))
```

```
## FALSE  TRUE
##  3136  1470
```

Of the 1602 people who have had a prior negative test, 78.9% of them have some missing date information in either their diagnosis date or their last negative test date.

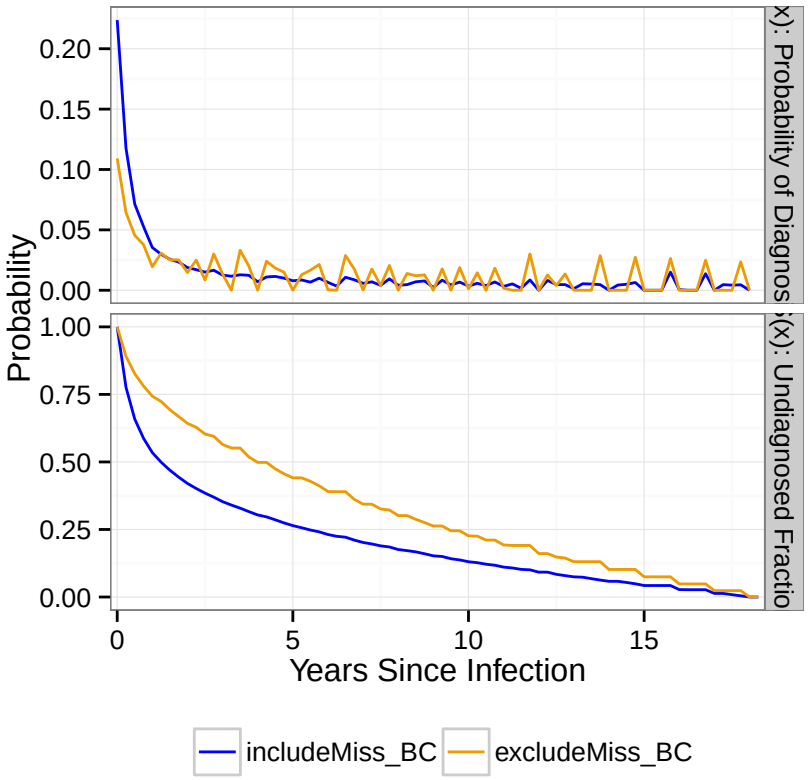
This means that when records with missing date information are not used to estimate the TID, 78.77% of the records used to estimate the TID are those with no prior negative test.

If we make some assumptions (e.g., impute day=15th when day is missing) to allow the use of the records with missing date info, the percentage of records used to estimate the TID who have no prior negative test goes down to 43.91%.

2.2 TID including vs excluding records with missing date info

```r
# TID excluding records having missing date info: just remove those cases who are
# everHadNegTest=TRUE but have missing date info
excludeInf <- dataS$infPeriod[!(missCases & !is.na(dataS$everHadNegTest) & dataS$everHadNegTest)]
excludeMissTID <- estimateTID(excludeInf, intLength = 0.25)

# TID including records having missing date info
includeMissTID <- estimateTID(dataS$infPeriod, intLength = 0.25)


# Combine both base cases and look at them
bothBC <- list(includeMiss_BC = includeMissTID[["base_case"]], excludeMiss_BC = excludeMissTID[["base_case"]])
class(bothBC) <- append(class(bothBC), "TID")
```

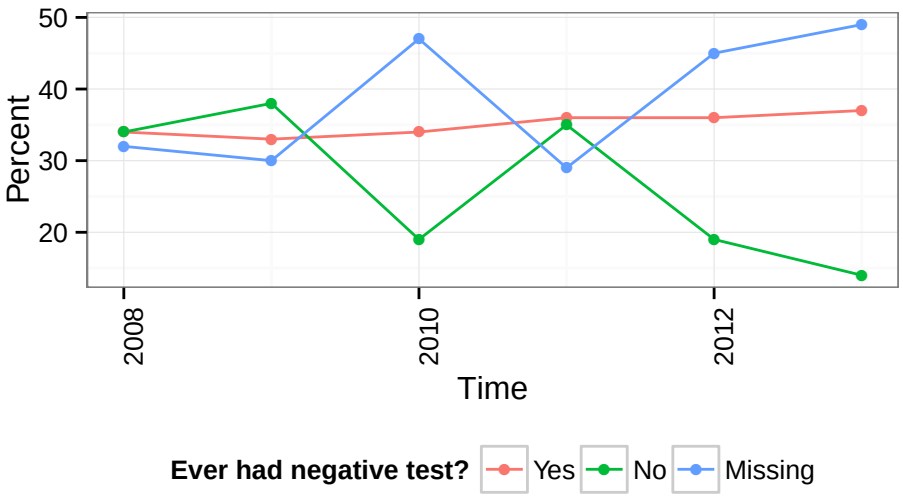
```r
plot(bothBC, intLength = 0.25)
```



3 Impact of time trends on TID

3.1 Trends in everHadNegTest

```
plotTestHist(dataS)
```



No's are being replaced by Missings. This will impact the TID via the assumption we make for the No's.

3.2 TID by year

3.2.1 Trend by year among all non-missing everHadNegTest

```
#################### First look at the infPeriod vector

# InfPeriod over time
by(dataS$infPeriod, dataS$yearDx, function(a) mean(a, na.rm = T))

## dataS$yearDx: 2008
## [1] 8.066162
## ------------------------------------------------------------
## dataS$yearDx: 2009
## [1] 8.28834
## ------------------------------------------------------------
## dataS$yearDx: 2010
## [1] 6.306386
## ------------------------------------------------------------
## dataS$yearDx: 2011
## [1] 8.181083
## ------------------------------------------------------------
## dataS$yearDx: 2012
## [1] 5.848472
## ------------------------------------------------------------
## dataS$yearDx: 2013
## [1] 4.823777

mean(dataS$infPeriod, na.rm = T)

## [1] 7.203842

oneway.test(infPeriod ~ yearDx, data = dataS)

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  infPeriod and yearDx
## F = 20.334, num df = 5.0, denom df = 1242.1, p-value < 2.2e-16

# Significant difference over time
```

There is a significant difference over time. However, this could be driven by the decrease in No's over time.

3.3 Trend by year among everHadNegTest=TRUE, i.e. those with an observed infPeriod

```
# Repeat the test for trend over time, but only use records where
# everHadNegTest=TRUE
dataS <- transform(dataS, infPeriodYES = ifelse(!is.na(everHadNegTest) & everHadNegTest,
    infPeriod, NA))
by(dataS$infPeriodYES, dataS$everHadNegTest, summary)

## dataS$everHadNegTest: FALSE
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      NA      NA      NA     NaN      NA      NA    1254
## ------------------------------------------------------------
## dataS$everHadNegTest: TRUE
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00274  0.43840  0.97810  2.05400  2.23400 17.98000

by(dataS$infPeriodYES, dataS$yearDx, function(a) mean(a, na.rm = T))

## dataS$yearDx: 2008
## [1] 2.176385
## ------------------------------------------------------------
## dataS$yearDx: 2009
```

```
## [1] 2.06833
## ------------------------------------------------------------
## dataS$yearDx: 2010
## [1] 2.225494
## ------------------------------------------------------------
## dataS$yearDx: 2011
## [1] 2.10448
## ------------------------------------------------------------
## dataS$yearDx: 2012
## [1] 1.785577
## ------------------------------------------------------------
## dataS$yearDx: 2013
## [1] 1.950721

mean(dataS$infPeriodYES, na.rm = T)

## [1] 2.054293

oneway.test(infPeriodYES ~ yearDx, data = dataS)

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  infPeriodYES and yearDx
## F = 0.84104, num df = 5.00, denom df = 732.75, p-value = 0.5208

# No longer a significant difference over time
```

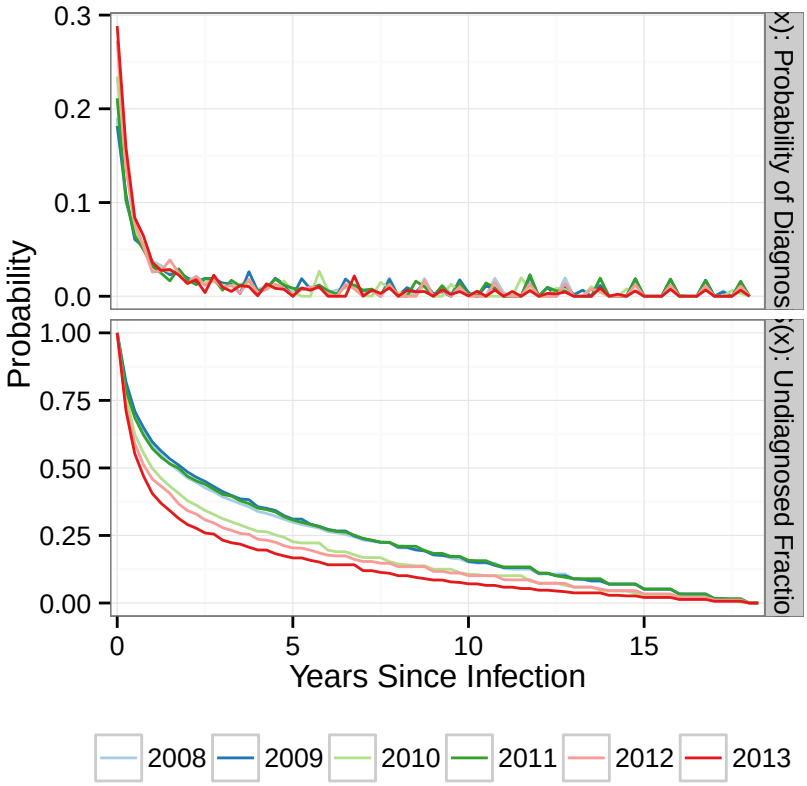
Now there is no longer a significant trend by year.

3.4 Impact on TID of trend in No's

If we estimate the TID by year using all non-missing everHadNegTest records, then we would expect to estimate faster times from infection to diagnosis as No's are replaced by Missings.

```
#################### Translate this to the TID over time

# Estimate TID by year and aggregate the Base Cases
yearlyTIDs <- lapply(unique(dataS$yearDx), function(yr) {
    dataY <- subset(dataS, yearDx == yr)
    # print(unique(dataY$timeDx))
    TID <- estimateTID(dataY$infPeriod, intLength = 0.25)
    return(TID[["base_case"]])
})
names(yearlyTIDs) <- unique(dataS$yearDx)
class(yearlyTIDs) <- append(class(yearlyTIDs), "TID")
```

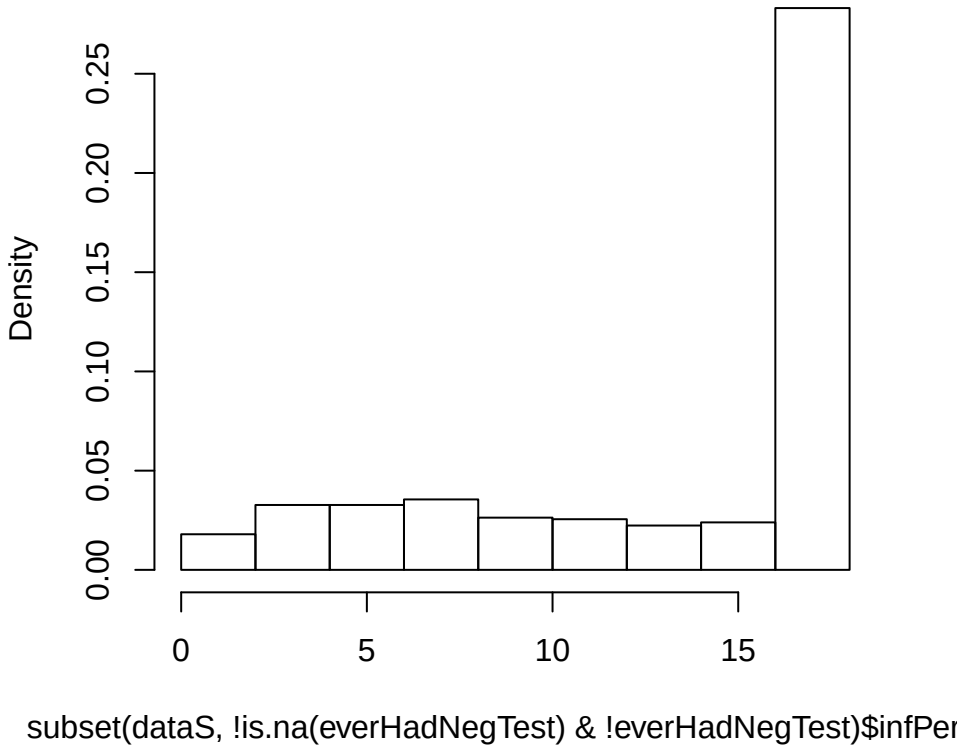
```
plot(yearlyTIDs, intLength = 0.25)
```



The trend is as expected, that in later years when there are more Missings than No's, the TID is shifted towards shorter times from infection to diagnosis.

4 Distribution of infPeriod for No's

```
# Summary
summary(subset(dataS, !is.na(everHadNegTest) & !everHadNegTest)$infPeriod)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   17.98   13.78   17.98   17.98

# Histogram
hist(subset(dataS, !is.na(everHadNegTest) & !everHadNegTest)$infPeriod, probability = TRUE,
    main = "Probability distribution of infPeriod among Nos")
```



5 Conclusions

5.1 Missing Dates

It seems reasonable to make some assumptions in order to not throw away dates that have missing day or month information. Otherwise, the TID will be overly influenced by the everHadNegTest=No records for whom we impute a LNT date.

5.2 TID over time

It is not possible to determine from the data whether the trend towards fewer No's and more Missing's is real, or whether it is a survey instrument issue. The testing history method currently pools records over all years in order to estimate a single TID that is then applied to each quarterly diagnosis count, regardless of what year. This actually seems reasonable given that we don't know whether the yearly differences in the data are real or artifical. Pooling the data gives us a sort of average over the years.