INSIGHT\_prelim\_analyses

Katie Gostic

10/2/2017

# Analysis of FLU003 birth year distributions

The provided summary data contain information on age and birth year distributions, but not on outcome measures. Without data on the age distributions of outcomes, I can't do power analyses (as below, for FLU002), but I can provide preliminary analyses of observed distributions of H1N1 and H3N2 hospitalized cases across birth years.

The general prediction based on HA imprinting patterns is that those born before 1968 will have imprinting protection against group 1 (H1N1), whereas those born after 1968 are more likely to have imprinting protection against group 2 (H3N2). See figures below.

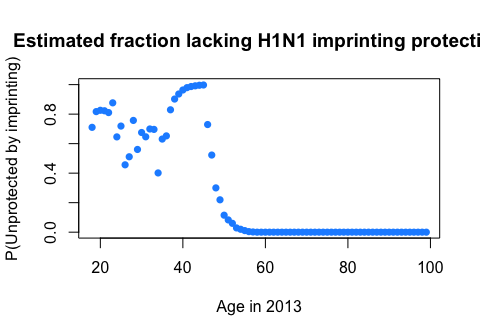
Thus, if we compare the age distributions of H1N1 vs. H3N2 cases in the FLU003 study (where all cases are severe enough to be hospitalized), we should expect to see a greater fraction H3N2 cases in pre-1968 birth years than the fraction of H1N1cases in pre-1968 birth years.

##### Estimated probability that each age group LACKS imprinting protection against the challenge strain:

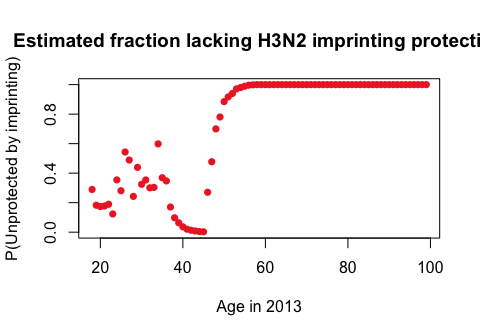
In other words, what fraction of each age group imprinted to the opposite group in childhood? Individuals lacking protective childhood imprinting are predicted to be at relatively high risk of severe disease and hospitalization.

Here, birth year is estimated from the ages given in summary data, assuming the median year of case detection, 2013. Reconstructions of childhood imprinting patterns follow methods developed in Gostic et al., 2016. In the final analysis, I will know individual patient birth years, which may improve power.

## Import imprinting reconstructions for USA  
load('../2017\_seasonal\_flu/Seasonal\_group\_weights.RData')  
# Keep only 2013 reconstructions  
wm.H1 = wm.H1['2013', as.character(18:99)]  
wm.H3 = wm.H3['2013', as.character(18:99)]  
wo.H1 = wo.H1['2013', as.character(18:99)]  
wo.H3 = wo.H3['2013', as.character(18:99)]  
# Plot the patterns  
plot(18:99, wo.H1, ylim = c(0, 1), xlab = 'Age in 2013', ylab = 'P(Unprotected by imprinting)', pch = 16, col = 'dodgerblue', main = 'Estimated fraction lacking H1N1 imprinting protection')



plot(18:99, wo.H3, ylim = c(0, 1), xlab = 'Age in 2013', ylab = 'P(Unprotected by imprinting)', pch = 16, col = 'firebrick2', main = 'Estimated fraction lacking H3N2 imprinting protection')



set.seed(11)

by.dist = read.csv('~/Dropbox/R/INSIGHT/FLU003\_birthyear.csv')  
by.table = matrix(NA, 2, 2, dimnames = list(c('H1N1', 'H3N2'), c('post.1968', 'pre.1968'))) # Empty matrix  
by.table[1,1] = sum(subset(x = by.dist, BirthYear >= 1968, select = H1N1))  
by.table[2,1] = sum(subset(x = by.dist, BirthYear >= 1968, select = H3N2))  
by.table[1,2] = sum(subset(x = by.dist, BirthYear < 1968, select = H1N1))  
by.table[2,2] = sum(subset(x = by.dist, BirthYear < 1968, select = H3N2))  
test = chisq.test(by.table)  
print('Proportion of H1N1 or H3N2 cases in pre- and post-1968 birth years')

## [1] "Proportion of H1N1 or H3N2 cases in pre- and post-1968 birth years"

prop.table(by.table, 1)

## post.1968 pre.1968  
## H1N1 0.3803245 0.6196755  
## H3N2 0.2322275 0.7677725

print('Observed counts')

## [1] "Observed counts"

by.table

## post.1968 pre.1968  
## H1N1 375 611  
## H3N2 196 648

print('Expected counts if birth year distribution is independent of subtype')

## [1] "Expected counts if birth year distribution is independent of subtype"

test$expected

## post.1968 pre.1968  
## H1N1 307.6536 678.3464  
## H3N2 263.3464 580.6536

print('Observed - expected')

## [1] "Observed - expected"

by.table-test$expected

## post.1968 pre.1968  
## H1N1 67.34645 -67.34645  
## H3N2 -67.34645 67.34645

test

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: by.table  
## X-squared = 45.775, df = 1, p-value = 1.326e-11

#### Conclusions:

As expected, there is a relative excess of FLU003 H3N2 cases in older birth years (born before 1968), and a relative dearth of H1N1 cases in the same birth cohorts. A chi square test indicates that differences in observed birth year distributions are not independent of subtype. These patterns are consistent with childhood imprinting patterns, although additional analyses are necessary to examine this potential link in a more rigorous way.

The proposed analyses of INSIGHT data will explore these patterns in greater detail, considering the effect of several variables (age, imprinting status, vaccination status, antiviral use), on several severity outcome measures (ICU admission, duration of hospitalization, etc.).

# Power analysis of FLU002 data

### FLU002 Logistic regression analyses

Use a logistic regression model to predict the probability of different severity outcomes from FLU002 (hospitalization, symptoms > 14 days), as a function of age and imprinting status.

Note: In the full analysis we will consider other factors such as vaccination status and antiviral use, which may also influence the probability of hospitalization, or of symptoms > 14 days. But since these preliminary data do not include information on case history, we omit these factors from the power analysis.

#### Basic strategy:

* Simulate data that assumes an individual's probability of hospitalization, or of symptoms lasting >= 14 days depends on age and imprinting status.
* Draw a number of simulated data points from each age group that matches the number of observations in the FLU002 data.
* Fit a logistic regression model to the simulated data and test whether the model is able to detect a significant relationship between imprinting and probability of hospitalization, or of symptoms >=14d, given the sample size and age distribution of available data.
  + Test three imprinting effect sizes.

# 1. Power analysis for effect of imprinting on probability of hospitalization

Assumptions:

* Assume the baseline probability of hospitalization is 0.039 for H1N1 and about 0.007 for H3N2. These values match the percentage of hospitalized cases in the provided summary data.
* Assume probability of hospitalization is roughly four times as high for adults in the 70+ age group, as for younger adults. This is consistent with patterns in the provided summary data:
  + H1N1: 3.72% of adults under age 70 hospitalized, vs 14.8% of adults over age 70 hospitalized
  + H3N2: 0.58% of adults under age 70 hospitalized, vs. 4.4% of adults over age 70 hospitalized
* Assume individuals who lack imprinting protection are X times as likely to become hospitalized, in comparison to individuals protected by their childhood exposures:
  + 1.25 times as likely (small effect size)
  + 1.75 times as likely (moderate effect size)
  + 2 times as likely (large effect size)

Set parameters:

# Set assumed baseline probabilities  
baseline.H1 = .039  
baseline.H3 = .007  
# Set assumed effects of age and imprinting protection  
age.factor = 4  
small.imp.factor = 1.25  
med.imp.factor = 1.7  
large.imp.factor= 2  
# How many simulations?  
n.sims = 1000

### Analysis using H1N1 parameters and sample sizes

Simulate 1000 datasets, fit a logistic model to each, and then assess what fraction of times the imprinitng coefficient is statistically significant at the p=0.05 level.

* Simulate data using the same number of cases in each age group as in the observed data.
* The fraction of times the imprinting coefficient is significant at p = 0.05 is equivalent to power.

# Initialize matrix of results  
H1.pvals = matrix(NA, n.sims, 3, dimnames = list(NULL, c('small', 'med', 'large')))  
  
  
## H1N1 simulation  
for(ss in 1:n.sims){  
no.pts.H1 = c(627, 526, 372, 260, 98, 27) # How many patients in each age group?  
bin.starts = c(18, 30, 40, 50, 60 ,70) # Define beginning and end of age bins  
bin.ends = c(29, 39, 49, 59, 69, 99)  
# Draw patient ages, using the numbers of observations in each bin  
simulated.ages = unlist(sapply(1:6, function(xx) sample(x = bin.starts[xx]:bin.ends[xx], size = no.pts.H1[xx], replace = TRUE)))  
  
# Get the probability that each simulated patient lacks imprinting, based on their age  
prob.lacks.imprinting = wo.H1[as.character(simulated.ages)]  
# Using the above probabilities, perform a Bernoulli trial. 0 indicates no imprinting, 1 indicates imprinting  
simulated.imprinting.status = sapply(prob.lacks.imprinting, function(pp) rbinom(n = 1, size = 1, prob = 1-pp))  
  
# Set imprinting-specific modifiers  
imp.small = imp.med = imp.large = rep(1, length(simulated.ages))  
# For small effect size, prob of hospitalization is 1.25 as high in individuals lacking imprinting protection  
imp.small[which(simulated.imprinting.status == 0)] = small.imp.factor  
# For med effect size, prob of hospitalization is 1.6 as high in individuals lacking imprinting protection  
imp.med[which(simulated.imprinting.status == 0)] = med.imp.factor  
# For large effect size, prob of hospitalization is 2 times as high in individuals lacking imprinting protection  
imp.large[which(simulated.imprinting.status == 0)] = large.imp.factor  
  
# Set age-specific modifier (prob of hospitalization is 4 times greater for those over age 70)  
age.mod = rep(1, length(simulated.ages))  
age.mod[which(simulated.ages >= 70)] = age.factor  
  
# Calculate probabilities of hospitalization based on age and imprinting status  
p.hosp = cbind(baseline.H1\*imp.small\*age.mod, baseline.H1\*imp.med\*age.mod, baseline.H1\*imp.large\*age.mod); colnames(p.hosp) = c('small', 'med', 'large')  
  
# Perform a Bernoulli trial for each simulated patient to determine if they are hospitalized or not  
sim.hosp.status = p.hosp\*0 # Initialize  
for(ii in 1:length(p.hosp)){  
 sim.hosp.status[ii] = rbinom(1, 1, p.hosp[ii])  
}  
  
# Fit a logistic model to simulated data  
small.fit = glm(sim.hosp.status[,1] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
med.fit = glm(sim.hosp.status[,2] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
large.fit = glm(sim.hosp.status[,3] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
  
H1.pvals[ss, 1] = summary(small.fit)$coefficients[3, 4]  
H1.pvals[ss, 2] = summary(med.fit)$coefficients[3, 4]  
H1.pvals[ss, 3] = summary(large.fit)$coefficients[3, 4]  
}

Power: what % of runs gave a significant result for small, med and large effect sizes?

apply(H1.pvals, 2, function(xx) sum(xx <= 0.05))/n.sims

## small med large   
## 0.115 0.297 0.497

#### Analysis using H3N2 parameters and sample sizes

no.pts.H3 = c(496, 521, 422, 307, 178, 90) # How many patients in each age group?  
bin.starts = c(18, 30, 40, 50, 60 ,70) # Define beginning and end of age bins  
bin.ends = c(29, 39, 49, 59, 69, 99)  
  
# Initialize matrix of results  
H3.pvals = matrix(NA, n.sims, 3, dimnames = list(NULL, c('small', 'med', 'large')))  
  
  
## H3N2 simulation  
for(ss in 1:n.sims){  
# Draw patient ages, using the numbers of observations in each bin  
simulated.ages = unlist(sapply(1:6, function(xx) sample(x = bin.starts[xx]:bin.ends[xx], size = no.pts.H3[xx], replace = TRUE)))  
  
# Get the probability that each simulated patient lacks imprinting, based on their age  
prob.lacks.imprinting = wo.H3[as.character(simulated.ages)]  
# Using the above probabilities, perform a Bernoulli trial. 0 indicates no imprinting, 1 indicates imprinting  
simulated.imprinting.status = sapply(prob.lacks.imprinting, function(pp) rbinom(n = 1, size = 1, prob = 1-pp))  
  
# Set imprinting-specific modifiers  
imp.small = imp.med = imp.large = rep(1, length(simulated.ages))  
# For small effect size, prob of hospitalization is 1.25 as high in individuals lacking imprinting protection  
imp.small[which(simulated.imprinting.status == 0)] = small.imp.factor  
# For med effect size, prob of hospitalization is 1.6 as high in individuals lacking imprinting protection  
imp.med[which(simulated.imprinting.status == 0)] = med.imp.factor  
# For large effect size, prob of hospitalization is 2 times as high in individuals lacking imprinting protection  
imp.large[which(simulated.imprinting.status == 0)] = large.imp.factor  
  
# Set age-specific modifier (prob of hospitalization is 4 times greater for those over age 70)  
age.mod = rep(1, length(simulated.ages))  
age.mod[which(simulated.ages >= 70)] = age.factor  
  
# Calculate probabilities of hospitalization based on age and imprinting status  
p.hosp = cbind(baseline.H3\*imp.small\*age.mod, baseline.H3\*imp.med\*age.mod, baseline.H3\*imp.large\*age.mod); colnames(p.hosp) = c('small', 'med', 'large')  
  
# Perform a Bernoulli trial for each simulated patient to determine if they are hospitalized or not  
sim.hosp.status = p.hosp\*0 # Initialize  
for(ii in 1:length(p.hosp)){  
 sim.hosp.status[ii] = rbinom(1, 1, p.hosp[ii])  
}  
  
# Fit a logistic model to simulated data  
small.fit = glm(sim.hosp.status[,1] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
med.fit = glm(sim.hosp.status[,2] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
large.fit = glm(sim.hosp.status[,3] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
  
H3.pvals[ss, 1] = summary(small.fit)$coefficients[3, 4]  
H3.pvals[ss, 2] = summary(med.fit)$coefficients[3, 4]  
H3.pvals[ss, 3] = summary(large.fit)$coefficients[3, 4]  
}

Power: what % of runs gave a significant result for small, med and large effect sizes?

apply(H3.pvals, 2, function(xx) sum(xx <= 0.05))/n.sims

## small med large   
## 0.034 0.037 0.049

#### Conclusion:

These results that we have well below 80% power to detect imprinting effects on the probability of hospitalization, given the available sample sizes. Thus, I will omit this aim from my proposal.

# 2. Power analysis for effect of imprinting on probability of symptoms lasting >= 14 d

Assumptions:

* Assume the baseline probability of symptoms lasting >= 14 d is 0.148 for H1N1 and about 0.125 for H3N2. These values are consistent with the percentage of cases with symptoms lasting >= 14 d in the provided summary data.
* Assume probability of symptoms >= 14d is roughly equal for adults in the 70+ age group and for younger adults. This goes against the conventional wisdowm that older adults are at high risk of severe infection, but the FLU002 summary data suggest the oldest patients are actually less likely to have symptoms after 14 days than younger adults:
  + H1N1: 14.8% of adults under age 70 had symptoms >= 14 days, vs 11.1% of adults over age 70 with symptoms >= 14d
  + H3N2: 12.6% of adults under age 70 had symptoms >= 14 days, vs. 10% of adults over age 70 with symptoms >= 14d
* Assume individuals who lack imprinting protection are X times as likely to have symptoms for 14 days or longer, in comparison to individuals protected by their childhood exposures:
  + 1.25 times as likely (small effect size)
  + 1.75 times as likely (moderate effect size)
  + 2 times as likely (large effect size)

Set parameters:

# Set assumed baseline probabilities  
baseline.H1 = .148  
baseline.H3 = .125  
# Set assumed effects of age and imprinting protection  
age.factor = 1  
small.imp.factor = 1.25  
med.imp.factor = 1.75  
large.imp.factor= 2  
# How many simulations?  
n.sims = 1000

## Simulate 1000 datasets, fit a logistic model to each, and then assess what fraction of times the imprinitng coefficient is significant

* Simulate data using the same number of cases in each age group as in the observed data.
* The fraction of itmes the imprinting coefficient is significant at p = 0.05 is equivalent to power.

#### Analysis using H1N1 parameters and sample sizes

# Initialize matrix of results  
H1.pvals = matrix(NA, n.sims, 3, dimnames = list(NULL, c('small', 'med', 'large')))  
  
no.pts.H1 = c(626, 524, 370, 259, 99, 27) # How many patients in each age group?  
bin.starts = c(18, 30, 40, 50, 60 ,70) # Define beginning and end of age bins  
bin.ends = c(29, 39, 49, 59, 69, 99)  
  
  
## H1N1 simulation  
for(ss in 1:n.sims){  
# Draw patient ages, using the numbers of observations in each bin  
simulated.ages = unlist(sapply(1:6, function(xx) sample(x = bin.starts[xx]:bin.ends[xx], size = no.pts.H1[xx], replace = TRUE)))  
  
# Get the probability that each simulated patient lacks imprinting, based on their age  
prob.lacks.imprinting = wo.H1[as.character(simulated.ages)]  
# Using the above probabilities, perform a Bernoulli trial. 0 indicates no imprinting, 1 indicates imprinting  
simulated.imprinting.status = sapply(prob.lacks.imprinting, function(pp) rbinom(n = 1, size = 1, prob = 1-pp))  
  
# Set imprinting-specific modifiers  
imp.small = imp.med = imp.large = rep(1, length(simulated.ages))  
# For small effect size, prob of hospitalization is 1.25 as high in individuals lacking imprinting protection  
imp.small[which(simulated.imprinting.status == 0)] = small.imp.factor   
# For med effect size, prob of hospitalization is 1.6 as high in individuals lacking imprinting protection  
imp.med[which(simulated.imprinting.status == 0)] = med.imp.factor  
# For large effect size, prob of hospitalization is 2 times as high in individuals lacking imprinting protection  
imp.large[which(simulated.imprinting.status == 0)] = large.imp.factor  
  
# Set age-specific modifier (prob of hospitalization is equal for those over age 70)  
age.mod = rep(1, length(simulated.ages))  
age.mod[which(simulated.ages >= 70)] = age.factor  
  
# Calculate probabilities of symptoms >= 14d based on age and imprinting status  
p.symp = cbind(baseline.H1\*imp.small\*age.mod, baseline.H1\*imp.med\*age.mod, baseline.H1\*imp.large\*age.mod); colnames(p.symp) = c('small', 'med', 'large')  
  
# Perform a Bernoulli trial for each simulated patient to determine if they are hospitalized or not  
sim.duration.status = p.symp\*0 # Initialize  
for(ii in 1:length(p.symp)){  
 sim.duration.status[ii] = rbinom(1, 1, p.symp[ii])  
}  
  
# Fit a logistic model to simulated data  
small.fit = glm(sim.duration.status[,1] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
med.fit = glm(sim.duration.status[,2] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
large.fit = glm(sim.duration.status[,3] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
  
H1.pvals[ss, 1] = summary(small.fit)$coefficients[3, 4]  
H1.pvals[ss, 2] = summary(med.fit)$coefficients[3, 4]  
H1.pvals[ss, 3] = summary(large.fit)$coefficients[3, 4]  
}

Power: what % of runs gave a significant result for small, med and large effect sizes?

apply(H1.pvals, 2, function(xx) sum(xx <= 0.05))/n.sims

## small med large   
## 0.205 0.845 0.958

#### Analysis using H3N2 parameters and sample sizes

no.pts.H3 = c(497, 518, 422, 307, 178, 90) # How many patients in each age group?  
bin.starts = c(18, 30, 40, 50, 60 ,70) # Define beginning and end of age bins  
bin.ends = c(29, 39, 49, 59, 69, 99)  
  
# Initialize matrix of results  
H3.pvals = matrix(NA, n.sims, 3, dimnames = list(NULL, c('small', 'med', 'large')))  
  
for(ss in 1:n.sims){  
# Draw patient ages, using the numbers of observations in each bin  
simulated.ages = unlist(sapply(1:6, function(xx) sample(x = bin.starts[xx]:bin.ends[xx], size = no.pts.H3[xx], replace = TRUE)))  
  
# Get the probability that each simulated patient lacks imprinting, based on their age  
prob.lacks.imprinting = wo.H3[as.character(simulated.ages)]  
# Using the above probabilities, perform a Bernoulli trial. 0 indicates no imprinting, 1 indicates imprinting  
simulated.imprinting.status = sapply(prob.lacks.imprinting, function(pp) rbinom(n = 1, size = 1, prob = 1-pp))  
  
# Set imprinting-specific modifiers  
imp.small = imp.med = imp.large = rep(1, length(simulated.ages))  
# For small effect size, prob of hospitalization is 1.25 as high in individuals lacking imprinting protection  
imp.small[which(simulated.imprinting.status == 0)] = small.imp.factor   
# For med effect size, prob of hospitalization is 1.6 as high in individuals lacking imprinting protection  
imp.med[which(simulated.imprinting.status == 0)] = med.imp.factor  
# For large effect size, prob of hospitalization is 2 times as high in individuals lacking imprinting protection  
imp.large[which(simulated.imprinting.status == 0)] = large.imp.factor  
  
# Set age-specific modifier (prob of hospitalization is equal for those over age 70)  
age.mod = rep(1, length(simulated.ages))  
age.mod[which(simulated.ages >= 70)] = age.factor  
  
# Calculate probabilities of symptoms >= 14d based on age and imprinting status  
p.symp = cbind(baseline.H3\*imp.small\*age.mod, baseline.H3\*imp.med\*age.mod, baseline.H3\*imp.large\*age.mod); colnames(p.symp) = c('small', 'med', 'large')  
  
# Perform a Bernoulli trial for each simulated patient to determine if they are hospitalized or not  
sim.duration.status = p.symp\*0 # Initialize  
for(ii in 1:length(p.symp)){  
 sim.duration.status[ii] = rbinom(1, 1, p.symp[ii])  
}  
  
# Fit a logistic model to simulated data  
small.fit = glm(sim.duration.status[,1] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
med.fit = glm(sim.duration.status[,2] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
large.fit = glm(sim.duration.status[,3] ~ simulated.ages + prob.lacks.imprinting, family = binomial)  
  
H3.pvals[ss, 1] = summary(small.fit)$coefficients[3, 4]  
H3.pvals[ss, 2] = summary(med.fit)$coefficients[3, 4]  
H3.pvals[ss, 3] = summary(large.fit)$coefficients[3, 4]  
}

Power: what % of runs gave a significant result for small, med and large effect sizes?

apply(H3.pvals, 2, function(xx) sum(xx <= 0.05))/n.sims

## small med large   
## 0.168 0.802 0.955

#### Conclusion: We have at least 80% power to detect moderate to large imprinting effects on probabilities that symptoms last 14 days or longer.