**INSIGHT GLM Analysis**

Methods

Data

We analyzed two datasets, INSIGHT 002 and INSIGHT 003. INSIGHT 002 enrolls adult patients 18 and over who present with fever, cough and/or sore throat, and who have a suspected influenza infection. On enrollment, all study participants reported influenza-like illness. Participating physicians recorded demographic information (age, country of enrollment), vaccination status and other relevant medical history for each patient. Infection status was determined to subtype (A/H1N1, A/H3N2, influenza B or no confirmed infection).

INSIGHT 003 enrolled hospitalized patients with a PCR-confirmed influenza diagnosis. The study tracks a number of outcomes, including duration of hospitalization, days in ICU and death.

We excluded five cases who were over the age of 90, and two cases who were under age 18. We chose 90 as an upper age cutoff because we cannot confidently reconstruct historical influenza circulation patterns before 1918. Excluding older cases insures that no one in our data set could have been born prior to the 1918 pandemic. We also excluded ## cases in whom cross-infection by multiple subtypes was detected, or in whom the infecting subtype was unknown.

**Multinomial model**

We obtained 11,679 records from the INSIGHT 002 study, which were collected from 2009 to 2017. Cases were reported from each of the following countries: Denmark, Spain, Germany, Estonia, USA, Belgium, Portugal, Poland, Austria, UK, Australia, Thailand, Argentina, Chile, Greece, Peru and Japan. We defined the northern hemisphere influenza season as spanning October-March, and the southern hemisphere season as spanning April-Sept. Tables 1 and 2 show a breakdown of cases by country and season. SHOULD I EXCLUDE CASES IF <25 WERE OBSERVED IN A GIVEN COUNTRY-SEASON COMBO? We excluded five cases who were over the age of 90. We chose 90 as an upper age cutoff because we cannot confidently reconstruct historical influenza circulation patterns before 1918. Excluding older cases insures that no one in our data set could have been born prior to the 1918 pandemic.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Argentina | Australia | Austria | Belgium | Chile | Denmark | Estonia | Germany | Greece |
| 3540 | 103 | 15 | 1322 | 43 | 202 | 217 | 375 | 428 |
| Poland | Portugal | Spain | Thailand | UK | USA | Japan | Peru |  |
| 362 | 6 | 150 | 3284 | 85 | 989 | 33 | 522 |  |

Table of cases by country.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| NH.09.10 | NH.10.11 | NH.11.12 | NH.12.13 | NH.13.14 | NH.14.15 | NH.15.16 | NH.16.17 |
| 700 | 237 | 281 | 511 | 1351 | 1111 | 1059 | 118 |
| SH.10 | SH.11 | SH.12 | SH.13 | SH.14 | SH.15 | SH.16 |  |
| 279 | 629 | 263 | 839 | 1102 | 1369 | 645 |  |

Table of cases by season.

EXCLUSIONS:

We excluded cases of influenza B, cases that were not identified to subtype, and cases that were classified as coinfections.

**Reconstruction of birth year-specific imprinting patterns**

We estimated the probability that each case in the data imprinted to an H1N1, H2N2 or H3N2 virus during childhood. Imprinting probabilities were reconstructed using previously described methods (CITE), except as noted below. To summarize briefly, we first estimated a case’s birth year based on their age in the year their case was observed. Then, we estimated the probability that the cases’s first childhood influenza exposure occurred 0, 1, 2, etc. years after the birth year. Finally, we cross-referenced these estimates with historical data on the influenza subtypes circulating during each possible year of first, childhood influenza exposure. Thus, we estimated the overall probability of first exposure to H1N1, H2N2 or H3N2.

The only notable modification to previously described methods (CITE), is that previously we estimated birth year as *y-a*, where *y* represents the year of case observation and *a* represents age. We have now updated the methods to better address uncertainty inherent to this approach*.* For example, consider an infant of age 0 observed in May 2000. Using our old method, this case would have been assigned a birth year of *y-a*=2000. However, in reality, this individual could have been born any time between late May 1999 and early May, 2000. This implies 7.5 possible birth months in 1999 [*y-a-1*], and only 4.5 possible birth months in 2000 [*y-a*].

SHIFT BEGINNING OF YEAR TO BEGINNING OF NH FLU SEASON?

We have generalized the above argument to estimate the probability that each case was born in year *y-a-1*, or in year *y-a* (Table ##). Then, we reconstructed imprinting probabilities for each of two possible birth years, and took a weighted average, where weights are determined by the month of observation (Fig. xx).

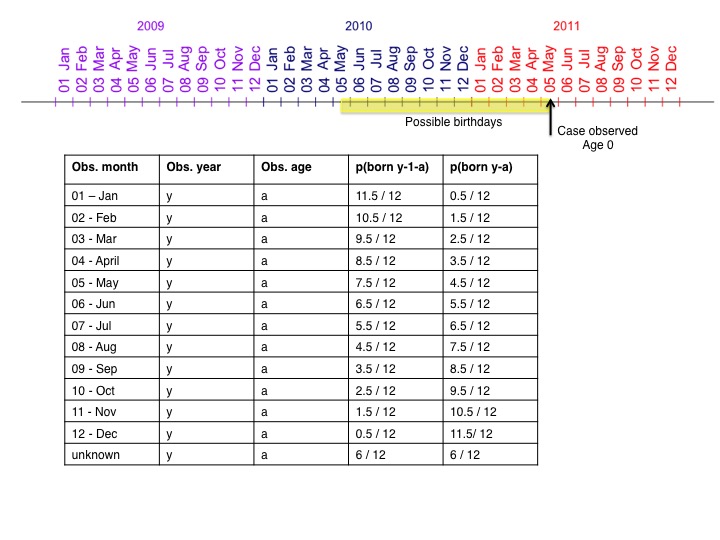


Figure xx or Table xx draft.

**Estimating probabilities of protection**

Using the above methods, we were able to estimate each individual’s probability of childhood imprinting to H1N1, H2N2, H3N2, or for children, the probability of remaining naïve to any influenza A virus. For each individual, pH1N1 +pH2N2 +pH3N2 = 1. From here, we used addition to calculate imprinting probabilities at the group level (Table xx).

|  |  |  |
| --- | --- | --- |
| **Hypothesis:** | **Challenge** | **p(protected)** |
| HA subtype level | H1N1 | *pH1N1* |
| H3N2 | *pH3N2* |
| NA subtype level | H1N1 | *pH1N1* |
| H3N2 | *pH2N2+ pH3N2* |
| HA group level | H1N1 | *p H1N1+ pH2N2* |
| H3N2 | *pH3N2* |

Table xx. Probabilities of protection against an H1N1 or H3N2 challenge, under different imprinting hypotheses.

In all analyses except one, the dataset included only patients with a confirmed influenza A infection. Here, it was clear from the lab results whether each patient had been challenged by H1N1 or H3N2, and probabilities of protection were assigned accordingly (Table xx). However, in one analysis where incidence was the outcome of interest, many cases necessarily lacked a confirmed influenza infection. To determine whether these individuals would most likely have been challenged by H1N1 or H3N2, we downloaded epidemiological data from WHO Flu Net for each country and season in our data set. We then calculated the fraction of influenza A circulation caused by H1N1 or H3N2 in the season and country of interest. If fewer than 30 observations were recorded in the season and country of interest, we used aggregate data from all countries in the relevant season. We then assigned each influenza negative individual in the outpatient data to an H1N1 or H3N2 challenge using a Bernoulli trial, with the probability of H1N1 challenge equal to the frequency of H1N1 circulation. To insure that this protocol was robust, we repeated random challenge assignment ## times, and refit the models using each of the ## resulting augmented data set. We tracked the fraction of times that…

**Models**

We used a generalized additive mixed model with a logit link to describe probabilities of infection in the outpatient data set. Here, Yics was a binary variable that took value 1 if patient i had a confirmed H1N1 or H3N2 infection, and value 0 otherwise. Subscripts c and s indicate the country and season in which the patient was observed. The baseline model included linear fixed effects from patient age, the presence of underlying conditions, and antiviral treatment. The baseline model also included the interaction between recent vaccination and the relevant seasonal challenge subtype (H1N1 or H3N2). We included this interaction because vaccine efficacy is known to differ between H1N1 and H3N2. Country and season were also included in the baseline model as random effects. We classified cases observed from October-March as falling into the corresponding year’s Northern Hemisphere (NH) influenza season, and cases observed from April-September into the corresponding Southern Hemisphere influenza season.

Preliminary exploration of the data suggested a nonlinear relationship between the probability of imprinting protection and probability of infection. We observed a similar nonlinear pattern whether we assumed protection acted at the HA subtype, HA group or NA subtype level (supplementary fig. ##). Thus, we tested used a penalized cubic regression spline on the protection variable of interest in the full model. We tested each full model against a reduced model in which there is a linear fixed effect of protection. We also included random intercepts for the season of observation and country of observation.

DO I NEED WEIGHTS FOR SAMPLE SIZES?

I don’t want to give to much weight to data in age groups where I only have one observation, but I’m not sure if including this data is problematic.

**Preliminary results**

**Data exploration**

We plotted the data against each independent variable of interest to examine patterns of interest.

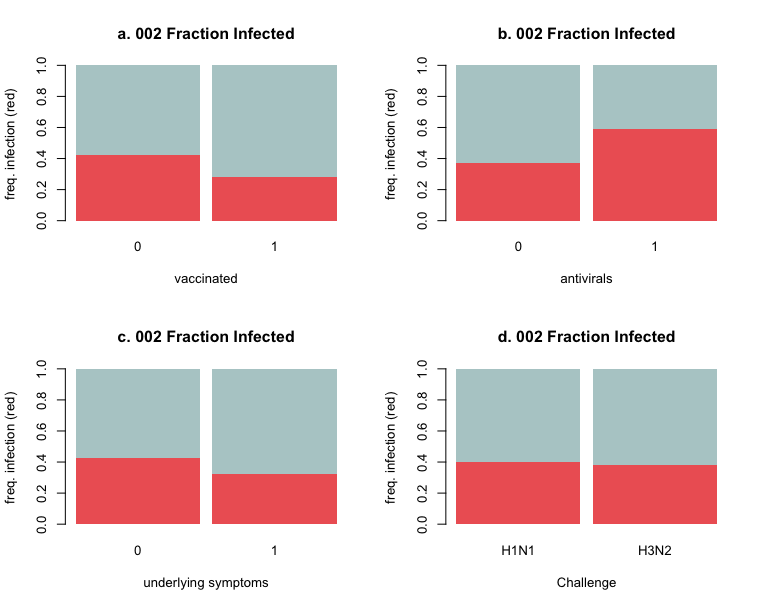


Fig. S1. Preliminary analysis of infection probabilities against binary, categorical variables. (a) It appears that vaccinated individuals have lower probabilities of infection. (b) It appears that individuals using antivirals have higher probabilities of infection. This may occur because these individuals have more severe infections, or other signs and symptoms that would prompt a doctor to recommend antiviral treatment. (c) Individuals with underlying symptoms are slightly less likely to be infected. Perhaps these individuals are more likely to visit the doctor given any resp. infection. (d) On average, those who we assume would have been challenged by an H3N2 virus are no more likely to test positive than those who we assume were challenged by an H1N1 virus.

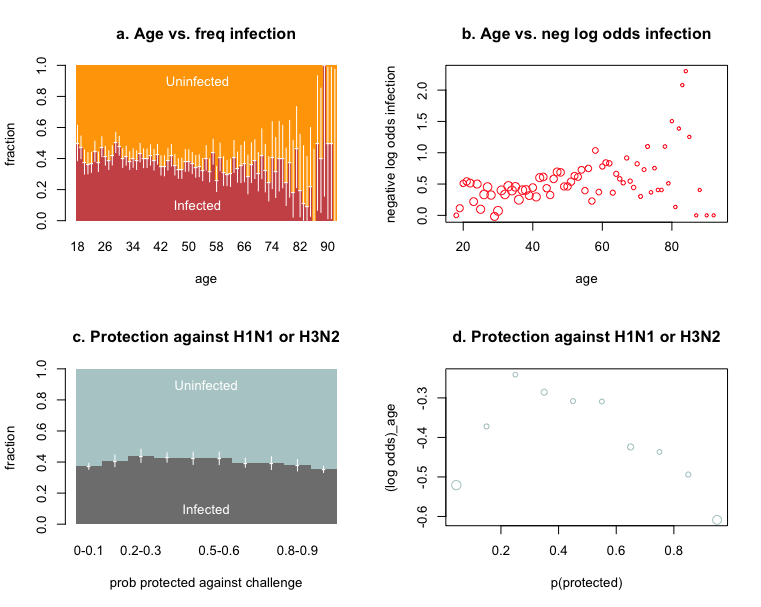


Fig. S2. Age vs. frequency of infection and probability of imprinting protection (binned into 10 groups) vs. frequency of infection. There seems to be a slight linear increase in prob. infection with age, but low case counts in older age groups are throwing off the variance. Can I correct this using weights? Meanwhile, there is a weird parabolic relationship between p(protection) and p(infection). Why?

**Results – 002 Incidence**

Best model contains nonlinear imprinting effects acting at the N subtype level. Effects are significant for an H1N1 challenge, but not for an H3N2 challenge. The next best model (deltaAIC ~=~ 5) contains imprinting at the subtype level for HA. So basically, H2 imprinting doesn’t seem to protect against H1N1, although it may protect against H3N2 a little bit? I’m still uncomfortable with the nonlinear effect shape here. Could this be because of biased imputation? If there is a protective effect, then p(protection) is not independent of infection status.

Formula:

infected ~ s(protected\_N, by = challenge, bs = "cr") + challenge +

age + anydx + anyav + anyvac \* challenge + s(country, bs = "re") +

s(season, bs = "re")

Parametric coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.663175 0.287715 -2.305 0.02117 \*

challengeH3N2 0.450148 0.070855 6.353 2.11e-10 \*\*\*

age 0.001693 0.001888 0.897 0.36998

anydx1 -0.204206 0.050417 -4.050 5.11e-05 \*\*\*

anyav1 0.901210 0.080629 11.177 < 2e-16 \*\*\*

anyvac1 -0.669848 0.088442 -7.574 3.62e-14 \*\*\*

challengeH3N2:anyvac1 0.309808 0.112688 2.749 0.00597 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:

edf Ref.df Chi.sq p-value

s(protected\_N):challengeH1N1 2.2451 9 20.393 0.00106 \*\*

s(protected\_N):challengeH3N2 0.4239 9 1.058 0.15289

s(country) 13.5533 15 6569.466 3.76e-11 \*\*\*

s(season) 13.7374 14 1544.758 1.24e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) = 0.125 Deviance explained = 10%

UBRE = 0.21041 Scale est. = 1 n = 10460