Time Series Analysis of Flu Cases

INTRODUCTION

Modeling & predicting positive flu cases within the United States

01

HISTORY OF THE FLU

Brief look at the flu since 1918 and milestones

02

DATASET SELECTION

Dataset description and details

03

TIME SERIES MODELING

Model seasonal flu trends from 2014 - 2020

04

RECOMMENDATIONS

Next steps to prepare healthcare provides and vaccination manufacturers

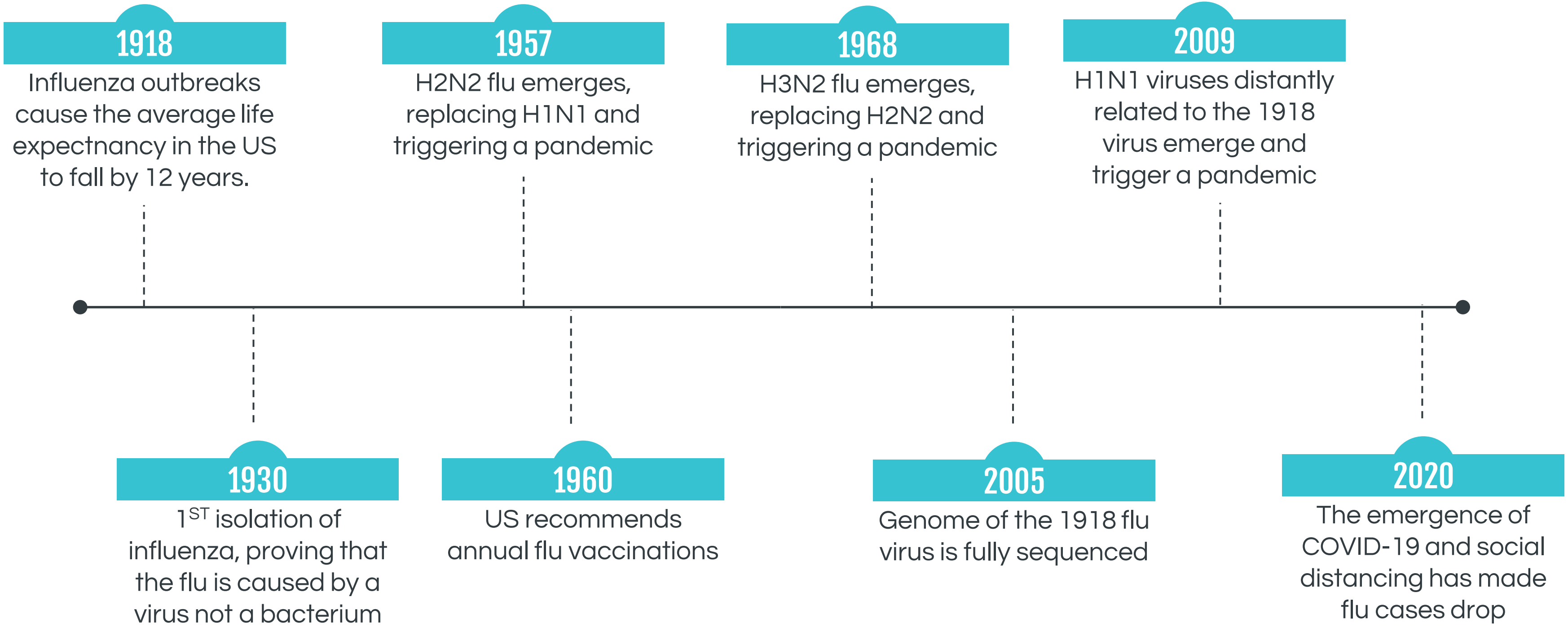
05

INTRODUCTION

The flu (influenza) virus has been teetering between pandemic status and global health crisis for over 100 years. Forecasting future positive cases and positivity ratios could help bolster preparedness for hospitals and vaccine producers.

Utilizing data from the World Health Organization (WHO), the U.S. is capable of forecasting future positive cases off historical data. While there may be a portion of undocumented positive cases, the WHO data is considered to be the most comprehensive.

BRIEF HISTORY OF THE FLU



1918
Influenza outbreaks cause the average life expectnancy in the US to fall by 12 years.

1930
1ST isolation of influenza, proving that the flu is caused by a virus not a bacterium

1957
H2N2 flu emerges, replacing H1N1 and triggering a pandemic

1960
US recommends annual flu vaccinations

1968
H3N2 flu emerges, replacing H2N2 and triggering a pandemic

2005
Genome of the 1918 flu virus is fully sequenced

2009
H1N1 viruses distantly related to the 1918 virus emerge and trigger a pandemic

2020
The emergence of COVID-19 and social distancing has made flu cases drop

Source: https://www.cdc.gov/flu/pandemic-resources/1918-commemoration/milestone-infographic.htm

THE DATASET

COUNTRY

The United States of America

DATES

2014 – 2020 Weekly Data

Data transformed into monthly data for better forecasting (ie. Forcasting 52 weeks can be less accurate than forecasting 12 months)

STRAINS TESTED

6

A H1N1
A H3
A No Type
B Yamagata
B Victoria
B No Type

RECORDS

8,418,471 Specimens Tested

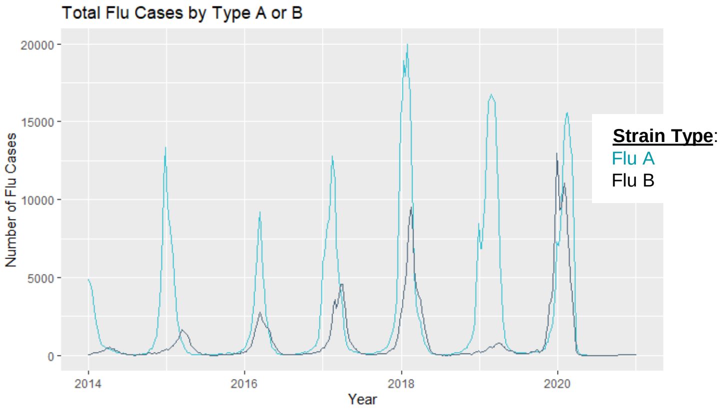
1,266,600 Tested Positive

15% positive

DATASET DETAILS

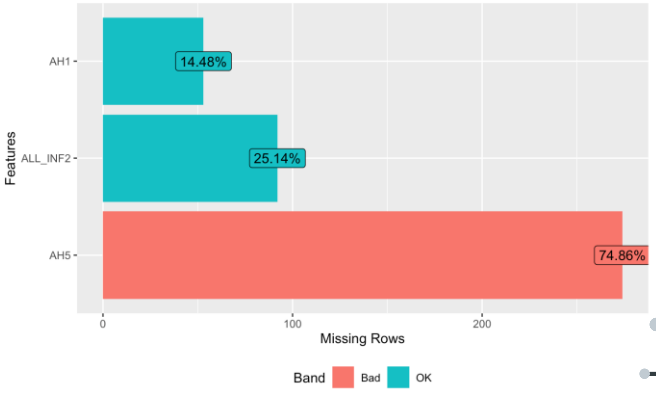
DATA EXTRACTION

- North American WHO Flu data
- Weekly data beginning in 2014 and ending at the end of 2020
- Number of specimens received and processed
- Flu strains H1, H1N12009, H3, H5, NoSubType and Influenza A
- Test results broken down into Influenza B and Influenza A
    - Final analysis conducted off the combination of Flu A & B
- Weekly title (i.e.. Widespread Outbreak, Sporadic, Local Outbreak, etc.)



MISSING VALUES

- Total Observations: 366
- AH1: 53, roughly 14% of the column
- AH5: 274, roughly 75% of the column
- ALL_INF2: 92, roughly 25% of the column
- Since the model relies on having consecutive dates within the analysis all values remained within the dataset



> All models are wrong,
> but some are useful.

George E.P. Box

MODELING TECHNIQUES

ARIMA MODEL:

An autoregressive integrated moving average model is used primarily used to understand time series data or to predict future points using past values in the series.

Any ‘non-seasonal’ time series that has clear pattern definition and is not white noise can be modeled with an ARIMA model.

Seasonality

$(1\text{-}B^{12})$

Time series data may include a seasonality component, which is a data pattern that is repeated cyclically (ie. Monthly or Quarterly)

Differencing (I)

$(1\text{-}B)$

Creating a “lagged” transformation by using the delta of current and previous value may normalize data to a more stationary realization.

Auto Regressive (AR)

A linear regression model that uses lags as predictors (i.e. number of past data points used to recognize trends)

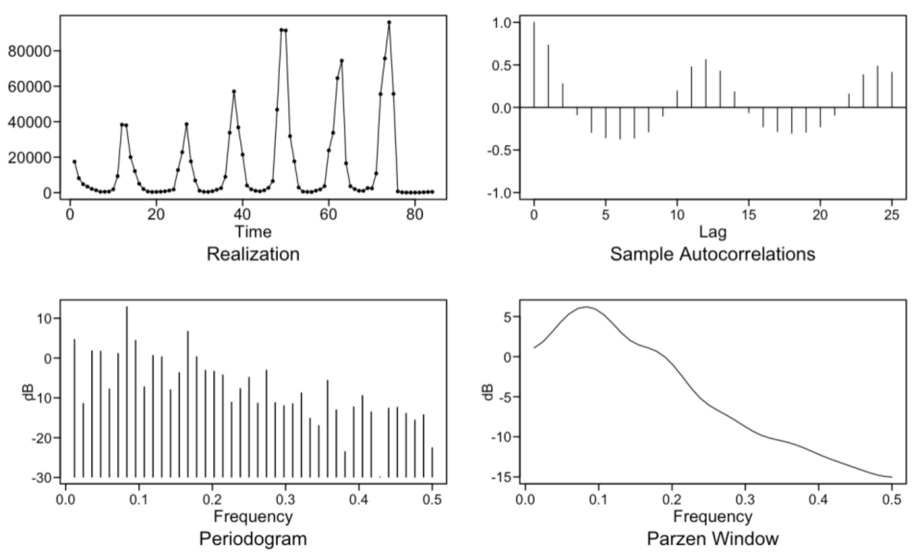
Moving Average (MA)

Refers to the portion of the ARIMA model that determines the number of lagged forecast errors

MODELING PROCESS

SEASONALITY

After aggregating the data on a monthly basis, the data shows strong seasonal behavior that needs to be modeled out. Looking at the Parzen window, there appears to be “AR”-like behavior with an initial peak at 0 and then a gradual decline in frequency. There is a strong peak here around 0.08, which is indicative of an annual frequency (1/12 months).



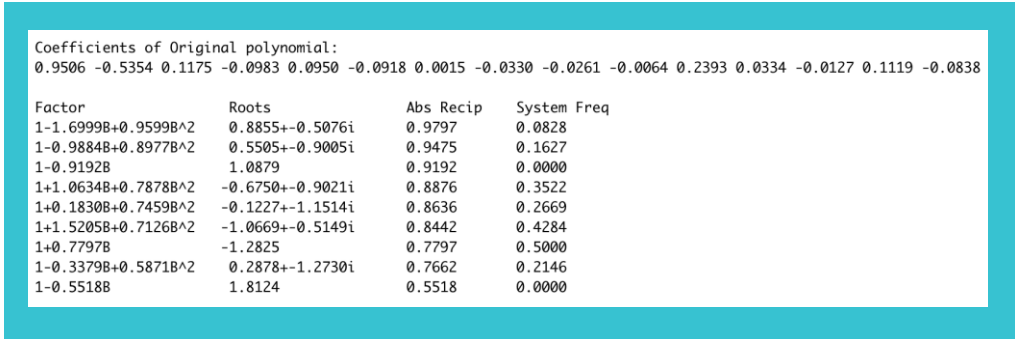
MODELING PROCESS

FREQUENCY ANALYSIS

By “overfitting” our model to an AR15 and generating frequency estimates, we see strong evidence of echoes of yearly frequencies. (divisible by 0.08).

Semi-annual, quarterly, and weekly seasonality were also run, in order to verify that monthly frequency was the best fit of seasonality for the data.

The absolute reciprocal indicates how strongly the data fits each frequency (perfect fit = 1)



```
Coefficients of Original polynomial:
0.9506 -0.5354 0.1175 -0.0983 0.0950 -0.0918 0.0015 -0.0330 -0.0261 -0.0064 0.2393 0.0334 -0.0127 0.1119 -0.0838

Factor                 Roots                Abs Recip    System Freq
1-1.6999B+0.9599B^2    0.8855+-0.5076i      0.9797       0.0828
1-0.9884B+0.8977B^2    0.5505+-0.9005i      0.9475       0.1627
1-0.9192B              1.0879               0.9192       0.0000
1+1.0634B+0.7878B^2   -0.6750+-0.9021i      0.8876       0.3522
1+0.1830B+0.7459B^2   -0.1227+-1.1514i      0.8636       0.2669
1+1.5205B+0.7126B^2   -1.0669+-0.5149i      0.8442       0.4284
1+0.7797B             -1.2825               0.7797       0.5000
1-0.3379B+0.5871B^2    0.2878+-1.2730i      0.7662       0.2146
1-0.5518B              1.8124               0.5518       0.0000
```

MODELING PROCESS

DIFFERENCING (I in ARIMA)

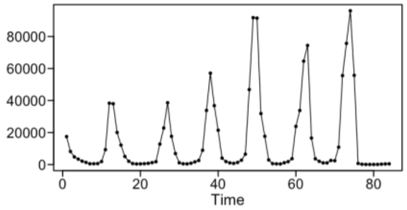
(1-B) was removed from the model in an attempt to convert the data to be more stationary for forecasting, although it was not ultimately included in the final model

SEASONALITY

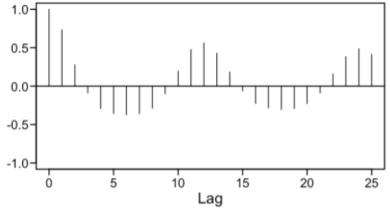
$(1-B^{12})$ was removed from the monthly data to account for annual seasonality that is seen in postive flu cases over time

ANALYSIS

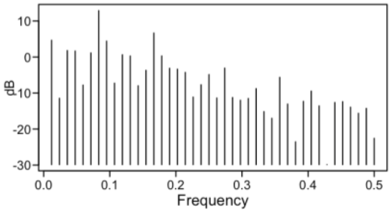
After transformations were made, most of the seasonality was modeled out, with much more stationarity in the remaining residuals.



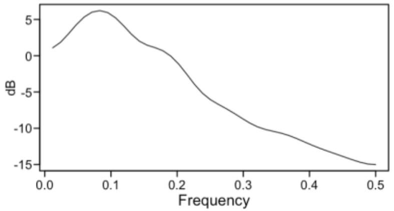
Realization



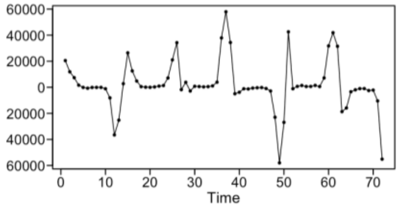
Sample Autocorrelations



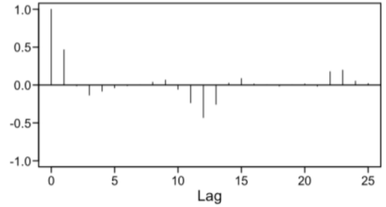
Periodogram



Parzen Window



Transformed Realization



Sample Autocorrelations

MODELING PROCESS

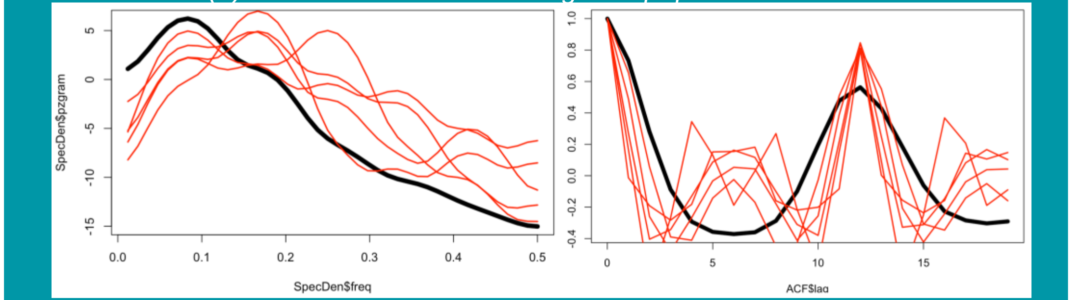
MODEL ESTIMATION

Using Bayesian Information Criterion (BIC) as the target model metrics, the optimal order estimates are provided in the table.

Exploration of the data shows that simulated spectral density and ACF plots (red lines) closely resemble that of the population of an AR2 model with an annual trend. While not a perfect fit, this can still prove to be a useful model.

This suggests AR2 is a good fit to our influenza data after inclusion of the seasonal component. Our final model is an AR(2) with a seasonal order of 12.

5 simulated AR(2) models were evaluated against population data to assess fit



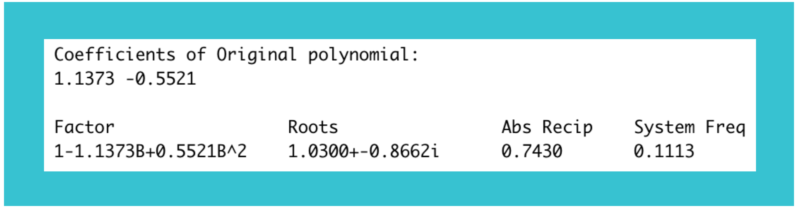
AIC5 Results (top 3)	AR – p	MA – q	BIC*
# 1	0	1	19.38891
# 2	0	2	19.40673
# 3	2	0	19.41115
# 4	1	1	19.42319
# 5	1	2	19.45841

**The lower the BIC the better*

MODELING PROCESS

MODEL COEFFICIENT ESTIMATIONS

By utilizing the Maximum Likelihood Estimate (MLE), the coefficients of the p’s are 1.137 and -0.552.



```
Coefficients of Original polynomial:
1.1373 -0.5521

Factor                 Roots                Abs Recip    System Freq
1-1.1373B+0.5521B^2    1.0300+-0.8662i      0.7430       0.1113
```

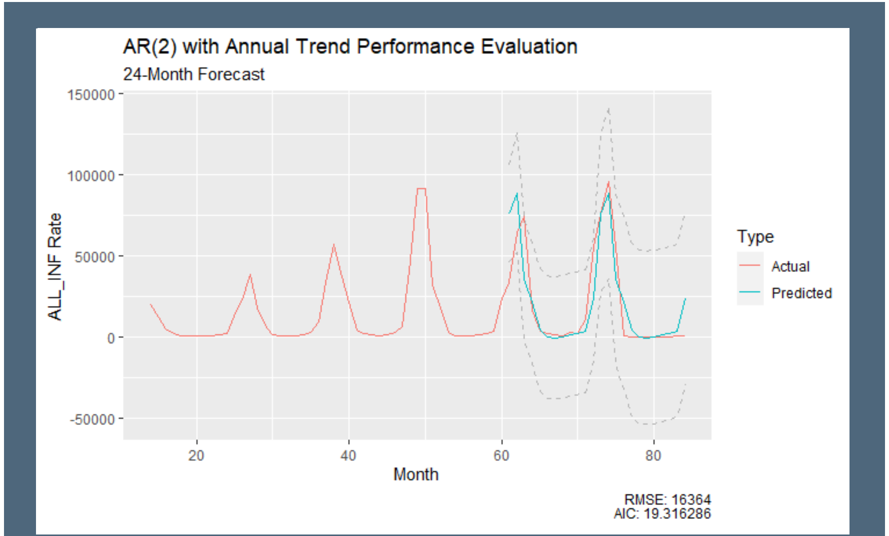
The resulting estimated formula for the time series model is:

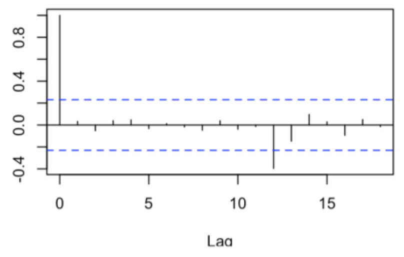
$$(1 + 1.137B - 0.552B^{2})Xt\ (1 - B^{12}) = \ a_{t}$$

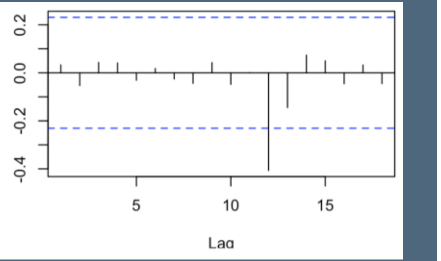
MODELING PROCESS

MODEL EVALUATION

By forecasting a known time interval, the model was evaluated against the ground truth. The seasonality and the peaks were well mimicked in the forecast. The autocorrelations of the residuals (error) proved to be mostly below statistical significance with one out of twenty observations having residuals beyond statistical significance. These findings suggest a useful model for predicting future influenza infection totals, with a root mean squared error (RMSE) of +/-**16,364** cases in a 24 month period.







Note the strong peaks at lag 12 in the auto-correlation and partial auto-correlation plots. We expect to see ~1 observation in 20 (5%) cross outside of the upper/lower limit boundaries (indicated by blue dotted lines)

MODEL APPLICATION

Provide adequate prep time for vaccine manufactures to prepare the necessary quantities.

Ensure that healthcare professionals are well preapred for the estimated rise in flu cases so they can be prepared to handle each case.

Potentially avoid manufacturing unneccesary vaccines that will not be used, thus reducing overall costs.

RECOMMENDATIONS

DOS

- Reevaluate the model on a semi-annual cadence considering the seasonality of flu cases
- Compare forecasted to future actuals to assess model's performance
- Run model with positivity rate as a predictor versus total flu cases (i.e.. Number of positive cases of those that tested for the flu)

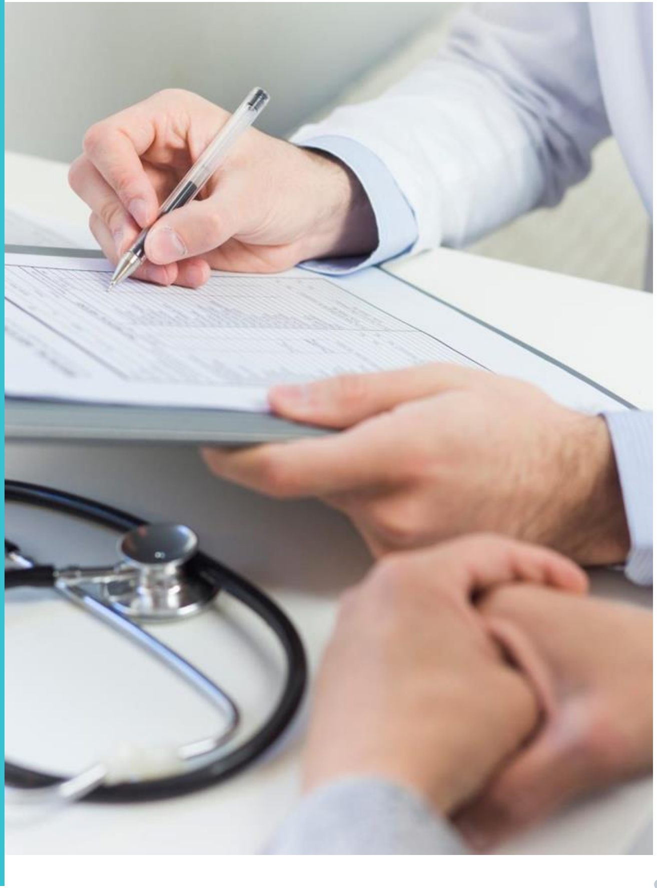
DON'TS

- Strictly apply model as only source of truth
- Forecast too far in advance: model accuracy decreases as forecast length increases

CONCLUSIONS

For best results in predicting flu cases in the U.S. use an AR(2) s=12 model on monthly aggregate data. This model can serve as an aid in understanding flu vaccine demand and the requirements necessary for healthcare services to treat the incoming population.

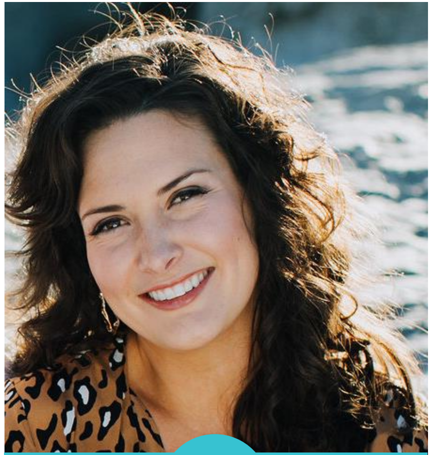
If multiple tests per individual are expected, you can run this model based on positivity rate (i.e. positive cases out of total flu tests) to more accurately account for multiple individual tests in the data.



OUR TEAM

DUSTIN BRACY

PAUL HUGGINS



GRACE LANG

BRANUM STEPHAN