# COVID-19 Predictive Analytics

This page describes our COVID-19 predictive analytics program, provides links to our code repository, and provide the actual model for free within your own EHR. The predictive analytics identifies people at highest risk of COVID-related adverse events such as emergency hospitalizations (our next update will have other adverse events such as likelihood of ICU and death). It is envisioned that medical and/or non-medical staff (such as social workers) at health systems and providers can outreach to these high-risk people proactively. It is expected that through that engagement that the staff can identify challenges the person may be facing that, if addressed, can reduce the likelihood of that person having an adverse event such as hospitalization and increase the likelihood of that the person can remain at home sheltered-in-place.

The engagement driven by these predictive analytics can be via various channels including phone calls, text messages, or messages via the EHR patient portal.

The analytics can be deployed within any EHR including Cerner’s Millennium, Athena, eClinicalWorks, and Epic’s Healthy Planet. Also, this first model release has analytics that were simplified such that most any authorized EHR user can enter the parameters; staff from the IT Department aren’t necessarily required.

All of the information here including in the associated support pages, including the information on Github, is being made freely-available under our [open source license](https://github.com/Lumeris-Health/Covid/blob/master/LICENSE).

## Instructions

This model was purposely made simple to use a points system analogous to Weight Watchers so that it can be typed into most any EHR manually by a person who is non-technical. For example, the model variables and points can be typed directly into Healthy Planet to create a patient registry. A PDF that provides a brief overview and instructions, including the variables and points to create the registry in your EHR, is [here](https://github.com/Lumeris-Health/Covid/blob/master/howTo.pdf).

The model variables and points in the PDF are also viewable/downloadable in Excel via this link [here](https://github.com/Lumeris-Health/Covid/blob/master/modelParameters.xlsx).

## Detail

### Model 1. Predicting COVID-19 related hospitalizations

This is a simple model with just 20 variables that performs well at predicting the likelihood of all-cause emergency hospitalizations in the next 90 days for people particularly susceptible to COVID-19. Emergency hospitalizations are defined as ED visits that are followed by an IP admission (the prediction is not just for ED visits that are followed by a discharge). It predicts the hospitalizations for people especially susceptible to COVID-19 based on the latest data published by the CDC [here](https://www.cdc.gov/coronavirus/2019-ncov/index.html) and on March 11 in The Lancet [here](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30566-3/fulltext).

### Applicable Population

Although the model was built on a small portion of the US population, in our judgment it can be applied to people of all ages (between 0 to 114 yo) and both males and females.

The number of people in the validation dataset is 789k. Data is from January 2018 thru March 24, 2020. The demographics of the people utilized in the validation dataset are below:

| Characteristic | Value |

| :-------------: |:-------------:|

| Age (median yo) | 48 (range 0 to 114 yo) |

| Female | 56.5% |

| COPD | 1.4% |

| Coronary Disease | 2.9% |

| Diabetes | 5.2% |

| Hypertension | 12.8% |

| Hospitalizations | 1.26% or 12.6 per 1000 |

Additional information on the demographics is available in Github [here](https://github.com/Lumeris-Health/Covid/blob/master/extendedDemographics.xlsx).

In preparing our work we noticed that there are other organizations creating and releasing free open source models. We did some basic due diligence on one model that seems to have gotten some traction in the marketplace. But when we tested it across our whole population we found that the accuracy of that model for people who are under age 65 yo is worse than random chance. Providers would be \*\*better off not using a model at all\*\*!!! We don’t know the exact reason for this issue, but we suspect that it is because their model was based on CMS claims data from 2015 and 2016, a dataset that primarily has people who are over age 65 yo. We notified this company privately but do not know their response yet.

### Model Details

This model for this release was trained and tested on a collection of contemporary (1/1/2018 to present) data from different health systems scattered across the US. It is intended to work well on data available either in the EHR or from administrative claims data (although all performance characteristics reported here are based on EMR data).

We trained models on both likelihood of hospitalization and likelihood of hospitalization due to COVID ([details of COVID data](https://github.com/Lumeris-Health/Covid/blob/master/covidCasesAlgorithm.pdf)).

Additional information about this model are available in our [public Github repository](https://github.com/Lumeris-Health/Covid).

### Model Performance

We evaluate performance using standard statistical tools applied to validation or “hold out” data sets – data that the computer algorithms did not have access to during the training of the model. We do this in order to avoid overly optimistic estimates of accuracy. Here we provide the overall accuracy (c-statistic, also known as area under the curve or AUC) together with sensitivity and positive predictive values (PPV) and sensitivities for the top 1% and 5% of the population for the prediction of COVID-19 related hospitalizations:

| Characteristic | Top 1% Sensitivity | Top 1% PPV | Top 5% Sensitivity | Top 5% PPV

| :-------------: |:-------------:|:-------------:|:-------------:|:-------------:|

| 0.833 | 0.141 | 0.138 | 0.387 | 0.075 |

The receiver operator characteristics (ROC) curve is [here](https://github.com/Lumeris-Health/Covid/blob/master/all.png). And additional information on the performance is available in Github [here](https://github.com/Lumeris-Health/Covid/blob/master/rocTable.csv).

The performance of our model is relatively strong. As a check we compared the performance of our model to one built from data published in that Lancet article mentioned [above](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30566-3/fulltext). These are the performance statistics for that model to predict hospitalizations:

| Characteristic | Top 1% Sensitivity | Top 1% PPV | Top 5% Sensitivity | Top 5% PPV

| :-------------: |:-------------:|:-------------:|:-------------:|:-------------:|

| 0.732 | 0.058 | 0.047 | 0.202 | 0.032 |

We performed a good faith effort to apply that model in the best way possible, but nevertheless in comparing the performance between those two models, our model outpeformed by a considerable amount. This table shows the percent outperformance between our model and the one published in The Lancet:

| Characteristic | Top 1% Sensitivity | Top 1% PPV | Top 5% Sensitivity | Top 5% PPV

| :-------------: |:-------------:|:-------------:|:-------------:|:-------------:|

| +13% | +131% | +259% | +84% | +186% |

There may be important details from that published paper that make such a comparison specious, but it nevertheless helped us believe that we were contributing to society by publishing our model.

Given the model performance above, economic modeling of the breakeven costs shows that providers will “come out ahead” economically if his/her engagement outreach to the top 1% is moderately effective and costs less than $1,382 (the breakeven for low and highly successful programs is $691 and $2,073, respectively). Said a different way, organizations using this predictive model to engage with the top 1% highest risk people in their population to reduce hospitalizations by 50%, can spend up to $1,382 per person to achieve that objective and still breakeven. Of course, this economic analysis does not include the humanistic value of helping people stay safely at home and avoid adverse events.

The breakeven point for the cost of interventions toward the Top 1% and Top 5% are shown in the table below:

| Characteristic | Top 1% | Top 5% |

| :------------- |:-------------:| :-------------:|

| Clinical Success^: | | |

| Low effectiveness (25% reduction in hospitalizations) | $691 | $225 |

| Moderate (50% reduction in hospitalizations) | $1,382 | $450 |

| High (75% reduction in hospitalizations) | $2,073 | $676 |

\* ^ Clinical success refers to the percentage reduction in hospitalizations due to the outreach, assumed to be a reduction in hospitalizations by either 25%, 50%, or 75% \*

\* Assumes COVID-19 hospitalization costs $20,000 based on approximately $2,200/day for 9 days \*

\* All values in US dollar\*

If you have questions or suggestions please send an email to: [Covid Predictive Analytics](mailto:info@lumeris.com?subject=[GitHub]%20Covid%20Predictive%20Analytics).