medRxiv preprint doi: https://doi.org/10.1101/2020.03.16.20036723; this version posted March 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Building a COVID-19 Vulnerability Index

Dave DeCapprio* Joseph Gartner† Carol J. McCall, FSA, MPH‡

Thadeus Burgess Sarthak Kothari Shaayaan Sayed
ClosedLoop.ai

March 15, 2020

Abstract

COVID-19 is an acute respiratory disease that has been classified as a pandemic by the World Health Organization. Information regarding this particular disease is limited, however, it is known to have high mortality rates, particularly among individuals with preexisting medical conditions. Creating models to identify individuals who are at the greatest risk for severe complications due to COVID-19 will be useful to help for outreach campaigns in mitigating the diseases worst effects. While information specific to COVID-19 is limited, a model using complications due to other upper respiratory infections can be used as a proxy to help identify those individuals who are at the greatest risk. We present the results for three models predicting such complications, with each model having varying levels of predictive effectiveness at the expense of ease of implementation.

I. Introduction

i. COVID-19 Virus

Coronaviruses (CoV) are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). CoV are zoonotic, meaning they are transmitted between animals and people. Coronavirus disease (COVID-19) is a new strain that was discovered in 2019 and has not been previously identified in humans[1].

COVID-19 is a respiratory infection with common signs of infection that include respiratory symptoms, fever, cough, shortness of breath and breathing difficulties. In more severe cases, infection can cause pneumonia, severe acute respiratory syndrome, kidney failure and death.

ii. Flattening the Curve

On March 11, 2020 the World Health Organization (WHO) declared COVID-19 to be a pandemic[2]. In their press conference, they were clear that pandemic was not a word they used lightly or carelessly or to cause unreasonable fear. They were also clear to highlight that this is the first

*Corresponding author: dave.decaprio@closedloop.ai
†Corresponding author: joseph.gartner@closedloop.ai
‡Corresponding author: carol.mccall@closedloop.ai

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.03.16.20036723; this version posted March 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

pandemic to ever be caused by a coronavirus and that all countries can still act to change its course.

Public health and healthcare experts agree that mitigation is required in order to slow the spread of COVID-19 and prevent the collapse of healthcare systems. On any given day, health systems in the United States run close to capacity[3] and so every transmission that can be avoided and every case that can be prevented has enormous impact.

iii. Identifying Vulnerable People

The risk of severe complications from COVID-19 is higher for certain vulnerable populations, particularly people who are elderly, frail, or have multiple chronic conditions. The risk of death has been difficult to calculate[4], but a small study[5] of people who contracted COVID-19 in Wuhan suggests that the risk of death increases with age, and is also higher for those who have diabetes, disease, blood clotting problems, or have shown signs of sepsis. With an average death rate of 1%, the death rate rose to 6% for people with cancer, high blood pressure and chronic respiratory disease, 7% for people with diabetes, and 10% for people with heart disease. There was also a steep age gradient; the death rate among people aged 80+ was 15%

Identifying who is most vulnerable is not necessarily straightforward. More than 55% of Medicare beneficiaries meet at least one of the risk criteria listed by the CDC[6]. People with the same chronic condition don't have the same risk, and simple rules can fail to capture complex factors like frailty[8] which makes people more vulnerable to severe infections.

II. Methods

i. datasets

Since real world data on COVID-19 cases is not readily available, the CV19 Index was developed using close proxy events. A person's CV19 Index is measured in terms of their near-term risk of severe complications from respiratory infections (e.g. pneumonia, influenza). Specifically, 4 categories of diagnoses were chosen from the Clinical Classification Software Refined (CCSR)[11] classification system:

- RSP002 - Pneumonia (except that caused by tuberculosis)
- RSP003 - Influenza
- RSP005 - Acute bronchitis
- RSP006 - Other specified upper respiratory infections

Machine learning models were created that use a patient's historical medical claims data to predict the likelihood that they will have an inpatient hospital stay due to one of the above conditions in the next 3 months. The data used was an anonymized 5% sample of the Medicare claims data from 2015 and 2016. This data spanned the transition from ICD-9 to ICD-10 on October 1st, 2016. The data set used to create the model was created by identifying all living members above the age of 18 on 9/30/2016. Only fee -for-service members were included because medical claims histories for other members are not reliably complete. We then excluded all members who had less than 6 months of continuous eligibility prior to 9/30/2016. We also excluded members who lost coverage within 3 months after 9/30/2016, except for those members who lost coverage due to death. Table 1 below summarizes the population selection.

The final data set is split 80/20% into train and test sets, with 1,481,654 people in the training set and 369,865 in the test set. The prevalence of the proxy event within the final population was 0.23%.

medRxiv preprint doi: https://doi.org/10.1101/2020.03.16.20036723; this version posted March 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Cumulative Population Step	**Selection Criteria**
3,114,713	Total members in the data set
1,867,879	Fee for service members with 6 months of continuous coverage prior to 9/30/2016
1,867,779	18 years old or older
1,861,868	Exclude members with a death date before 9/30/2016
1,851,519	Exclude members who lose coverage in the next 3 months not due to death.

Table 1: *Features used associated with risk factors identified by CDC and their corresponding CCSR codes.*

The labels for the prediction task were created by identifying all patients who had an inpatient visit with an admission date from 10/1/2016 through 12/31/2016 with a primary diagnosis from one of the listed categories. A 3 month delay was imposed on the input features to the model, so that no claims after 6/30/2016 were used to make the predictions. This 3 month delay simulates the delay in claims processing that usually occurs in practical setting and enables the model to be used in realistic scenarios.

ii. models

We highlight a few approaches to building models to help identify individuals who are vulnerable to complications to respiratory infections. All 3 approaches described are machine learning methods created using the same data set. We have chosen 3 different approaches that represent a tradeoff between accuracy and ease of implementation. For individuals who have access to data, but not the coding background to adopt our model, we hope that the simple model can be easily ported to other systems. For a more robust model, we create a Gradient boosted tree leveraging age, gender, and medical diagnosis history. This model has been made open sourced, and can be obtained from github (https://github.com/closedloop-ai/cv19index). Finally, we have created a third model that uses an extensive feature set generated from Medicare claims data along with linked geographical and social determinants of health data. This model is being made freely available through our hosted platform. Information about accessing the platform can be found at https://cv19index.com.

iii. Logistic Regression

The first approach is aimed at reproducing the high level recommendations coming from the Center for Disease Control (CDC) website[7] for identifying those individuals who are at risk. They identify risk features as:

- Older adults
- Individuals with Heart disease
- Individuals with Diabetes
- Individuals with Lung disease

To turn this into a model, we extract International Classification of Diseases, Version 10 (ICD-10) diagnosis codes from the claims and aggregate them using the Clinical Classification Codes Revised (CCSR) categories. We create indicator features for the presence of any code in the CCSR category. The mapping between the CDC risk factors and the CCSR codes are described in

medRxiv preprint doi: https://doi.org/10.1101/2020.03.16.20036723; this version posted March 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Feature name	**Coefficient**	**CCSR Code**
Intercept	-5.49	
Gender male	-0.212	n/a
# of Admissions (12M)	0.186	n/a
Age	-0.014	n/a
Pneumonia except that caused by tuberculosis	0.117	CCSR:RSP002
Other and ill-defined heart disease	-0.111	CCSR:CIR015
Heart failure	-0.271	CCSR:CIR019
Acute rheumatic heart disease	-0.008	CCSR:CIR002
Coronary atherosclerosis and other heart disease	-0.529	CCSR:CIR011
Pulmonary heart disease	-0.005	CCSR:CIR014
Chronic rheumatic heart disease	-0.014	CCSR:CIR001
Diabetes mellitus with complication	-0.273	CCSR:END003
Diabetes mellitus without complication	-0.496	CCSR:END002
Chronic obstructive pulmonary disease and bronchiectasis	-0.269	CCSR:RSP008
Other specified and unspecified lower respiratory disease	-0.020	CCSR:RSP016
age X Pneumonia	0.010	n/a
age X Other and ill-defined heart diseas	0.003	n/a
age X Heart failure	0.009	n/a
age X Acute rheumatic heart disease	0.003	n/a
age X Coronary atherosclerosis and other heart disease	0.011	n/a
age X Pulmonary heart disease	-0.000	n/a
age X Chronic rheumatic heart disease	-0.001	n/a
age X Diabetes mellitus with complication	0.007	n/a
age X Diabetes mellitus without complication	0.009	n/a
age X Chronic obstructive pulmonary disease and bronchiectasis	0.013	n/a
age X Other specified and unspecified lower respiratory disease	0.006	n/a

Table 2: *Features used associated with risk factors identified by CDC and their corresponding CCSR codes.*

Table 2. We start with these features as they give us an ability to quantify the portion of the at risk population that are encapsulated by the high level CDC recommendations. In addition to the conditions coming from the recommendations of the CDC, we will look at features that our other modeling efforts surfaced as important and avail those features to the model as well. We also provide gender, age in years, as well as an interaction term between age and the diagnostic features. This simple data set is used to train a logistic regression model[9]. In addition to the CCSR Codes, Table 2 additionally includes

iv. Gradient Boosted Trees

Our more robust approach uses gradient boosted trees. Gradient boosted trees are a machine learning method that use an ensemble of simple models to create highly accurate predictions[9]. The resulting models demonstrate higher accuracy. The drawback to these models are that they are significantly more complex, however, "by hand" implementations of such models is impractical. Here, we create two variations of the models. The first, is a model that leverages similar information as our logistic regression model. A nice feature of Gradient Boosted Trees is they are fairly robust against learning features that are excentricities of the training data, but do

medRxiv preprint doi: https://doi.org/10.1101/2020.03.16.20036723; this version posted March 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

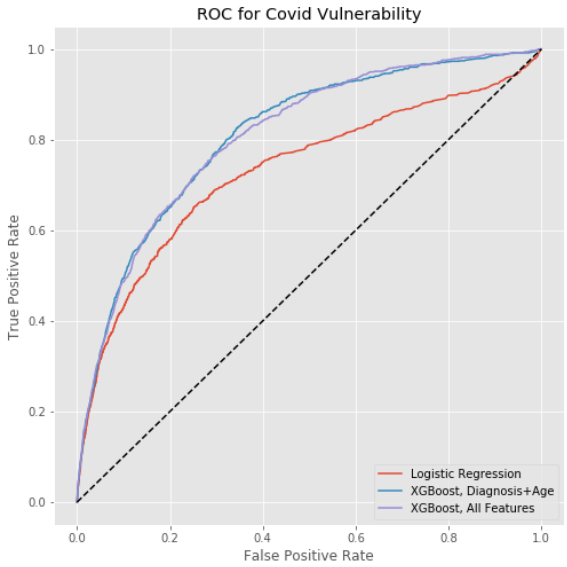

Figure 1: *A Receiver Operating Characteristic graph depicting the performance of the three models.*

not extend well to future data. As such, we allow full diagnosis histories to be leveraged within our simpler XGBoost model. In this approach, every category in the full CCSR is converted into an indicator feature, resulting in 559 features. Details about how to connect the full diagnosis history with the open sourced model are provided with the open sourced version of the mode

We additionally built a model within the ClosedLoop platform. The ClosedLoop platform is a software system designed to enable rapid creation of machine learning models utilizing healthcare data. The full details of the platform are outside the bounds of this paper, however, using the platform allows us to leverage engineered features coming from peer reviewed studies. Examples are social determinants of health, and the Charlson Comorbidity Index[12]. We chose not to include these features within the open sourced model, because the purpose of the open sourced version is to be as accessible as possible for the greater healthcare data science community.

III. Results and Model Interpretation

We quantify the performance of our models using metrics that are standard within the data science community. In particular, we visualize the performance of our model using a Receiver Operation Characteristic graph, see Figure 1. Additionally, the metrics quantifying the effectiveness of our models are in Table 3. As you can see, the performance of both Gradient Boosted Tree models are very similar. The ROC curve demonstrates that even as the decision threshold increases, the percentage of the potentially affected population increases at roughly the same rate. Similarly, the Logistic Regression model has similar performance at low alert rates. We can see that at a 3% alert rate, the difference in sensitivity is only .02. The performance at higher alert rates experiences a significant performance disadvantage, however, for most interventions this would be at alert rates higher than is practical.

medRxiv preprint doi: https://doi.org/10.1101/2020.03.16.20036723; this version posted March 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Model Type	ROC AUC	Sensitivity at 3% Alert Rate	Sensitivity at 5% Alert Rate
Logistic Regression	.731	.214	.314
XGBoost			
Diagnosis History + Age	.810	.234	.324
XGBoost			
Full Feature Set	.810	.231	.327

Table 3: *Measures of effectiveness for the three models.*

IV. Accessing Models

There are two ways of accessing the models that we are providing. The first, is to access our open sourced version of our model. As stated, we have released an open sourced version of the model, available at https://github.com/closedloop-ai/cv19index. This model is written in the Python programming language. We have included synthetic data for the purpose of walking individuals through the process of going from tabular diagnosis data to the input format specific for our models. We encourage the healthcare data science community to fork the repository, and adapt it to their own purposes. We encourage collaboration from the open sourced community, and pull requests will be considered for inclusion in the main branch of the package. For those wishing to use our models within our platform, we are providing access to the COVID-19 model free of charge. Please visit https://closedloop.ai/cv19index for instructions on how to gain access.

V. Conclusions

This pandemic has already claimed thousands of lives, and sadly, this number is sure to grow. As healthcare resources are constrained by the same scarcity constraints that effect us all, it s important to empower intervention policy with the best information possible. We have provided several models and means of access for those individuals with varying levels of technical expertise. It is our hope that by providing these models quickly to the healthcare data science community, widespread adoption will lead to more effective intervention strategies, and ultimately, help to curtail the worst effects of this pandemic.

References

[1] World Health Organization. "Coronavirus." Who.Int, 2019, www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed 15 Mar. 2020.

[2] World Health Organization. "WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020." Who.Int, 11 Mar. 2020, www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020. Accessed 15 Mar. 2020.

[3] Specht, Liz. "Simple Math Offers Alarming Answers about Covid-19, Health Care." STAT, 10 Mar. 2020, www.statnews.com/2020/03/10/simple-math-alarming-answers-covid-19/. Accessed 15 Mar. 2020.

medRxiv preprint doi: https://doi.org/10.1101/2020.03.16.20036723; this version posted March 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

[4] Page, Michael Le. "Why Is It so Hard to Calculate How Many People Will Die from Covid-19?" New Scientist, 11 Mar. 2020, www.newscientist.com/article/mg24532733-700-why-is-it-so-hard-to-calculate-how-many-people-will-die-from-covid-19/. Accessed 15 Mar. 2020.

[5] Hamzelou, Jessica. "Coronavirus: Risk of Death Rises with Age, Diabetes and Heart Disease." New Scientist, 9 Mar. 2020, www.newscientist.com/article/2236846-coronavirus-risk-of-death-rises-with-age-diabetes-and-heart-disease/. Accessed 15 Mar. 2020.

[6] Centers for Disease Control and Prevention. "Coronavirus Disease 2019 (COVID-19)." Centers for Disease Control and Prevention, 11 Feb. 2020, www.cdc.gov/coronavirus/2019-ncov/specific-groups/high-risk-complications.html.

[7] Centers for Disease Control and Prevention. "Coronavirus Disease 2019 (COVID-19)." Centers for Disease Control and Prevention, 11 Feb. 2020, www.cdc.gov/coronavirus/2019-ncov/specific-groups/high-risk-complications.html.

[8] Hubbard, Ruth E, et al. "Frailty Status at Admission to Hospital Predicts Multiple Adverse Outcomes." Age and Ageing, vol. 46, no. 5, 2017, pp. 801–806, www.ncbi.nlm.nih.gov/pubmed/28531254, 10.1093/ageing/afx081. Accessed 11 Jan. 2020.

[9] James, Gareth, et al. An Introduction to Statistical Learning, with Applications in R. Springer, 2013.

[10] "Measures and Data Sources | County Health Rankings and Roadmaps." Countyhealthrankings.Org, University of Wisconsin Population Health Institute, 2019, www.countyhealthrankings.org/explore-health-rankings/measures-data-sources. Accessed 1 Feb. 2020.

[11] The Healthcare Cost and Utilization Project. "Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses." Www.Hcup-Us.Ahrq.Gov, www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs refined.jsp. Accessed 15 Mar. 2020.

[12] Charlson, M E, et al. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation." Journal of Chronic Diseases, vol. 40, no. 5, 1987, pp. 373–83, 10.1016/0021-9681(87)90171-8.

VI. Appendix: Full Feature List

We include a full list of features available within our platform. The majority of features are binary variables indicating if a patient has had one type of medical event 15 moths prior to the date of prediction, excluding the 3 most recent months.

medRxiv preprint doi: https://doi.org/10.1101/2020.03.16.20036723; this version posted March 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Gender
Age
Social Factor: Current County Health Rankings Analytics
Social Factor: Current County Health Rankings Outcome and Factors
Social Factor: Current Social Vulnerability Index
Social Factor: Current USDA Food Atlas Access
Social Factor: Current USDA Food Atlas Restaurants
Social Factor: Current USDA Food Atlas Socioeconomic
Social Factor: Current Area Deprivation Index
Prediction Month
State of Residence
Medicaid
Race & Ethnicity
Prior termination
Medical Eligibility Months
Different Residences (County level)
Medicare ESRD Status
Medicare Disability Status
Distinct DME Categories (12M)
Hospice
Days Since Last Hospital Discharge
Acute Fall-Related Injuries (12M)
Use of Preventative Services
Procedure History
DME
ICU Stays (12M)
'Frailty Indicator
# of Distinct Providers seen in Office Visits (12M)
# of Distinct Providers seen in E&M Visits (12M)
PCP Visit After Discharge
Discharge Disposition
OutlierAdmits
Monthly OOP Cost
Diagnosis History
Monthly Medical Cost
# of Admissions (12M)
Inpatient Days
# of Observation Stays
# of ER Visits (12M)
# of ER Visits (6M)
# of Office Visits (12M)
# of Evaluation and Management Visits (12M)
Admission Cause
Charlson Comorbidity Index (CCI)
Evidence of tobacco smoking

Table 4: *Full feature list available within platform.*

medRxiv preprint doi: https://doi.org/10.1101/2020.03.16.20036723; this version posted March 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Evidence of smokeless tobacco
Evidence of tobacco non-use
Falls risk assessment performed
Abnormal BMI
Normal BMI
LDL
LVEF
Systolic Blood Pressure
Diastolic Blood Pressure
Continuity of Care Index (12M)
PQI Any
PQI Diabetes
PQI 5
PQI 7
PQI 8
PQI 11
PQI 12
Distinct CCSR Body Systems
Distinct CCSR Diagnosis Categories (12M)
PCS Procedure History
Prior Respiratory Infections
Prior Hospital Acquired Infections
Diagnosis History High-Level

Table 5: *Full feature list available within platform (continued).*