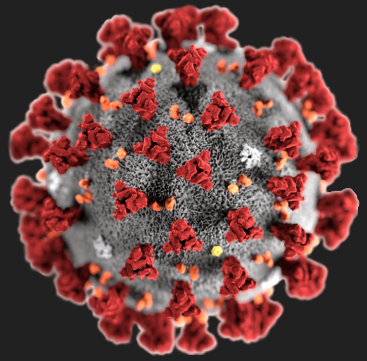
Who does viral pneumonia hit the hardest?

Using patient features to predict 90-day mortality risk

Yuan Meng

Background

- On December 31st, 2019, a novel coronavirus (2019-nCoV) outbreak took China by storm
- Similar to SARS and MERS, typical symptoms of 2019-nCoV infections include pneumonia, fever, cough, and shortness of breath
- As of February 9th, 37,612 cases have been confirmed, among which 815 died and 2,990 recovered
- Mortality rate in China: 2.1% nationwide, 4.9% in Wuhan, 3.1% in Hubei, and 0.16% in other provinces



2019-nCoV

A hospital in Wuhan

Coronavirus Cases:

37,612

of which **6,198**
in **severe condition**

Deaths:

815

Recovered:

2,990

Live updates on Worldometer

How can we identify patients at risk?

Where to get data?

- Only two peer-review studies (Huang et al., 2020, Chen et al., 2020) examined patient features; *neither* shared data
- Pneumonia caused by 2019-nCoV has **similar symptoms** as that cause by other viruses, which may present similar challenges to respiratory and immune systems and **put similar patients at risk**
- **Data on viral pneumonia:** A previous study (Guo et al., 2019) collected data on 528 pneumonia patients infected by known viruses and shared this data

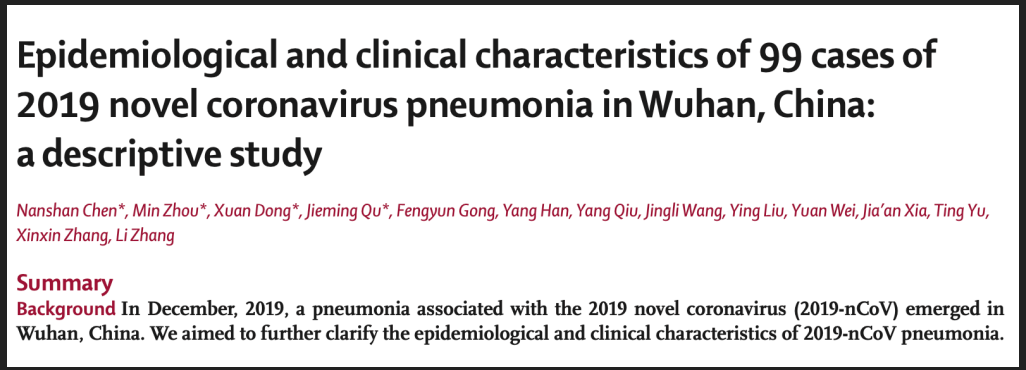
Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China

Chaolin Huang, Yeming Wang*, Xingwang Li*, Lili Ren*, Jianping Zhao*, Yi Hu*, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jiaan Xia, Yuan Wei, Wenjuan Wu, Xuelei Xie, Wen Yin, Hui Li, Min Liu, Yan Xiao, Hong Gao, Li Guo, Jungang Xie, Guangfa Wang, Rongmeng Jiang, Zhancheng Gao, Qi Jin, Jianwei Wang†, Bin Cao†*

Summary

Background A recent cluster of pneumonia cases in Wuhan, China, was caused by a novel betacoronavirus, the 2019 novel coronavirus (2019-nCoV). We report the epidemiological, clinical, laboratory, and radiological characteristics and treatment and clinical outcomes of these patients.

Huang et al. (2020), *The Lancet*



Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study

Nanshan Chen, Min Zhou*, Xuan Dong*, Jieming Qu*, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, Jia'an Xia, Ting Yu, Xinxin Zhang, Li Zhang*

Summary

Background In December, 2019, a pneumonia associated with the 2019 novel coronavirus (2019-nCoV) emerged in Wuhan, China. We aimed to further clarify the epidemiological and clinical characteristics of 2019-nCoV pneumonia.

Chen et al. (2020), *The Lancet*

Data from Guo et al. (2019)

- **Outcome variable:** whether patients died (1) or survived (0) within 90 days in hospital
- **61 explanatory variables**
  - **Socio-demographic:** age (5 year bins), body mass index, occupation, etc.
  - **Clinical:** CURB-65 score, lymphocyte count, community acquired vs. hospital acquired, virus species, use of ventilation machines, clinical test results (e.g., TB), etc.
  - **Others:** hospital fee, days in hospital, whether patient was put in ICU, etc.

Excerpt of dataset in Guo et al. (2010)

	C	D	E	F	G
	species type of virus	multivirus infection	inpatient day	department	deathinhospital (0survival 1death)
1	FluA	0	3/3/18	九舍七病区	0
2	FluA	0	2/4/19	呼吸八楼病区	0
3	RSVB	0	10/30/17	呼吸科监护	1
4	FluB	0	1/8/18	呼吸科监护	1
5	FluA	0	2/4/19	感染三楼病区	0
6	FluB	0	1/3/18	肾脏病区	0
7	CoV(OC43	0	12/31/17	9舍5病区	1
8	RSVB	0	3/13/18	九舍七病区	0
9	CoV(OC43	0	3/24/17	呼吸科监护	1
0	PIV3	0	4/17/17	九舍三病区	0
1	HMPV	0	1/18/19	感染四楼病区	0
2	FluB	0	1/29/18	九舍三病区	0
3	RSVA/FluB	1	3/7/19	肾脏病区	0
4	HRV	0	11/17/17	肾脏病区	0
5	AdV	0	5/5/15	消化二病区	0
6	HRV	0	11/25/17	肾脏病区	1
7	EV	0	2/19/18	呼吸八楼病区	0
8	RSVB	0	2/5/17	呼吸七楼病区	0
9	FluA	0	1/23/19	呼吸科监护(RIC	0
0	PIV3	0	4/9/16	呼吸七楼病区	0
1	FluA	0	2/12/19	呼吸科监护(RIC	0
2	FluA	0	2/18/19	呼吸科监护(RIC	0
3	AdV	0	11/06/2018	呼吸六楼病区	0
4	CoV(OC88/CMV	1	6/1/18	肾脏病区	0
5	FluA	0	2/13/19	内科急症病区	0
6	HMPV	0	2/1/16	急诊ICU	0
7	HRV	0	2/14/17	感染三楼病区	0
8	FluA	0	1/29/19	九舍三病区	0
9	FluA	0	3/12/18	肾脏病区	0
0	CoV(OC43	0	7/6/16	呼吸六楼病区	1
1	FluA	0	2/1/19	内科急症病区	0
2	AdV	0	5/20/18	急诊ICU	0
3	HMPV	0	6/10/15	呼吸七楼病区	0

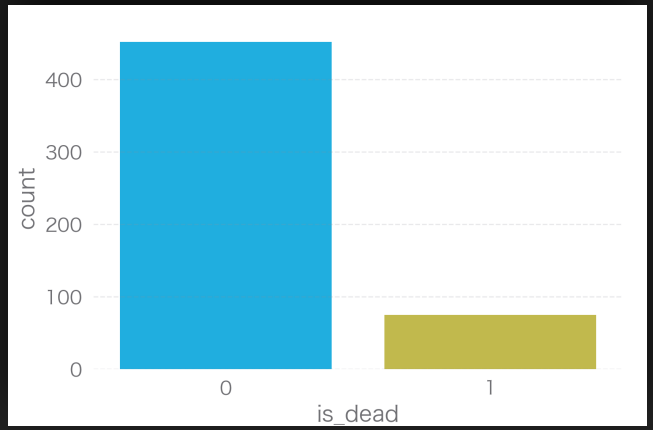
Goals and plans

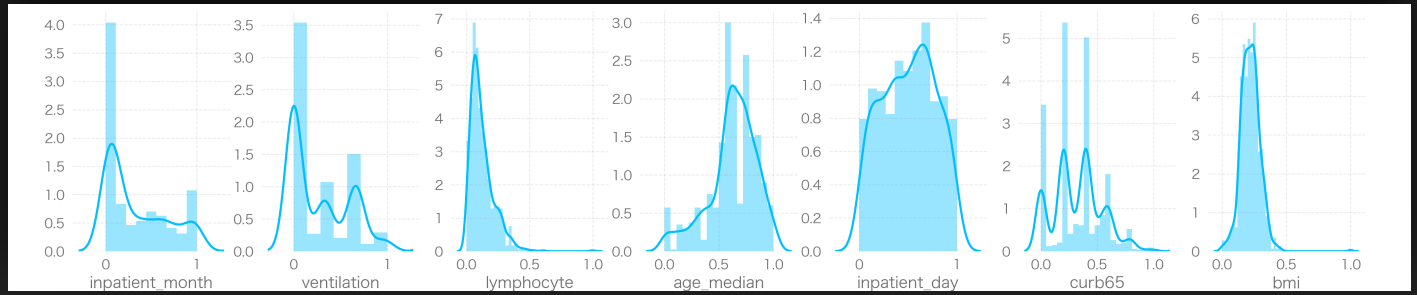
- Research questions
  - Predictive modeling: Given patients' socio-demographic, clinical, and other features, what's their 90-day mortality risk?
  - Feature importance: Which features contribute the most to mortality risk?
- Project plans
  - "Shotgun approach": Train and compare 5 classifiers

  - Select best model according to precision, recall, and F1 score
  - Extract feature importances viewed by the best performing model

Data exploration and feature engineering

- Outcomes are imbalanced: more survived (0) than died (1)
  - Oversample the minority class using synthetic minority over-sampling technique (SMOTE)
- Some variables have strings values
  - Use one-hot encoding to convert each unique string value into a binary feature (e.g., “age_60_65”: 0 or 1)
- Most continuous variables are not normally distributed
  - Use MinMaxScaler ($\frac{\text{value} - \min}{\text{range}}$) to normalize data



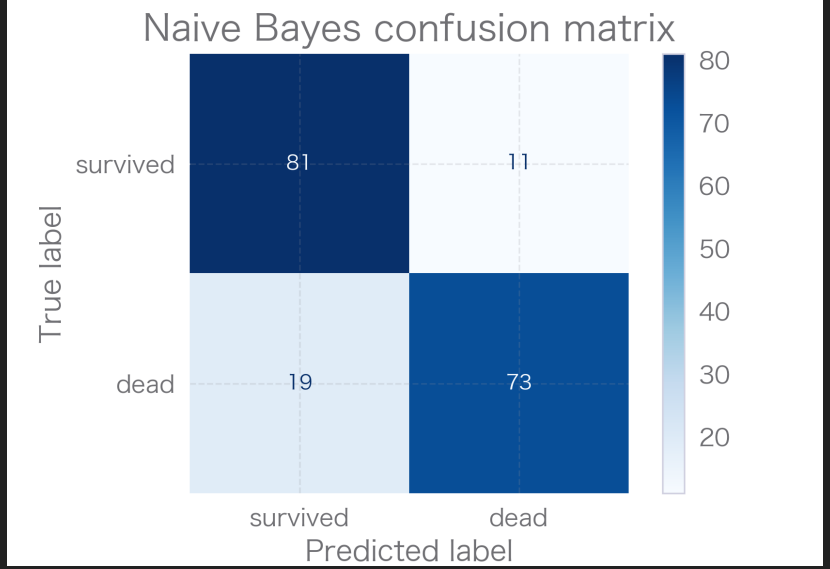


Saved 20% data for testing and trained 5 classifiers on the other 80%…

Naïve Bayes

- **Training:** “hybrid” Naive Bayes
  - Gaussian Naïve Bayes on continuous features
  - Bernoulli Naïve Bayes on binary features
  - Multiplied class assignment probabilities to obtain final results
- **10-fold cross validation** (on testing data): average F1 score across 10 folds is 0.87
- **Testing:** F1 score is 0.84

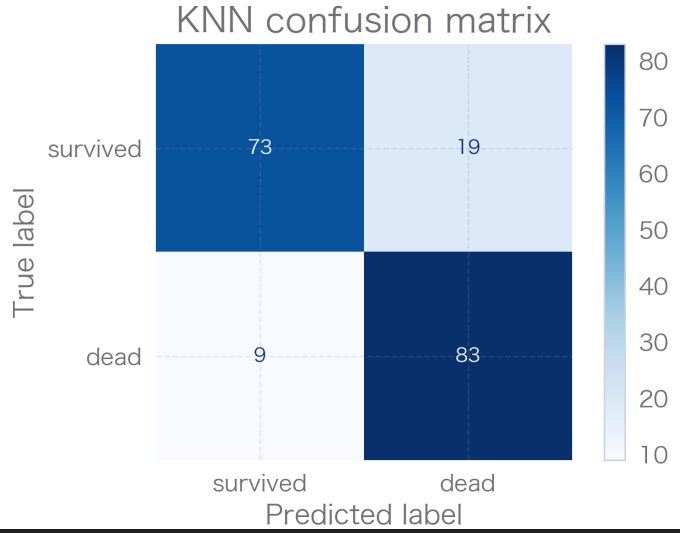
	precision	recall	f1-score	support
0	0.81	0.88	0.84	92
1	0.87	0.79	0.83	92
accuracy			0.84	184
macro avg	0.84	0.84	0.84	184
weighted avg	0.84	0.84	0.84	184



KNN

- **Hyper-parameter tuning:** tried from *k* = 1 to 26; best value is 3
- **Training:** used the best estimator with 3 nearest neighbors
- **10-fold cross validation** (on testing data): average F1 score across 10 folds is 0.92
- **Testing:** F1 score is 0.85
  - Potentially overfitting

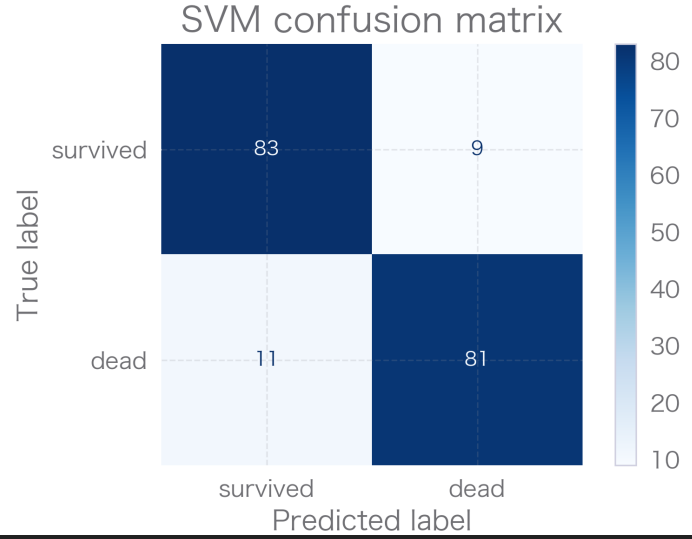
	precision	recall	f1-score	support
0	0.89	0.79	0.84	92
1	0.81	0.90	0.86	92
accuracy			0.85	184
macro avg	0.85	0.85	0.85	184
weighted avg	0.85	0.85	0.85	184



SVM

- **Hyper-parameter tuning:** used grid search through all combinations of C (cost of misclassification: [0.1, 1, 10, 100, 1000]) and γ (influence of single example: [1, 0.1, 0.01, 0.001, 0.0001])
- **Training:** used the best estimator (C = 10, γ = 0.1) to train model
- **10-fold cross validation** (on testing data): average F1 score across 10 folds is 0.93
- **Testing:** F1 score is 0.89
  - Potentially overfitting

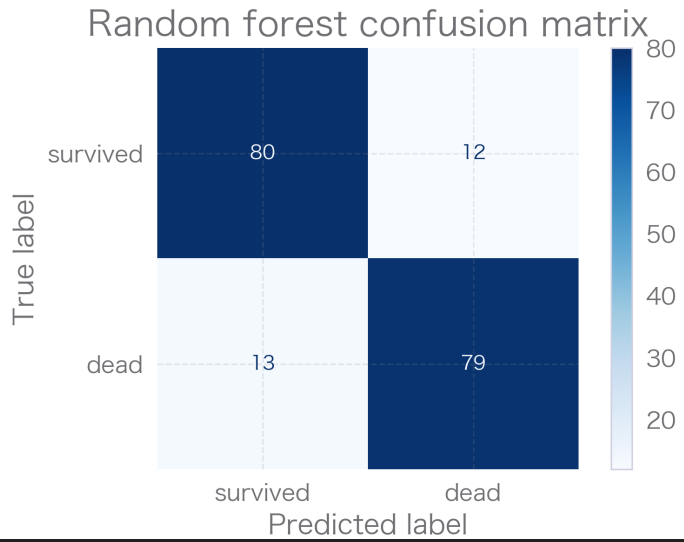
	precision	recall	f1-score	support
0	0.88	0.90	0.89	92
1	0.90	0.88	0.89	92
accuracy			0.89	184
macro avg	0.89	0.89	0.89	184
weighted avg	0.89	0.89	0.89	184



Random forest

- **Hyper-parameter tuning:** used randomized search through the parameter space (n_estimators: [100, 300, 500, 800, 1200], max_depth: [5, 8, 15, 25, 30], min_samples_split: [2, 5, 10, 15, 100], min_samples_leaf: [1, 2, 5, 10], and max_features: [“sqrt”, 0.25, 0.5, 0.75, 1.0])
- **Training:** used the best estimator (n_estimators = 500, min_samples_split = 5, min_samples_leaf = 1, max_features = 1.0, max_depth = 8) to train model
- **10-fold cross validation** (on testing data): average F1 score across 10 folds is 0.91
- **Testing:** F1 score is 0.86
  - Potentially overfitting

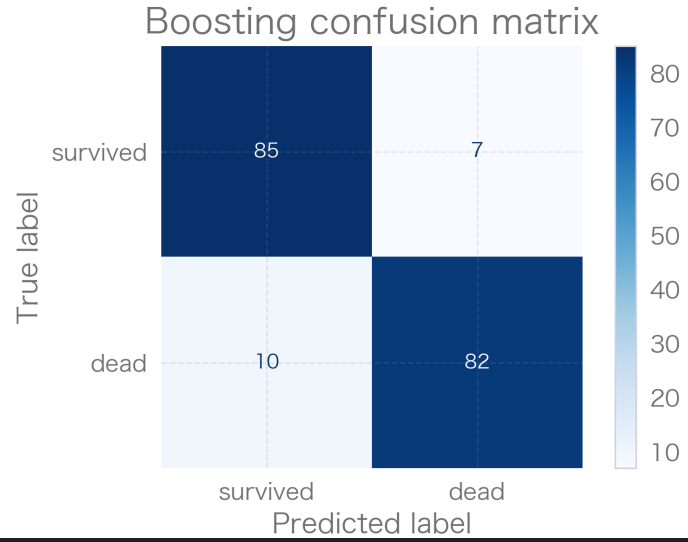
	precision	recall	f1-score	support
0	0.86	0.87	0.86	92
1	0.87	0.86	0.86	92
accuracy			0.86	184
macro avg	0.86	0.86	0.86	184
weighted avg	0.86	0.86	0.86	184



Gradient boosting

- **Hyper-parameter tuning:** used randomized search through the parameter space (n_estimators: [100, 300, 500, 800, 1200], max_depth: [5, 8, 15, 25, 30], min_samples_split: [2, 5, 10, 15, 100], min_samples_leaf: [1, 2, 5, 10], max_features: [“sqrt”, 0.25, 0.5, 0.75, 1.0], and learning_rate: [0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5])
- **Training:** used the best estimator (n_estimators = 800, min_samples_split = 100, min_samples_leaf = 5, max_features = 0.25, max_depth = 30, learning_rate = 0.1) to train model
- **10-fold cross validation** (on testing data): average F1 score across 10 folds is 0.95
- **Testing:** F1 score is 0.91
  - Potentially overfitting

	precision	recall	f1-score	support
0	0.89	0.92	0.91	92
1	0.92	0.89	0.91	92
accuracy			0.91	184
macro avg	0.91	0.91	0.91	184
weighted avg	0.91	0.91	0.91	184



Which classifier is the best?

Best model: Gradient boosting

- Highest precision, recall, and F1 score
- Fewest misclassified cases

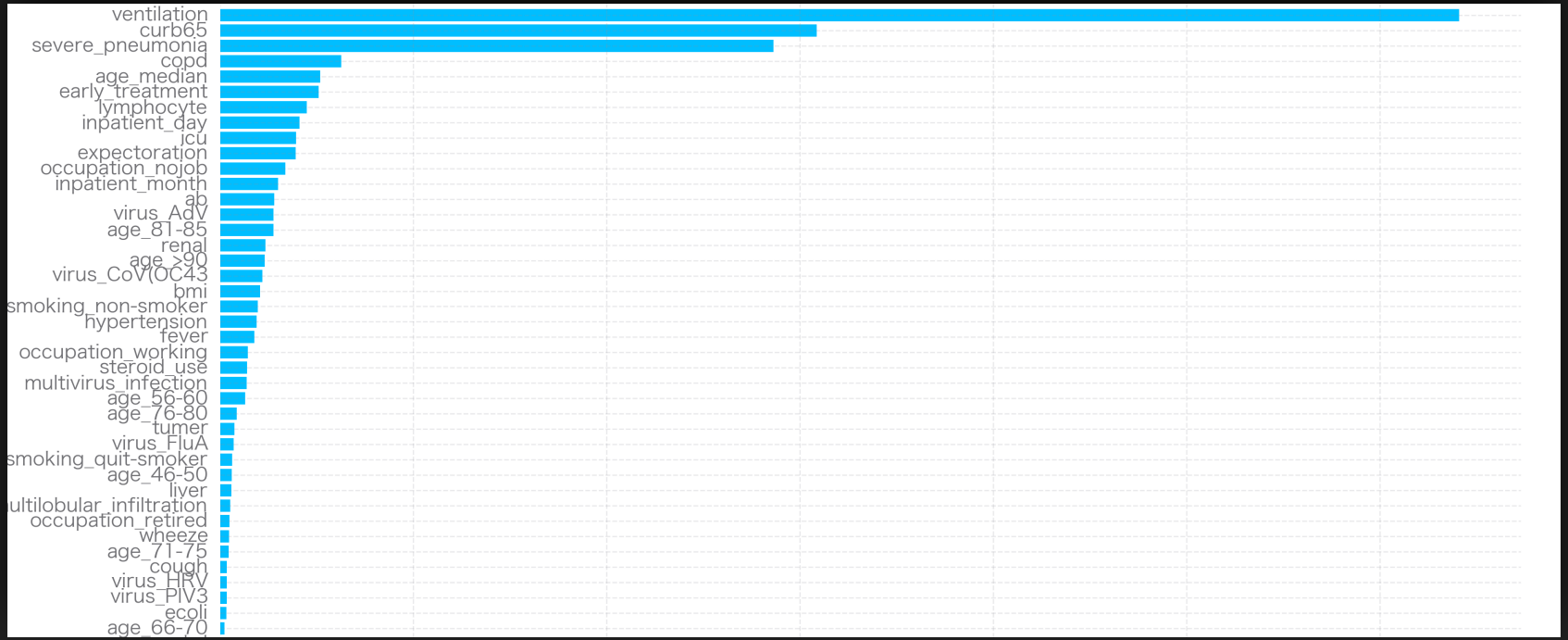
Second best: SVM

Worst model: Naïve Bayes

- Followed by KNN and random forests

classifier	precision	recall	f1_score
bst	0.908043	0.907609	0.907584
knn	0.851985	0.847826	0.847375
nb	0.839524	0.836957	0.836648
rf	0.864173	0.864130	0.864126
svm	0.891489	0.891304	0.891292

Which features are most important?



Gradient boosting classifier

Socio-demographic features that mattered the most are patients' age and BMI; clinical features that mattered the most are lymphocyte counts, CURB-65 scores, whether patients had severe pneumonia or chronic obstructive pulmonary disease, and whether they were using a ventilation machine.

Limitations

- **No direct data** on 2019-nCoV patients — had to assume because of similar symptoms, the same type of patients are at highest risk for both 2019-nCoV infections and other viral pneumonia
- **Overfitting:** apart from Naïve Bayes, all classifiers had higher F1 scores in the training data than in the testing data
- **Overlapping features:** e.g., CURB-65 already includes age information (“age≥65”), which may be redundant if we also use age
- **Further improvement:** current highest F1 score is 0.91, which may be further improved using more advanced techniques