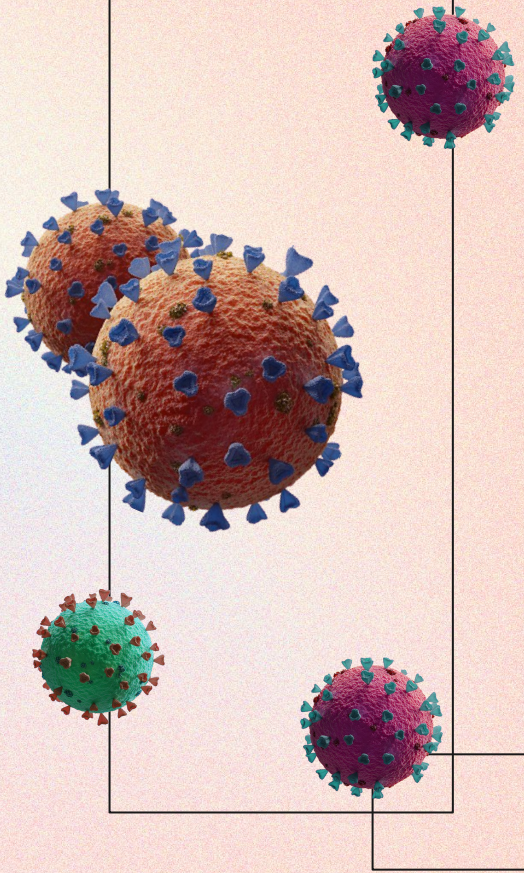
A Novel Diagnosis

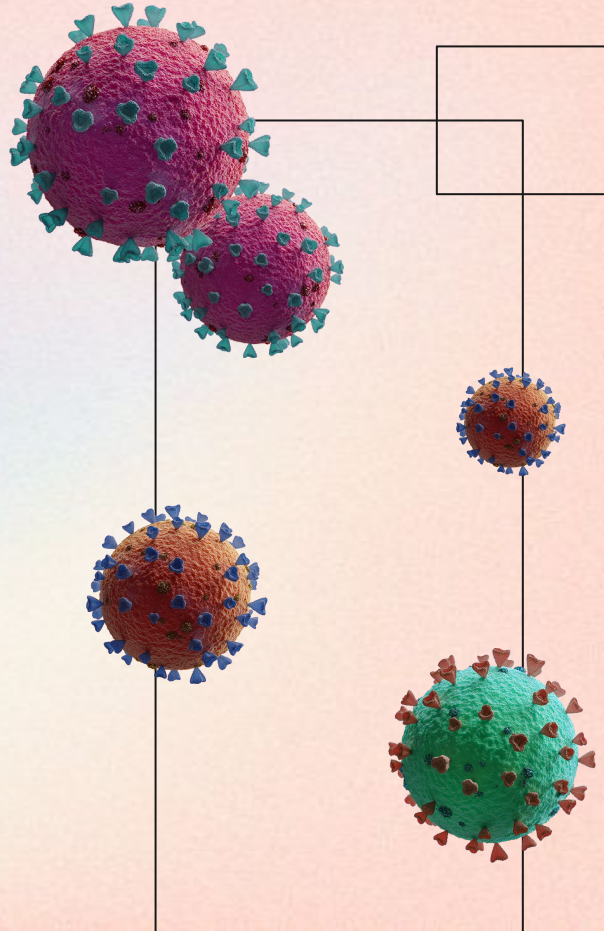
PM511B Spring 2022
Group 4:
Mario Gastelum
Naghmeh Aminzadeh
Bryan Queme



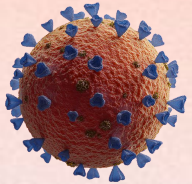
Background

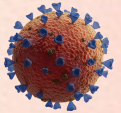
Aim: To build a prediction model to determine which demographic and symptom variables are related to COVID-19 diagnosis.

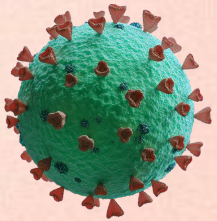
Source: Data obtained from the Israeli Ministry of Health.



Questions







1. Which symptom most highly predict COVID-19 diagnosis?
2. What if we count only 1 symptom, regardless of how many they have?
3. What if we only looked at < 60 y.o.?
4. Can we correctly identify 99.5% of all individuals who actually have COVID-19?
5. Does predictive power of “test_indication” increase or decrease as time goes on?

Data Set Content

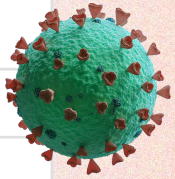
Outcome Variable: Corona_Result

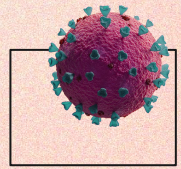
Almost **300k** observations

*cough: n=252- None
*fever: n=252 - None
*sore_throat: n=1 - None
*shortness_of_breath: n=1- None
*head_ache: n=1 - None
*corona_result: n=3,892 - "other"
*age_60_and_above: n=127,320 - "None"
*gender: n=19,563 - "None"

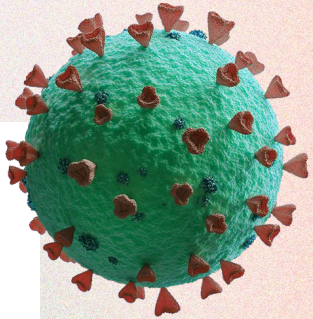
After data cleaning:
136,294 observations

Variable	Definition	Values
Gender		Male/female
Age_60_and_above	Age >= 60 years	1=yes, 0=no
Cough	Presence of cough	1=yes, 0=no
Fever	Presence of fever >100F	1=yes, 0=no
Sore_throat	Presence of sore throat	1=yes, 0=no
Head_ache	Presence of recent headache	1=yes, 0=no
Shortness_of_breath	Shortness of breath	1=yes, 0=no
Test_indication	History of possible exposure	Abroad = traveled abroad, Contact with confirmed = contact with confirmed case, Other = other
Corona_result	Result of COVID-19 test	Negative/positive



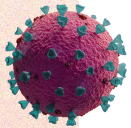


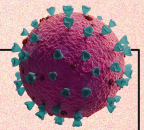
Exploratory Univariate Analysis



Univariate analysis:

Variable	P-value
Gender	<.001
Age_60_and_above	<.001
Cough	<.001
Fever	<.001
Sore_throat	<.001
Head_ache	<.001
Shortness_of_breath	<.001
Test_indication	<.001
OUTCOME: **Corona_result**	Result of COVID-19 test





Final Prediction Model #1

Fitted Final Model: includes all variables

Symptom that most strongly predicts Corona result:
Headache: The odds of having a positive coronavirus result increases by 137 times if there is a headache. P-value <.001
95%CI(104.4987,179.8033)

Methods:
85% training set (n=115,850)
15% testing set
Univariate
Stepwise Method
Interaction
GOF

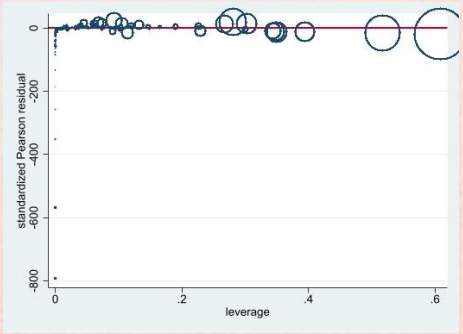
Logistic regression — Number of obs = 115,850; LR chi2(9) = 28793.93; Prob > chi2 = 0.0000; Log likelihood = -17259.458; Pseudo R2 = 0.4548

corona_result_c	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
cough_r	1.083006	.0422961	2.04	0.041	1.0032	1.16916
fever_r	5.489767	.2257757	41.41	0.000	5.064619	5.950603
sore_throat_r	58.463	7.520404	31.63	0.000	45.43459	75.22733
shortness_of_breat~r	65.4781	9.537138	28.71	0.000	49.21701	87.11177
head_ache_r	137.0738	18.97699	35.54	0.000	104.4987	179.8033
age60_C	1.652949	.0629116	13.20	0.000	1.534131	1.780969
gender_c	1.442147	.044746	11.80	0.000	1.35706	1.532569
test_indication_c						
1	.0283835	.0014738	-68.60	0.000	.025637	.0314242
2	.0220151	.0008678	-96.81	0.000	.0203783	.0237834
_cons	.7557731	.0299107	-7.08	0.000	.6993651	.8167306

Note: _cons estimates baseline odds.

Collinearity Diagnostics

Variable	VIF	SQRT VIF	Tolerance	R-Squared
cough_r	1.49	1.22	0.6707	0.3293
fever_r	1.31	1.14	0.7635	0.2365
sore_throat_r	1.16	1.08	0.8600	0.1400
shortness_of_breath_r	1.08	1.04	0.9221	0.0779
head_ache_r	1.20	1.10	0.8319	0.1681
age60_C	1.01	1.00	0.9942	0.0058
gender_c	1.00	1.00	0.9970	0.0030
test_indication_c	1.39	1.18	0.7177	0.2823
Mean VIF	1.21			



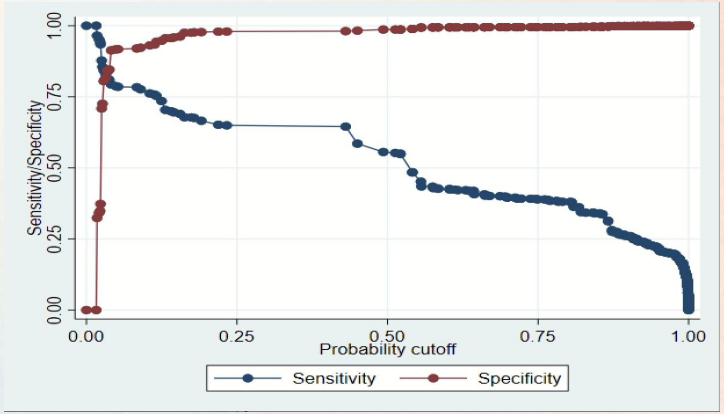
Model#1 Validation: Training Set

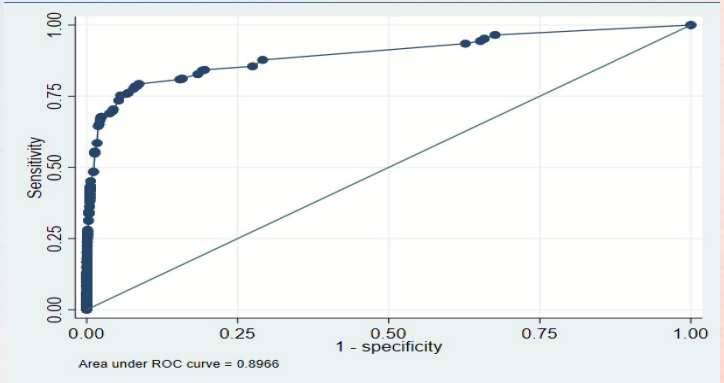
Logistic model for corona_result_c

Classified	True D	True ~D	Total
+	7144	9244	16388
-	1864	97598	99462
Total	9008	106842	115850

Classified + if predicted Pr(D) >= .0396676
True D defined as corona_result_c != 0

Sensitivity	Pr(+| D)	79.31%
Specificity	Pr(-|~D)	91.35%
Positive predictive value	Pr(D| +)	43.59%
Negative predictive value	Pr(~D| -)	98.13%
False + rate for true ~D	Pr(+|~D)	8.65%
False - rate for true D	Pr(-| D)	20.69%
False + rate for classified +	Pr(~D| +)	56.41%
False - rate for classified -	Pr(D| -)	1.87%
Correctly classified		90.41%





Area under ROC: 89.66%

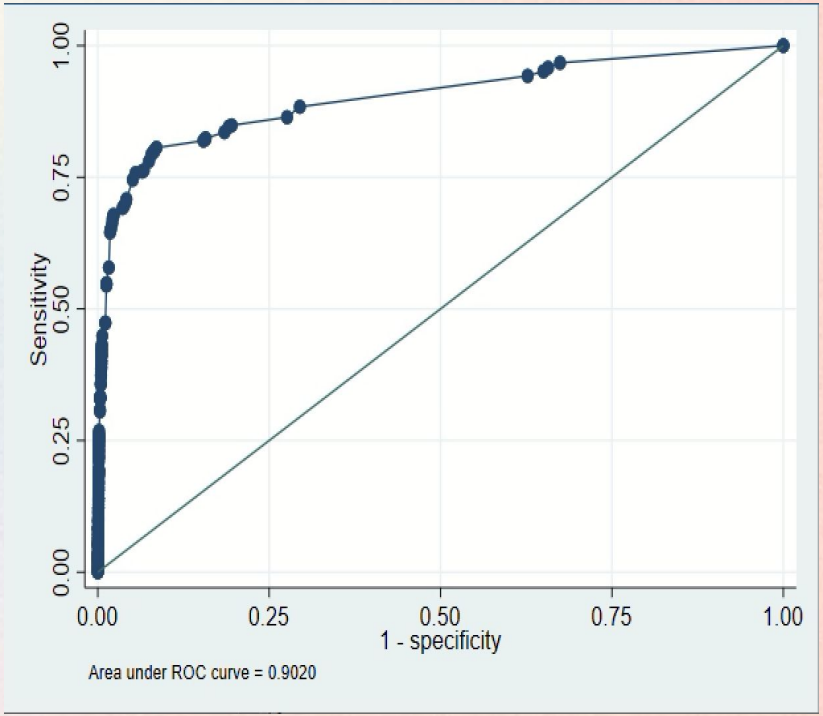
Model #1 Validation: Testing Set

Logistic model for corona_result_c

Classified	True D	True ~D	Total
+	1304	1611	2915
-	314	17215	17529
Total	1618	18826	20444

Classified + if predicted Pr(D) >= .0396676
True D defined as corona_result_c != 0

Sensitivity	Pr(+| D)	80.59%
Specificity	Pr(-|~D)	91.44%
Positive predictive value	Pr(D| +)	44.73%
Negative predictive value	Pr(~D| -)	98.21%
False + rate for true ~D	Pr(+|~D)	8.56%
False - rate for true D	Pr(-| D)	19.41%
False + rate for classified +	Pr(~D| +)	55.27%
False - rate for classified -	Pr(D| -)	1.79%
Correctly classified		90.58%

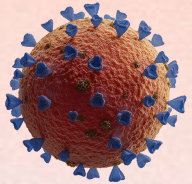


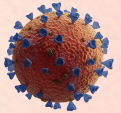
Area under ROC: 90.20%

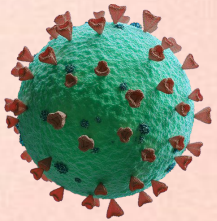
Questions

1. Which symptom most highly predict COVID-19 diagnosis?
2. What if we create a new variable for any symptom expressed?
3. What if we only looked at < 60 y.o.?
4. Can we correctly identify 99.5% of all individuals who actually have COVID-19?
5. Does predictive power of “test_indication” increase or decrease as time goes on?

Headache





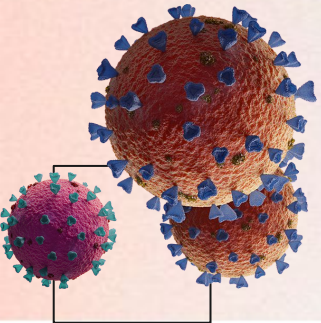


Prediction Model #2: Symptomatic?

Final Model:
All symptoms combined

Methods:
Generated variable if symptomatic

Results:
Correctly Classified: 85.80%
AUROC: 0.8835



Logistic regression

Number of obs = 115,850
LR chi2(5) = 22501.60
Prob > chi2 = 0.0000
Pseudo R2 = 0.3554

Log likelihood = -20405.625

corona_result_c	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
symptomatic	8.106773	.2506193	67.69	0.000	7.630154	8.613165
age60_C	1.678042	.057378	15.14	0.000	1.569269	1.794355
gender_c	1.392978	.0388299	11.89	0.000	1.318915	1.471201
test_indication_c						
1	.0244277	.0011092	-81.75	0.000	.0223477	.0267014
2	.0363858	.0013198	-91.36	0.000	.0338889	.0390666
_cons	.4670471	.018143	-19.60	0.000	.4328075	.5039954

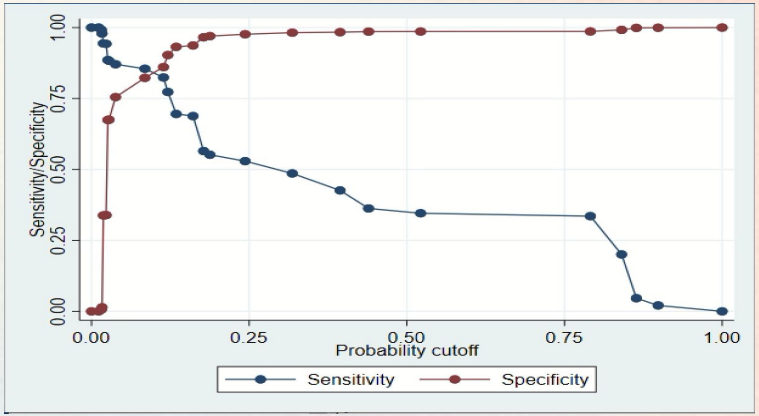
Note: _cons estimates baseline odds.

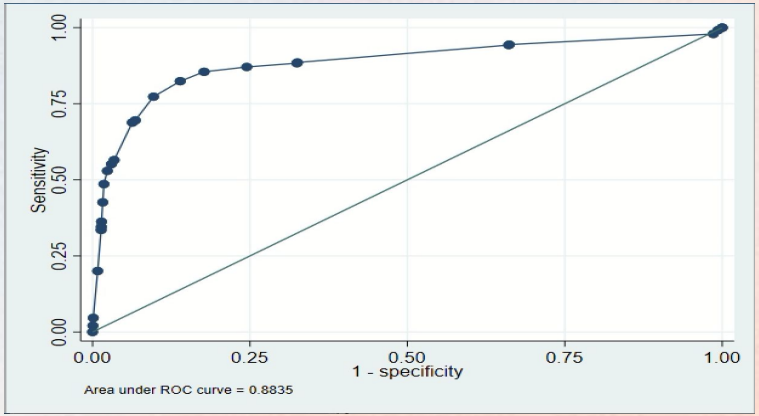
Model #2 Validation: Training Set

Classified	True D	True ~D	Total
+	7423	14867	22290
-	1585	91975	93560
Total	9008	106842	115850

Classified + if predicted Pr(D) >= .0993954
True D defined as corona_result_c != 0

Sensitivity	Pr(+| D)	82.40%
Specificity	Pr(-|~D)	86.09%
Positive predictive value	Pr(D| +)	33.30%
Negative predictive value	Pr(~D| -)	98.31%
False + rate for true ~D	Pr(+|~D)	13.91%
False - rate for true D	Pr(-| D)	17.60%
False + rate for classified +	Pr(~D| +)	66.70%
False - rate for classified -	Pr(D| -)	1.69%
Correctly classified		85.80%





Area under ROC:
.8835

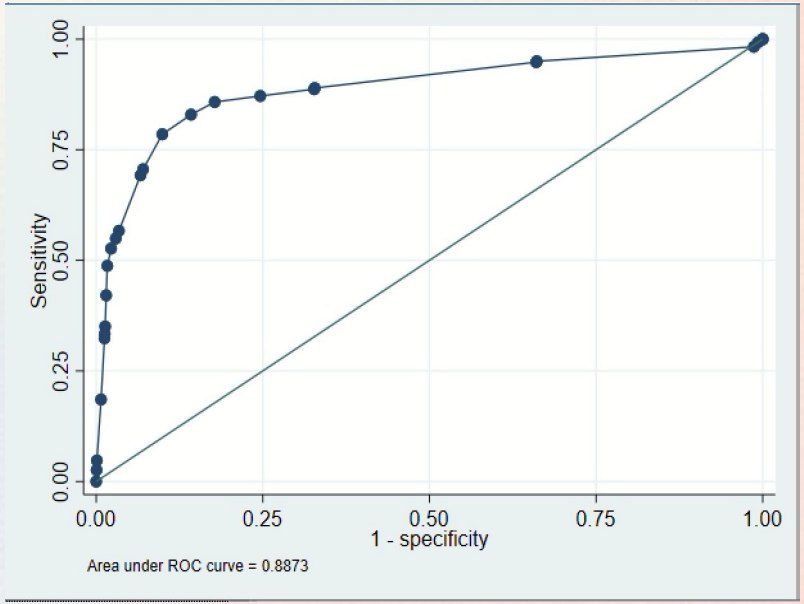
Model #2 Validation: Testing Set

Logistic model for corona_result_c

Classified	True D	True ~D	Total
+	1342	2680	4022
-	276	16146	16422
Total	1618	18826	20444

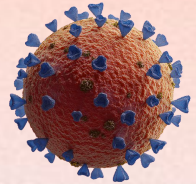
Classified + if predicted Pr(D) >= .0993954
True D defined as corona_result_c != 0

Sensitivity	Pr(+| D)	82.94%
Specificity	Pr(-|~D)	85.76%
Positive predictive value	Pr(D| +)	33.37%
Negative predictive value	Pr(~D| -)	98.32%
False + rate for true ~D	Pr(+|~D)	14.24%
False - rate for true D	Pr(-| D)	17.06%
False + rate for classified +	Pr(~D| +)	66.63%
False - rate for classified -	Pr(D| -)	1.68%
Correctly classified		85.54%



Area under ROC: .8873

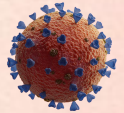
Questions

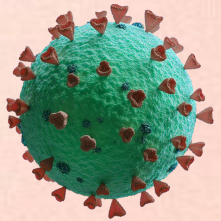


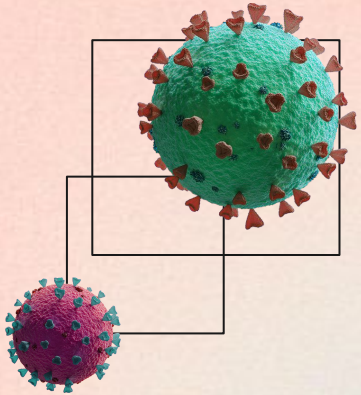
1. Which symptom most highly predict COVID-19 diagnosis?
2. What if we count only 1 symptom, regardless of how many they have?
3. What if we only looked at < 60 y.o.?
4. Can we correctly identify 99.5% of all individuals who actually have COVID-19?
5. Does predictive power of “test_indication” increase or decrease as time goes on?

Headache

85% correct class







Prediction Model #3: Age<60 years?

Final Model: only < 60 y.o.

Methods:
Separated by age groups < 60 & >=60
Most of our data set is <60 years old

Results:
Cough was removed from the model.

Logistic regression

Number of obs = 95,704
LR chi2(7) = 23687.34
Prob > chi2 = 0.0000
Log likelihood = -13961.263
Pseudo R2 = 0.4590

corona_resul~c	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
fever_r	4.832609	.208756	36.47	0.000	4.440296	5.259583
sore_throat_r	56.1456	7.517349	30.08	0.000	43.18652	72.99336
shortness_of~r	49.11548	8.180022	23.38	0.000	35.43673	68.07431
head_ache_r	149.7506	22.34038	33.58	0.000	111.7849	200.6107
gender_c	1.367548	.0472449	9.06	0.000	1.278015	1.463353
test_indicat~c						
1	.0301724	.0016819	-62.80	0.000	.0270496	.0336558
2	.0219496	.0008891	-94.29	0.000	.0202744	.0237631
_cons	.833365	.0322884	-4.70	0.000	.7724241	.8991138

Note: _cons estimates baseline odds.

Prediction Model #3: Age<60 years?

Final Model:
All variables except cough

Methods:
Separated by age groups < 60 & >=60

Results:
Cough was removed from the model.

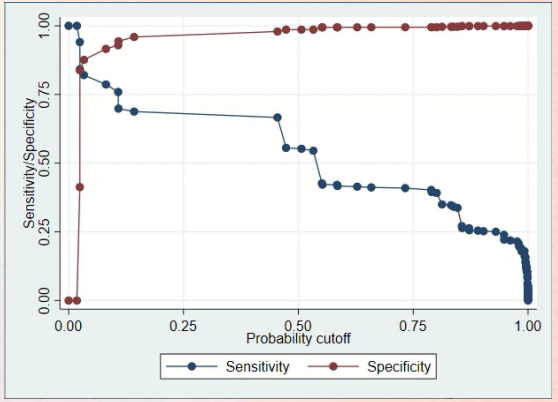
ROC: 0.8950

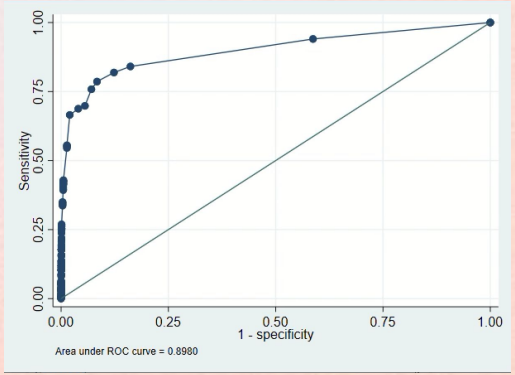
Logistic model for corona_result_c

Classified	True D	True ~D	Total
+	5740	7400	13140
-	1562	81002	82564
Total	7302	88402	95704

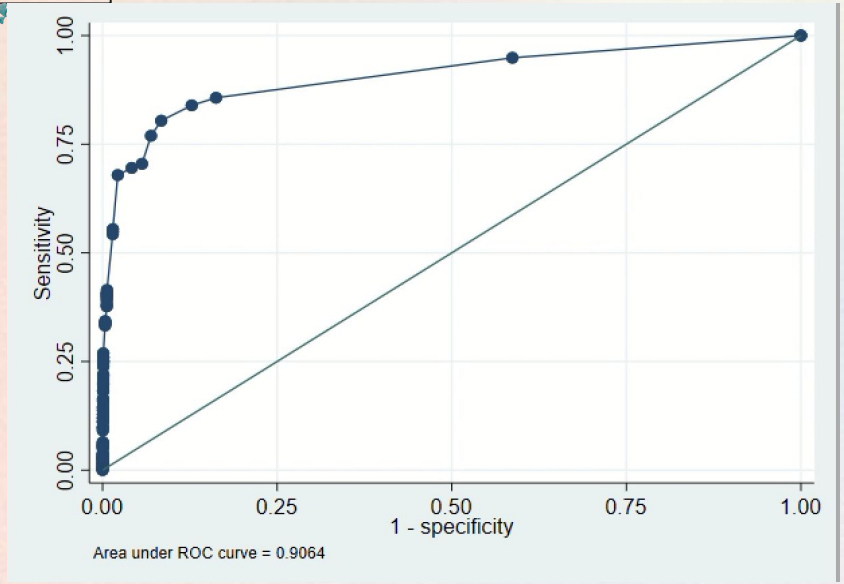
Classified + if predicted Pr(D) >= .057231
True D defined as corona_result_c != 0

Sensitivity	Pr(+| D)	78.61%
Specificity	Pr(-|~D)	91.63%
Positive predictive value	Pr(D| +)	43.68%
Negative predictive value	Pr(~D| -)	98.11%
False + rate for true ~D	Pr(+|~D)	8.37%
False - rate for true D	Pr(-| D)	21.39%
False + rate for classified +	Pr(~D| +)	56.32%
False - rate for classified -	Pr(D| -)	1.89%
Correctly classified		90.64%





Prediction Model #3: Age<60 years?



Logistic model for corona_result_c

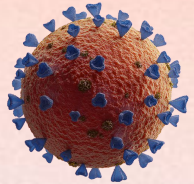
Classified	True D	True ~D	Total
+	1022	1313	2335
-	249	14305	14554
Total	1271	15618	16889

Classified + if predicted Pr(D) >= .057231
True D defined as corona_result_c != 0

Sensitivity	Pr(+| D)	80.41%
Specificity	Pr(-|~D)	91.59%
Positive predictive value	Pr(D| +)	43.77%
Negative predictive value	Pr(~D| -)	98.29%
False + rate for true ~D	Pr(+|~D)	8.41%
False - rate for true D	Pr(-| D)	19.59%
False + rate for classified +	Pr(~D| +)	56.23%
False - rate for classified -	Pr(D| -)	1.71%
Correctly classified		90.75%

ROC: 0.9064

Questions

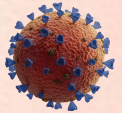


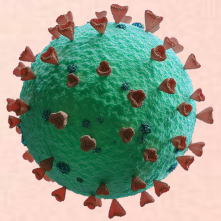
1. Which symptom most highly predict COVID-19 diagnosis?
2. What if we count only 1 symptom, regardless of how many they have?
3. What if we only looked at < 60 y.o.?
4. Can we correctly identify 99.5% of all individuals who actually have COVID-19?
5. Does predictive power of “test_indication” increase or decrease as time goes on?

Headache

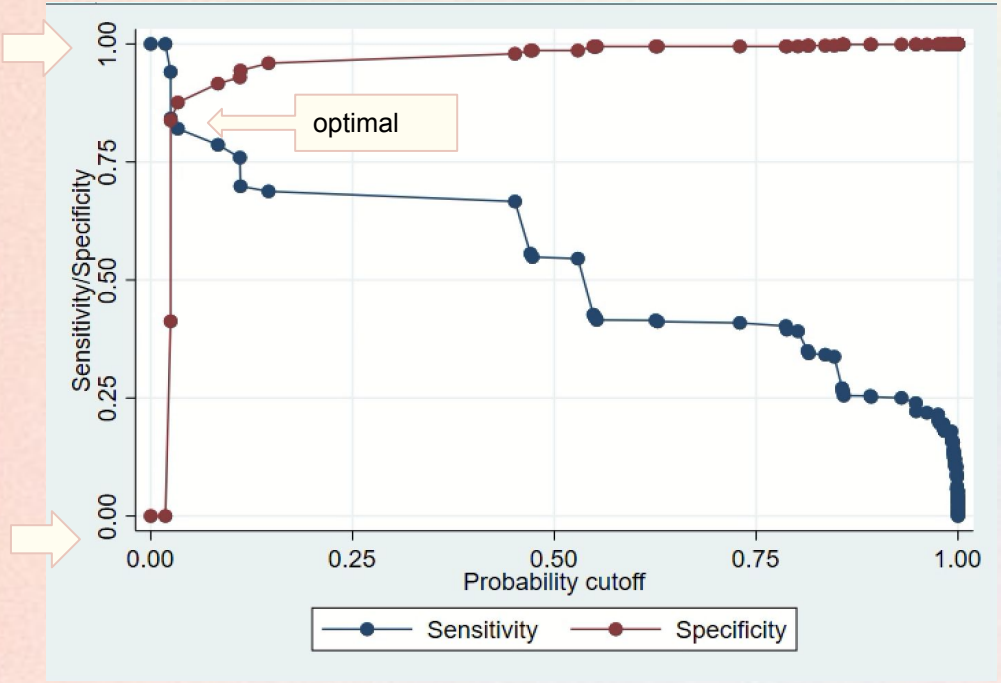
85% correct class

Cough var removed





Can we correctly identify 99.5% of all individuals who actually have COVID-19?

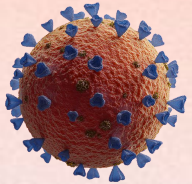


Sensitivity - true positives

Specificity - true negatives

Yes. However, by increasing our sensitivity cutoff we lose specificity, making our model useless.

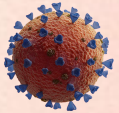
Questions



1. Which symptom most highly predict COVID-19 diagnosis?
2. What if we count only 1 symptom, regardless of how many they have?
3. What if we only looked at < 60 y.o.?
4. Can we correctly identify 99.5% of all individuals who actually have COVID-19?
5. Does predictive power of “test_indication” increase or decrease as time goes on?

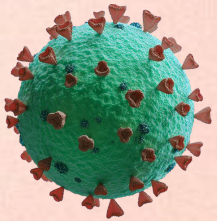
Headache

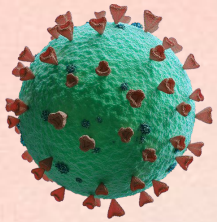
85% correct class



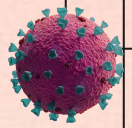
Cough not significant

Yes, but useless model





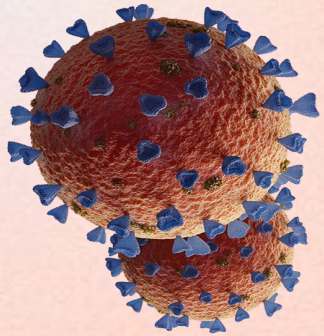
Does predictive power of “test_indication” increase or decrease as time goes on?



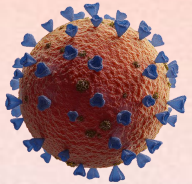
Method:
Test date: 4 equal size groups

	Time 1	Time 2	Time 3	Time 4
Pseudo R^2	0.3372	0.3131	0.2684	0.2589

Results:
Test Indication loses power slightly as time goes.



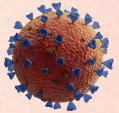
Questions



1. Which symptom most highly predict COVID-19 diagnosis?
2. What if we count only 1 symptom, regardless of how many they have?
3. What if we only looked at < 60 y.o.?
4. Can we correctly identify 99.5% of all individuals who actually have COVID-19?
5. Does predictive power of “test_indication” increase or decrease as time goes on?

Headache

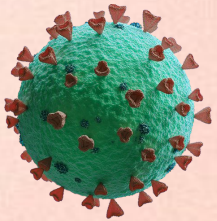
85% correct class



Cough not significant

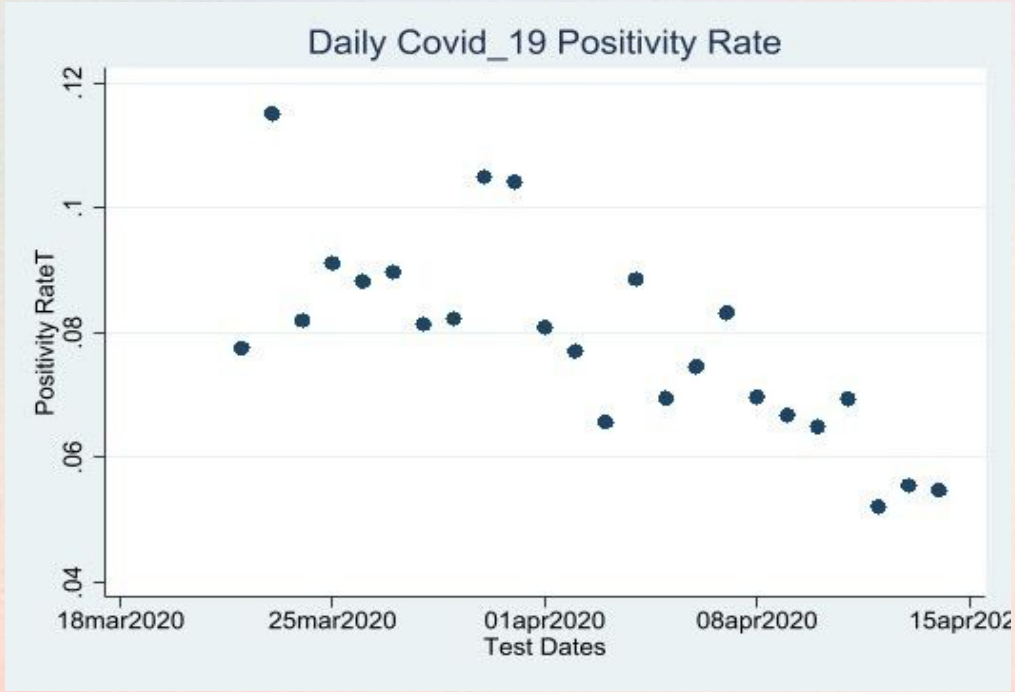
Yes, but useless model

It decreases slightly



Bonus

- Graph daily test positivity rate



THANK YOU!