The Lancet Public Health

Developing a Pre-Testing Diagnostic Tool for COVID-19 Using Big Data Predictive Analytics

--Manuscript Draft--

Manuscript Number:	
Article Type:	Article (Original Research)
Keywords:	COVID-19; SARS-COV-2; PCR tests; Machine Learning; Predictive Analytics; Pre-Testing; Big Data
Corresponding Author:	Ramy Elitzur, Ph.D. University of Toronto Toronto, Ontario CANADA
First Author:	Ramy Elitzur, Ph.D.
Order of Authors:	Ramy Elitzur, Ph.D.
	Dmitry Krass, Ph.D.
	Eyal Zimlichman, M.D.
Manuscript Region of Origin:	CANADA
Abstract:	Background Standard Polymerase Chain reaction (PCR) tests for SARS-COV-2 are in short supply to meet demand in many countries presenting a need to improve testing efficiency. Pre-testing tools can be used to ensure continued public safety as systems move through the pandemic. In this study we set out to create an instrument based on big data predictive tools to assess pre-test probability for COVID-19. Methods We analyzed data reported by the Israeli Ministry of Health (IMOH) for standard PCR tests done for SARS-COV-2 from March to April, 2020, overall 108,852 cases. Demographics and symptoms of the patients were collected at time of testing. Four supervised machine learning algorithms were used to analyze 20,537 test results of cases who presented with symptoms. Model results were used to develop efficient pre-test diagnostic tool. Findings Of symptomatic patients tested, 6,427 (31.3%) tested positive for SARS-COV-2, and 14,110 (68.7%) tested negative. In all models used headache, shortness of breath, sore throat, fever, and having contact with an infected person came up as most predictive of a positive test. The area under the curve of the receiver operating characteristic curve for the test sample was found to be 0.88 and the misclassification rate was between 4.7% and 6.5% for all predictive models, demonstrating effective classification ability. Using our pre-test probability screening tool with conventional PCR testing can potentially increase efficiency by 141%. Interpretation We suggest a simple diagnostic pre-test tool for assessing the probability of infection can increase efficiency of testing and effectiveness of public health COVID-19 programs. Funding None

Developing a Pre-Testing Diagnostic Tool for COVID-19 Using Big Data Predictive Analytics

Prof. Ramy Elitzur*

Prof. Dmitry Krass*

Dr. Eyal Zimlichman*

* **Contributed equally**

Rotman School of Management, University of Toronto (Prof. Ramy Elitzur PhD and Prof. Dmitry Krass PhD) and Sheba Medical Center and Sackler School of Medicine, Tel Aviv University (Dr. Eyal Zimlichman, MD)

Correspondence to: Dr. Ramy Elitzur, the Edward Kernaghan Professor of Financial Analysis. 105 St. George St., Toronto, Ontario, Canada, M5S 3E6
ramy.elitzur@rotman.utoronto.ca

Acknowledgements: The authors thank Nina Ahuja MD, David N. Fisman MD, David Naylor MD, Alexander Forstner, Paul Heakes, and Karen Steinmann for their insightful comments.

Abstract

Background Standard Polymerase Chain reaction (PCR) tests for SARS-COV-2 are in short supply to meet demand in many countries presenting a need to improve testing efficiency. Pre-testing tools can be used to ensure continued public safety as systems move through the pandemic. In this study we set out to create an instrument based on big data predictive tools to assess pre-test probability for COVID-19.

Methods We analyzed data reported by the Israeli Ministry of Health (IMOH) for standard PCR tests done for SARS-COV-2 from March to April, 2020, overall 108,852 cases. Demographics and symptoms of the patients were collected at time of testing. Four supervised machine learning algorithms were used to analyze 20,537 test results of cases who presented with symptoms. Model results were used to develop efficient pre-test diagnostic tool.

Findings Of symptomatic patients tested, 6,427 (31.3%) tested positive for SARS-COV-2, and 14,110 (68.7%) tested negative. In all models used headache, shortness of breath, sore throat, fever, and having contact with an infected person came up as most predictive of a positive test. The area under the curve of the receiver operating characteristic curve for the test sample was found to be 0.88 and the misclassification rate was between 4.7% and 6.5% for all predictive models, demonstrating effective classification ability. Using our pre-test probability screening tool with conventional PCR testing can potentially increase efficiency by 141%.

Interpretation We suggest a simple diagnostic pre-test tool for assessing the probability of infection can increase efficiency of testing and effectiveness of public health COVID-19 programs.

Funding None

Research in context

Evidence before this study

The COVID-19 crisis is still unfolding but it is already clear that it has far reaching consequences beyond the disease itself, such as the related economic fallout. As such, it is crucial for public health agencies to be able to have reliable data on the presence of SARS-COV-2. We searched the University of Toronto Libraries database for journal articles (964 results) and Google Scholar (16 results) for articles searching for the terms "COVID-19" and "SARS-COV-2". We also conducted searches for the terms "COVID-19"and "SARS-COV-2" on the websites of *British Medical Journal* (19 results), *Journal of the American Medical Association* (188 results), *Lancet* (845 results), *Nature* (532 results), and *New England Journal of Medicine* (218 results). Removing duplicate references and focusing on symptoms screening, clinical characteristics, digital technology, big-data, artificial intelligence and predictive analytics resulted in 30 cited articles. Adding the two references to the Israeli Ministry of Health (IMOH) database yielded the 32 references used in this study. The first search was conducted on April 1, 2020 and the last one on June 20, 2020. The only studies that we were able to identify dealt with other aspects of big data tools in the context of the COVID-19, such as why such tools are needed, or predicting severe outcomes for hospitalized patients. No studies have explicitly analyzed big data to assess pre-test probability for a positive PCR test. One possible reason for the lack of studies on this subject is that the only large dataset on the PCR tests' results (both positive and negatives), as far as we know, is the IMOH one and some of the data fields appear in Hebrew, which might present a language barrier.

Added value of this study

In this study we have utilized data collected at time of testing for SARS-COV2 to develop a predictive tool able to assess pre-test probability for a positive PCR test. We found that the highest risk factors are, in order: Headache, Sore Throat, Shortness of Breath, Contact with a Known Carrier, and Fever. It is also notable that Abroad and

Cough both have negative coefficients – indicating that, all else being equal, individuals with these factors are less likely to test positive. Based on these risk factors, we have tested several big data analytic models and developed a tool based on the strongest model. This tool is included as an Appendix to this manuscript. We are also showing how use of this tool for pre-test screening can boost test efficiency significantly.

Implications of all the available evidence

We propose that such a tool can be used to improve testing efficiency when used as a pre-test screening tool. With the often shortage in PCR testing capacity, public health agencies should consider utilizing such screening for symptomatic patients sent for diagnostic testing. Such a tool can also be used to prioritize which case to handle first when a batch arrives at the lab, with a purpose of minimizing turnaround time for cases more likely to be positive. We also suggest that pre-test screening can serve as a decision aid when re-testing is considered in high risk cases for COVID-19 that test negative on PCR. Further research is needed to provide evidence for this utility.

Introduction

One of the main challenges confronting healthcare systems in the COVID-19 pandemic is accurate and efficient diagnostic strategies. While large-scale testing has been implemented by many countries, the supply of test kits is often insufficient to meet the demand. The supply shortages are likely to be exacerbated as the economies begin to reopen and mass testing of patients will be required in a variety of settings ranging from outpatient clinics to manufacturing facilities. Another constraint on testing capacity is that a proper administration of the test requires training – as noted below, incorrect sample collection is thought to account for the low sensitivity of the Polymerase Chain Reaction (PCR) test in the field. Finally, there is often a significant turn-around time until the test results are known.

Diagnostic tools based on predictive modeling can be clinically useful in several ways. Perhaps the most direct use is in using pre-test screening. A tool with sufficient accuracy is able to improve efficiency (calculated as true positive divided by number tested) considerably and allow to better utilize PCR test kits. Furthermore, such pre-test tools can be useful in obtaining instantaneous risk assessment when SARS-COV2 tests are either unavailable or the results are delayed.

The notion that big-data and machine learning tools can help curb the COVID-19 pandemic has been gaining traction lately[1-5]. Specifically, this paper aims to use data analytics to enhance efficiency of PCR testing[6-11]. The study uses predictive analytics tools to identify symptoms and characteristics of those people infected with SARS-COV-2 and, in turn, develop pre-testing diagnostic tools, which can be used by health professionals.

Methods

Data Source

We analyzed data on 126,067 COVID-19 initial standard PCR tests were conducted by the Israeli Ministry of Health (IMOH) from March 22, 2020 to April 15, 2020[12]. The tests were done using both an oral and a nasal swab. We dropped 17,215 of the tests because they were missing data. This, in turn, led to a sample of 108,852 observations for this period. We used 20,537 observations, related to symptomatic individuals (symptomatics) to generate predictive analytics for symptomatics.

The data that was collected by the IMOH included the results of the tests (positive or negative) as well as information collected from tested individuals including symptoms (cough, fever, shortness of breath, headache), gender, whether the tested individual was over 60 years old, came into contact with a known infected individual, or recently returned from a trip abroad[13]. All these variables were coded as binary. According to the IMOH, symptoms are based on self-reporting by the patients. The IMOH reports that a comprehensive epidemiological investigation was conducted by the IMOH, and specific inquiries were made about the symptoms, for individuals who tested positive. For negatively tested individuals the data was not consistently collected by the IMOH and no direct questions were asked.

While individuals presenting with at least one symptom were always tested, the criteria for testing asymptomatic individuals evolved over time. In the early stages of the testing policy, the focus was on those returning from a trip to certain countries or having a confirmed contact with an infected person. IMOH also conducted surveys in selected populations (e.g., healthcare workers), some of whom were asymptomatic and tested positive.

The symptomatic cases were randomly partitioned into training, validation and test sets; we used the standard 50%/25%/25% partition. The training set, containing 10,209 (50%) observations, was used for model calibration, i.e. tuning model parameters (such as logistic regression coefficients). Since most big data algorithms also contain a number

of hyper-parameters that determine the model structure (e.g., number of layers in the Neural Network), these hyper-parameters are tuned using the validation sample. Once all parameters are tuned, the model accuracy is assessed on the test sample (also known as the "holdout sample"). This emulates the expected out-of-sample performance of the model.

Modelling analysis

The dependent variable, Test Result, is binary and, therefore, assumes the value of one when the test is positive and zero when it is negative. The following symptoms were assigned one when they occurred (and zero otherwise): cough, fever, sore throat and shortness of breath. One was also used if the following applied (and zero otherwise): female patients, a recent trip abroad, contact with a known infected person, and patients that were over sixty years old.

The following logistic regression model was run using the training sample data:

$$\begin{aligned} Test\ Result = \alpha + \beta_1 cough + \beta_2 fever + \beta_3 sore\ throat \\ + \beta_4 shortness\ of\ breath + \beta_5 headache + \beta_6 Gender \\ + \beta_7 Abroad + \beta_8 Contact + \beta_9 SixtyPlus \end{aligned}$$

In addition, using JMP® Pro 14.3.0 software, the following supervised machine learning algorithms were conducted on the training sample data to predict test results, using Test Result as the response variable, Y, and the independent variables from (1) as factors (X):

(1) Gradient Boosted Decision Trees (denoted as Boosted Tree)

(2) Random Forests (denoted as Bootstrap Forest)

(3) Neural Networks

We note that the analysis presented below was approved by the IMOH Data Sharing Institutional Review Board.

Results

Table 1 depicts the summary statistics of the total sample. The overall proportion of positive test results was 8%. However, the percent testing was positive was quite different for symptomatic and asymptomatic cases: asymptomatic individuals ("asymptomatics") account for 28% of all positive tests and 86% of all negative tests. This translates to 2.9% positive test rate for asymptomatic and 31.3% positive rate for symptomatic individuals ("symptomatics").

To develop the predictive models, for each model class, we tested two modeling approaches: developing one model equation for the whole dataset, or a separate equation for symptomatic individuals. In all cases, the second approach performed slightly better, which is hardly surprising given that only a subset of predictors is available for asymptomatic individuals. We also developed predictive analytics for the asymptomatics but the results, which were robust in terms of the diagnostics, simply told us that contact with a known carrier is highly predictive of a positive PCR test, hardly a surprising result. All results presented below correspond to the separate symptomatic model.

Table 2 provides the summary statistics for the symptomatics sample, including the breakdown into the training, validation and test datasets. As previously discussed, the symptomatics sample, was randomly partitioned into a training, validation and test sets using the standard 50%/25%/25% partition, as Table 2 shows. The data is made of 9,538 symptomatic females (46%) and 10,999 symptomatic males (54%). 17,421 (85%) of the tests involved people under sixty, and 3,116 (15%) involved people over sixty. 14,110 (69%) of these tests were negative and 6,427 (31%) were positive.

Model accuracy comparisons (for the Test dataset) on a variety of standard measures are given on Table 3. It can be seen that the performance of all four models is very close, with Neural Network achieving a slightly higher accuracy than the other approaches. It is interesting to note that Logistic Regression achieves similar accuracy to the other models, while being significantly more interpretable.

What factors account for the high accuracy of the models? Since logistic regression performs nearly as well as the other method and, unlike other approaches, is quite transparent, we can examine the regression coefficients to get some insights. These are presented on Table 4. The z-statistic can be used as a rough measure of variable importance. For symptomatic cases, Table 4 indicates that the highest risk factors are, in order: Headache, Sore Throat, Shortness of Breath, Contact with a Known Carrier, and Fever. It is also notable that Abroad and Cough both have negative coefficients – indicating that, all else being equal, individuals with these factors are less likely to test positive.

To estimate the sensitivity and specificity of a pre-test screen based on the predictive models we can use another common accuracy visualization, the ROC curves, which are presented on Figure 1. A related numerical measure, the Area Under the Curve (AUC) is also presented (AUC values are also presented in the last column of Table 3). All models seem to perform well for symptomatic cases (AUC of 0.877-0.879).

Before applying the results of our predictive models to design a more efficient pre-testing procedure, we briefly review the testing process, which is illustrated on Figure 2. A randomly drawn patient considered for testing is assumed to be infected with probability q (the disease prevalence rate). A probabilistic pre-test test with sensitivity p_{SE} and specificity p_{SP} is administered to decide whether a patient should be tested or not: an infected patient is tested with probability p_{SE} and (wrongly) rejected with probability (1-p_{SE}). For non-infected cases, the pre-test (correctly) rejects the case for testing with probability p_{SP} and (wrongly) admits for testing with probability (1-p_{SP}). Following the standard notation, we denote, for the test, type I error as α and type II error as β, which implies sensitivity of (1- β) and specificity (probability of detecting no infection given the patient is not infected) of (1- $\propto$). The four possible outcomes of testing are true positive (TP), false negative (FN), true negative (TN) and false positive (FP). The equations for the probabilities of various outcomes are given on Figure 2; if no pre-test is used, the corresponding expressions are obtained by just deleting the terms containing p from the equations on the chart (of course, the proportion of

Untested is 0% in this case). Recall that our goal is to design a pre-test with high efficiency, defined as the proportion of TP per performed test, with the formula given by:

$$\text{Efficiency}=\frac{qp_{SE}(1-\beta)}{1-q(1-p_{SE})-p_{SP}(1-q)}.$$

If no pre-test is used, efficiency is given by $q(1-\beta)$. *Efficiency gain* is defined as the ratio of the efficiency with and without the pre-test.

Each point on the ROC curve on Figure 1 provides a combination of p_{SE} and (1-p_{SP}) values, which can be used directly in the efficiency equation above. We assume sensitivity of 75% (i.e., $\beta = 25\%$) and specificity of 99.9% ($\alpha = .1\%$) for the PCR. The final required parameter is *q*. From Table 2, the average positive rate for the symptomatic cases is 31%. However, this is composed of both TP and FP cases, and thus must be corrected for the sensitivity and specificity of the PCR test. The proper correction is: $q = \frac{0.312-\beta}{1-\alpha-\beta} = 41.64\%$.

We can now compute the Efficiency Gain as a function of pre-test sensitivity, p_{SE} (2a) and as a function of the percentile at which the pre-test will be conducted (2b); these are depicted on Figure 3. For example, as Figure 1(c) shows, the sensitivity of 0.9 on the curve corresponds to a specificity of 0.55. If we decide to apply the PCR test at this level it will result in testing of up to 59.1$^{\text{th}}$ percentile of symptomatics, achieving an expected 41% Efficiency Gain over the no-screen option (with expected detection of 441 true positives per 1000 tests). We see that Efficiency Gain is decreasing with p_{SE} (this is because ROC curve is increasing and higher pre-test sensitivity p_{SE} comes at the cost of lower specificity p_{SP}), with Efficiency Gain of nearly 140% over the no-test option when only 12.9% of symptomatics are admitted for testing, falling to 8% when just under 90% are admitted.

To implement this pre-test policy, we need a real-time evaluation of the presenting individual's model score and percentile. The score (probability of testing positive), can be easily calculated from the logistic regression as follows: $\text{Prob(positive test)} = \frac{1}{1+e^{-x}}$,

where x is the calculated as by multiplying the coefficients from Table 4 by the value of each variable for the individual under evaluation, summing these values up and adding the constant term. The resulting score is converted to the model percentile. For example, a symptomatic individual with the model score of 0.9 must belong to the first decile, i.e., fall below the 10^{th} percentile. These calculations are automated with the simple diagnostic instrument, available as a supplement to this manuscript, that can be used by healthcare workers to assess the risk level of an individual, both in absolute and relative terms.

Discussion

In this study we have analyzed data reported by the IMOH regarding routine SARS-COV-2 PCR tests performed for three weeks during the COVID-19 pandemic in Israel.

An interesting result that is consistent with other studies is the large incidence of asymptomatic individuals who tested positive[14-21]. Specifically, 28% of the people who tested positive in the overall sample were asymptomatic. For symptomatic patients, we have identified that the symptoms that correlate most with a positive test are headache, shortness of breath, sore throat and fever (in this order). We also found that contact with a known carrier is a much stronger predictor than a recent trip abroad. Using both a logistic regression model as well as machine learning algorithms we have developed a predictive analytic tool able to predict PCR test result.

Our finding demonstrating a large incidence of asymptomatic cases who tested positive is of great importance and suggests that measures for screening such as taking the temperature or symptom questionnaires are probably ineffective for preventing the spread of the virus. In addition, it further demonstrates the need for healthcare workers to protect with all patients, as there is a good chance that some of these people are asymptomatic carriers capable of silent transmission.

Regarding specific symptoms correlated with a positive test, our results are somewhat surprising. However, it is worth mentioning that most other studies have looked at symptoms for hospitalized patients rather than all positive tested patients[22-27]. For

example, in contrast with initial studies[24], coughing is much less important as a symptom than, for example, a headache. Moreover, gender and old age are far less important than common wisdom[23]. These differences could be explained through the difference in study design and population (i.e. all positive patients vs. hospitalized). It can also theoretically be explained through variability between base-line characteristics of populations as our results might reflect differences related to ethnicity, demographics, etc. Further research on larger and more diverse data sets will likely shed light on the cause of these difference.

Our predictive modeling can play an important role in pre-test probability for symptomatic patients. Such screens are common where testing kits are in a shortage: to date most jurisdictions set one or more admission criteria that the patient must meet to be tested. The goal of such pre-test screens is to increase test efficiency, i.e., the number of "true positive" cases that are detected pre administered test (i.e., patients who are infected and are correctly detected by the test). It is clear that the higher the infection prevalence rate among the tested group, the higher the testing efficiency. While the present pre-test screening criteria are based on rather crude decision rules (e.g., the presence of at least one symptom), our predictive models can provide a significantly more accurate estimate of infection risk, allowing for much greater testing efficiency, particularly when testing capacity is constrained.

Our predictive models can also be used to overcome some of the shortcomings of the current testing methodology. The standard PCR test for SARS-COV-2 has excellent in vitro sensitivity and specificity, however it has been repeatedly documented to suffer from low in vivo sensitivity. This is due to several factors: the sampling procedure requires deep nasal swab, if done incorrectly, can miss virus-bearing mucous. Another factor is that, as the infection progresses, the amount of virus present in the nasal passage varies, making it harder to detect during certain stages of the infection[28]. This has led to estimates of in vivo sensitivity of between 70% and 80%[29-30]. On the other hand, the specificity of the test in vivo is thought to be high, 99% or above. The low sensitivity leads to many false negatives. In fact, the standard advice to clinicians is

that a patient who is displaying clear clinical signs of COVID-19 should be treated as infected, even if the test result is negative[31]. It is quite common to re-test patients displaying some COVID-19 symptoms when the initial PCR test is negative. However, who should be re-tested? If the tested group has low infection prevalence, the vast majority will (correctly) test negative. Thus re-testing all negatives will drastically lower test efficiency. On the other hand, if we have an accurate pre-test screening tool that indicates high probability that a given patient is infected, re-testing after a negative outcome may be quite efficient; in fact, with a sufficiently high prior probability, multiple rounds of re-testing may be justified. Predictive models that quantify the probability of infection prior to test allow us to develop exact thresholds for who should and who should not be re-tested after each negative outcome.

This study has several limitations. First, while the data set that we use is one of the largest in the analysis of COVID-19, it is, at the same time, quite shallow. For example, its classification into only two age groups, with the sixty being the threshold, is simplistic. As such we were not able to look at other age groups as covariate in our analysis. Another limitation of the data is the lack of outcomes for the people tested. The data is comprised of initial standard PCR tests conducted but we do not know whether this ensued in the person having no symptoms whatsoever afterwards, mild symptoms, or severe symptoms that led to their hospitalization, or worse. Furthermore, we are provided with data on symptoms presented but not their severity. For example, if a person suffered from shortness of breath, we have no idea of the level of the person's oxygen saturation. Another limitation is that we have no data on the pre-conditions for the tested persons. Lastly, the data on symptoms in the tests is incomplete and includes only the four symptoms provided by the IMOH. For example, there are documented cases of other symptoms, such as sudden and complete loss of the olfactory function without any nasal obstruction[32].

To our knowledge, this study is the first to use big data predictive analytics tools for the pre-screening of COVID-19. The results of the predictive analytics models that were used in the study demonstrate their effectiveness as predictors of COVID-19.

Specifically, we developed a simple diagnostic instrument, which provides a fast and costless means for the calculation of the probability of a patient being infected with SARS-COV-2 and which can be used by healthcare workers as a pre-testing mechanism for patients, or health lines in assessing callers. Such predictive tool can enhance efficiency of PCR testing, which can be important when testing kits are limited or to prioritized which samples need to be tested first. Our tool might also help in decision for retesting cases found to be negative when the pre-test probability was high. Still, this needs to be further studied to support the use for re-testing. Further research is also needed to validate or adjust this tool for use in other countries.

Contributors

All authors worked closely to design and write the manuscript. RE led the data management and analysis.

Declaration of interests

We declare no competing interests.

Data sharing

The Israeli Ministry of Health data on the characteristics of tested individuals for SARS-COV-2 is publicly available. The regression model used is provided in Model (1) in the manuscript. The details of the machine learning models are given in Appendix A.

References

1. Ienca, M., Vayena, E. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nat Med* **26**, 463–464 (2020). https://doi.org/10.1038/s41591-020-0832-5
2. Ting, D.S.W., Carin, L., Dzau, V. et al. Digital technology and COVID-19. *Nat Med* **26**, 459–461 (2020). https://doi.org/10.1038/s41591-020-0824-5
3. Park S, Choi GJ, Ko H. Information Technology–Based Tracing Strategy in Response to COVID-19 in South Korea—Privacy Controversies. *JAMA.* Published online April 23, 2020. doi:10.1001/jama.2020.6602
4. Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing. *JAMA*. 2020;323(14):1341–1342. doi:10.1001/jama.2020.3151
5. Ting, D.S.W., Carin, L., Dzau, V. et al. Digital technology and COVID-19. *Nat Med* 26, 459–461 (2020). https://doi.org/10.1038/s41591-020-0824-5
6. Evgeniou, T., Hardoon, D.R., Ovchnnikov, A. Leveraging AI to Battle This Pandemic – And The Next One, *Harvard Business Review*, April 20, 2020, https://hbr.org/2020/04/leveraging-ai-to-battle-this-pandemic-and-the-next-one
7. R. Vaishya, M. Javaid, I.H. Khan, A. Haleem Artificial intelligence (ai) applications for COVID-19 pandemic *Diabetes Metab Syndrome*, 14 (4) (2020), pp. 337-339, 10.1016/j.dsx.2020.04.012
8. Wynants Laure, Van Calster Ben, Bonten Marc M J, Collins Gary S, Debray Thomas P A, De Vos Maarten et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal *BMJ* 2020; 369 :m1328
9. Chow EJ, Schwartz NG, Tobolowsky FA, et al. Symptom Screening at Illness Onset of Health Care Personnel With SARS-CoV-2 Infection in King County, Washington. *JAMA*. Published online April 17, 2020. doi:10.1001/jama.2020.6637
10. Tagarro A, Epalza C, Santos M, et al. Screening and Severity of Coronavirus Disease 2019 (COVID-19) in Children in Madrid, Spain. *JAMA* Pediatr. Published online April 08, 2020. doi:10.1001/jamapediatrics.2020.1346

11. Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., . . . Huang, Y. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Computers, Materials, & Continua, 63(1), 537-551. doi:http://dx.doi.org.myaccess.library.utoronto.ca/10.32604/cmc.2020.010691
12. Israeli Ministry of Health. COVID-19 Database. https://data.gov.il/dataset/covid-19 retrieved on April 15, 2020.
13. Tested individuals' characteristic data README. https://data.gov.il/dataset/covid-19/resource/3f5c975e-7196-454b-8c5b-ef85881f78db. Retrieved April 15, 2020.
14. Arons MM, Hatfield KM, Reddy SC, et al. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *N Engl J Med*. DOI: 10.1056/NEJMoa2008457.
15. Gandhi Monica, Yokoe Deborah S., Havlir Diane V.. (2020) Asymptomatic Transmission, the Achilles' Heel of Current Strategies to Control Covid-19. *N Engl J Med* DOI: 10.1056/NEJMe2009758.
16. He, X., Lau, E.H.Y., Wu, P. et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med (2020). https://doi.org/10.1038/s41591-020-0869-5
17. Bai Y, Yao L, Wei T, et al. Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA*. 2020;323(14):1406–1407. doi:10.1001/jama.2020.2565
18. Ip DKM, Lau LLH, Leung NHL, et al. Viral shedding and transmission potential of asymptomatic and paucisymptomatic influenza virus infections in the community. *Clin Infect Dis* 2017;64:736-742.
19. M. Lipsitch, D.L. Swerdlow, L. Finelli. Defining the epidemiology of covid-19 - studies needed. *N Engl J Med* (2020), 10.1056/NEJMp2002125
20. Fauci, Anthony & Lane, H. & Redfield, Robert. (2020). Covid-19 — Navigating the Uncharted. *N Engl J Med* (2020), 382. 10.1056/NEJMe2002387.
21. Kimball A, Hatfield KM, Arons M, et al. Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility — King County, Washington, March 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:377–381. DOI: http://dx.doi.org/10.15585/mmwr.mm6913e1

22. Kujawski, S.A., Wong, K.K., Collins, J.P. et al. Clinical and virologic characteristics of the first 12 patients with coronavirus disease 2019 (COVID-19) in the United States. *Nat Med* (2020). https://doi.org/10.1038/s41591-020-0877-5
23. Bertocchi, G. COVID-19 susceptibility, women, and work. *VOX CEPR Policy Portal*. https://voxeu.org/article/covid-19-susceptibility-women-and-work
24. Wang D, Hu B, Hu C et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA*. 2020
25. Huang C, Wang Y, Li X. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497.
26. Guan WJ, Ni ZY, Hu Y. et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med*. 2020
27. Chen N, Zhou M, Dong X. et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*. 2020;395(10223):507.
28. Kucirka LM, Lauer SA, Laeyendecker O, Boon D, Lessler J. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction-Based SARS-CoV-2 Tests by Time Since Exposure [published online ahead of print, 2020 May 13]. *Ann Intern Med*. 2020;M20-1495.
29. Gage, J. COVID-19 tests likely only accurate 70 per cent of the time, health experts warn. *National Post*. April 5, 2020. https://nationalpost.com/news/world/covid-19-tests-likely-only-accurate-70-per-cent-of-the-time-health-experts-warn
30. Wikramaratna P, Paton RS, Ghafari M, Lourenco J. Estimating false-negative detection rate of SARS-CoV-2 by RT-PCR. *BMJ* 2020;(epub ahead of print). https://doi.org/10.1101/2020.04.05.200533
31. Krumholz, H.M. If You Have Coronavirus Symptoms, Assume You Have the Illness, Even if You Test Negative. *The New York Times*. April 1, 2020 (updated April 27, 2020). https://www.nytimes.com/2020/04/01/well/live/coronavirus-symptoms-tests-false-negative.html

32. Eliezer M, Hautefort C, Hamel A, et al. Sudden and Complete Olfactory Loss Function as a Possible Symptom of COVID-19. *JAMA Otolaryngol Head Neck Surg.* Published online April 08, 2020. doi:10.1001/jamaoto.2020.0832

Table 1 – Summary Statistics Total Sample

Total Sample No.(%)				
Demographic information				
		Total Sample		
Total No.		108,852 (100)		
Gender				
	Female	53,261 (49)		
	Male	55,591 (51)		
Age				
	Sixty and under	91,304 (84)		
	Over sixty	17,548 (16)		
Symptoms				
	Symptomatic	20,537 (19)		
	Asymptomatic	88,315 (81)		
Test Results			Symptomatic	Asymptomatic
	Positive	8,956 (8)	6,427 (72)	2,529 (28)
	Negative	99,896 (92)	14,110 (14)	85,786 (86)

Table 2 – Summary statistics for the symptomatic cases sample

		Total Sample	Training	Validation	Test
Total No.		20,537 (50)	10,209 (50)	5,141 (25)	5,187 (25)
Gender					
	Female	9,538 (46)	4,760 (50)	2,402 (25)	2,376 (25)
	Male	10,999 (54)	5,449 (50	2,739 (25)	2,811 (25)
Age					
	Sixty and under	17,421 (85)	8,635 (50)	4,361 (25)	4,425 (25)
	Over sixty	3,116 (15)	1,574 (51)	780 (25)	762 (24)
Test Results					
	Positive	6,427 (31)	3,170 (49)	1,670 (25)	1,587 (25)
	Negative	14,110 (69)	7,039 (50)	3,471 (25)	3,600 (25)

Table 3 – Measures of Fit for the Test Set, for Symptomatic Cases

Creator	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Fit Nominal Logistic	0.3887	0.5372	0.3765	0.3342	0.2244	0.1442	5187	0.8775
Neural	0.4060	0.5556	0.3658	0.3315	0.2178	0.1413	5187	0.8794
Bootstrap Forest	0.4022	0.5516	0.3681	0.3318	0.2180	0.1427	5187	0.8768
Boosted Tree	0.3998	0.5490	0.3696	0.3322	0.2200	0.1425	5187	0.8770

AUC –The area under the ROC curve

RMSE - Root Mean Square Error

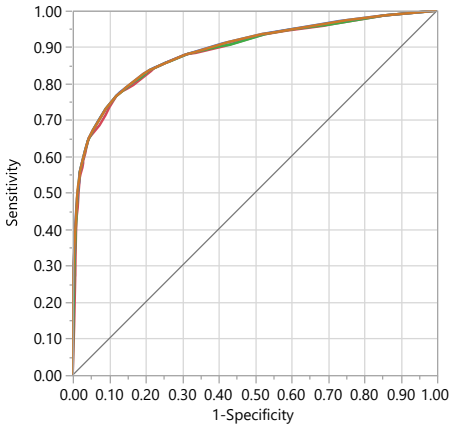
Table 4 – Results of the Logistic Regression

	Symptomatic
	(1)
VARIABLES	TestResults
cough	-0.815***
	(-10.40)
fever	0.887***
	(13.71)
sore throat	3.045***
	(17.91)
shortness of breath	2.911***
	(16.25)
headache	3.685***
	(20.60)
Gender	-0.431***
	(-7.166)
Abroad	-0.714***
	(-10.28)
Contact	1.771***
	(23.06)
SixtyPlus	0.691***
	(9.225)
Constant	-1.232***
	(-12.34)
Observations	10,209
Log pseudolikelihood	-3741
McFadden Pseudo R-squared	0.409

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

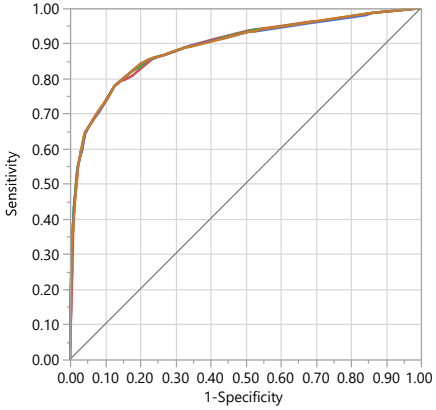
Figure 1 - ROC Curve for the Training, Validation and Test Datasets for All Predictive Analytics Tools

(a) Training Set



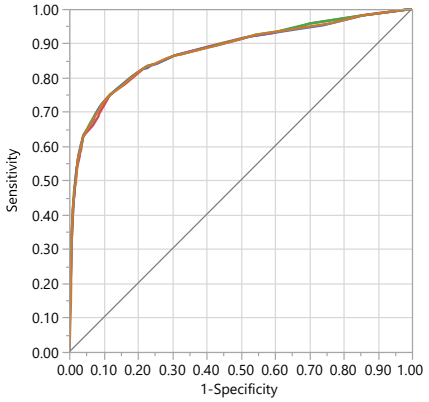
Predictor	AUC
Fit Nominal Logistic	0.8877
Neural	0.8904
Boostrap Forest	0.8932
Boosted Tree	0.8929

(b) Validation Set



Predictor	AUC
Fit Nominal Logistic	0.8907
Neural	0.8916
Boostrap Forest	0.8902
Boosted Tree	0.8910

(c) Test Set



Predictor	AUC
Fit Nominal Logistic	0.8775
Neural	0.8794
Boostrap Forest	0.8768
Boosted Tree	0.8770

Figure 2 - Pre-screening process model

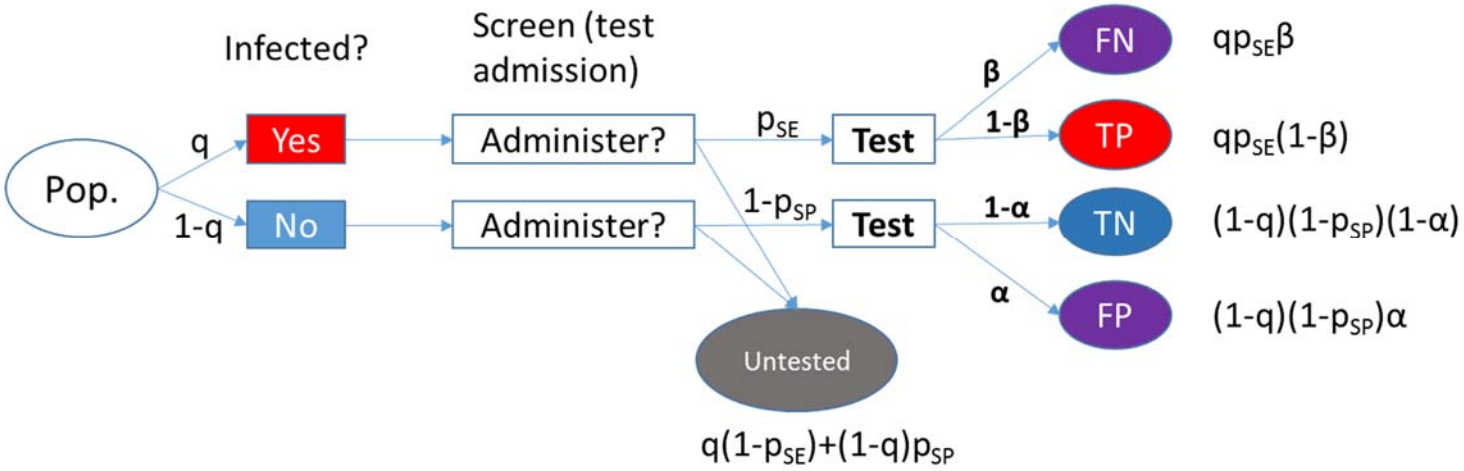
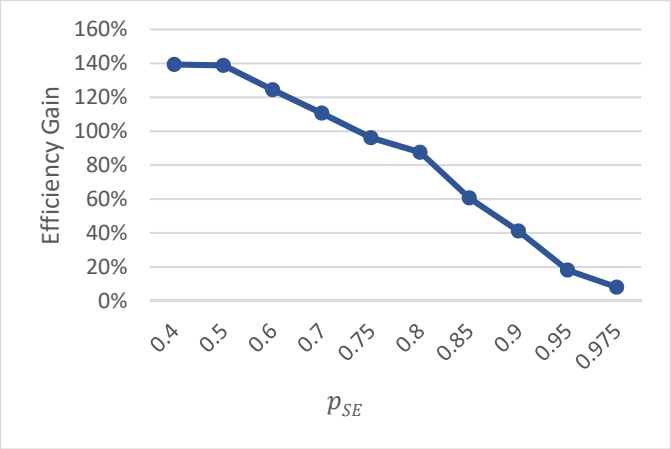
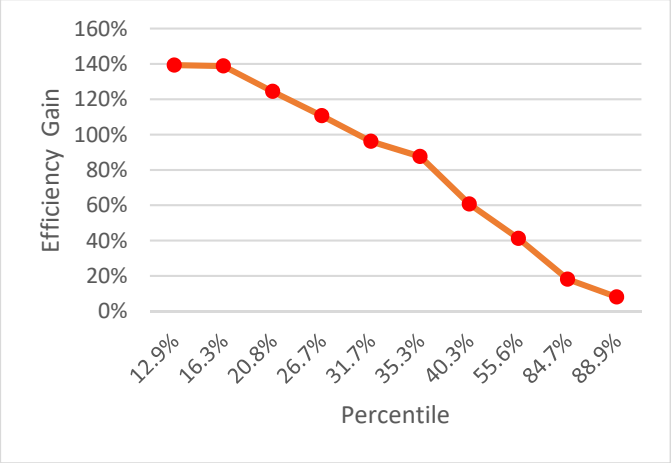


Figure 3 – Efficiency Gain from the Pre-Test Screening

a) As a function of p_{SE}



b) As a function of percentile of symptomatics tested



Click here to access/download

Necessary Additional Data

Pre-screening Diagnostic Template.xlsx