MEDJ, Volume *1*

Supplemental Information

A Prediction Model to Prioritize Individuals for a SARS-CoV-2 Test Built from National Symptom Surveys

Saar Shoer, Tal Karady, Ayya Keshet, Smadar Shilo, Hagai Rossman, Amir Gavrieli, Tomer Meir, Amit Lavon, Dmitry Kolobkov, Iris Kalka, Anastasia Godneva, Ori Cohen, Adam Kariv, Ori Hoch, Mushon Zer-Aviv, Noam Castel, Carole Sudre, Anat Ekka Zohar, Angela Irony, Tim Spector, Benjamin Geiger, Dorit Hizi, Varda Shalev, Ran Balicer, and Eran Segal

Supplementary Method

COVID-19 survey

a. Online version questions

Age: _____

Gender:
- ❑ Male
- ❑ Female

City, Street:

I am:
- ❑ Feeling well
- ❑ Not feeling well

Are you experiencing any of the following symptoms?

Cough
- ❑ Dry cough (no sputum)
- ❑ Wet cough (with sputum)

- ❑ Fatigue
- ❑ Muscle pain
- ❑ Shortness of breath
- ❑ Rhinorrhea (runny nose) and/or nasal congestion
- ❑ Diarrhea
- ❑ Nausea and/or vomiting
- ❑ Sore throat
- ❑ Headache
- ❑ Chills
- ❑ Confusion
- ❑ Loss of taste or smell

Have you been diagnosed with any of the following conditions:
- ❑ Diabetes mellitus
- ❑ Hypertension
- ❑ Cardiovascular disease or stroke
- ❑ Chronic lung disease including Asthma (with the exception of childhood Asthma)
- ❑ Chronic kidney disease
- ❑ Malignancy (cancer)
- ❑ Immunodeficiency (including consumption of drugs which cause immunodeficiency)

I am currently:
- ❑ Not in isolation
- ❑ In isolation (including from family members, staying in a separate room) from the date of _____ due to:
  - ❑ A recent international travel

- ❑ A contact with an individual who was infected with coronavirus
- ❑ Experiencing disease symptoms
- ❑ Voluntary isolation

❑ I have a confirmed infection with COVID-19 (by a lab test) and currently:

- ❑ In home isolation
- ❑ Staying in a hotel
- ❑ Hospitalized in a hospital
- ❑ I recovered from COVID-19 infection

Cigarette smoking habits:

- ❑ I currently smoke
- ❑ I used to smoke and stopped more than 5 years ago
- ❑ I used to smoke and stopped less than 5 years ago
- ❑ I have never smoked

What is your current body temperature?

- ❑ I did not measure my temperature in the last 24 hours
- ❑ I measured my temperature and it was _____

How many individuals have you been in contact with in the last 24 hours?
(within approximately 2 meters (6 ft 7 in) for more than 15 minutes)
Adults (age above18 years old _____)
Children (age below 18 years old _____)

b. IVR version questions

Age group:
- ❑ 0-20
- ❑ 21-30
- ❑ 31-40
- ❑ 41-50
- ❑ 51-60
- ❑ 61-70
- ❑ > 70

Gender:
- ❑ Male
- ❑ Female

Were you confirmed with COVID-19 (by a lab test)?
- ❑ Yes
- ❑ No

Are you currently in isolation?
- ❑ Yes
- ❑ No

I am:
- ❑ Feeling well
- ❑ Not feeling well

Are you experiencing any of the following symptoms?
- ❑ Sore throat
- ❑ Cough
- ❑ Shortness of breath
- ❑ Loss of taste or smell

Did you measure your body temperature today?
- ❑ No
- ❑ Yes:
  - ❑ I had a fever (body temperature >38℃)
  - ❑ I did not have a fever (body temperature <38℃)

Have you been diagnosed with any of the following conditions: Diabetes mellitus, Hypertension, Cardiovascular disease, Chronic lung disease, Chronic kidney disease, Malignancy (cancer) or Immunodeficiency?
- ❑ Yes
- ❑ No

Are you currently staying at your own home?
- ❑ Yes
- ❑ No

Supplementary Figures

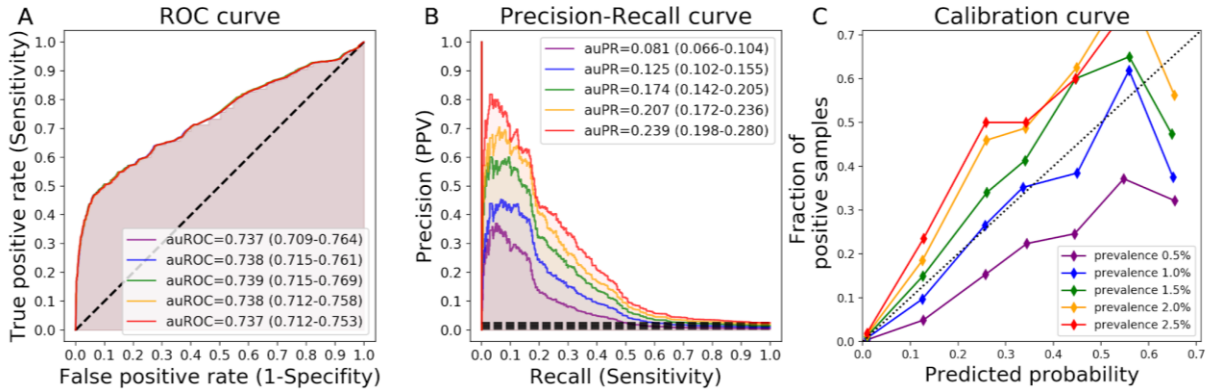

Figure S1. relates to Figure 2 and Table S5. Primary model performance in different COVID-19 diagnosed prevalences. Purple 0.5% COVID-19 diagnosed prevalence, Blue 1.0%, Green 1.5%, Yellow 2.0%, Red 2.5%. Abbreviations: auROC/auPR - area under the ROC/PR curve, ROC - Receiver-Operator-Characteristic, PR - Precision-Recall. Parenthesis - Confidence Interval. **A:** ROC curve of our model consisting of 9 simple questions. **B:** Precision-Recall curve of our model. **C:** Calibration curve. Colored dots represent deciles of predicted probabilities. Dotted diagonal line represents an ideal calibration.

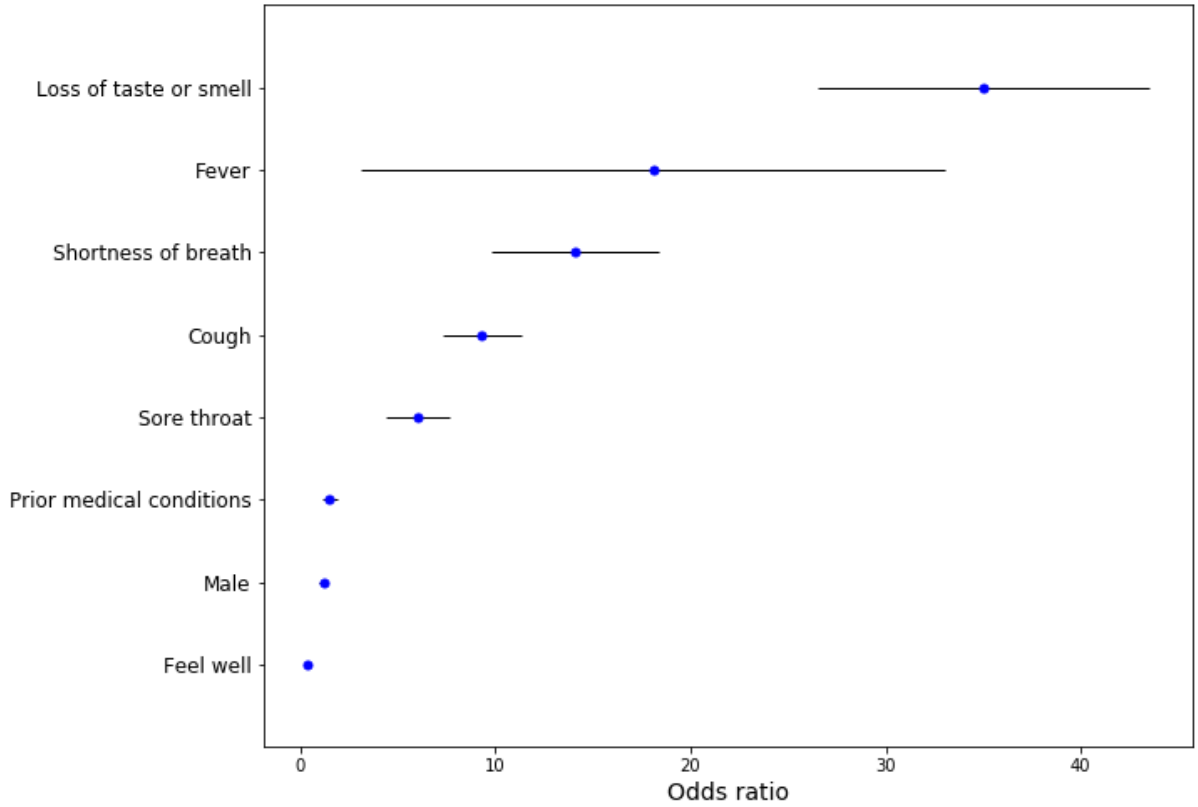

Figure S2. relates to Figure 5. Odds ratio (unadjusted) of the primary model population. Error bars represent CI.

Supplementary Tables

Table S1, relates to Figure 1. COVID-19 diagnosis prevalence and response rate in the IVR cities

City	Total population*	Target population* (age > 20 years)	Number of COVID-19 diagnosed individuals[#]	COVID-19 diagnosis prevalence	Number of responders	Response rate
HAZOR HAGELILIT	9K	6K	3	0.03%	1428	15.58%
NETIVOT	36K	20K	89	0.25%	4462	12.52%
ZEFAT	36K	20K	66	0.18%	3935	11.02%
EL'AD	48K	18K	371	0.78%	3924	8.20%
IMMANU'EL	4K	2K	not published	not published	239	6.47%
LOD	76K	48K	86	0.11%	4789	6.32%
REKHASIM	12K	6K	37	0.31%	754	6.23%
BENE BERAQ	199K	96K	2747	1.38%	12335	6.20%
BET SHEMESH	119K	54K	436	0.37%	7152	6.03%

* Number of residents in each city and age group are according to the Israeli Central Bureau of Statistics for 2018.

[#] Number of COVID-19 diagnosed individuals is according to the Israeli Ministry of Health on April 26, 2020.

Table S2, relates to Figure 1. Baseline characteristics of the extended features model population

Characteristic, mean (SD) or %	All individuals n=131,166 (100%)	COVID-19 undiagnosed n=131,052 (99.913%)	COVID-19 diagnosed n=114 (0.087%)
Age in years	48.479 (15.836)	48.485 (15.836)	41.842 (15.011)
Gender - male	58,545 (44.634%)	58,484 (44.627%)	61 (53.509%)
COVID-19 diagnosed	114 (0.087%)	0 (0.0%)	114 (100.0%)
Diabetes	9,894 (7.543%)	9,886 (7.544%)	8 (7.018%)
Hypertension	19,959 (15.217%)	19,951 (15.224%)	8 (7.018%)
Ischemic heart disease	6,844 (5.218%)	6,843 (5.222%)	1 (0.877%)
Lung disease	6,712 (5.117%)	6,710 (5.12%)	2 (1.754%)
Kidney disease	801 (0.611%)	800 (0.61%)	1 (0.877%)
Cancer	6,050 (4.612%)	6,046 (4.613%)	4 (3.509%)
Immune system suppression	2,658 (2.026%)	2,658 (2.028%)	0 (0.0%)
Feeling well	124,961 (95.269%)	124,864 (95.278%)	97 (85.088%)
Shortness of breath	1,667 (1.271%)	1,659 (1.266%)	8 (7.018%)
Runny nose	6,932 (5.285%)	6,924 (5.283%)	8 (7.018%)
Fatigue	5,839 (4.452%)	5,807 (4.431%)	32 (28.07%)
Nausea or vomiting	889 (0.678%)	885 (0.675%)	4 (3.509%)
Muscle pain	1,201 (0.916%)	1,194 (0.911%)	7 (6.14%)
Sore throat	5,092 (3.882%)	5,079 (3.876%)	13 (11.404%)
Cough dry	4,581 (3.493%)	4,557 (3.477%)	24 (21.053%)
Cough moist	4,821 (3.675%)	4,814 (3.673%)	7 (6.14%)
Diarrhea	1,698 (1.295%)	1,692 (1.291%)	6 (5.263%)
Fever (body temperature above 38 °C)	243 (0.185%)	237 (0.181%)	6 (5.263%)
Chills	954 (0.727%)	948 (0.723%)	6 (5.263%)
Confusion	399 (0.304%)	397 (0.303%)	2 (1.754%)
Loss of taste or smell	520 (0.396%)	488 (0.372%)	32 (28.07%)

Table S3, relates to Figure 2. Prediction models area under the curve of the Main dataset

	Primary model		**Extended features model**	
	auROC	**auPR**	**auROC**	**auPR**
Logistic Regression				
Baseline model	0.556 (CI: 0.527-0.581)	0.013 (CI: 0.012-0.015)	0.613 (CI: 0.560-0.654)	0.001 (CI: 0.001-0.002)
Full model	0.737 (CI: 0.712-0.759)	0.144 (CI: 0.119-0.177)	0.767 (CI: 0.724-0.811)	0.030 (CI: 0.018-0.047)
Gradient Boosting Decision Trees				
Baseline model	0.553 (CI: 0.530-0.585)	0.015 (CI: 0.013-0.017)	0.606 (CI: 0.545-0.645)	0.001 (CI: 0.001-0.002)
Full model	0.700 (CI: 0.668-0.728)	0.150 (CI: 0.125-0.187)	0.709 (CI: 0.630-0.774)	0.047 (CI: 0.025-0.088)

Abbreviations: auROC/auPR - area under the ROC/PR curve, ROC - Receiver-Operator-Characteristic, PR - Precision-Recall.

Table S4, relates to Figures 2, 4. Prediction models precision, recall and NPV in different decision probability thresholds

	Logistic Regression Israeli dataset			**Gradient Boosting Decision Trees Israeli dataset**			**Logistic Regression U.S., U.K. and Sweden dataset**		
Decision probability thresholds	Precision	Recall	NPV	Precision	Recall	NPV	Precision	Recall	NPV
0	0.011391	1.000000	1.000000	0.011382	1.000000	1.000000	0.065332	1.000000	1.000000
0.1	0.225989	0.240964	0.991253	0.195620	0.269076	0.991548	0.294324	0.234899	0.947265
0.2	0.398104	0.168675	0.990514	0.369748	0.176707	0.990578	0.363760	0.149329	0.942893
0.3	0.447368	0.136546	0.990159	0.445161	0.138554	0.990160	0.357309	0.086130	0.939340
0.4	0.485149	0.098394	0.989713	0.564103	0.088353	0.989605	0.358491	0.053132	0.937535
0.5	0.507692	0.066265	0.989401	0.630435	0.058233	0.989269	0.504274	0.032998	0.936553
0.6	0.500000	0.042169	0.989155	0.625000	0.020080	0.988842	0.576271	0.019016	0.935772
0.7	0.500000	0.022088	0.988886	0.250000	0.002008	0.988639	0.789474	0.008389	0.935171
0.8	0.285714	0.004016	0.988706	0.000000	0.000000	0.988617	1.000000	0.002237	0.934805

Table S5, relates to Figure S1. Primary model precision, recall and NPV in different decision probability thresholds and in different COVID-19 diagnosed prevalences

COVID-19 prevalence	**0.5%**			**1.0%**			**1.5%**			**2.0%**			**2.5%**		
Decision probability thresholds	Precision	Recall	NPV	Precision	Recall	NPV	Precision	Recall	NPV	Precision	Recall	NPV	Precision	Recall	NPV
0	0.0050	1.0000	1.0000	0.0100	1.0000	1.0000	0.0150	1.0000	1.0000	0.0200	1.0000	1.0000	0.0250	1.0000	1.0000
0.1	0.1099	0.2410	0.9962	0.1990	0.2410	0.9923	0.2857	0.2410	0.9885	0.3499	0.2410	0.9846	0.4096	0.2410	0.9808
0.2	0.2366	0.1687	0.9958	0.3636	0.1687	0.9917	0.4670	0.1707	0.9875	0.5638	0.1687	0.9833	0.5944	0.1707	0.9791
0.3	0.2776	0.1365	0.9957	0.4048	0.1365	0.9914	0.5231	0.1365	0.9870	0.6018	0.1365	0.9827	0.6355	0.1365	0.9784
0.4	0.3063	0.0984	0.9955	0.4298	0.0984	0.9910	0.5833	0.0984	0.9865	0.6622	0.0984	0.9819	0.7101	0.0984	0.9774
0.5	0.3367	0.0663	0.9953	0.4533	0.0683	0.9907	0.5763	0.0683	0.9861	0.6800	0.0683	0.9814	0.7727	0.0683	0.9768
0.6	0.3281	0.0422	0.9952	0.3889	0.0422	0.9905	0.5385	0.0422	0.9857	0.6176	0.0422	0.9809	0.7778	0.0422	0.9762
0.7	0.2973	0.0221	0.9951	0.3667	0.0221	0.9903	0.5500	0.0221	0.9854	0.5500	0.0221	0.9805	0.7333	0.0221	0.9757
0.8	0.1250	0.0040	0.9950	0.1429	0.0040	0.9901	0.3333	0.0040	0.9851	0.4000	0.0040	0.9802	0.4000	0.0040	0.9753

Table S6, relates to Figure 5. Primary model feature importance

	Mean absolute Shapley value	Odds-ratio
Loss of taste or smell	0.051686	35.011023
Cough	0.050932	9.311583
Age	0.01556	
Shortness of breath	0.014809	14.07521
Sore throat	0.011866	6.054205
Male	0.008462	1.199557
Feel well	0.001609	0.406353
Prior medical conditions	0.001331	1.53291
Fever	0.000326	18.088595