npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms

Matteo Gadaleta[1], Jennifer M. Radin[1], Katie Baca-Motes[1], Edward Ramos[1,2], Vik Kheterpal[2], Eric J. Topol[1], Steven R. Steinhubl[1] and Giorgio Quer[1]✉

Individual smartwatch or fitness band sensor data in the setting of COVID-19 has shown promise to identify symptomatic and pre-symptomatic infection or the need for hospitalization, correlations between peripheral temperature and self-reported fever, and an association between changes in heart-rate-variability and infection. In our study, a total of 38,911 individuals (61% female, 15% over 65) have been enrolled between March 25, 2020 and April 3, 2021, with 1118 reported testing positive and 7032 negative for COVID-19 by nasopharyngeal PCR swab test. We propose an explainable gradient boosting prediction model based on decision trees for the detection of COVID-19 infection that can adapt to the absence of self-reported symptoms and to the available sensor data, and that can explain the importance of each feature and the post-test-behavior for the individuals. We tested it in a cohort of symptomatic individuals who exhibited an AUC of 0.83 [0.81–0.85], or AUC = 0.78 [0.75–0.80] when considering only data before the test date, outperforming state-of-the-art algorithm in these conditions. The analysis of all individuals (including asymptomatic and pre-symptomatic) when self-reported symptoms were excluded provided an AUC of 0.78 [0.76–0.79], or AUC of 0.70 [0.69–0.72] when considering only data before the test date. Extending the use of predictive algorithms for detection of COVID-19 infection based only on passively monitored data from any device, we showed that it is possible to scale up this platform and apply the algorithm in other settings where self-reported symptoms can not be collected.

npj Digital Medicine (2021)4:166 ; https://doi.org/10.1038/s41746-021-00533-1

INTRODUCTION

Frequent monitoring to quickly identify, trace, and isolate cases of SARS-CoV-2 is needed to help control the spread of the infection as well as improve individual patient care through the earlier initiation of effective therapies[1]. Frequent diagnostic testing is one important option but suffers from implementation challenges and a lack of accessibility for individuals affected most by COVID-19[2]. Self-reporting of symptoms has been found to be predictive of a positive test[3], and could be used to encourage individuals to get tested earlier. However, such an approach not only requires active engagement of the individual, but also misses the approximately one-third of asymptomatic infected individuals completely, and delays diagnosis in those who are infected but pre-symptomatic[4]. On the other hand, passive monitoring is possible with commercial sensor devices measuring biometrics such as resting heart rate[5], sleep[6] or activity, which have been shown to be effective in the detection of COVID-19 versus non-COVID-19 when incorporated in combination with self-reported symptoms[7].

Individual sensor data in the setting of COVID-19 has also shown promise in identifying pre-symptomatic infection[8], the need for hospitalization[9], correlations between peripheral temperature and self-reported fever[10], differences in the changes in wearable data between individuals with COVID-19 versus influenza-like-illnesses[11], and an association between changes in heart-rate-variability and infection[12]. These studies focused on a specific device brand, or on a predefined set of signals. However, for a broader use of personal health technologies it is important to design algorithms that are device agnostic and can adapt to the specific data collected by any sensor, including the less costly devices.

Our prospective app-based research platform DETECT (Digital Engagement and Tracking for Early Control and Treatment) allows participants to enter self-reported symptoms or COVID-19 test results, and to share data from any wearable device that is connected to Google Fit or Apple Health Kit platform. In a previous study, we developed a deterministic algorithm to discriminate between symptomatic individuals testing positive or negative for COVID-19, analyzing changes in daily values of resting heart rate, length of sleep and amount of activity, together with self-reported symptoms[7].

In order to provide the most accurate early warnings for COVID-19 to all participants for a wide variety of wearable devices, we proposed and validated a machine learning algorithm that ingests all available sensor data for the detection of COVID-19 infection. The algorithm can outperform our previously proposed algorithm in similar conditions (AUC = 0.83, IQR = [0.81, 0.85]), and more importantly, it can automatically adapt to the specific sensor used, exploiting all the information collected from the more advanced sensors or focusing on a smaller set of signals from more basic sensors, and explaining the feature importance and the post-test behavioral changes for the individual. The algorithm uses self-reported symptoms when they are available, or otherwise makes its inference based on sensor data only, thus adapting to different engagement levels of the individuals in the study.

RESULTS

Machine learning model for COVID-19 detection

In this study, we investigated the accuracy of a machine learning model in the detection of COVID-19 infection based on the available data acquired from wearable devices and self-reported

[1]Scripps Research Translational Institute, 3344N Torrey Pines Ct Plaza Level, La Jolla, CA 92037, USA. [2]CareEvolution, 625N Main Street, Ann Arbor, MI 48104, USA. ✉email: gquer@scripps.edu

surveys. We analyzed the accuracy of the detection algorithm for individuals who self-reported at least one symptom prior to the COVID-19 test (named the "symptomatic cohort" in what follows) and not-reporting any symptom prior to the COVID-19 test (the "no-symptom-reported cohort"). We also separately investigated the accuracy obtained using only data collected before a COVID-19 test versus including pre- and post-test data, in order to explore the effect of behavioral changes just the act of testing for COVID-19 might have on individuals.

Participant characteristics

A total of 38,911 individuals (61% female, 15% over 65) have been enrolled between March 25, 2020 and April 3, 2021. Among these participants, 1118 (66% female, 8% over 65) reported at least one positive and 7032 (63% female, 14% over 65) at least one-negative COVID-19 nasal swab test. The total number of COVID-19 swab tests reported during the same period was 18,175, with 1360 (7.5%) positives, 16,398 negatives and 417 with non-reported results. Among the positive tests, 539 (48% of the considered cases) reported at least one symptom in the 15 days preceding the test date, 592 (52%) did not report any symptom, and 229 have been excluded from the analysis for lack of sufficient data or for being too close to a prior test.

Data set description

The participants of the study shared their personal device data (including historical data collected prior to enrollment), self-reported symptoms and diagnostic test results during the data collection period. We divided the measures into four categories: symptom features, including all self-reported symptoms; sensor features, including all measures related to activity, heart rate or sleep; anthropometrics; and demographics (Table 1).

Detection of COVID-19

The normalized deviations from the baseline for a subset of representative features are reported (Table 2), highlighting the difference between positive and negative COVID-19 individuals, both excluding and including data in the 5 days after the test date, based on gender and age. As expected, we observed larger variation from the baseline, in terms of heart rate, sleep and activity related features, for individuals who tested positive for COVID-19 with respect to individuals who tested negative. This observation held for all the demographic groups, both excluding or including post-test data (Fig. 1). Based on these features, a prediction model was trained and tested in different conditions. The model provides an output between 0 and 1, indicating the risk of infection. In general, a higher output value denotes a higher risk of infection. For the symptomatic cohort, we observed a significant difference in the model's output between participants who tested positive or negative, showing that the two groups can be effectively separated (Fig. 1b), even if we consider only the days preceding the test date (Fig. 1a), thus excluding any behavioral bias potentially caused by taking the test and awaiting results or knowledge of the test outcome. We showed also the predictions for the no-symptom-reported cohort, considering the data before the test date or all the available data, respectively (Fig. 1c, d). As expected, while a significant difference between the individuals testing positive or negative could still be observed, it is harder to clearly separate the two groups.

Symptomatic cases exhibited an area under the receiver operating characteristic (ROC) curve (AUC) of 0.83 [0.81–0.85], while when considering only data before the test date the performance slightly decreased, with AUC = 0.78 [0.75–0.80]. For the no-symptom-reported cohort, we observed an AUC of 0.74 [0.72–0.76], or AUC = 0.66 [0.64–0.68] when considering only data before the test date (Fig. 2).

Importance of each feature

For the symptomatic cohort, self-reported symptoms were of crucial importance for the most accurate diagnosis of the disease. Considering only data before the test, self-reported symptoms accounted for 60% of the relative contribution to the predictive model (Fig. 3a), while considering all peri-test data, the importance of the self-reported symptoms decreased to a relative contribution of 46% (Fig. 3b).

For both the symptomatic and no-symptom-reported cohorts, we observed a consistent change in the importance of the activity sensor features, if we consider only data before the test. For the no-symptom-reported cohort (Fig. 3c, d), the importance of the activity sensor features increased from 46 to 54% when all peri-test data were considered—potentially as a consequence of precautionary measures imposed after testing and awaiting results or receiving a positive test outcome. Sleep sensor features importance did not change significantly when post-test data were included for either the cohort reporting symptoms, or those not reporting symptoms, potentially because sleep was less affected by the knowledge of a test result. Sensor features in the heart rate category had a small relative contribution (6%) for the symptomatic cohort (Fig. 3a), while their contribution increased (18%) in the no-symptom-reported cohort (Fig. 3c), acquiring more importance in the absence of information about symptoms and when only pre-test data was considered. Anthropometrics, such as height or weight, provided only a small relative contribution, while the contribution of demographic features, such as age or gender, was negligible.

Finally, we provided more details about specific symptoms, and how each of them, on average, affects the model's prediction (Fig. 4). We identified highly discriminative symptoms (cough and decrease in taste and smell, with ≥10% relative contribution), medium discriminative symptoms (congestion or runny nose, fever, chills or sweating and congestion or runny nose with <10% and ≥5% relative contribution) and low discriminative features (e.g., body aches, headache, fatigue, with < 5% relative contribution).

DISCUSSION

Our machine learning model based on decision trees can discriminate between individuals who tested positive or negative for COVID-19 based on multiple data types collected by wearable devices, demographic information and self-reported symptoms when available. The adaptability of the algorithm to the available data allows us to also study the performance of the algorithm for individuals in the absence of self-reported symptoms, who may account for almost half of COVID-19 positive individuals[4]. In order to estimate the effects of the behavioral changes due to the act of testing and/or receiving a positive COVID-19 test, we performed a temporal analysis dividing the data collected before and after the date of COVID-19 testing. The model has been shown to perform well for the identification of COVID-19 infection when incorporating data from symptomatic individuals that includes the five days following the date of testing, with an AUC of 0.83 (IQR: 0.81–0.85). By considering only data preceding the test date, we achieved an AUC of 0.78 (IQR: 0.75–0.80) for people who reported symptoms. When available, self-reported symptoms remain the predominant feature category considered by the model in all our test scenarios, demonstrating the importance of an engaging system that allows participants to easily report this information at any time. Among participants with symptoms, we identified cough and decrease in taste and smell as the most highly discriminative symptoms for a COVID-19 infection, followed in order of importance by fever, chills or sweating, and congestion or runny nose.

Using the same model, we also investigated individuals who did not report any symptoms. Despite the lack of self-reported

Table 1. Description and categorization of the feature set.

Feature category	Feature description	Total COVID-19 individual tests	Number of individuals	Fitbit users [%]	Available days median [IQR]
Symptoms features	Fatigue	1149	1091	—	—
	Headache	1104	1061	—	—
	Difficulty breathing	208	206	—	—
	Diarrhea or vomiting	378	368	—	—
	Decrease in taste smell	247	247	—	—
	Cough	892	859	—	—
	Fever chills or sweating	660	645	—	—
	Congestion Or runny nose	1152	1097	—	—
	Neck pain	409	396	—	—
	Body aches	823	795	—	—
	Sore throat	973	922	—	—
	Stomach ache	312	302	—	—
Sensor features (activity)	Total number of daily steps	9348	6983	82%	673 [507–707]
	Total daily distance traveled on foot	9281	6938	82%	676 [554–708]
	Calories burned from periods above sedentary level	7629	5679	100%	676 [554–708]
	Calories burned inclusive of BMR	8279	6142	100%	676 [554–708]
	Minutes spent fairly active	7291	5450	100%	676 [554–708]
	Minutes spent sedentary	8196	6082	100%	676 [554–708]
	Minutes spent very active	7171	5370	100%	676 [554–708]
	Minutes spent lightly active	7628	5678	100%	676 [554–708]
Sensor features (Heart)	Daily resting heart rate	9105	6810	81%	492 [341–653]
	Maximum daily heart rate variability	1812	1407	0%	363 [185–499]
	Minimum daily heart rate variability	1812	1407	0%	363 [185–499]
	Average daily heart rate variability	1812	1407	0%	363 [185–499]
Sensor features (Sleep)	Total daily sleep time	7473	5603	94%	445 [240–629]
	Total daily time spent in bed	7473	5603	94%	445 [240–629]
	Sleep efficiency of the main sleep	7473	5603	94%	445 [240–629]
	Sleep time of the main sleep	7473	5603	94%	445 [240–629]
Anthropometrics	Body mass index	8478	6303	—	—
	Self-reported height	1673	1277	—	—
	Body weight	9240	6896	—	—
	Body fat percentage	4594	3431	—	—
	Basal metabolic rate (BMR) only calories	8279	6142	—	—
Demographic features	Self-reported gender	10,494	7853	—	—
	Age at the time of test	10,479	7841	—	—

The set of features for each category is reported in the table. The same individual may report multiple tests. The total number of individual COVID-19 tests is the number of tests with the corresponding feature value available for the analysis, after exclusion criteria. The number of corresponding unique individuals is also reported. The available days for Sensor Features represent the median number of days available for all the participants (IQR reported in brackets).

information about the sickness, the model achieved an AUC of 0.74 (IQR: 0.72–0.76) when considering the period following the test, and an AUC of 0.66 (IQR: 0.64–0.68) excluding post-test data. Looking at the importance of the features used by the algorithm, we noticed that the importance of sensor-based Activity substantially increases when considering also post-test data, likely reflecting a potential behavioral change for the participants due to imposed precautionary measures[13]. On the other hand, the importance of the heart rate features, which are less likely to be affected by short term behavioral changes, is higher when the model consider only data before the COVID-19 test. Moreover, since heart rate elevation might serve as an indicator of inflammatory conditions[14], its relative importance increases significantly in the absence of self-reported symptoms.

These results build on our prior retrospective work on resting heart rate[5] and sleep[6], which when aggregated at the population level, have been shown to significantly improve real-time predictions for influenza-like illness[15]. In an early study, using the initial data from DETECT, we demonstrated the potential of using self-reported symptoms and wearable data for the discrimination of positive and negative cases of COVID-19[7], which has been validated by several subsequent independent studies

Table 2. Normalized deviation from baseline values for a selected number of representative features.

Feature	Subset	*Z*-score: Median [IQR] (*N* = number of valid samples)							
		Excluding data after COVID-19 Test				Including data after COVID-19 Test			
		COVID-19 negative		COVID-19 positive		COVID-19 negative		COVID-19 positive	
Steps	Overall	−0.08 [−0.45, 0.34]	(*N* = 8290)	−0.44 [−0.84, 0.01]	(*N* = 977)	−0.10 [−0.44, 0.27]	(*N* = 8360)	−0.68 [−1.10, −0.24]	(*N* = 988)
	Male	−0.08 [−0.44, 0.34]	(*N* = 2977)	−0.42 [−0.84, −0.01]	(*N* = 324)	−0.10 [−0.43, 0.26]	(*N* = 3004)	−0.65 [−1.02, −0.28]	(*N* = 326)
	Female	−0.07 [−0.46, 0.34]	(*N* = 5313)	−0.46 [−0.85, 0.03]	(*N* = 653)	−0.10 [−0.44, 0.27]	(*N* = 5356)	−0.70 [−1.13, −0.22]	(*N* = 662)
	Young (<40)	−0.05 [−0.44, 0.35]	(*N* = 2759)	−0.39 [−0.76, 0.17]	(*N* = 348)	−0.06 [−0.40, 0.31]	(*N* = 2793)	−0.65 [−1.00, −0.10]	(*N* = 354)
	Middle age (40–65)	−0.08 [−0.45, 0.33]	(*N* = 4543)	−0.47 [−0.88, −0.04]	(*N* = 557)	−0.11 [−0.45, 0.24]	(*N* = 4575)	−0.72 [−1.15, −0.31]	(*N* = 561)
	Old Age (>65)	−0.10 [−0.47, 0.35]	(*N* = 975)	−0.41 [−0.80, 0.04]	(*N* = 71)	−0.14 [−0.50, 0.26]	(*N* = 979)	−0.51 [−1.14, −0.01]	(*N* = 72)
Resting heart rate	Overall	0.04 [−0.38, 0.49]	(*N* = 8061)	0.23 [−0.26, 0.78]	(*N* = 945)	0.04 [−0.33, 0.44]	(*N* = 8150)	0.16 [−0.28, 0.70]	(*N* = 955)
	Male	0.05 [−0.33, 0.48]	(*N* = 2922)	0.27 [−0.21, 0.77]	(*N* = 318)	0.05 [−0.29, 0.44]	(*N* = 2955)	0.20 [−0.20, 0.75]	(*N* = 321)
	Female	0.04 [−0.42, 0.50]	(N = 5139)	0.22 [−0.27, 0.77]	(*N* = 627)	0.04 [−0.36, 0.44]	(*N* = 5195)	0.12 [−0.32, 0.65]	(*N* = 634)
	Young (<40)	0.06 [−0.44, 0.56]	(*N* = 2685)	0.26 [−0.29, 0.79]	(*N* = 333)	0.06 [−0.38, 0.49]	(*N* = 2725)	0.11 [−0.36, 0.68]	(*N* = 337)
	Middle age (40–65)	0.04 [−0.36, 0.48]	(*N* = 4403)	0.23 [−0.26, 0.76]	(*N* = 542)	0.04 [−0.31, 0.42]	(*N* = 4446)	0.17 [−0.25, 0.69]	(*N* = 548)
	Old age (>65)	0.01 [−0.33, 0.36]	(*N* = 960)	0.17 [−0.15, 0.73]	(*N* = 69)	0.02 [−0.28, 0.36]	(*N* = 966)	0.10 [−0.09, 0.79]	(*N* = 69)
Average daily heart rate variability	Overall	−0.02 [−0.41, 0.39]	(*N* = 1591)	−0.25 [−0.65, 0.19]	(*N* = 196)	0.02 [−0.33, 0.32]	(*N* = 1614)	−0.11 [−0.46, 0.29]	(*N* = 198)
	Male	−0.02 [−0.38, 0.38]	(*N* = 743)	−0.27 [−0.68, 0.14]	(N = 84)	0.02 [−0.30, 0.31]	(*N* = 757)	−0.13 [−0.61, 0.18]	(*N* = 84)
	Female	−0.00 [−0.43, 0.40]	(*N* = 848)	−0.21 [−0.60, 0.20]	(*N* = 112)	−0.00 [−0.36, 0.34]	(*N* = 857)	−0.10 [−0.43, 0.37]	(*N* = 114)
	Young (<40)	−0.04 [−0.43, 0.45]	(*N* = 469)	−0.20 [−0.60, 0.39]	(*N* = 71)	0.01 [−0.36, 0.34]	(*N* = 473)	−0.04 [−0.41, 0.43]	(*N* = 72)
	Middle age (40–65)	−0.01 [−0.38, 0.37]	(*N* = 885)	−0.29 [−0.67, 0.13]	(*N* = 110)	0.01 [−0.32, 0.32]	(*N* = 901)	−0.10 [−0.46, 0.17]	(*N* = 111)
	Old age (>65)	0.01 [−0.39, 0.40]	(*N* = 235)	−0.19 [−0.39, −0.01]	(*N* = 15)	0.02 [−0.32, 0.28]	(*N* = 238)	−0.20 [−0.51, −0.03]	(*N* = 15)
Daily sleep time	Overall	0.01 [−0.36, 0.36]	(*N* = 6575)	0.23 [−0.20, 0.71]	(*N* = 791)	0.00 [−0.28, 0.30]	(*N* = 6672)	0.38 [−0.01, 0.85]	(*N* = 801)
	Male	0.00 [−0.36, 0.35]	(*N* = 2299)	0.25 [−0.20, 0.71]	(*N* = 255)	−0.00 [−0.29, 0.29]	(*N* = 2340)	0.34 [−0.03, 0.83]	(*N* = 258)
	Female	0.01 [−0.36, 0.37]	(*N* = 4276)	0.23 [−0.19, 0.70]	(*N* = 536)	0.00 [−0.28, 0.30]	(*N* = 4332)	0.40 [−0.00, 0.86]	(*N* = 543)
	Young (<40)	0.02 [−0.36, 0.40]	(*N* = 2259)	0.21 [−0.17, 0.66]	(*N* = 270)	0.01 [−0.27, 0.32]	(*N* = 2294)	0.39 [0.03, 0.84]	(*N* = 276)
	Middle age (40–65)	0.00 [−0.36, 0.35]	(*N* = 3555)	0.26 [−0.20, 0.73]	(*N* = 463)	0.00 [−0.28, 0.30]	(*N* = 3606)	0.39 [−0.03, 0.88]	(*N* = 467)
	Old age (>65)	−0.04 [−0.39, 0.29]	(*N* = 750)	0.11 [−0.17, 0.67]	(*N* = 57)	−0.06 [−0.31, 0.25]	(*N* = 761)	0.30 [−0.02, 0.66]	(*N* = 57)

		BMI: Median [IQR] (*N* = number of valid samples)			
		COVID-19 negative		COVID-19 positive	
BMI	Overall	26.56 [23.52, 31.00]	(*N* = 7561)	27.60 [23.92, 32.03]	(*N* = 917)
	Male	26.65 [24.11, 30.28]	(*N* = 2679)	27.80 [24.50, 30.57]	(*N* = 303)
	Female	26.50 [23.13, 31.51]	(*N* = 4882)	27.46 [23.48, 32.46]	(*N* = 614)
	Young (<40)	26.07 [22.97, 30.93]	(*N* = 2635)	26.93 [23.45, 31.74]	(*N* = 346)
	Middle age (40–65)	27.18 [23.96, 31.43]	(*N* = 4075)	27.99 [24.38, 32.29]	(*N* = 501)
	Old age (>65)	25.64 [23.14, 28.91]	(*N* = 840)	27.53 [23.46, 30.29]	(*N* = 69)

Values for positive and negative COVID-19 individuals, including or excluding the period after the test date, are reported. The results are stratified among gender and age groups. Median, interquartile range (IQR), and number of COVID-19 cases analyzed are reported.

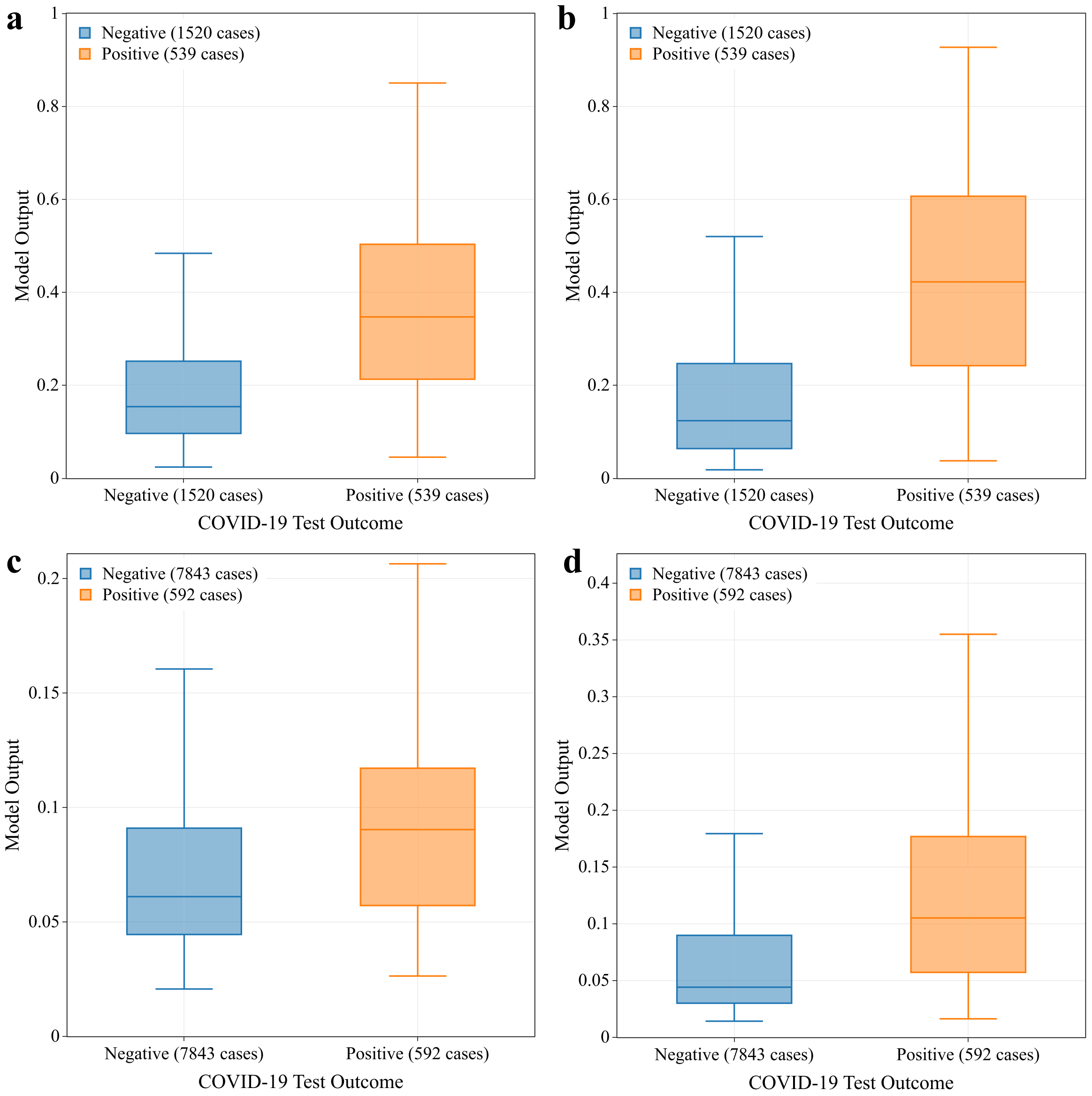

Fig. 1 Output of the prediction models. The model's output statistic is reported for symptomatic cases, excluding (**a**) and including the data after the test date (**b**), and for no-symptom-reported cases, excluding (**c**) and including the data after the test date (**d**). The boxes represent the IQR, and the horizontal lines are the median values. The number of cases considered for the analysis are reported in the legend.

evaluating detection of COVID-19 from wearable devices[16]. The availability of high-frequency intra-day data has shown promise to identify pre-symptomatic infection[8], even if additional studies with a larger number of individuals are needed to prove this point. Several studies focused on the specific data provided by a single sensor brand, showing that an increase in respiratory rate[17] and heart rate[9], or a decrease in heart rate variability[12], are significant during an illness, and that the changes in these physiological signals are more severe for COVID-19-positive cases relative to those affected by other influenza-like-illnesses[11].

We believe a strength of this research program is that anyone, with any wearable sensor, can participate. As wearable sensors continue to evolve and increase in number, predictive algorithms not dependent on a specific device or data type are needed to optimize the value of continuous, individual data. The algorithm proposed in this work is designed to ingest all available data, exploiting the information provided by the most advanced sensors, while detecting the presence of a COVID-19 infection for everybody owning any type of wearable sensor. The algorithm recognized the importance of self-reported symptoms in the prediction accuracy, but it is also designed to work in the absence of them, thus extending its applicability to the asymptomatic, pre-symptomatic or just a less engaged population who may not want to bother with reporting symptoms.

The analysis of individuals without self-reported symptoms extends the use of the algorithms for a fully passive monitoring of the pandemic and provides the possibility of applying the algorithm in other settings that collect wearable sensor data but are not equipped to collect and analyze self-reported symptoms. (Supplementary Information) Among them, the largest is Corona-Dataspende, a project developed by the Robert Koch Institute to collect sensor data from more than 500,000 individuals, monitoring the course of the pandemic in Germany[18].

The negligible importance given by our algorithm to the demographic features may be explained by observing that the physiological features we consider are changes with respect to an individual baseline. While an individual's baseline differs based on their demographic features, the changes with respect to the

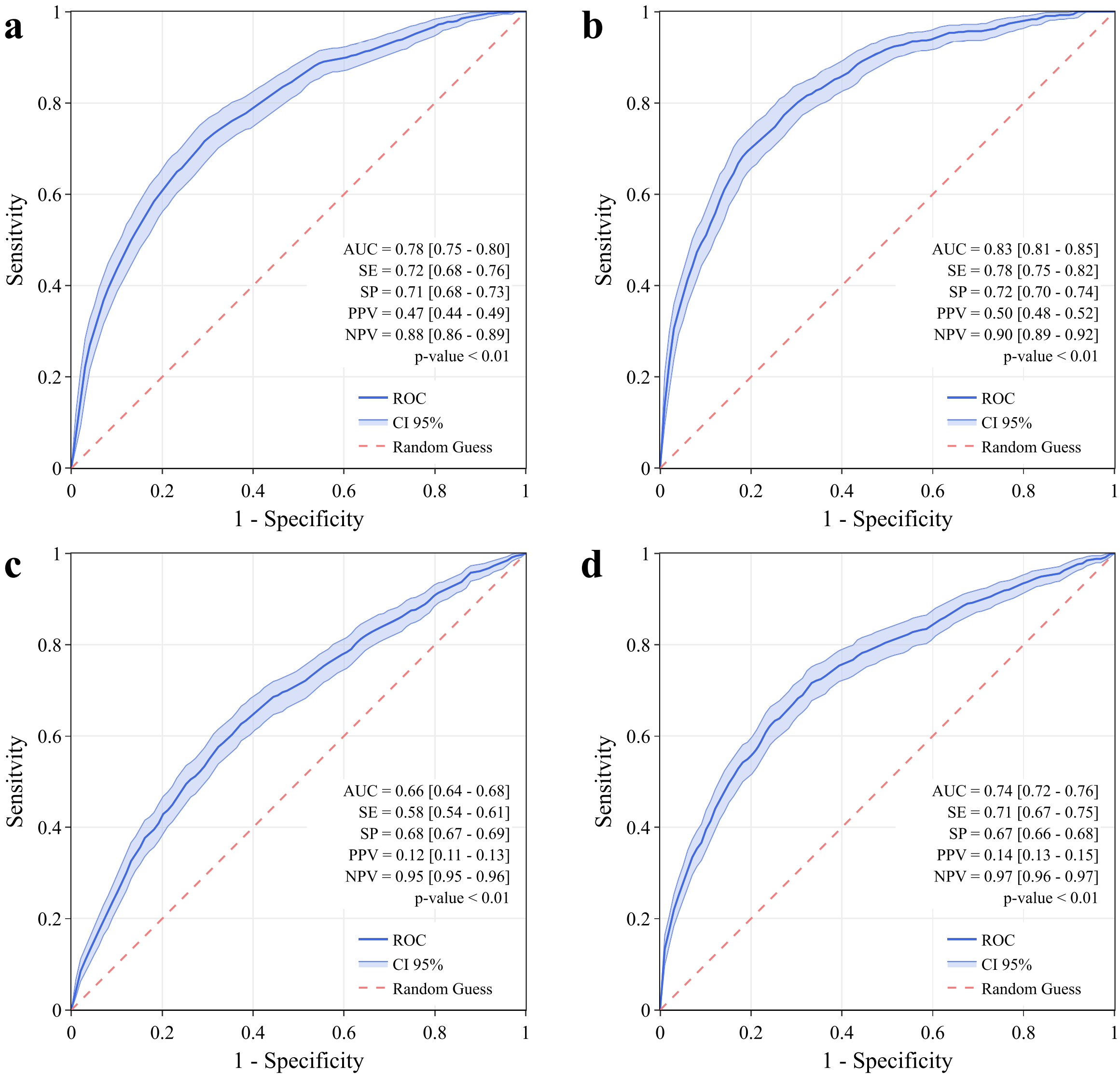

Fig. 2 Evaluation results for the discrimination between COVID-19 positive and COVID-19 negative. Receiver operating characteristic curves (ROCs) for the discrimination between COVID-19 positive and COVID-19 negative. Performance for symptomatic cases, excluding (**a**) and including the data after the test date (**b**), and for no-symptom-reported cases, excluding (**c**) and including the data after the test date (**d**), are reported. The model is a gradient boosting prediction model based on decision trees. Median values and 95% confidence intervals (CIs) for sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) are reported, considering the point on the ROC with the highest average value of sensitivity and specificity. Error bars represent 95% CIs. *p*-values of the one-sided Mann-Whitney U test are reported.

baseline that we use in our algorithm are not much affected by the demographic characteristics of the individual.

While the use of machine learning in the detection and prognostication for COVID-19 based on chest radiographs and CT scans have been questioned in a systematic review that discussed how none of the current studies are of potential clinical use due to biases or methodological flaws[19], the use of machine learning to enable a continuous and passive COVID-19 early detection is both very promising—for the potential to be scaled up effectively to a large fraction of the population—and repeatable—since we used a strictly separated test set for each of the cross-validation folds. Furthermore, the prediction algorithms developed as part of the DETECT system could be adapted to study the long term health problems due to COVID-19[20–25], or the effects of COVID-19 vaccine on vital signs and individual behavior[26–28]. For future infectious pathogen epidemics and pandemics, the new machine learning algorithms developed from the DETECT data can be adapted and re-used for early detection of various types of infections, towards the development of a new system to monitor the spread of future viral illness and prevent future outbreaks or pandemics.

In DETECT, all data is participant reported with no validation of the accuracy of self-reported symptoms, test dates or results. While we were able to collect continuous data, the amount of sensor data collected, or the accuracy of self-reported symptoms, depends entirely on the willingness of the participants to wear the sensor and accurately report how they feel. Despite the fact that the information collected may not be as accurate as in a controlled laboratory setting, previous work has demonstrated the value of participant-reported outcomes[29–31]. In the data analysis, among the people who reported the COVID-19 test outcome (active participants), we separated participants who reported at least one symptom from those who did not report any symptoms. The app indeed did not have an explicit way to report the absence of symptoms, so potentially some symptomatic individual may have not reported their symptoms.

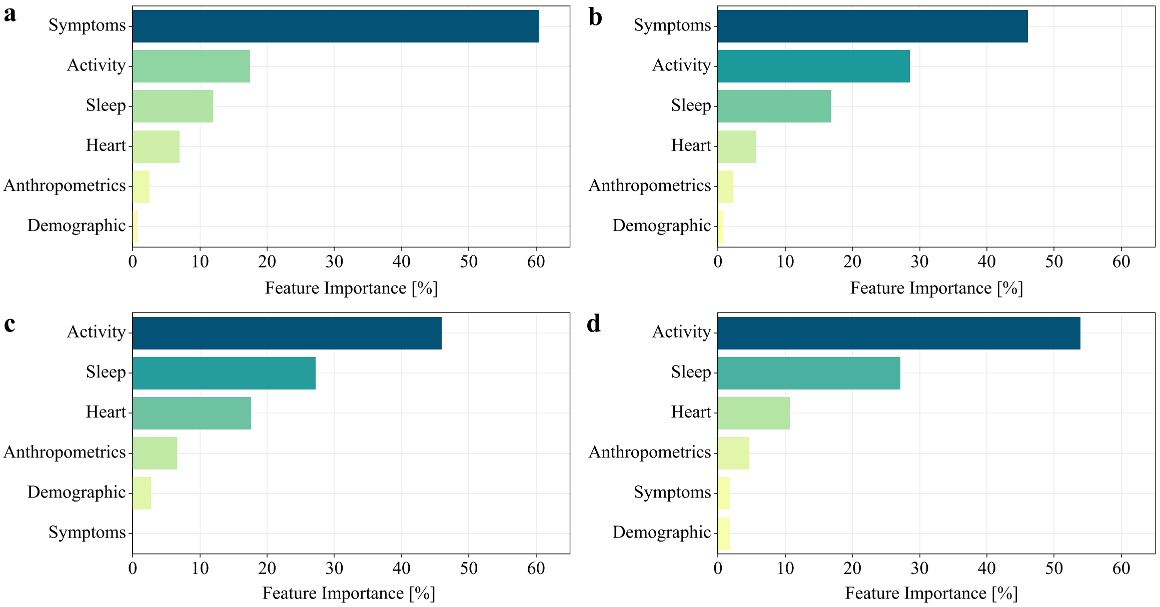

Fig. 3 Overall feature importance. Overall feature importance based on the average prediction changes when the feature value is perturbed. Values are normalized as percentages. Features have been aggregated into macro categories. Results for symptomatic cases, excluding (**a**) and including data after test date (**b**), and for no-symptom-reported cases, excluding (**c**) and including data after test date (**d**), are reported.

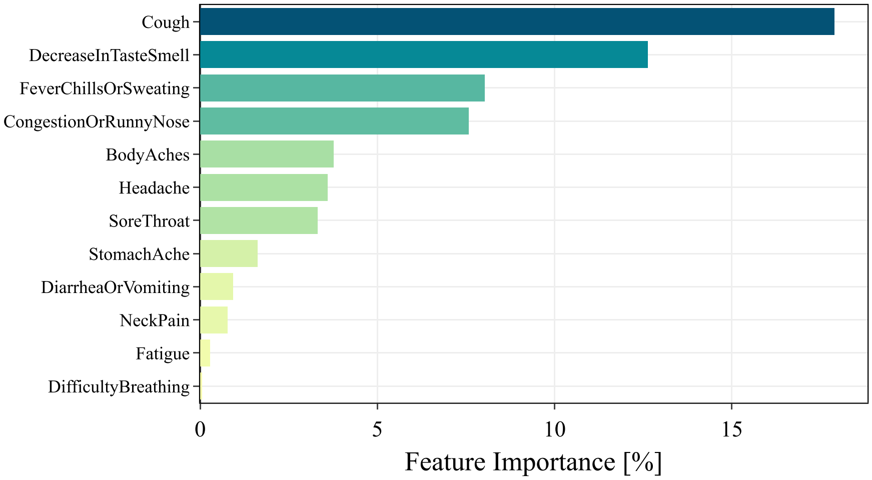

Fig. 4 Feature importance associated with specific symptoms. Only symptoms reported before the test date have been considered. Values are normalized as percentages. The results refer to symptomatic cases only.

Furthermore, this study is based on the aggregation of continuously monitored data into a finite number of daily features. A recent study has provided new insights about the analysis of intra-day changes for monitoring physiological variations[32], that may be used in future studies. Changes in more advanced metrics, like respiratory rate[17], peripheral temperature[10] or HRV[12], may also prove to add to the prediction of a COVID-19 infection, even if they have been marginally considered in our work since only a small fraction of participants were providing this type of data.

While previous studies have shown the importance of remote monitoring of individuals, extending health research beyond the limits of brick and mortar health systems[33,34], additional disparities are introduced when the study relies on wearable sensors, due to reduced accuracy for certain skin tones[35] and unequal access to this digital technology[36]. The decreasing cost of wearable sensors (some now less than $35) and the inner adaptability of our detection algorithm to any sensor and any given level of engagement of the participant with the in-app system will hopefully help in decreasing the barriers for underserved and underrepresented populations.

METHODS

Study population

Individuals living in the United States and at least 18 years old are eligible to participate in the DETECT study. After downloading the iOS or Android research app, MyDataHelps, and consenting into the study, participants are asked to share their personal device data (including historical data collected prior to enrollment) from any wearable device connected through direct API (for Fitbit devices), or via Apple HealthKit or GoogleFit data aggregators. A participant is invited to report symptoms, diagnostic test results, vaccine status, and connect their electronic health records, but they can opt to share as much or as little data as they would like. The recruitment of participants happens via the study website (www.detectstudy.org), several media reports, or outreach from our partners at Walgreens, CVS/Aetna, Fitbit and others.

Ethical considerations

All individuals participating in the study provided informed consent electronically. The protocol for this study was reviewed and approved by the Scripps Office for the Protection of Research Subjects (IRB 20–7531).

Data collection, aggregation, and group definition

All the participants with at least one self-reported result for a COVID-19 swab test during the entire data collection period have been considered in this study. Based on the reported data, an individual is considered Negative if the test resulted negative and no other positive tests have been reported in the period from 60 days before to 60 days after the test date. A minimum distance of 60 days is guaranteed between tests from the same individual considered in the analysis. This ensures that, if multiple tests are reported in the same period, only the first one is considered in our analysis, and the ones reported in the following 60 days will be ignored.

For each participant, we collect the data preceding and following the test date from all the connected devices, including, among others, detailed sleep intervals, number of steps and daily resting heart rate values. All the considered metrics are reported and detailed (Table 1). If multiple values per day are available for the same data type, a specific pre-processing has been applied to obtain a single representative daily value. Data has been collected from all the devices synchronized with the Fitbit or HealthKit application available on the smartphone. If data of the same type is available from multiple devices, only the most used device in the monitored period is considered.

Along with device data, we also analyze the reported surveys looking for self-reported symptoms. We considered all the symptoms reported from 15 days before to the day of test, further dividing the participants into two groups: symptomatic cohort, if we observe at least one reported symptom before the day of test, and no-symptom-reported cohort, if no symptom has been reported during this period. The frequency of each reported symptom for positive and negative cases are also reported (Fig. 5).

Baseline evaluation

Behavioral and physiological data acquired from wearable devices are highly idiosyncratic. The intrinsic inter-individual variability of physiological metrics, the different habits of the users, and the multiple purposes of the wearable devices requires a careful definition of the baseline value for each of the considered metrics. Also, the different hardware and software used by each manufacturer is an additional source of discrepancies. In this study, we limited both the inter-subject and inter-device variability considering the deviation from a dynamic daily baseline, which makes the feature values independent from the actual measures recorded by the device, and only dependent on the difference with respect to historical records.

Thus, the daily baseline is calculated using an exponentially weighted moving average:

$$\text{Baseline}[d] = \sum_{n=0}^{60} \text{Weight}[n] \times \text{DailyValue}[d-n] \quad (1)$$

where d is the current day and n represent the number of days before d, with a maximum of 60 days before the current date, while *DailyValue* can be any of the daily data measures among the ones considered.

The oscillation of the measures during the baseline period also changes over time. To measure the daily baseline variability, we evaluate the weighted standard deviation using the same weights of the baseline as in Eq. (2).

$$\text{Baselinevariability}[d] = \sqrt{\sum_{n=0}^{60} \text{Weight}[n] \times (\text{DailyValue}[d-n] - \text{Baseline}[d])^2} \quad (2)$$

The weights (Fig. 6) decrease exponentially as $e^{-\alpha n}$ with $\alpha = 0.05$. We exclude the first 7 days ($n < 7$) from the computation to avoid recent changes to affect the baseline value.

Many behavioral habits present strong weekly patterns, such as an increased sleep duration during the weekend, or weekly physical activities. To take into account these behaviors in the baseline evaluation, and to reduce the chances of false positives, we consider weekly patterns by increasing three times all the weights corresponding to the same day of the week.

During the course of a temporary disease, physiological measures may be different. In order not to affect the baseline value, which should only depend on the normal behavior, we exclude the 10 days following any reported symptoms from the baseline evaluation.

Finally, the weights are normalized to sum to 1 using Eq. (4).

$$w[n] = \begin{cases} 0 & \textit{if } n < 7 \textit{ or } n > 60 \\ e^{\alpha n} & \textit{if } n > 7 \\ 3e^{\alpha n} & \textit{if } n = 7m, \; m = 1, 2, \ldots, 8 \\ 0 & \textit{if } n \in [s, s+10], s = \text{symtpom date} \end{cases} \quad (3)$$

$$\text{Weight}[n] = \frac{w[n]}{\sum_i w[i]} \quad (4)$$

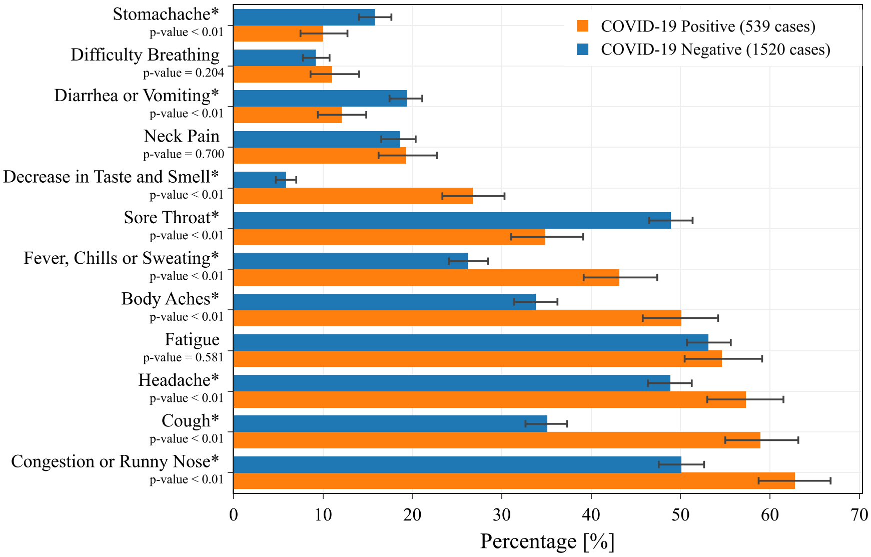

Fig. 5 Percentage of reported symptoms. Percentage of reported symptoms for participants who reported at least one symptom from 15 days before to 5 days after the test date. The frequencies of the indicated symptoms are shown for positive and negative cases. The error bars represent 95% percent confidence intervals. The *p*-values of a two-sided Fisher's exact test applied to COVID-19 positive (539 individuals) and negative participants (1,520 individuals) are reported. Symptoms with a significant difference (*p*-value < 0.05) are marked with an asterisk.

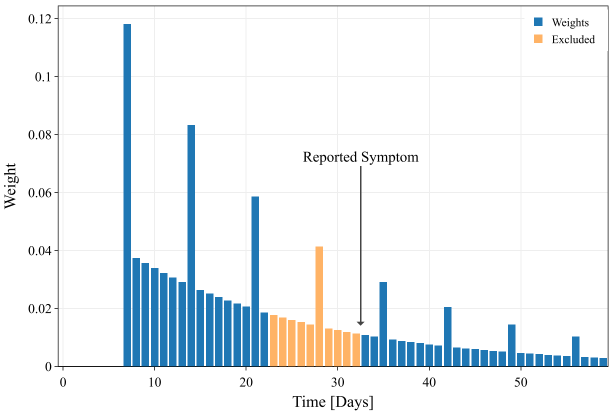

Fig. 6 Weights for the evaluation of the baseline data. Exponentially decreasing weights for the evaluation of the baseline data, with weekly patterns. The abscissa represents the temporal distance preceding the analyzed day considered for the baseline evaluation. The first 6 days have been excluded to avoid recent changes to affect the baseline. Additionally, if a symptom has been reported in this time frame, we set the weights to zero from the day of symptom to the next 10 days.

The deviation from the baseline values is then evaluated as:

$$\text{DailyValueDev}[n] = \frac{\text{DailyValue}[n] - \text{Baseline}[n]}{\text{BaselineVariability}[n]} \tag{5}$$

This value represents how far the specific metric is from the expected normal value, day by day. Values are considered to be valid only if at least 50% of corresponding data are available during the baseline period.

Feature extraction

We propose two analyses, one considering all available data (5 days before and 5 days after the test date), and one considering only the period before the test date (5 days before test), in order to further analyze the impact of the test outcome on the individual behavior and the natural course of the disease. The choice of a 5 day period is related to the dynamic nature of the baseline. Considering an extended period, potentially altered measures may be included in the baseline evaluation, and a short abnormal status may not be detected if averaged over a long time period.

We consider four different macro categories of feature (Table 1).

Sensor features: all the features acquired or derived from the device measurements belong to this group. In this study, we consider the minimum, average, and maximum deviation values from the baseline in the days considered. This category is further divided into three subcategories, including activity, heart and sleep related features.

Symptom features: a separate binary feature is considered for each of the reported symptoms. If the corresponding symptom has been reported in the considered period its value is set to 1, otherwise 0.

Anthropometrics: if available from the monitored devices, several anthropometric features are also considered like body weight, height, body mass index, fat percentage, and basal metabolic rate.

Demographic features: this category includes age and gender self-reported by the participants.

Using the aforementioned features, we developed a gradient boosting prediction model based on decision trees[37]. The sparse nature of the data set, resulting from the availability of only a subset of features for most of the participants, and the absence of many daily values due to a low wearing time, makes decision tree-based algorithm a natural choice for this study.

The model has been trained and tested in four different conditions, using data from the symptomatic or no-symptom-reported cohort, and preceding the reported test date or considering all available data around the test date. Normalized deviation (*Z*-score) from an idiosyncratic and dynamic baseline value was evaluated daily for each metric and each individual. A weighted average based on past data was defined as the baseline estimation, whose weights are reported (Fig. 6).

The entire data set has been randomly divided into 5 separate non-overlapping test sets. For each test set, a model is trained using all the remaining data, ensuring an equal percentage of positive cases between train and test sets. For each model, we also ensure that the test set remains strictly separate from the training, so training data are not involved in the test.

To analyze the intrinsic variability of the model due to data availability, we estimate 95% confidence intervals for the presented results. Bootstrap method has been utilized for this purpose, with 10,000 independent random iterations from the test outcomes.

To have a better understanding of the effect of COVID-19 on physiological and behavioral aspects, we consider symptomatic and no-symptom-reported cases separately. Additionally, different models have been analyzed considering sensor features evaluated including and excluding sensor data after the reported test date. Comparative results are presented in terms of AUC of the ROCs. Sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV), associated with an optimal operating point, are also reported. The optimal operating point is defined as the point with the highest average value between SE and SP.

The interpretable nature of the decision tree model allows for the evaluation of feature importance estimates[38,39]. To this end, we evaluate, for each feature, the average prediction changes when a variation is applied to the feature value. The bigger the value of the importance, the bigger is the change to the prediction value if this feature is changed. To have a more comprehensive overview of the feature importance, we further aggregated the importance associated with features in the same category. Feature importance values are normalized so that they can be expressed as percentages.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

All interested investigators will be allowed access to the analysis data set following registration and pledging to not re-identify individuals or share the data with a third party. All data inquiries should be addressed to the corresponding author.

CODE AVAILABILITY

Further information on research design and on all software packages adopted in the analysis is available in the Life Sciences Reporting Summary linked to this article.

Received: 15 June 2021; Accepted: 19 October 2021;
Published online: 08 December 2021

REFERENCES

1. NIH. COVID-19 Treatment Guidelines. https://www.covid19treatmentguidelines.nih.gov/whats-new/ (NIH, 2021).
2. Manabe, Y. C., Sharfstein, J. S. & Armstrong, K. The need for more and better testing for COVID-19. *JAMA* **324**, 2153–2154 (2020).
3. Menni, C. et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat. Med.* **26**, 1037–1040 (2020).
4. Oran, D. P. & Topol, E. J. Prevalence of asymptomatic SARS-CoV-2 Infection. *Ann. Int. Med.* **173**, 362–367 (2020).
5. Quer, G., Gouda, P., Galarnyk, M., Topol, E. J. & Steinhubl, S. R. Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults. *PLoS ONE* **15**, e0227709 (2020).
6. Jaiswal, S. J. et al. Association of sleep duration and variability with body mass index: sleep measurements in a large US population of wearable sensor users. *JAMA Int. Med.* **180**, 1694–1696 (2020).
7. Quer, G. et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nat. Med.* **27**, 73–77 (2021).
8. Mishra, T. et al. Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat. Biomed. Eng.* **4**, 1208–1220 (2020).
9. Natarajan, A., Su, H.-W. & Heneghan, C. Assessment of physiological signs associated with COVID-19 measured using wearable devices. *NPJ Digital Med.* **3**, 156 (2020).

10. Smarr, B. L. et al. Feasibility of continuous fever monitoring using wearable devices. *Sci. Rep.* **10**, 21640 (2020).
11. Shapiro, A. et al. Characterizing COVID-19 and influenza illnesses in the real world via person-generated health data. *Patterns* **2**, 100188 (2021).
12. Hirten, R. P. et al. Physiological data from a wearable device identifies SARS-CoV-2 infection and symptoms and predicts COVID-19 diagnosis: observational study. *J. Med. Internet Res.* **23**, e26107 (2021).
13. Cleary, J. L., Fang, Y., Sen, S. & Wu, Z. A caveat to using wearable sensor data for COVID-19 detection: the role of behavioral change after receipt of test results. Preprint at *medRxiv*, 2021.2004.2017.21255513 (2021).
14. Whelton, S. P. et al. Association between resting heart rate and inflammatory biomarkers (high-sensitivity C-reactive protein, interleukin-6, and fibrinogen) (from the Multi-Ethnic Study of Atherosclerosis). *Am. J. Cardiol.* **113**, 644–649 (2014).
15. Radin, J. M., Wineinger, N. E., Topol, E. J. & Steinhubl, S. R. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. *Lancet Digit. Health* **2**, e85–e93 (2020).
16. Radin, J. M., Quer, G., Jalili, M., Hamideh, D. & Steinhubl, S. R. The hopes and hazards of using personal health technologies in the diagnosis and prognosis of infections. *Lancet Digit. Health* **3**, e455–e461 (2021).
17. Miller, D. J. et al. Analyzing changes in respiratory rate to predict the risk of COVID-19 infection. *PLoS ONE* **15**, e0243693 (2020).
18. Robert Koch-Institut. *Corona Datenspende*, https://corona-datenspende.de/science/en (Robert Koch-Institut, 2020).
19. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
20. Dani, M. et al. Autonomic dysfunction in 'long COVID': rationale, physiology and management strategies. *Clin. Med. (Lond.)* **21**, e63–e67 (2021).
21. Puntmann, V. O. et al. Outcomes of cardiovascular magnetic resonance imaging in patients recently recovered from coronavirus disease 2019 (COVID-19). *JAMA Cardiol.* **5**, 1265–1273 (2020).
22. Sudre, C. H. et al. Attributes and predictors of long COVID. *Nat. Med.* **27**, 626–631 (2021).
23. Logue, J. K. et al. Sequelae in adults at 6 months After COVID-19 infection. *JAMA Netw. Open* **4**, e210830–e210830 (2021).
24. Radin, J. M. et al. Assessment of prolonged physiological and behavioral changes associated with COVID-19 infection. *JAMA Netw. Open* **4**, e2115959–e2115959 (2021).
25. Radin, J. M. et al. Assessment of prolonged physiological and behavioral changes associated with COVID-19 infection. *JAMA Netw. Open* **4**, e2115959 (2021).
26. Voysey, M. et al. Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet* **397**, 99–111 (2021).
27. Benedict, C. & Cedernaes, J. Could a good night's sleep improve COVID-19 vaccine efficacy?. *Lancet Respir. Med.* **9**, 447–448 (2021).
28. Quer, G. et al. The Physiologic Response to COVID-19 Vaccination. Preprint at *medRxiv* 2021.2005.2003.21256482 (2021).
29. Basch, E. et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. *JAMA* **318**, 197–198 (2017).
30. Bell, S. K. et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA Netw. Open* **3**, e205867–e205867 (2020).
31. Rivera, S. C. et al. The impact of patient-reported outcome (PRO) data from clinical trials: a systematic review and critical analysis. *Health Qual. Life Outcomes* **17**, 156 (2019).
32. Mishra, T. et al. Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat. Biomed. Eng.* **4**, 1208–1220 (2020).
33. Steinhubl, S. R. et al. Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: The mSToPS randomized clinical trial. *JAMA* **320**, 146–155 (2018).
34. Radin, J. M. et al. Pregnancy health in POWERMOM participants living in rural versus urban zip codes. *J. Clin. Transl. Sci.* **4**, 457–462 (2020).
35. Colvonen, P. J., DeYoung, P. N., Bosompra, N.-O. A. & Owens, R. L. Limiting racial disparities and bias for wearable devices in health science research. *Sleep* **43**, zsaa159 (2020).
36. Beaunoyer, E., Dupéré, S. & Guitton, M. J. COVID-19 and digital inequalities: Reciprocal impacts and mitigation strategies. *Comput. Hum. Behav.* **111**, 106424 (2020).
37. Dorogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. Preprint at https://arxiv.org/abs/1810.11363 (2018).
38. Lundberg, S. M. & Lee, S.-I. Consistent feature attribution for tree ensembles. Preprint at https://arxiv.org/abs/1706.06060 (2017).
39. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. Preprint at https://arxiv.org/abs/1802.03888 (2018).

ACKNOWLEDGEMENTS

This work was funded by grant number UL1TR002550 from the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH) (E.J.T., S.R.S., G.Q.) and by grant number OIA-2040727 (Convergence Accelerator) at the National Science Foundation (NSF) (M.G., G.Q.).

AUTHOR CONTRIBUTIONS

M.G. and G.Q. made substantial contributions to the study conception and design. M.G., J.M.R., K.B.-M., E.R., S.R.S., and G.Q. made substantial contributions to the acquisition of data. M.G. and G.Q. conducted statistical analysis. M.G., S.R.S., and G.Q. made substantial contributions to the interpretation of data. M.G. and G.Q. drafted the first version of the manuscript. M.G., J.M.R., K.B.-M., E.R., V. K., E.J.T., S.R.S., and G.Q. contributed to critical revisions and approved the final version of the manuscript. M.G. and G.Q. take responsibility for the integrity of the work.

COMPETING INTERESTS

S.R.S. is employed by PhysIQ. The other authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-021-00533-1.

Correspondence and requests for materials should be addressed to Giorgio Quer.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021