Prediction of SARS-CoV-2 PCR positivity in UK Fenland study participants using smartphone-submitted vital sign measurements

Alex Moore (✉ alex.moore@huma.com )
Huma Therapeutics Ltd

Nikola Dolezalova
Huma Therapeutics Ltd.

Davide Morelli
Huma Therapeutics Ltd.

Effrossyni Gkrania-Klotsas
University of Cambridge  https://orcid.org/0000-0002-0930-8330

Adam Cunningham
Huma Therapeutics  https://orcid.org/0000-0002-9791-7813

Adam Booth
Huma Therapeutics Ltd.

David Plans
Huma Therapeutics Ltd.

Angus Reed
Huma Therapeutics Ltd.

Mert Aral
Huma Therapeutics

Kirsten Rennie
MRC Epidemiology Unit, University of Cambridge

Nick Wareham
MRC Epidemiology Unit  https://orcid.org/0000-0003-1422-2993

Article

Keywords: SARS-CoV-2, telemedicine, prediction model, machine learning, smartphone data

Posted Date: August 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1885504/v1
Abstract

Introduction

Mobile health applications are increasingly being used in health and clinical research. SARS-CoV-2 has proven to have high infectivity, making outbreaks difficult to contain. Early detection can help prevent spread, but there is a need to develop easy-to-use screening tools that can help identify potential infection as early as possible. Here, we describe the development of a machine learning classifier that can predict SARS-CoV-2 PCR positivity using smartphone-submitted vital sign measurements.

Methods

The Fenland App study followed 2,199 UK participants using a smartphone application from August 2020 and for a minimum of six months. Participants completed a baseline questionnaire and then monthly questionnaires about SARS-CoV-2 status and vaccinations. Three times a week, participants provided measurements of their blood oxygen saturation, body temperature, and resting heart rate via a pulse oximeter, digital thermometer, and their smartphone. The participants participated in self initiated SARS-CoV-2 testing as per concurrent public health guidelines. We built predictive models SARS-CoV-2 PCR positivity status as obtained from national surveillance PCR test results.

Results

A total of 77 positive and 6,339 negative SARS-CoV-2 tests were recorded during the study. The final model achieved an ROC AUC of 0.695 ± 0.045. There was no difference in model performance when using 4, 8 or 12 weeks of baseline data before a SARS-CoV-2 test (F(2) = 0.80, p = 0.472). Addition of demographic or symptom information had no impact on model performance.

Conclusions

Using only three smartphone collected vital sign measurements, it is possible to predict SARS-CoV-2 PCR positivity, using a four week baseline period. Smartphone based remote monitoring of patient vital signs could allow for earlier screening for potential infections. This method could be applicable to any infectious disease that causes physiological changes in vital signs.

Introduction

The coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to a great loss of human life, affected the long-term health of countless people, and resulted in unprecedented pressure on healthcare systems worldwide. The use of telemedicine and mobile health technologies increased rapidly over the course of the pandemic in order to reduce person-to-person contact. Remote patient monitoring and telehealth have been used in patients with COVID-19, and other conditions, including remote triaging for patients on surgical waiting lists, or
simply for delivering routine medical appointments where possible. Smartphone applications have also been used to allow scientific research to take place remotely\textsuperscript{5}, and to support ongoing clinical trials\textsuperscript{6,7}.

The SARS-CoV-2 non-delta variants that were dominant in the UK for the duration of the study were characterised by an incubation period with a median length of five days\textsuperscript{8,9} and a peak of infectivity 1–4 days prior to the onset of symptoms\textsuperscript{10–12}. Early detection of SARS-CoV-2 infection would allow for timely isolation and testing of infected individuals and thus a potential reduction in disease transmission. Regular asymptomatic testing can help identify cases, but the virus can be detected in the upper respiratory tract only about two days before the onset of symptoms\textsuperscript{13}. Using vital sign data for early detection of physiological changes after infection could alleviate the need for asymptomatic testing \textsuperscript{14–16}.

Infection-triggered early inflammatory response has been shown to provoke changes in heart rate\textsuperscript{14,15,17–19}, heart rate variability\textsuperscript{15,20–23}, respiratory rate\textsuperscript{15}, peripheral temperature\textsuperscript{16,18}, and behavioural changes in sleep\textsuperscript{14,19} and activity\textsuperscript{14,19}. With the rise of mobile health technologies, these vital signs can now be readily collected by smartphones and wearables, making it simpler than ever before to longitudinally monitor individuals' health status and alert the individual to the need of further investigation or tests if necessary. Using machine learning models leveraging this vital sign data, early detection models can be designed, including for SARS-CoV-2. While a number of risk prediction or prognostic prediction models have been developed for SARS-CoV-2\textsuperscript{14,15,22,24}, there are a much smaller number of published diagnostic models.

The Fenland App cohort study followed 2,199 UK participants for a minimum of six months and up to nine months from August 2020 using a smartphone app. High engagement from the participants allowed for the generation of a rich dataset containing smartphone-collected vital signs, regular health questionnaires, symptom information, and SARS-CoV-2 test results from the Second Generation Surveillance System (SGSS). In this study, we aim to leverage this unique dataset to develop a model for the prediction of SARS-CoV-2 PCR positivity.

**Methods**

**Participant recruitment and data collection**

Participants were recruited from an ongoing observational cohort, the Fenland study. The Fenland study is a population-based cohort study of 12,435 participants born between 1950 and 1975. The study was set up with the aim of examining the interaction between genetic and environmental factors in determining risk of diabetes, obesity, and related health conditions. Participants were recruited from General Practice sampling frames in Cambridgeshire. Further details are published elsewhere\textsuperscript{25,26}. The data used in the present study were collected through the Fenland App sub-study of the Fenland COVID-19 study. The main aim of the Fenland COVID-19 study was to determine the prevalence of previous
infection with COVID-19 in this known population-based cohort using three-monthly blood sample measures of SARS-CoV-2 IgG antibodies. The Fenland App study aimed to investigate the progression of SARS-CoV-2 from the pre-symptomatic to the symptomatic stage via smartphone-acquired digital measures, using the Fenland COVID-19 App, which was developed by Huma Therapeutics. Ethical approval was provided by the South West - Cornwall & Plymouth Research Ethics Committee (REC reference 20/SW/0100).

**Collected data**

SARS-CoV-2 PCR testing data from the Second Generation Surveillance System (SGSS), the national reporting system across England, were obtained for all Fenland COVID-19 study participants during the study period. These contained all routine laboratory tests for SARS-CoV-2 infections from hospitals (patients and NHS staff) and community testing in the general population before and during the study period. In this analysis, we used the first confirmed positive SARS-CoV-2 PCR test result to classify if participants had a SARS-CoV-2 virus infection either before or during the study using the date of the PCR test.

Dried blood spot samples were collected remotely by participants every three months during the study to determine the presence of SARS-CoV-2 antibodies. These were analysed for SARS-CoV-2 IgG antibodies using a commercial enzyme linked-immunosorbent assay (ELISA) targeting Spike (S2) and Nucleoprotein (N) from SARS-CoV-2 (Omega Diagnostics, UK), interfacing with the semi-quantitative Omega/Mologic SARS-CoV-2 IgG assay. Results were classified into two categories; positive or negative (negative or borderline results). Baseline IgG antibody results were used to ascertain the exclusion criteria described below.

All participants were asked to complete a baseline questionnaire at the onset of the study containing information about any previous SARS-CoV-2 infections. Following that, on a monthly basis, participants completed questionnaires about changes in their health status and whether they had received a SARS-CoV-2 vaccination in the prior month, including the date and type of vaccination.

Three times a week, participants were also asked to provide measurements of oxygen saturation levels, body temperature, and resting heart rate using a provided pulse oximeter (ChoiceMMed MD300C29), digital thermometer (Genial Digital Thermometer T12L), and their smartphone camera, respectively. Participants were asked to manually enter the results from the pulse oximeter and thermometer into the app. Resting heart rate was captured by the participant placing their finger over the camera on their smartphone for approximately 60 seconds.

At each measurement time point, participants were also asked to record whether they were experiencing any symptoms from a list provided or select “no symptoms”. The option to add other symptoms not in the list was also given. The list of symptoms was updated regularly during the study as further symptoms were reported. For this analysis, the presence of any of the three confirmed SARS-CoV-2
symptoms (fever, cough, and loss of taste and/or smell) identified at the time of the study were
categorised as yes or no.

**Participant inclusion/exclusion and data censoring**

Participants of the Fenland App Study who completed the baseline questionnaire were included in this
study. Exclusion criteria were a previous self-reported SARS-CoV-2 infection or a baseline positive SARS-
CoV-2 antibody test. Furthermore, participants with insufficient longitudinal vital sign data were excluded.
Sufficient longitudinal vital sign data were defined as a minimum of one record of heart rate, oxygen
saturation, and temperature during the week before a SARS-CoV-2 PCR test and at least two records in the
three weeks prior to that time point.

Data collection on the smartphone app started on 6th August 2020 and the study closed on 30th April
2021. Participant-specific censoring was applied from the time of withdrawal until the end of the study,
from 90 days before a positive SARS-CoV-2 antibody test until the end of the study, from the day after a
positive SARS-CoV-2 test until the end of the study, and from the day of SARS-CoV-2 self-reported
vaccination for a duration of five days.

**Data pre-processing**

Raw longitudinal vital sign data were cleaned as follows: non-physiological values were removed (< 89%
and > 100% for oxygen saturation, < 40 and > 180 BPM for resting heart rate, and temperature
measurements were restricted to < 35℃). Resting heart rate values further than five standard deviations
from the population mean were removed. For all three vital signs, if a participant took multiple
measurements in a day, their mean was considered as the daily vital sign value.

The longitudinal data was then up-sampled on a daily basis and linear interpolation was applied to fill in
missing values between individual time points and values were forward-filled after the last provided time
point. Time points of interest were then filtered relative to the date of the SARS-CoV-2 test. All features
were scaled by subtracting the mean and divided by the standard deviation calculated on the training
dataset prior to model training.

**Feature transformations**

In addition to using the raw vital sign data, we performed several transformations on the data. These
were based on splitting the pre-processed longitudinal data into baseline data (e.g. -28 days to -7 days
prior to test) used to calculate the baseline/normal representation for each individual and transformed
data (e.g. -6 days to 0 days prior to test) used as transformed inputs for classification.

The first transformation was a z-score: for each day of the transformed data period, a z-score was
calculated using the mean and standard deviation of baseline data (separate for each vital sign). For the
next two transformations, we used anomaly/novelty detection algorithms.
The Isolation Forest algorithm, implemented via the python scikit-learn library\textsuperscript{28}, was fitted on the baseline data and for each day of the transformed data period, an anomaly score was predicted and used as the multivariate feature.

Next, a Vector AutoRegression (VAR) model, implemented via the python statsmodels library\textsuperscript{29}, was fitted on the baseline data for each participant and data was forecasted for each day of the transformed period. This forecasted data was then compared to the actual data and reconstruction errors were used as the transformed features (one for each vital sign and a summed multivariate feature).

Finally, we created an additional feature based on each of the above which took a maximum value of each transformed feature over the transformed time period (e.g. seven days). The list of all transformed features used in the model can be found in Supplementary Table 1.

**Definition of model outcome**

The machine learning task was designed as a binary classification of positive and negative SARS-CoV-2 PCR tests using longitudinal records of vital signs as inputs. The positive class consisted of participants who had a positive SARS-CoV-2 PCR test during the study period and the longitudinal data were capped at the date of their first positive test. The negative class was sampled from participants who never had a positive antigen test during the study period and had a negative PCR test during the study period. To remove seasonality effects and achieve a more balanced dataset, four negative examples were sampled for each positive example based on having a date of negative test within three days before or after the corresponding positive example test date.

**Model selection and evaluation pipeline**

Each experiment was configured based on the input features (e.g. heart rate, oxygen saturation, and temperature), the decision on whether to use raw or transformed features, the number of weeks of data provided to the model (e.g. four, eight, or twelve weeks) and the end of the longitudinal data stream (e.g. \(-3, -2, -1, 0\) days before the SARS-CoV-2 test). All further parameters were optimised by the pipeline described below.

If raw features were used in the model, the feature space was reduced by recursive feature elimination (using the RFECV class of scikit-learn library, to create a support vector classifier with linear kernel and balanced class weights, evaluated over three stratified folds by the area under ROC curve, with a minimum of five features to be selected). For transformed features, features were selected in sets based on transformation operation by Optuna\textsuperscript{30} hyperparameter optimization procedure described below.

The optimisation pipeline was further comparing three classification algorithms: logistic regression, random forest, and support vector classification, all implemented in the scikit-learn library. Balanced class weights were used for all three models and the maximum number of iterations was increased to 4000 for logistic regression. Other parameters of the models were either optimised during the hyperparameter search or kept as default.
Optimisation was performed using Tree-Structured Parzen Estimator (TPE) from the Optuna library and the details of the search space are provided in Supplementary Table 2. Based on the size of the search space, 100 and 500 Optuna trials were performed for raw or transformed features, respectively. Optimisation was based on maximisation of the area under the ROC curve score (AUC ROC score). Random seeds were set to allow for reproducibility of the results.

To obtain an unbiased estimate of the model performance, a nested cross-validation procedure was used. The dataset was split into five outer folds (stratified on outcome), for each of which the best model was selected after a separate feature selection and hyperparameter optimization were evaluated on the five inner folds. The performance of this best model was then evaluated on the unseen holdout set for the respective outer fold31.

In order to explore if it was possible to predict PCR test positivity in the period before the test date, models were trained using the previously determined optimal model hyperparameters and data from different days relative to the positive test.

Finally, we also investigated if it was possible to detect vaccination events, under the assumption that a vaccination would cause a change in vital signs. Models were constructed as described previously, but used the date of first COVID-19 vaccination instead of the positive test date for the positive class. Participants used as negative examples must not have been vaccinated during the window. Positive and negative examples were matched on age and sex.

Besides the AUC ROC score, accuracy, precision, and recall were recorded for each result of the cross-validation pipeline. Further, we obtained results through the generation of confusion matrices, ROC AUC curves and precision-recall curves (PR curves). All of these were calculated according to the implementation in the scikit-learn library.

**Statistical analysis and visualisations**

Demographic comparison of positive and negative classes was performed with the help of the python tableone library32. Statistical comparisons were obtained by performing a two-sample t-test for continuous variables and a chi-squared test for categorical variables, evaluated on a significance level of 0.05.

All visualisations were generated using the matplotlib33 and seaborn34 libraries. A full list of all python libraries and their versions used in this study can be found in Supplementary Table 3.

**Results**

**Study population**

In total, there were 2,199 participants in the Fenland COVID-19 Huma App study who completed the baseline questionnaire. After participant onboarding, vital sign and symptom data was collected over the
course of the study. The majority of positive SARS-CoV-2 PCR tests were recorded in the winter period while the number of negative tests steadily rose over the course of the study as testing became more widely available (Fig. 1a). In this participant pool, there were a total of 77 positive and 6,339 negative SARS-CoV-2 antigen tests during the study period. After applying inclusion and exclusion criteria and sampling negative participants, 33 positive and 113 negative participants were included in the analysis here (Supplementary Figs. 1a and b). Further to this, some data points were censored after withdrawal from the study (7), a positive SARS-CoV-2 antigen test, a positive SARS-CoV-2 antibody test (124), or vaccination (1355) (Supplementary Fig. 1c).

There were no statistically significant differences between the average vital sign values and demographics of positive (pre-infection) and negative groups (Supplementary Table 4). Visual comparison of data collected on positive and negative groups around the time of SARS-CoV-2 antigen test can be found in Fig. 1b.

**Feature transformations improve model performance**

As primary input features, we used resting heart rate, oxygen saturation, and body temperature. While daily step data was available, we decided not to use this as the self-isolation policies during the course of the study would not enable us to distinguish between a drop in physical activity due to physiology and the effect of having to self-isolate.

To begin with, we supplied the features to the model raw, i.e. one non-transformed record per vital sign per day. This resulted in a cross-validated ROC AUC of 0.538 ± 0.124, precision of 0.305 ± 0.172 and recall of 0.304 ± 0.168. Looking closer at the features which were selected using recursive feature elimination in the five folds, there is no obvious pattern to which features (heart rate vs. oxygen saturation vs. temperature) nor time point before testing that the model found useful for the predictions (Supplementary Fig. 2a).

Because we used simple binary classifiers, we considered the importance of feature transformations on the longitudinal vital sign data. We implemented four different feature transformations: z-score, Isolation Forest, Vector AutoRegression, and maximum over transform days (see Methods section for details). Using these transformed features instead of the raw inputs resulted in a significant increase of predictive performance, with ROC AUC rising from 0.538 ± 0.124 to 0.695 ± 0.045 (Fig. 2a, p = 0.045, two-tailed t-test). It should be noted that the feature selection process (outlined in the Methods section) results in the model only using the most predictive subset of these transformed features. The precision of the model with transformed features was 0.465 ± 0.104 and recall 0.601 ± 0.138.

**A reliable personal baseline can be estimated from four weeks of collected data**

Larger baseline windows increase the amount of data available to the model when calculating a baseline. However, the increased number of entries is also associated with a much higher variance. Table 1 below
summarises the number of oxygen saturation data entries submitted for different baseline window sizes. It should be noted that the vast majority of data submissions involved the participant entering readings for all three vital signs. Other types of data submissions followed the same pattern of increasing variance with longer baseline periods.

Table 1
Mean and standard deviation of number of data entries for oxygen saturation.

<table>
<thead>
<tr>
<th>Baseline Window Size</th>
<th>Number of Data Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>4 weeks</td>
<td>11.42</td>
</tr>
<tr>
<td>8 weeks</td>
<td>20.87</td>
</tr>
<tr>
<td>12 weeks</td>
<td>25.76</td>
</tr>
</tbody>
</table>

In terms of accuracy of the baselines built on four, eight or twelve weeks of data, it seems that four weeks provided sufficient data to create a reliable baseline. For example, Fig. 2b indicates that there is an average error of 0.1°C in temperature between the baseline data collected over four weeks, while the average increase in temperature in SARS-CoV-2 positive participants was over 0.2°C. This suggests that even with the shortest baseline duration, the error in measurement is small enough to not be lost by noise. As baseline durations increase, the error in temperature measurements decreases.

Indeed, the performance evaluated in the cross-validation pipeline was comparable for the three input window lengths, as a one way ANOVA revealed no significant difference in model performance (Fig. 2c, Supplementary Fig. 4b) (F(2) = 0.80, p = 0.472), suggesting that four weeks of data is sufficient to provide a reliable baseline for anomaly detection transformations. Confusion matrices are presented in Fig. 2d.

Addition of demographic features or symptom information has no impact on model performance

We further hypothesised that if the physiological response to SARS-CoV-2 infection differs depending on the age and sex of the infected individual, the addition of these demographic features may improve the discriminative ability of the model. We also had access to self-reported symptom records during the week of the PCR test for 76% of the example participants. We evaluated whether the addition of symptom information could further boost model performance. Symptom information consisted of four binary features for reported SARS-CoV-2 symptoms (breathlessness, fever, persistent or non-persistent cough) in the seven days before the PCR test, the number of times SARS-CoV-2 symptoms were reported in the seven days before the test and number of times any symptoms were reported in the seven days before the test (Supplementary Fig. 3a). Full lists of symptoms can be found in Supplementary Table 1.
There was no statistical difference in the ROC AUC score for the original model 0.695 ± 0.045 and the model which included demographic features (0.660 ± 0.105, p = 0.56, two-tailed t-test, Supplementary Fig. 3b) or symptom features (0.705 ± 0.141, p = 0.94, two-tailed t-test, Supplementary Fig. 3c).

**Smartphone-collected vital signs can be used to predict SARS-CoV-2 PCR test positivity**

Based on the previous analyses, the pipeline was set up with the use of transformed features for resting heart rate, oxygen saturation, and body temperature, while supplying four weeks of data. As described in the Methods section, the unbiased performance was calculated using the pipeline with five-fold cross-validation while the final model was generated using the whole dataset (without any hold-out test set).

The final model selected by the optimal parameter search pipeline was a support vector classifier with linear kernel and regularisation parameter C of 1,000. During the cross-validation, the models selected in the five folds varied (Table 2). Furthermore, the pipeline selected seven days to be the optimal number of transformed feature days (i.e. of the four weeks of total data provided, 21 days were used to generate the baseline, seven days transformed as input features for the model). This was also the most popular selection in the five-fold cross-validation.

Regarding transformed features selected as inputs, the final model included only the eight features from the “maximum over transform days”. In the cross-validation, this was also the most commonly used feature set, with four out of five folds selecting it, while VAR was selected three times and the Isolation Forest feature twice. Z-score, which was the simplest transformation of all, was never selected (Table 2). Final model parameters and performance are presented in Table 2.
Table 2

Summary of the final model parameters, features and performance. Five-fold cross-validation results and final model results are shown alongside in columns. LR = Logistic Regression, RF = Random Forest, SVC = Support Vector Classifier, VAR = Vector AutoRegression.

<table>
<thead>
<tr>
<th></th>
<th>Cross-validation</th>
<th>Final model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>3 x LR, 1 x RF, 1 x SVC</td>
<td>SVC</td>
</tr>
<tr>
<td>Transform days</td>
<td>1 x 3 days, 3 x 7 days, 1 x 10 days</td>
<td>7 days</td>
</tr>
<tr>
<td><strong>Features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-score</td>
<td>5 x exclude</td>
<td>Exclude</td>
</tr>
<tr>
<td>Isolation Forest</td>
<td>1 x include, 4 x exclude</td>
<td>Exclude</td>
</tr>
<tr>
<td>VAR</td>
<td>3 x include, 2 x exclude</td>
<td>Exclude</td>
</tr>
<tr>
<td>Maximum over transform days</td>
<td>4 x include, 1 x exclude</td>
<td>Include</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROC AUC in training</td>
<td>0.762 ± 0.068</td>
<td>0.782</td>
</tr>
<tr>
<td>ROC AUC on test</td>
<td>0.695 ± 0.045</td>
<td>-</td>
</tr>
<tr>
<td>Recall on test</td>
<td>0.601 ± 0.138</td>
<td>-</td>
</tr>
<tr>
<td>Precision on test</td>
<td>0.465 ± 0.104</td>
<td>-</td>
</tr>
<tr>
<td>Accuracy on test</td>
<td>0.751 ± 0.052</td>
<td>-</td>
</tr>
</tbody>
</table>

The model can detect positive SARS-CoV-2 tests up to 3 days in advance of positive test date

In order to explore how model performance would change over the course of infection, we trained new models using the previously determined optimal model hyperparameters (Table 2) and data censored at different number of days prior to the positive test. Performance accuracy across the five outer folds is summarised in Fig. 3a. The models performed significantly better than random when censored from up to three days before a positive test. We also found that performance did not improve when using data from the days following a confirmed infection.

Model can detect other events accompanied by similar vital sign changes

We further investigated if our models could predict vaccination events since the vital sign changes after vaccination follow a similar pattern to a real infection, albeit smaller in magnitude (data not shown). We applied the same inclusion criteria as in the main study and used the date of first COVID-19 vaccination
instead of the positive test date; participants used as negative examples must not have been vaccinated during the window. Positive and negative examples were matched on age and sex. The ROC AUC score of the final model was 0.754, with 0.818 accuracy. As shown in Table 3 summarising precision and recall, the model performs particularly well when predicting negative cases. Vaccination elicited similar changes in vital signs as infection (Fig. 4).

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unvaccinated</td>
<td>0.92</td>
<td>0.85</td>
</tr>
<tr>
<td>Vaccinated</td>
<td>0.47</td>
<td>0.66</td>
</tr>
</tbody>
</table>

**Discussion**

In this study, we present a classification method for predicting SARS-CoV-2 upper respiratory PCR test positivity from vital sign data in the days preceding the SARS-CoV-2 test. While other studies have achieved similar aims, they have focused on classifying from continually worn specialised wearable devices\textsuperscript{14,22,35} that collect considerably more data points; rather than short measurements taken three times a week with standard devices. At a population level, the method used in this study may be more feasible to apply and has comparable performance to those using continually-worn wearables. Addition of demographic or symptom data did not improve model performance, suggesting that intermittent vital sign information alone is sufficient for early detection of potential infection.

The primary outcome was a classification model with a mean 5-fold cross-validated ROC AUC of 0.70 indicating good performance. Models based on raw longitudinal vital sign data performed more poorly than models built with transformed longitudinal vital sign data. One of the reasons for this is that the actual date of infection is unknown and PCR tests are not taken on the same day relative to this, resulting in misalignment of the timeline for the vital sign data changes. Simple classification models cannot deal with such complex longitudinal data. Simpler transformations, such as the calculation of z-scores, resulted in features that were less useful than more complex anomaly detection models such as the isolation forest and VAR. Future studies on small datasets should consider preprocessing of longitudinal vital sign data using anomaly detection models before use in simple classification models. For large sample sizes, models such as neural networks, or autoencoders might be able to handle the raw untransformed data. Unsurprisingly, the most useful transformation was the maximum over transform days, which results in a better alignment of the different participants in the time domain.

Interestingly, we demonstrated that just four weeks of intermittent baseline data is sufficient as input for the anomaly detection transformations. There was no difference in performance of classifiers built on
four, eight or twelve weeks worth of vital sign data, which is promising for adoption in real-world settings. However, our dataset was relatively sparse in terms of data points per participant. While increasing measurement density may allow for shorter baseline periods, it should be considered carefully. More frequent monitoring may increase user burden and cause poorer user engagement. By using a longer baseline, and imputing missing timepoints we can keep participant inconvenience to a minimum. Continuous, passive monitoring, such as through wearable devices, also has problems. Wearables have different acceptability ratings by different populations, and when individuals are feeling ill they may simply not wear their device\textsuperscript{14,36}. Additionally, wearable devices may have different compatibility levels with iOS and Android devices, where certain features may only be available on certain operating systems. This study was implemented on both iOS and Android devices increasing the applicability of the study to the general population, rather than being restricted to certain device types due to compatibility issues.

The addition of age and sex did not improve the model performance. This was surprising as SARS-CoV-2 is known to affect males and older individuals more severely\textsuperscript{37}. It may be that these variables have no or little effect on the likelihood of progressing from a pre-symptomatic to a symptomatic state. However, due to the low number of positive examples in the dataset, we might not have sufficient statistical power to identify such effects. Importantly, including self-reported symptoms also did not increase model performance, possibly due to the quality of the collected symptom data or the participant’s perception of experiencing relevant symptoms. Changes in vital signs may occur before the occurrence of clinical symptoms, and symptom records are more subjective than vital sign measurements. Not requiring participants to input symptom data is an additional strength of the method, reducing uncertainty around symptom severity, or subjectivity of symptom input.

A particular strength of the study is that physical activity data, in the form of step counts, was not required, unlike previous research using smart or wearable devices for data collection\textsuperscript{14,15,22,35}. Step data could be influenced by periods of governmental social restriction or shelter-in-place instructions that were widespread as part of the response to SARS-CoV-2\textsuperscript{38,39}. It would be difficult to distinguish between a decrease in physical activity due to individuals remaining at home due to these restrictions or due to a physiological response to an infection or illness. Here, we present models with a good predictive performance that require only a smartphone, thermometer, and pulse oximeter. Importantly, inputting the measurements used in this study takes approximately five minutes, only three times a week, and does not require expensive additional devices, such as a wearable. This increases the applicability of the model to large-scale use in order to benefit public health.

Further strength of this study is that it accounted for the seasonality of the data. There are known seasonal variations during the year within individuals’ vital signs. To account for this, we ensured that the negative class is distributed across the year in the same manner as the positive class by matched sampling.

While we were able to predict PCR test positivity in the days leading up to a positive test, it should be noted that at this stage of the pandemic, displaying one of the following symptoms was a requirement...
for access to a PCR test in the UK: persistent cough, fever, and change in smell or taste. As a result, predictions made in the days running up to a positive test cannot be considered pre-symptomatic. However, prediction in advance of a positive test is still useful for alerting individuals to consider having a test, particularly at the beginning of symptom presentation when symptoms are mild and not necessarily indicative of COVID-19 infection. Such early alert systems may in turn lead to earlier testing and detection, reducing the spread of the disease and in the future potentially supplementing testing as a method of population surveillance.

The main limitation of this study is the small number of recorded positive cases. Due to this small sample size, all of these were included in the nested cross-validation without leaving out an unseen test set for validation. To best leverage the available data for validation, we considered the change in participants’ vital signs in response to vaccination, which elicited similar physiological responses to the COVID-19 infection (Fig. 3b). Our model was able to distinguish between vaccinated and non-vaccinated participants. This suggests that the method of using changes in vital signs to detect infection events is applicable not only to this condition or cohort but could potentially be used to detect other events with similar physiological signatures, such as other influenza-like illnesses, or responses to medical treatments.

Another limitation is that while pulse oximeters and thermometers are widely available, they still add extra devices that are required for data collection. Inaccurate data collection using these devices is also a potential issue, demonstrated by the fact that a proportion of body temperature records were at the lower limit of the digital thermometer of 35°C. Symptom data is also subject to bias. While the participants were asked to complete the symptom questionnaire irrespective of whether they were experiencing symptoms, we found that participants were more likely to complete the symptom questionnaire when experiencing symptoms. Furthermore, the exact symptom onset time is unknown as it could have happened at any time between the symptom questionnaire timepoints. This means we were unable to explore symptom progression as participants move from asymptomatic to symptomatic infection. Some bias may have been introduced into our engineered features through calculation of means and standard deviations on resampled data. Individuals who have less data are likely to have smaller standard deviations.

A possible way to improve the performance of the model would be to better align the participant data with regards to the actual infection date. While this date is unknown, there are several potential strategies to estimate it. Firstly, we attempted to align the data based on data shape (data not shown) but due to the low number of time points per participant this did not lead to improvement in alignment. Another approach would be a probabilistic estimation of the infection date based on known incubation times of the virus and symptom onset, similar to the approach utilised in Hellewell et al. However, due to the low availability of accurate symptom data for this cohort, we did not attempt this approach.

In conclusion, we present a method for predicting symptomatic SARS-CoV-2 with corresponding SARS-CoV-2 test positivity from vital sign baseline data only without the need for step data, self-reported...
symptom data, or additional demographic information. The model has moderate performance and requires only data collected through a smartphone, thermometer, and pulse oximeter. It could be useful in remote settings for early detection of SARS-CoV-2 infection, and provides evidence for the utility of smartphone collected vital sign data in detecting physiological changes associated with infection.

Declarations

Acknowledgements

We would like to thank all the participants who kindly took part in the study.

Funding

The study and NJW are supported by the Medical Research Council (grant MC_UU_00006/1). KR and EGK are supported by the National Institute of Health Research (NIHR) Biomedical Research Centre in Cambridge (IS-BRC-1215-20014). EGK was also supported by the NIHR via a Greenshoots Award. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. Additional support was provided by Huma Therapeutics.

Author contributions

ND, DM, AM, AB, KR, EGK, DP, MA, NJW, conceived and designed the study. ND, AM, DM, AB conducted modelling and statistical analysis. ND, DM, EGK, AM, ACC, AB, DP, ABR, MA, KR, NJW, aided in writing the manuscript and reviewed for intellectual content. All authors approved the final version.

Data availability

Data is available on reasonable request to the authors.

Code availability

Code is available on reasonable request to the authors.

Conflict of interest

ND, DM, AM, AC, AB, MA ABR and DP are or were employees of Huma Therapeutics Ltd at the time of writing.

References


**Figures**

**Figure 1**

**Summary of collected data. a)** Timeline of data collection for PCR and antibody/serological testing, baseline questionnaire, symptom questionnaire, and vital signs. Each bar represents the number of records in a particular week of the study. Data for all 2199 participants is shown. **b)** Seven-day moving average of vital sign records for positive (n=33) and negative (n=304) example participants included in the study around the time of the positive/negative SARS-CoV-2 PCR test (±12 weeks).

**Figure 2**

**Optimisation of feature pre-processing. a)** ROC-AUC curves for models using raw vs. transformed features. **b)** Absolute differences between baselines calculated using 4, 8, and 12 weeks of measurements and a long-term baseline calculated from all available values for heart rate, oxygen saturation, and body temperature. Each data point corresponds to a participant included in the study. **c)** ROC-AUC curves for models using 4 weeks vs. 8 weeks vs. 12 weeks of input data. **d)** Confusion matrices
with summed values from 5 folds, along with percentages of total cases included in the test sets. In a) and c) the solid line shows the mean of the ROC-AUC curves from 5 folds, and the filled area covers ± one standard deviation of the ROC-AUC curves from 5 folds.

**Figure 3**

*Model performance in predicting PCR test positivity at different days in advance of the positive test date.* Day 0 corresponds to the day of the PCR test. ROC AUC score of 0.5 corresponds to a random guess prediction. Mean ± SD shown.

**Figure 4**

*Three-day moving average of vital sign records for participants with a positive SARS-CoV-2 PCR test (orange) and first COVID-19 vaccination (blue) around the time of the test/vaccination (±4 weeks).*

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx