Supplementary Materials for

Symptom clusters in Covid19: A potential clinical prediction tool from the COVID Symptom study app

Authors: Carole H. Sudre[1]*†& Karla A. Lee[2]† & Mary Ni Lochlainn[2]†, Thomas Varsavsky[1], Benjamin Murray[1], Mark S. Graham[1], Cristina Menni[2], Marc Modat[1], Ruth C E Bowyer[2], Long Nguyen[3], David Drew[3], Amit D. Joshi[3], Wenjie Ma[3], Chuan-Guo Guo[3], Chun-Han Lo[3] , Sajaysurya Ganesh[4], Abubakar Buwe[4], Joan Capdevila Pujol[4], Julien Lavigne du Cadet[4], Alessia Visconti[2], Maxim B Freydin[2], Julia S. El-Sayed Moustafa[2], Mario Falchi[2], Richard Davies[4], Maria F. Gomez[5], Tove Fall[5], M.Jorge Cardoso[1], Jonathan Wolf[4], Paul W. Franks[2,5], Andrew T. Chan[3], Tim D. Spector[2], Claire J Steves[2]†, Sébastien Ourselin[1*]†

.

Correspondence to: carole.sudre@kcl.ac.uk / sebastien.ourselin@kcl.ac.uk

This PDF file includes:

Materials and Methods
Figs. S1 to S4

Material and Methods

Assessment of exposure, Ascertainment of outcomes, Ascertainment of covariates

Exposure, outcome and covariates were all ascertained via the app as previously described *(3,4).* A subset of individuals reported being tested for COVID-19. BMI was calculated as kg/m^2. Visit to hospital was recorded if the location was ever recorded as hospital or "back from hospital".

The analysis of the disease course was separated into the unsupervised clustering and a predictive analysis for need of respiratory support using the projected clusters.

Subject Selection

All participants included in the analysis were required to:

1) Report a hospital visit or show sign of recovery (recovery was defined as a significant drop (at least 2) in number of reported symptoms at the day of last report compared to the day where sum of symptom was maximal)
2) Record their symptoms on the app at least three times over four days or more between the time their symptoms start and either a hospital visit or the start of symptom decline.
3) For the recovery group: be tested positive
4) For the group visiting hospital: either a) be tested positive, b) be imputed positive from the day where sum of symptoms was maximal following the imputation method described in Menni et al *(4)* c) Reported classic symptoms of COVID19 (2 days or more of fever and cough)

All participants fulfilling the criteria of selection were included with no selection done on the country (UK, US or Sweden) in which they reported.

Statistical analysis

Data from the app were downloaded into a server and only records where the self-reported characteristics fell within the following ranges were utilized for further analyses: age between 16 (18 in the US) and 100 years; BMI between 16 and 55 kg/m^2.

14 symptoms were recorded: abdominal pain, chest pain, sore throat, shortness of breath, fatigue, hoarse voice, headache, loss of smell or taste, confusion, diarrhea, fever, persistent cough, unusual muscle pains or skipped meals.

For the assessment of risk factors, reported respiratory support was used as dependent variable in a logistic regression, using age, BMI, as continuous variables and sex, frailty score, diabetes, lung, heart or kidney disease as binary variables. The frailty score was binarized at a threshold of 3 points.

Clustering analysis

Unsupervised time series clustering was performed using Mc2PCA *(5)* with 6 dimensions of projection from the 14 recorded binarized symptom course. This method allows for the clustering of time series with non-equal duration, using the covariance matrix of the time series. Optimization of the clustering is performed using a K-means iterative process as follows: for each cluster a singular value decomposition is performed over the average of covariance matrices and the first n (here 6) dimensions are used to calculate the projection. Attribution to each cluster is then chosen to minimize the residual error after projection. The process was iterated until the change in error ratio was below 10^{-4} (convergence criteria)

Linear interpolation between time points was used in the case of missing data and a limit of 5 days of interrupted record was filled in. Where more than one record was present for a single day, the latest record was considered.

To determine the optimal number of clusters to consider over the disease continuum, for each number of clusters, the K-Means clustering Mc2PCA was run with 20 random initialization and the attempt with final minimal average distortion was selected. The Bayesian information criteria (BIC) was applied to balance model fit and model complexity leading to a choice of 6 clusters. Supplementary figure 2 presents the output of the BIC selection:

After separation of the different clusters, an average symptom course was calculated as the interpolated frequency of reported symptom for each given day over the mean duration for the given cluster. Ability to predict final cluster classification with a reduced number of days of reporting was assessed via weighted precision and recall score in the 6-class problem and reported for the training and test set.

Predictive analysis

To demonstrate the relevance of the clustering to predict need for respiratory support, the performance of two random forest models on the testing set were compared. The first one used only the demographic characteristics while the second used the projected cluster at 5 days and the associated aggregated sum of symptoms over the five days. Youden Index obtained from the training set (0.066 and 0.059) for the model including symptomatology or only personal characteristics was applied on the test set for binary classification.

The random forests were trained to optimize the receiver operating curve area in a five-fold cross-validation setting with randomized grid search over hyperparameters.

A bootstrap analysis with 1000 samples was used to provide confidence intervals in the reported precision and false positive rates.

Figure S1. Flowchart showing entry of participants into analysis

Abbreviations: T0 no test; T+/-: Tested positive/negative; H+/H-: Attended hospital or did not attend hospital; I+: imputed COVID-19; SR+: self-reported classical symptoms of COVID-19 (2 days or more of fever and cough)

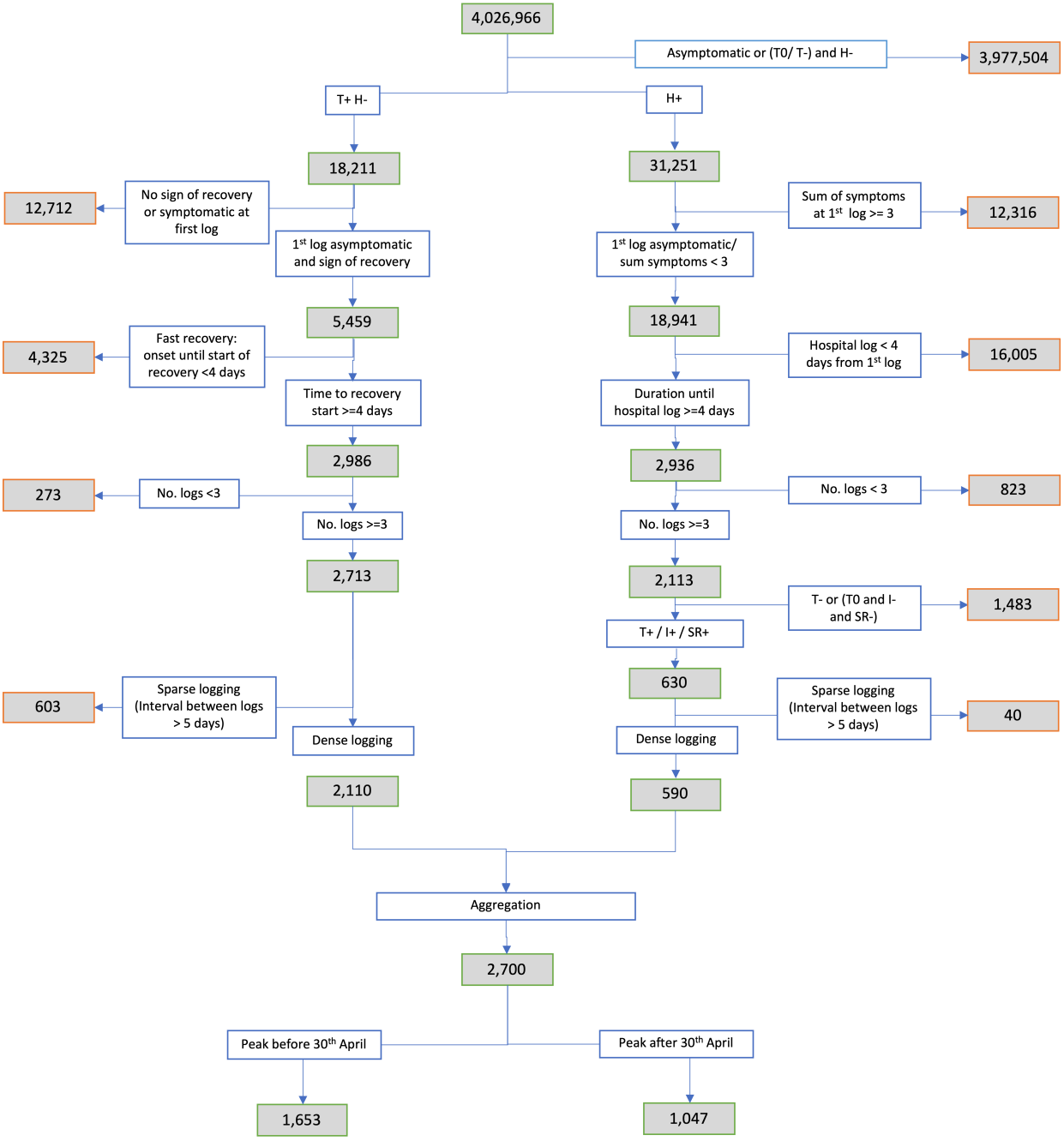
4,026,966
Asymptomatic or (T0/ T-) and H-
3,977,504
T+ H-
H+
18,211
31,251
12,712
No sign of recovery or symptomatic at first log
Sum of symptoms at 1st log >= 3
12,316
1st log asymptomatic and sign of recovery
1st log asymptomatic/ sum symptoms < 3
5,459
18,941
4,325
Fast recovery: onset until start of recovery <4 days
Hospital log < 4 days from 1st log
16,005
Time to recovery start >=4 days
Duration until hospital log >=4 days
2,986
2,936
273
No. logs <3
No. logs < 3
823
No. logs >=3
No. logs >=3
2,713
2,113
T- or (T0 and I- and SR-)
1,483
T+ / I+ / SR+
630
603
Sparse logging (Interval between logs > 5 days)
Sparse logging (Interval between logs > 5 days)
40
Dense logging
Dense logging
2,110
590
Aggregation
2,700
Peak before 30th April
Peak after 30th April
1,653
1,047

Figure S2 Bayesian information criterion for the choice of number of clusters

Figure S3

Top: associated Z-Score of presentation of symptoms over overall symptom distribution (red = reported more than average; blue = reported less than average)

Bottom: Frequency of positive answers per symptoms across days for each cluster (darker = reported more frequently)

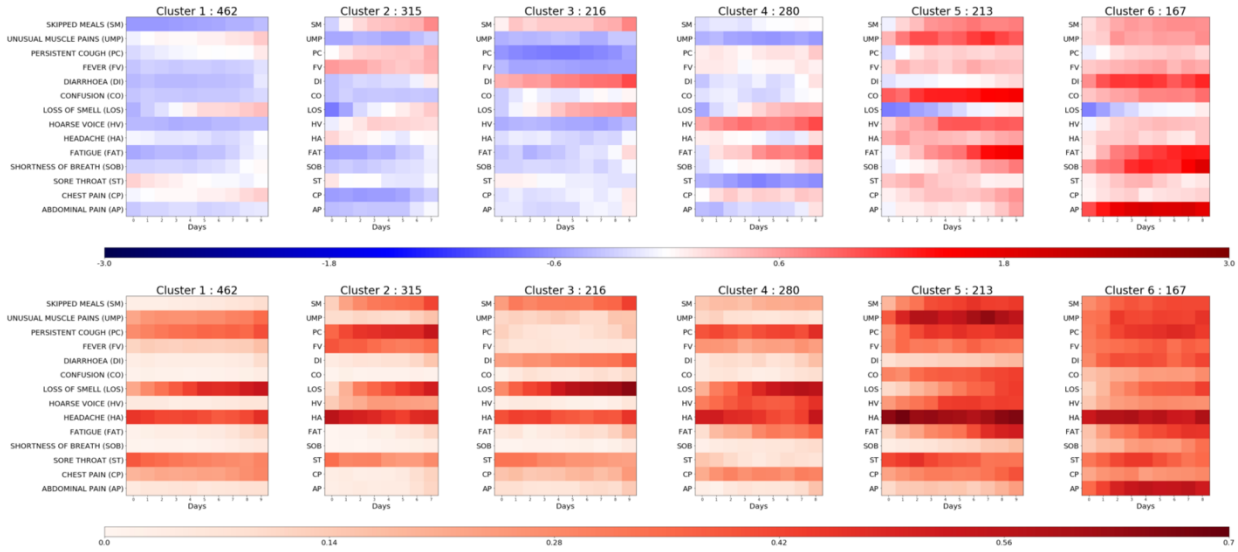

Fig S4 Confusion matrix for cluster assignment on the independent replication set using the projections learnt with limited number of days from the training set