Validation of Machine Learning (ML) Models as Diagnostic Tools to Predict Infection with SARS-CoV-2

May 26, 2020

Machine learning models & Statistical analysis plan

Study Phases
The study will have multiple phases, each phase will consist of 14 days. This short time period reflects the urgent nature of the pandemic and declining infection rate in the UK, and a need to get validation data as soon as possible in an unpredictable environment.

Before the start of each phase (day 0), a set of machine learning models will be frozen and submitted for validation on data collected during this and subsequent phases. These will be numbered numerically according to the phase at which they were submitted; e.g. V1; V2 and so forth.

After the end of each phase (day 14), we will report on the performance of all models submitted for validation in this and previous phases, each on the complete dataset collected across all phases since that given model was submitted for validation.

Machine learning models improve with accumulation of data. Therefore it is necessary to continue data collection to train improved models which will be submitted for validation in later phases. Therefore 14-day phases will continue as long as tests are available and app users consent to joining the study.

Machine learning models
Four model types will be validated:

a. Binary classifier to predict if infection is present/absent based on reported symptoms and phenotypic data. Variant tuned for high specificity, so with a bias towards false negative prediction
b. Binary classifier variant tuned for high sensitivity, so with a bias towards false positive prediction
c. Binary classifier variant tuned for a balance between specificity and sensitivity, so with a balance between false negative and false positive prediction
d. Four-category classifier using the three classifiers a, b and c above to divide predicted chance of infection into a four point scale

For each model type, variants will be validated which produce outputs based on either symptoms reported during the first 2 days or during the first 3 days of symptoms.

These models will have been trained on a dataset of test results, reported symptoms and phenotypic data collected prior to the phase at which they are submitted for validation. The models will be frozen and described on GitHub at https://github.com/zoe/covid-validation-study before they are submitted for validation.

As described above, 14 day phases will continue whilst it is deemed of value to public health, and additional models may be submitted for validation before the start of each 14 day phase.

Statistical analysis plan
For each of the models, the primary validation metric will be the measured performance of the model's predictions compared to the 'gold standard' swab PCR tests on the entire unadjusted study population for the phases since they were submitted for validation.

In addition to models being frozen before they are submitted for validation, we will also record hashes of each day's data extracts when they are taken, to ensure integrity.

For each of the three binary classifiers, we will report sensitivity and specificity for each model on the entire unadjusted study population. We will estimate confidence intervals on these measures by bootstrapping from within the study population.

For the four-category classifier, we will report the measured percentage of the entire unadjusted study population falling in each pairing of a swab test result and one of the four categories as predicted by the model. We will estimate confidence intervals for these percentages by bootstrapping from within the study population.

In addition to these primary validation metrics on the entire unadjusted study population, we will also report performance for the same analyses on:

- Subgroups of the study population divided by age category (20-39, 40-59, 60-79), sex (male, female), ethnicity (white, black, asian), BMI category (underweight, normal, overweight, obese, severely obese), IMD (quintiles), healthcare worker (yes, no)
- A first rebalanced study population, where each participant has been weighted according the subgroups listed above (age category, sex, ethnicity, BMI category, IMD quintile, healthcare worker) such that the total weighting of each subgroup reflects its weighting within the UK adult population below 80 years old.
- A second rebalanced study population, skewed to overweight more vulnerable groups to reflect the highest regional levels of these vulnerabilities within the UK. Vulnerabilities to be considered are level of dependency (PRISMA7), history of heart disease, history of diabetes, history of lung disease and asthma.

Each of these analyses will be carried out only on those participants who have completed the corresponding phenotype fields.