**Supplementary Materials**

# **Participants profile**

The breakdown of participants is shown in Table 1.

Table 1. Participant numbers in the trial.

|  |  |
| --- | --- |
| **Total number of participants**  | **100** |
| Covid-19 PCR swab test positive | 74 |
| Covid-19 PCR swab negative but with symptoms | 1 |
| Covid-19 PCR swab test negative | 24 |

For the trial, the age of participants ranged from 24 to 99. The median age was 60.5.

Table 2. Quartiles of participant ages.

|  |  |
| --- | --- |
| **Participant Ages** |  |
| Youngest | 24 |
| 1st Quartile | 53 |
| Median | 60.5 |
| 3rd Quartile | 72 |
| Oldest | 99 |
| Male Median | 60.5 |
| Female Median | 60.5 |

A breakdown of the ethnicity of the participants is shown in Table 3.

Table 3. Ethnicities of the participants.

|  |  |
| --- | --- |
| Black/African/Caribbean | 46 |
| White | 41 |
| Mixed Race | 6 |
| Did not specify | 6 |
| Sri Lankan | 1 |

For the analysis, some cases were excluded. The total number of participants used in the data analysis was 90.

Table 4. Patient numbers used for analysis.

|  |  |  |
| --- | --- | --- |
| **Total number of participants**  | **90** | **Repeats:** |
| Covid-19 PCR swab test positive | 66 | 0 |
| Covid-19 PCR swab test negative | 24 | 0 |

**Data Collection**

The GC-IMS device was configured for manual sample injection, Ruszkiewicz *et al*., (2020). Each identical instrument has a 30 m, 0.53 mm inner diameter MTX-WAX capillary GC column (Restek, Bellefonte, PA, US). The GC column is interfaced to a linear IMS. During each ten-minute sample run, the temperature was kept at 45 °C. The internal flow rate was 5 ml/min for the first 30 seconds, then linearly increased to 30 ml/min by the end of the program. The temperature of the IMS drift tube was held at 45 °C.

The performance of the instrument was checked by taking background measurements at the start and end of each day. This was to ensure the nominal performance of the instrument and to check for contamination, degradation in filter performance, etc. The instruments were put on a higher temperature and flow rate overnight to clean out any contaminants.

This consisted of a plastic tube with a 5 ml syringe inserted perpendicular to the direction of flow. The internal configuration of the breath probe is shown in Fig. 1. At the end of the exhalation process, the syringe was drawn to obtain the sample. The sample was then immediately taken to the location of the sampling instruments for processing.



Figure 1. The internal configuration of the breath probe.

# **Data Quality Assessment**

A data quality rating was assigned to each patient measurement. This was done after the sampling process. The spectra were rated on a scale from Q0 (best quality) to Q3 (worst). Three aspects were considered. Firstly, the spectra were reviewed in the LAV© IMS viewing software. An assessment was made of whether any contamination (in the form of lines in the GC dimension) or artefact/carryover from previous measurements (in the form of spurious markers in the spectra) was present. These are identifiable by appearing out of position on the Y-axis. The strength of these anomalous signals was rated as non-existent (Q0), mild (Q1), moderate (Q2), and strong (Q3). The instrument background spectra recorded at the start of each day were also used to identify contaminants in the patient spectra.

Secondly, the delay between arrival in the instrumentation room and sampling by the instrument was noted. Measurements carried out before the study showed a drop in relative peak height in the breath signals of up to 10% over ten minutes. In the same period, signals originating from the syringe materials increased by a factor of five. Therefore, the aim was to measure the breath sample within ten minutes. The following ratings were used: 0–5 minutes (Q0), 5–10 minutes (Q1), 10–20 minutes (Q2), and >20 minutes (Q3).

Thirdly, any comment logged by the instrument operator that suggested a problem with the sampling process was considered. In the event that this was expected to impact the sample quality (e.g., a critically ill patient could only breathe weakly), the spectra would be downgraded according to the impact. This was a qualitative assessment based on the severity of the issue. The final quality rating for each patient would be the lowest of the three aspects considered. A histogram of the number of cases by quality rating is shown in Figure 2.

Figure 2. Histogram of cases by quality rating.

**Machin Learning Statistical Data Analysis**

The breath data set was processed by the ML algorithms as follows: The breath sample spectra and patient data are imported into the model. Then the model is trained using a proportion of the data, keeping the remainder for testing. For the data shown here, 80% of the trial data was used for training. The remaining 20% of the data was used to test the predictions made by the algorithm. This process was repeated four more times, such that each 20% chunk of data was tested once. The number of iterations used needs to be large enough for the model to converge.

A conversion analysis was undertaken to optimise the acceptable number of iterations. During the analysis, data were processed with different iteration numbers from 500 to 3,000. The results are shown in Fig. 3. It was found that an increase in the iteration number above 2000 did not improve output. Therefore, this study used 2000 iterations for the three trials.

The ML training can be further optimised by using cross-validation and choosing the number of folds (k) that the data are split into prior to training. For the dataset, three folds were applied due to the number of participants. Double cross-validation was used for the training.

Once the model has been trained to learn features in the spectra that correspond to COVID-19-infected and non-COVID-19 patients, it can be tested. For each of the remaining test cases, a prediction is made using the model. A score is given to each case based on the accuracy of the prediction. After each case has been scored, the individual scores can be used to give a metric for the accuracy of the model. One such metric is the ROCAUC (Bzdok *et al*., 2018; Ruszkiewicz *et al*., 2020). Using this metric, a perfect set of predictions would give a ROCAUC of 100, while a random test dataset with no correlation to the training set would give a ROCAUC of 50.

In operation, a breath biomarker spectrum and a set of clinical data are introduced to the trained software (DeepBreath), where the information flows through several layers of neural networks (Cover and Thomas, 2006). During this information flow, more and more features relevant to positive and negative elements are accumulated. And finally, the flow of information is coming out of the neural network with predominantly positive or negative diagnostic results.

Figure 3. An example of ROCAUCs as a function of iterations.

# The training and testing phase for the full data set required 1 to 3 days of machine time (at the top-of-the range PC workstation). The information was analysed, including 2D spectra and clinical data, with reference diagnostic data. Once the first phase is completed and the software is validated, a single GC-IMS spectra can be processed in a few seconds.

**Two approaches in VOC metabolomics studies**

Current VOC metabolomics studies can be placed into two general categories: those that aim to understand biological processes and those that aim to identify biomarkers. There are different statistical and machine learning strategies that can be used to design multi-metabolite biomarker models and explain how these models can be assessed using ROC curves (Xia *et al*., 2012).

Studies in the first group focus primarily on gaining improved biological understanding through the analysis of metabolite profiles. Data analysis is usually performed using multivariate statistical methods such as principal component analysis (PCA) or partial least squares discriminant analysis (PLS-DA) (Trygg *et al*., 2007).

Performing biomarker selection based on univariate statistical significance is equally inappropriate, as often metabolites that are not significant in isolation can, when combined into a single multivariate model, produce clear and reproducible discrimination.

Perhaps the most important difference to remember is that biomarker models are not intended to help explain biology. Rather, they are designed to discriminate with optimal sensitivity and specificity without regard to biological cause or biological interpretation. In other words, biological understanding is not an absolute prerequisite for biomarker development (Bzdok *et al*., 2018). However, understanding the underlying biological pathways certainly can give some rationale to support an assay or give some direction to develop a treatment.

There are many potentially useful data projection and machine learning methodologies available for this task. Some of the most popular methods that have been applied to metabolomic studies are: linear discriminant analysis, PLS-DA, decision trees (e.g., CART), random forests, artificial neural networks, and support vector machines (Cortes and Vapnik, 1995; Barker and Rayens, 2003; Breiman, 2001; Eriksson *et al*., 2001; Trygg *et al*., 2007).

**PCR teats**

PCR tests and their role in aiding clinical diagnosis or therapeutic monitoring have become increasingly important as advances in the field are made. For one, there are known sampling issues with nasopharyngeal PCR tests. While PCR itself is incredibly robust, it relies on collecting samples of actively amplifying viral genetic material. Though uncommon, it is possible to “miss” swabbing an area with active viral loads, which leads to a false-negative test result. There have been many more issues with operational logistics and product supply chains that have strained testing systems during the lockdown public health crisis. The liquid reagents needed for the PCR test and the nasal swabs are in high demand, thus limiting availability in some locations and causing alterations to planned testing protocols. Finally, although PCR is very reliable, there can be a significant time delay between sampling and when the results are available, depending on the processing capabilities of the test site (Davis *et al*., 2021).