ROYAL SOCIETY OF CHEMISTRY

PAPER

Snapshot of long COVID in young adults: fast screening using electronic noses

José Alberto Ulloa-Rosales, [iD] [a] Sofía Bernal-Silva,[a] Mariana Palau-Vázquez,[a] Sandra Cadena-Mota,[a] Luz Eugenia Alcántara-Quintana,[b] Jan Mitrovics,[g] Fátima Rosales-Arellano,[e] Luis Daniel Becerra-Montes,[e] Gabriela Flores-Rangel,[c] Andreu Comas-García,*[ae]
Boris Mizaikoff [iD] [cd] and Lorena Díaz de León-Martínez [iD] *[cf]

Background: Long COVID (LC) is a multisystemic condition characterized by persistent symptoms following SARS-CoV-2 infection. Although most research focuses on older or hospitalized individuals, young adults are frequently overlooked despite significant effects on their academic, professional, and social functioning. *Methods*: This cross-sectional study evaluated 78 university students (median age 20 years; 56.4% female) with prior COVID-19 infection, classified according to the WHO Delphi consensus definition. Sociodemographic, clinical, and spirometry data were collected, and exhaled breath samples were analyzed using an electronic nose system (e-Nose) under controlled conditions. Chemometric and machine learning techniques—Principal Component Analysis (PCA), Partial Least Squares–Discriminant Analysis (PLS-DA), Canonical Analysis of Principal Coordinates (CAP), and Random Forest (RF)—were applied to identify LC-associated volatile organic compound (VOC) patterns. *Findings*: LC prevalence was 29.5%. Acute-phase fatigue, (odds ratio) (OR = 3.22), dyspnea (OR = 6.09), nausea (OR = 3.57), and vomiting (OR = 11.37) were significantly associated with LC. Post-acute anosmia (OR = 3.65), sleep disturbances (OR = 4.34), and bradycardia were also more frequent among LC cases. All participants exhibited normal spirometry. e-Nose data revealed distinct group-associated VOC patterns and demonstrated promising discriminatory potential between LC and control participants (PCA variance 94.1%; CAP R^2 = 0.95; PLS-DA accuracy 97.4%, Q^2 = 0.534). The RF model achieved an out-of-bag error of 3.42% and receiver operating characteristic curve (ROC) area under the curve (AUC) of 0.966. *Interpretation*: Nearly one-third of young adults experienced LC despite normal pulmonary function, suggesting substantial subclinical and systemic alterations. e-Nose breath analysis represents a promising, non-invasive, and rapid approach for LC screening; while these findings support the feasibility of breath-based screening for LC, further validation in larger and independent cohorts is required before clinical implementation.

Introduction

Long COVID (LC) refers to a spectrum of symptoms that develop in individuals with a history of probable or confirmed SARS-CoV-2 infection.[1] These symptoms typically appear within three months of the initial illness and persist for at least two months, with no alternative diagnosis to explain them. Symptoms may be new-onset after recovery from acute COVID-19 or persist from the initial infection, often fluctuating or relapsing over time. LC can severely impair daily functioning, affecting work, household tasks, social participation, and overall quality of life.[2,3]

LC symptomatology is heterogeneous and can be grouped into physical, psychological, and cognitive clusters, often overlapping and resulting in physical and financial disability.[4] Fatigue is frequently reported as the primary symptom, followed by post-exertional malaise, dyspnea, myalgia, insomnia, and anosmia. Other symptoms may include joint pain, headaches, gastrointestinal disturbances (*e.g.*, diarrhea, dysgeusia, loss of appetite), dizziness, hoarseness, sweating,

[a] Department of Microbiology, School of Medicine, Universidad Autónoma de San Luis Potosí, 78210, San Luis Potosí, Mexico. E-mail: andreu.comas@uaslp.mx

[b] Unidad de Innovación en Diagnóstico Celular y Molecular, Coordinación para la Innovación y la Aplicación de la Ciencia y Tecnología, Universidad Autónoma de San Luis Potosí, 78120 San Luis Potosí, Mexico

[c] Institute of Analytical and Bioanalytical Chemistry, Ulm University, 89081, Ulm, Germany. E-mail: lorena.diaz-de-leon@uni-ulm.de

[d] Hahn Schickard, 89077, Ulm, Germany

[e] School of Medicine, Universidad Cuauhtémoc San Luis Potosí, 78290, San Luis Potosí, Mexico

[f] Breathlabs GmbH, 89077, Ulm, Germany. E-mail: lorena@breathlabs.com

[g] JLM Innovation GmbH, D-72070, Tübingen, Germany

alopecia, and reduced libido. Psychological complaints range from anxiety and depression to post-traumatic stress, while cognitive symptoms, such as short-term memory loss, difficulty concentrating, reduced processing speed, language fluency problems, and executive dysfunction, further interfering with daily and occupational functioning.[5]

According to the World Health Organization, about 6% of individuals infected with COVID-19 develop LC globally.[1] Other studies estimate that 10–20% of patients with mild or moderate COVID-19 develop persistent symptoms, and this prevalence increases to 50–70% among hospitalized patients.[6] Meta-analyses from 2024 report a global prevalence between 33% and 40%, with an estimated overall prevalence of 36% (95% CI)[7] (confidence interval). Despite growing research, the university-student-aged population remains understudied. A 2022 study at Washington State University involving 1338 COVID-19 cases found higher LC risk among females, unvaccinated individuals, and those with preexisting conditions, with 28.6% reporting symptoms lasting beyond six months.[8] Recent findings also show that LC can cause long-term impairments in adolescents (10–18 years) and young adults (18–44 years), especially those in education or early professional stages. Studies suggest that 5–10% of young adults experience persistent symptoms following SARS-CoV-2 infection.[9–11]

The impact of LC on young adults is particularly concerning because it disrupts education, professional development, and social life. Persistent fatigue, cognitive deficits, and emotional disturbances can severely limit their ability to work, study, or engage in daily activities. Emerging evidence indicates that young and middle-aged adults may experience more severe or prolonged neurological and psychological symptoms—such as headaches, depression, anxiety, cognitive difficulties, and sleep or mood disturbances—than older adults, independent of the initial severity of infection. While many recover within two years, a significant proportion continue to experience symptoms for over a year. The multisystemic nature of LC complicates diagnosis, posing major challenges for public health systems and worsening patients' quality of life.[12–14]

Currently, LC diagnosis relies mainly on the duration and persistence of symptoms, typically those lasting beyond 4–12 weeks after acute infection. Since LC is a diagnosis of exclusion, individuals presenting with multiple unexplained symptoms are evaluated only after ruling out other conditions. This complexity highlights the urgent need for precise and timely diagnostic tools that can support clinical decisions and enable individualized treatment. Consequently, researchers have focused on identifying reliable biomarkers and innovative diagnostic technologies to detect LC more effectively. Among these emerging tools, volatile organic compounds (VOCs) have gained significant attention. VOCs are a diverse group of molecules that volatilize at room temperature and can be detected in biological matrices such as urine, blood, sweat, faeces, and exhaled breath.[15,16] They can originate endogenously from normal metabolism, be produced by pathogens, or result from the host's inflammatory response. Transported via the bloodstream, VOCs reach the lungs and are exhaled through the alveoli, providing a non-invasive window into physiological and pathological processes.[16] VOC analysis can reflect both healthy and diseased metabolic states and has been investigated for diagnosing infectious diseases such as COVID-19,[17] influenza,[18] tuberculosis,[19] and Helicobacter pylori infection,[20] as well as chronic conditions like lung cancer,[21] gastric cancer,[22] breast cancer,[23] chronic obstructive pulmonary disease,[24] and asthma.[25] Recent studies suggest that VOC patterns may also serve as indicators for LC, offering a promising avenue for its monitoring.

A particularly effective, low-cost, and non-invasive approach to detect VOCs in complex matrices such as breath is through electronic noses (e-Noses). These artificial olfactory systems mimic the mammalian sense of smell and typically consist of sensor arrays capable of detecting various volatile molecules, along with microprocessors and signal processing hardware. E-Noses generate chemical "fingerprints" that correspond to the composition of gaseous mixtures, detecting disease-specific VOC patterns at trace concentrations, typically in the parts-per-billion by volume (ppbv) range and, for selected compounds and sensor configurations, approaching parts-per-trillion by volume (pptv). It should be noted that while some advanced sensor technologies may approach pptv-level sensitivity under controlled conditions, most commercially available e-Nose systems primarily operate with meaningful signal resolution in the ppbv range rather than at true pptv concentrations. When coupled with machine learning and deep learning algorithms, e-Noses can analyze complex data in real time, providing rapid diagnostic insights within minutes. Their advantages, non-invasiveness, speed, sensitivity, and affordability, make them suitable for point-of-care applications where fast decision-making is critical. Early and accurate detection can improve patient outcomes, optimize treatment strategies, and reduce healthcare burdens.[16]

In this context, the present research aimed to assess the feasibility of using e-Nose technology as a screening method for LC in a university-student-aged population. This approach seeks to contribute to the development of innovative diagnostic tools capable of identifying metabolic signatures of LC, thereby enabling early intervention and improving quality of life among affected young adults.

Experimental

Study population

This observational, cross-sectional, and analytic study was conducted at the School of Medicine of Cuauhtémoc University, San Luis Potosí, Mexico, between January and May 2025, with approval from the Institutional Ethics Committee. The required sample size was calculated using an estimated LC prevalence of 9.7% among university-aged individuals, based on previously published epidemiological

data in similar populations, with a 5% margin of error and 80% statistical power, resulting in 126 participants. This estimated frequency refers to the expected proportion of students with persistent post-COVID symptoms within the target population. Using the formula to calculate the sample size (n) for a population in a finite population[26] (eqn (1)), where, n is the required sample size, *EDFF* stands for the effective design factor; N is the total population size; p the estimated proportion of the characteristic being in the studied population; d as the desired precision; $Z_{1-\alpha/2}$ is the Z-score corresponding to the desired confidence level, α being the tail probability (2 sided Z value). After recruiting 62% of the target, statistically significant differences were observed between students with and without LC, and the main clinical variables achieved over 80% statistical power.

$$n = \frac{[\text{EDFF} \cdot Np(1-p)]}{\left[\left(\frac{d^2}{Z_{1-\frac{\alpha}{2}}^2}\right) \cdot (N-1) + p \cdot (1-p)\right]} \tag{1}$$

Eligible participants were medical students aged 18 years or older who could provide information on prior COVID-19 infection, LC status, and vaccination history. Exclusion criteria included autoimmune or oncologic diseases and pregnancy during or before COVID-19 infection, while elimination criteria involved withdrawal or incomplete data. Participants completed a detailed questionnaire on sociodemographic factors, medical and COVID-19 history, LC symptoms, and vaccination records, followed by signing informed consent. Each participant underwent anthropometric measurements, digital spirometry, and exhaled breath sampling using an e-Nose under controlled environmental conditions to ensure data reliability and minimize bias.

Electronic nose analysis

Exhaled breath analysis was performed using a custom-developed e-Nose system designed specifically for biomedical breath analysis. The device was developed by the authors' research teams in collaboration with engineering partners and has been described in detail in previous publications.[27,28] It is not a commercial off-the-shelf instrument. The system consists of a temperature-controlled aluminum sensing chamber (3 mL), maintained at 45 ± 1 °C to prevent condensation of volatile organic compounds, and is equipped with a buffered end-tidal sampling module to selectively collect alveolar breath. The sensor array comprises 21 digital metal-oxide semiconductor (MOX) sensors, 8 analog MOX sensors targeting complementary VOC sensitivities, and 6 auxiliary sensors for monitoring temperature, humidity, and pressure.

Signal acquisition, preprocessing, and data export were performed using proprietary software developed for this platform. Detailed hardware architecture, sensor specifications, calibration procedures, and system validation have been reported previously. A generalized description of the MOX

sensor materials and operating principles is provided in the SI, without attribution to individual sensor models, to improve transparency while respecting intellectual property constraints.

Sample collection

The sensor boards were consistently maintained at 40 °C throughout the analysis. First, ambient air was measured to establish a baseline for the sensors, ensuring that any subsequent changes detected were due to compounds in the exhaled breath rather than environmental factors. Then, participants exhaled through the mouthpiece of the breath sampler, allowing their breath to be analyzed. After each breath measurement, the system executed an automated purge sequence designed to remove residual VOCs and restore the sensor baseline. This purge comprised evacuation of the sampling pathway and sensing chamber, followed by flushing with filtered ambient air. This procedure was intended to minimize chemical carryover between measurements and stabilize sensor responses and should not be interpreted as a sterilization step. To support hygienic operation and reduce the risk of microbial contamination, a new disposable mouthpiece was used for every participant, and the external and accessible internal surfaces of the sampling pathway were wiped with isopropanol swabs between measurements. This chemical cleaning step was implemented to reduce potential bacterial or viral residues on surfaces. Each one of the participants was required to blow twice; therefore, the data was analyzed by duplicate.

Chemometric and statistical data analysis

All demographic and clinical data were compiled anonymously in Microsoft Excel 2016 and statistically analyzed using SPSS Statistics for Mac (version 30.0) and Epi Info™ 7.2. A two-tailed p-value <0.05 was considered statistically significant. Categorical variables were summarized as frequencies and percentages and compared between groups using contingency tables analyzed with Mantel–Haenszel chi-squared or Fisher's exact tests, as appropriate. Continuous variables were evaluated using parametric (Student's t-test) or non-parametric (Mann–Whitney U test) methods, depending on data distribution, to compare differences in central tendency (mean or median) between participants with LC and those without LC. The null hypothesis tested in each case was that no difference existed between groups. Odds ratios (OR), defined as the ratio of the odds of an outcome occurring in one group compared with another, were calculated to assess associations between acute COVID-19 symptoms and the subsequent development of LC.

Given the high dimensionality and multivariate nature of e-Nose sensor data, a stepwise chemometric strategy was adopted. Analyses were structured to i) explore global variance and data structure, ii) build a supervised classification model aligned with the study objective of screening for LC, and iii) assess robustness and feature

relevance. Partial Least Squares-Discriminant Analysis (PLS-DA) was selected as the primary classification method, while Principal Component Analysis (PCA), Canonical Analysis of Principal Coordinates (CAP), and Random Forest (RF) were used as complementary and confirmatory approaches by using a custom-developed Python script (version 3.13.6) and MetaboAnalyst 6.0 (https://www.metaboanalyst.ca/). PCA was used as an unsupervised exploratory tool to visualize intrinsic data structure, detect outliers, and assess whether breath profiles exhibited group-related trends without imposing class labels. PCA was not used for classification, but rather to evaluate overall variance distribution and data quality. As for PLS-DA it was employed as the primary supervised classification method since it is well suited for correlated sensor data and binary outcomes. PLS-DA identifies latent variables that maximize covariance between predictor variables (sensor responses) and class membership (LC *vs.* healthy), thereby enhancing group separation while reducing dimensionality. Model performance was evaluated using cross-validation, reporting accuracy, R^2 (explained variance), and Q^2 (predictive ability). Receiver operating characteristic (ROC) analysis was used to assess classification behavior across probability thresholds. CAP was applied as a constrained ordination method to confirm group separation observed with PLS-DA, using class membership as a constraining variable to evaluate whether between-group variation exceeded within-group variability in a distance-based framework. Finally, RF analysis was used as a complementary, non-linear classification approach to assess model robustness and identify the most informative sensors contributing to group discrimination. RF outputs were interpreted in conjunction with PLS-DA results rather than as an independent primary classifier.

Results

Characteristics of the study population

Among 78 participants, 23 (29.5%) met the criteria for LC. Over half were female, with a median age of 20 years. The mean body mass index (BMI) was 26.1 ± 5.2, with most participants classified as normal weight, followed by overweight and obesity. Twenty-five percent reported at least one comorbidity, and 33% experienced reinfection. Nearly all (98%) were vaccinated; 68.8% had infection before vaccination, and 54.5% after. Pulmonary parameters – peak expiratory flow (PEF), forced expiratory volume in one second (FEV_1), forced vital capacity (FVC), FEV_1/FVC ratio – were similar between groups, with all participants showing normal lung function (>70% predicted values) (Table 1). Overall, sociodemographic and clinical characteristics were comparable between LC and healthy groups. Although a higher absolute number of females was observed in both groups, the proportion of participants with LC was comparable between females and males, and no statistically significant association between sex and LC status was detected.

Table S1 summarizes the clinical characteristics of the first COVID-19 episode. The most common symptoms were headache, myalgia, fever, nasal discharge, sore throat, fatigue, sneezing, and dysgeusia. Compared with those

Table 1 Sociodemographic characteristics of the study participants. The Mann–Whitney U-test, T-test, chi-square, and Fisher's exact test were performed

Parameters	All	Long COVID	Healthy	p
N, (%)	78	23 (29.5%)	55 (70.5%)	N.C.
Sex, n (%)				
Female[b]	44 (56.4%)	14 (60.9%)	30 (54.4%)	0.608
Male[b]	34 (43.6%)	9 (39.1%)	25 (45.5%)	
Age, median (p25–p75)	20 (20–22)	20 (20–23)	21 (20–22)	0.441
BMI, kg m^{-2}, mean (±SD)	26.1 (±5.2)	26.6 (±5.2)	25.8 (±5.2)	0.812
Low weight, n (%)	2 (2.6%)	0 (0.0%)	2 (3.6%)	1.000
Normal weight, n (%)	37 (47.4%)	9 (34.8%)	29 (52.7%)	0.148
Overweight, n (%)	22 (28.2%)	10 (43.5%)	12 (21.8%)	0.053
Obesity, n (%)	17 (21.8%)	5 (21.7%)	12 (21.9%)	0.994
Comorbidities,[a] n (%)	21 (27.6%)	8 (34.8%)	13 (24.5%)	0.358
Reinfections, n (%)	27 (34.6%)	10 (43.5%)	17 (30.9%)	0.287
COVID-19 vaccine, n (%)	77 (98.7%)	23 (100.0%)	54 (98.2%)	1.000
COVID-19 before vaccination, n (%)	53 (68.8%)	17 (73.9%)	36 (66.7%)	0.520
COVID-19 after vaccination, n (%)	42 (54.5%)	14 (60.9%)	29 (51.9%)	0.467
PEF l s^{-1}, average (±SD)	6.03 (±2.15)	6.58 (±1.8)	6.2 (±2.3)	0.461
FEV_1 l, mean (±SD)	3.3 (±1.15)	3.1 (±0.9)	3.4 (±1.3)	0.141
FVC l, mean (±SD)	3.77 (±1.27)	3.5 (±1.0)	3.9 (±1.4)	0.174
PEF/FVC, mean (±SD)	87.73 (±6.11)	87.1 (±5.6)	88.0 (±6.4)	0.378
FEV_1 predicted, median (p25–p75)	91.8 (77.8–98.0)	86.0 (74.0–105.3)	92.1 (78.6–96-4)	0.727
FVC predicted, median (p25–p75)	89.9 (76.0–100.9)	90.4 (74.0–105.3)	89.4 (75.7–96.5)	0.815
FEV_1/FVC predicted, median (p25–p75)	100.4 (94.1–105.9)	99.2 (91.7–105.6)	101.2 (94.6–106.5)	0.417

[a] Comorbidities: asthma, polycystic ovary syndrome, anxiety, endometriosis, allergic rhinitis, ADHD, dyslipidemia, vasovagal syndrome, esophagitis, and arterial hypertension. [b] Percentages for sex are shown relative to the total number of participants within each study group. Statistical comparisons were performed using contingency-table analysis comparing the prevalence of long COVID between females and males across the entire cohort.

without LC, participants with LC reported higher frequencies of fatigue, dyspnea, nausea, and vomiting during acute infection. Overall, 59% experienced post-COVID sequelae, with a higher proportion in the LC group (82.6% *vs.* 49.1%). Significant risk factors for LC included fatigue (OR 3.22, p = 0.041), dyspnea (OR 6.09, $p < 0.001$), nausea (OR 3.57, p = 0.042), and vomiting (OR 11.37, p = 0.041). Table S2 presents persistent symptoms after the acute phase. The most frequent LC manifestations were fatigue, hair loss, headache, dyspnea, anxiety, reduced lung function, eating disorders, and anosmia. Post-acute symptoms significantly associated with LC included anosmia (OR 3.65, p = 0.030) and sleep disturbances (OR 4.34, p = 0.017); bradycardia was also observed exclusively in the LC group.

Chemometric data analysis

Data obtained from e-Nose sensors, were analysed through chemometric approaches. The average response from the sensors to the exhaled breath of each one of the study groups is shown in Fig. 1, differences in the sensors' response can be clearly observed, indicating a higher overall response in the LC group. For all chemometric analyses, only the sensors with the highest importance index by RF and PLS-DA, were selected for analysis to enhance model's performance (Fig. S2).

The PCA plots (Fig. S1) show separation between the LC and healthy participants with an overall model variance of 94.11%, through 5 principal components (PC1 = 52.67%; PC2 = 16.37%; PC3 = 13.23%; PC4 = 6.41% and PC5 = 5.43%).

The PCA indicates that separation is most evident when considering PC1 and PC2. While there's some overlap, the LC group (blue circles) tends to cluster differently from the healthy group (orange circles). The PCA results suggest that the sensors can distinguish between LC participants and healthy participants.

The CAP analysis revealed clear separation along the first canonical axis (CAP1) between the two groups. The CAP

presented an R^2 value of 0.95. As shown in Fig. S3, the LC group cluster distinctly from those in the healthy group. CAP1 accounted for the full variability explained in the constrained model, with all additional axes contributing negligible variance (nearly zero), suggesting a unidimensional separation primarily driven by group differences. This is supported quantitatively by the CAP1 score distribution, where healthy group predominantly had positive CAP1 values, ranging from 0 to 2.5, and LC group exhibited negative CAP1 scores ranging from −5.2 to 0. This polarity in CAP1 values indicates a discriminant function that effectively differentiates between the clinical states, validating the presence of group-structured variation in the dataset.

The PLS-DA results confirm the separation between LC participants and healthy individuals, revealing key differences in the exhaled breath patterns. The PLSDA (Fig. 2) presents an overall model variance of 97.5% through 5 PLS components (PLS1 = 67.5%; PLS2 = 8.7%; PLS3 = 13.3%; PLS4 = 5.9% and PLS5 = 2.1%). PLS1 exhibits a wide range of loadings (−0.6 to 7.3), suggesting that specific sensors strongly differentiate LC from healthy, while PLS2 shows more subtle variations (−2.8 to 5.2). The cross-validation results (Fig. S4a) confirm that model performance improves through 5 PLS components, reaching 0.974 accuracy with five components, alongside increasing R^2 (0.606) and Q^2 (0.534) values.

This indicates that the model effectively captures the biological differences between individuals with LC and healthy individuals, with higher component numbers enhancing classification reliability.

The permutation test further validates the model's significance, yielding a p-value of 0.96 (0/100), indicating that the observed separation between the LC and healthy groups is highly unlikely due to random chance (Fig. S4b). The frequency distribution of permuted statistics confirms that the model's performance exceeds random expectations.
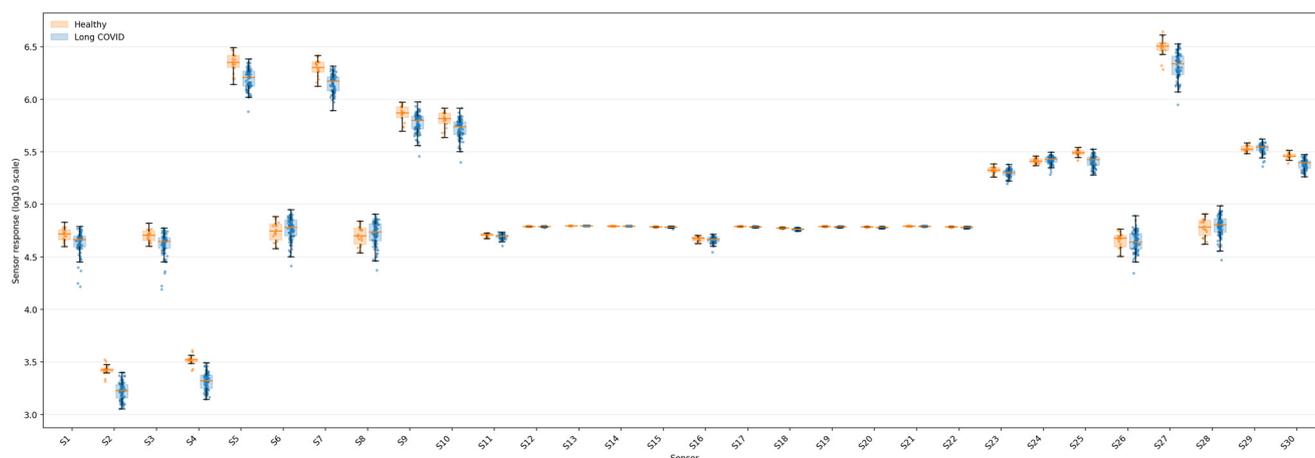
Fig. 1 Distribution of e-Nose sensor responses for healthy and long COVID participants. Boxplots show median and interquartile range (IQR), with whiskers indicating dispersion; individual measurements are overlaid as jittered points. Responses are shown on a log 10 scale to accommodate the large dynamic range across sensors.
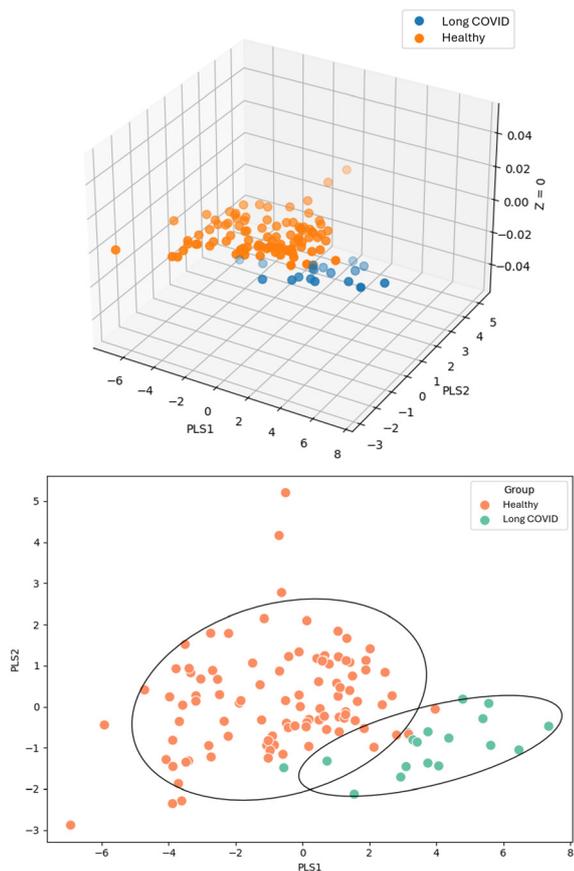
Fig. 2 Partial Least Squares-Discriminant Analysis (PLS-DA) score plot of exhaled breath sensor responses. Each point represents one participant, with long COVID shown in blue and green and healthy participants in orange. PLS-DA is a supervised multivariate method that highlights group-associated VOC patterns; partial overlap reflects shared metabolic features and inter-individual variability.

Together, these findings suggest that PLS-DA successfully identifies discriminative patterns in participants with LC, through five components providing the optimal balance of accuracy and predictive power.

Although group separation was evident across multiple chemometric approaches, partial overlap between LC and healthy participants was observed, with a subset of samples occupying intermediate regions of the multivariate space. This overlap likely reflects biological heterogeneity within LC as well as shared metabolic features between groups.

Regarding the RF classifier, it demonstrates strong classification performance. With 500 trees and seven predictors, the model achieves an out-of-bag error (OOB) of 3.42%, indicating high overall accuracy. The RF indicates a higher misclassification rate for LC individuals with a 16.7% classification error than healthy classification rate of 1%. This suggests that while the model excels at identifying healthy, it is slightly less precise in confirming LC, possibly due to overlapping features or a smaller sample size in the LC group. The error rate stabilizes after ~100–200 trees, confirming that the ensemble is well-optimized.

The variable importance plot for the RF analysis (Fig. S2) highlights the key biomarkers and clinical measures that drive the classification. Sensors S4 and S2 were the top three contributors to model accuracy, suggesting these markers are strongly associated with LC group. Spirometry metrics, such as FEV1/FVC predicted, FVC, and PEF, also play significant roles, aligning with the known respiratory and inflammatory effects of LC. The low error rate for LC classification, combined with the biological relevance of the top predictors, supports the RF model's utility in clinical or research settings for identifying LC cases with high confidence. Further refinement, such as balancing the dataset or incorporating additional healthy samples, could improve specificity for healthy individuals.

The receiver operating characteristic curve (ROC) demonstrates an excellent performance in distinguishing LC from healthy participants, with an AUC of 0.966 (95% CI = 0.891–1) (Fig. 3). This high AUC indicates strong sensitivity (true positive rate) and specificity (low false positive rate), meaning the model reliably identifies true LC cases while minimizing misclassification of healthy individuals.

Using a fixed probability threshold of 0.5 for class assignment (LC = positive class), sensitivity and specificity were calculated directly from the cross-validated confusion matrix. At this threshold, the model achieved a sensitivity of 98.9% (92/93), indicating a high true positive rate for identifying LC cases, while specificity was 65.4% (17/26), reflecting a moderate false positive rate among healthy participants. This operating point corresponds to a Youden index of 0.64, highlighting a classification balance that prioritizes sensitivity over specificity. The corresponding ROC coordinate was located at a false positive rate of approximately 0.35 and a true positive rate of approximately 0.99. The model maintains high accuracy even with fewer features, suggesting clinical utility for streamlined diagnostics.
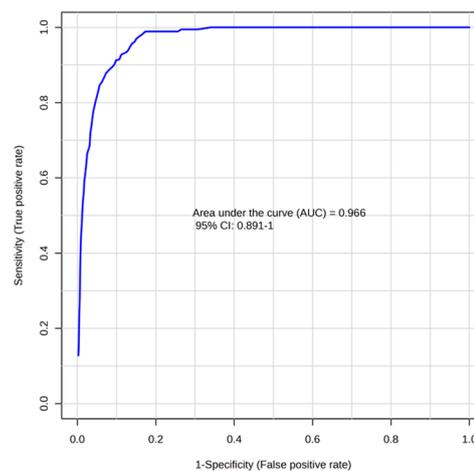
Fig. 3 Receiver operating characteristic curve (ROC) from groups of study.

Discussion

This study provides new evidence on the prevalence, clinical characteristics, and breath-based diagnostic potential of LC in a university-student-aged population—an age group often overlooked in post-acute COVID-19 research. Nearly one-third (29.5%) of participants met the WHO Delphi definition for LC, a proportion slightly higher than that reported in other non-hospitalized young adult cohorts (10–30%). These findings highlight that LC is not confined to older or high-risk individuals and can affect otherwise healthy young adults, underscoring the need for early detection and targeted interventions in this demographic.

The impact of LC on young adults extends beyond health, affecting academic and occupational performance. Our results align with those from the National Autonomous University of Mexico (UNAM), where over 30% of students reported fatigue, non-restorative sleep, and cognitive impairment after COVID-19.[29] Persistent symptoms such as chronic fatigue, "brain fog", and sleep disturbances can severely impair concentration, memory, and learning, leading to reduced academic performance and delayed progress. Additionally, the abrupt transition between in-person and remote education during the pandemic likely exacerbated these cognitive and emotional challenges, contributing to reduced motivation and engagement. From an occupational standpoint, LC has been linked to decreased work capacity in young adults. Studies from Canada and other countries report that individuals with persistent post-COVID symptoms experience higher unemployment rates, reduced working hours, and diminished participation in the work force.[30,31] The lack of official recognition of LC as a distinct medical condition further complicates access to workplace accommodations, rehabilitation programs, and social support. Consequently, many affected individuals, especially students and young professionals, struggle to regain normal functioning and maintain productivity.

In our study, acute-phase fatigue, dyspnea, nausea, and vomiting were significantly associated with LC, with dyspnea showing the strongest link (OR = 6.09). These findings are consistent with prior reports suggesting that respiratory distress during the acute infection predicts a higher likelihood of persistent symptoms. Interestingly, spirometry results were normal across all participants, yet post-COVID symptoms such as anosmia, sleep disturbances, and bradycardia were significantly more frequent among LC cases. This suggests that persistent symptoms in young adults may reflect functional, autonomic, or neurologic dysregulation not detectable by standard pulmonary testing. The normal spirometry findings emphasize that conventional lung function tests may fail to capture subclinical dysfunctions associated with LC. These could involve microvascular, inflammatory, or neural mechanisms not reflected in standard pulmonary parameters. Which reinforces the need for alternative screening approaches

capable of detecting metabolic or functional changes in non-hospitalized populations.

Delayed or missed LC diagnoses carry major social and economic consequences. Without formal recognition, patients may be denied medical leave, insurance coverage, and rehabilitative care, leading to prolonged disability and emotional distress. Misunderstanding and stigmatization—especially in academic or workplace settings—can intensify social isolation and psychological strain. From a healthcare perspective, the absence of standardized diagnostic criteria results in repeated consultations, redundant testing, and ineffective symptomatic treatments, increasing costs for both patients and healthcare systems. Economically, LC-related absenteeism and reduced productivity impose a growing burden on national economies and individual livelihoods.

Our chemometric analyses revealed a distinct and reproducible separation between LC and non-LC participants, demonstrating that VOC patterns in exhaled breath can effectively distinguish affected individuals even in a young, healthy cohort. These findings suggest that LC produces metabolic or inflammatory alterations detectable through breath analysis, supporting the feasibility of e-Nose technology as a diagnostic tool.

Previous studies have reported similar applications. To this extent, Nidheesh V. R. et al. (2022) reported that VOC profiles could differentiate post-COVID patients from controls with accuracies exceeding 95%.[32] Our findings align with theirs but extend the evidence to younger adults with normal lung function, highlighting that e-Nose devices can achieve comparable accuracy in populations without significant comorbidities or pulmonary sequelae. This is crucial because early, non-invasive breath-based screening could be implemented in universities or workplaces to facilitate rapid identification, monitoring, and referral of LC cases. Zamora-Mendoza et al. (2023) also demonstrated that e-Nose technology could discriminate between post-COVID patients with and without lung damage using chemometric models such as PCA, PLS-DA, and RF.[33] Their results achieved classification accuracies between 80–90%. In contrast, our analysis in a university-student-aged cohort achieved even stronger performance: PLS-DA accuracy of 97.4% (R^2 = 0.606, Q^2 = 0.534), CAP R^2 of 0.95, and an RF classifier with an out-of-bag error of only 3.42% and ROC AUC of 0.966. These outcomes suggest that VOC breath signatures may be more distinct—and thus easier to classify—among young, non-hospitalized individuals, possibly due to fewer comorbidities and less physiological variability. The reduced heterogeneity of our sample likely enhanced the signal-to-noise ratio, contributing to the high accuracy of machine learning models.[33]

Despite the promising findings, this study has limitations. The relatively small sample size restricts the generalizability of results and limits the capacity to explore subgroup differences. Its cross-sectional design prevents causal inference, and reliance on self-reported symptoms introduces potential recall bias. Nonetheless, the integration of

objective, sensor-based data strengthens confidence in the observed distinctions between LC and healthy groups. Future studies should validate these findings within larger, multi-center cohorts and longitudinally track VOC signatures to assess recovery or chronicity over time. Investigating how VOC profiles correlate with biological markers of inflammation, autonomic dysfunction, or mitochondrial impairment could further elucidate LC pathophysiology. Additionally, integrating e-Nose analysis with clinical parameters obtained *via* spirometry and correlating with heart rate variability could enhance the diagnostic precision and enable early identification of subclinical cases. Importantly, despite the high classification metrics obtained, overlap between LC and healthy participants was observed in multivariate analyses, indicating the presence of ambiguous breath profiles. This overlap may reflect inter-individual variability, differing LC phenotypes, or transitional metabolic states. As such, the present findings should be interpreted as proof-of-concept evidence supporting the feasibility of e-Nose-based LC screening rather than definitive diagnostic discrimination.

Conclusion

This study represents a significant advancement on investigating multisystemic and multifactorial diseases such as LC by utilizing biomarkers present in exhaled breath analyzed *via* advanced olfactory technologies such as the electronic nose. This non-invasive methodology demonstrates promising potential to differentiate individuals with LC from healthy controls, even in underrepresented populations such as young adults. Despite their generally lower risk of severe complications, this age group can experience a wide range of persistent symptoms that impact their physical, mental, and functional health.

This work also serves as a motivation for health authorities to formally recognize LC within the diagnostic frameworks of public and private health systems such that early identification becomes a priority, enabling timely access to appropriate treatments and improving the quality of life of those affected.

Finally, this study reinforces the potential of emerging technologies, as e-Noses as complementary tools for the diagnosis and monitoring of multisystemic and multifactorial diseases, whereby the early differentiation from other conditions is crucial for guiding effective therapeutic interventions and optimizing patient outcomes.

Author contributions

JAUR: sampling, epidemiologic and questionnaires data assessment, writing, editing. SBS: conceptualization, sampling, ethical clearance procedures, data handling; comprehensive review and editing of manuscript. MPV: sampling, epidemiologic and questionnaires data assessment. SCM: sampling, epidemiologic and questionnaires data assessment. LEAQ: conceptualization, writing, editing and data handling. JM: device design and assembly, conceptualization. FRA: sampling, epidemiologic and questionnaires data assessment. LDBM: sampling, epidemiologic and questionnaires data assessment. GFR: laboratory device testing, writing, editing, data handling. ACG: conceptualization, sampling, ethical clearance procedures, epidemiologic and questionnaires data assessment. BM: conceptualization, writing, editing, funding acquisition and data handling. LDLM: conceptualization, chemometric data analysis, laboratory device testing, writing, editing.

Conflicts of interest

ACG has been speaker for Astra Zeneca; LDLM, serves as Chief Scientific Officer of Breathlabs GmbH. Breathlabs GmbH declares no conflict of interest. The remaining authors declare no conflict of interest.

Data availability

Data for this article are available at OPARU (Open Access Repositorium der Universität Ulm und Technischen Hochschule Ulm) through this link https://doi.org/10.18725/OPARU-58411.

Supplementary information (SI): data on the characteristics of past COVID-19 infections among the study participants are presented in Table S1. Information regarding Long COVID–related symptoms reported by the participants is provided in Table S2. Potential sources of bias, along with detailed information on the metal oxide (MOX) sensors used in the electronic nose, are described in the SI. In addition, a representative schematic of the electronic nose employed in this study is shown in Fig. S5. The results of the additional chemometric analyses are presented in Fig. S1–S4. See DOI: https://doi.org/10.1039/d5sd00204d.

Acknowledgements

Notes and references

1 W. H. O. (WHO), Post COVID-19 condition (long COVID), https://www.who.int/news-room/fact-sheets/detail/post-covid-19-condition-(long-covid), (accessed 12.08.2025, 2025).

2 J. B. Soriano, S. Murthy, J. C. Marshall, P. Relan and J. V. Diaz, *Lancet Infect. Dis.*, 2022, **22**, e102–e107, DOI: 10.1016/S1473-3099(21)00703-9.

3 T. Greenhalgh, M. Sivan, A. Perlowski and J. Ž. Nikolich, *Lancet*, 2024, **404**, 707–724, DOI: 10.1016/S0140-6736(24)01136-X.

4 C. Cha and G. Baek, *J. Clin. Nurs.*, 2024, **33**, 11–28, DOI: 10.1111/jocn.16150.

5 E. Aretouli, M. Malik, C. Widmann, A. M. Parker, E. S. Oh and T. D. Vannorsdall, *BMJ*, 2025, **390**, e081349, DOI: 10.1136/bmj-2024-081349.

6 H. E. Davis, L. McCorkell, J. M. Vogel and E. J. Topol, *Nat. Rev. Microbiol.*, 2023, **21**, 133–146, DOI: 10.1038/s41579-023-00896-0.

7 Y. Hou, T. Gu, Z. Ni, X. Shi, M. L. Ranney and B. Mukherjee, *medRxiv*, 2025, preprint, DOI: 10.1101/2025.01.01.24319384.

8 M. Landry, S. Bornstein, N. Nagaraj, G. A. Sardon, Jr., A. Castel, A. Vyas, K. McDonnell, M. Agneshwar, A. Wilkinson and L. Goldman, *Emerging Infect. Dis.*, 2023, **29**, 519–527, DOI: 10.3201/eid2903.221522.

9 S. Lopez-Leon, T. Wegman-Ostrosky, N. C. Ayuzo del Valle, C. Perelman, R. Sepulveda, P. A. Rebolledo, A. Cuapio and S. Villapol, *Sci. Rep.*, 2022, **12**, 9950, DOI: 10.1038/s41598-022-13495-5.

10 S. Ekström, I. Mogensen, M. Ödling, A. Georgelis, A.-S. Merritt, S. Björkander, E. Melén, A. Bergström and I. Kull, *BMC Public Health*, 2025, **25**, 1330, DOI: 10.1186/s12889-025-22522-9.

11 National Institute for Public Health and the EnvironmentMinistry of Health, Welfare and Sport, (RIVM). More than 3% of adults and 5% of young people have persistent symptoms after COVID-19, https://www.rivm.nl/en/news/more-than-3-of-adults-and-5-of-young-people-have-persistent-symptoms-after-covid-19, (accessed 01.08.2025, 2025).

12 N. A. Choudhury, S. Mukherjee, T. Singer, A. Venkatesh, G. S. Perez Giraldo, M. Jimenez, J. Miller, M. Lopez, B. A. Hanson, A. P. Bawa, A. Batra, E. M. Liotta and I. J. Koralnik, *Ann. Neurol.*, 2025, **97**, 369–383, DOI: 10.1002/ana.27128.

13 I. Mogensen, S. Ekström, J. Hallberg, A. Georgelis, E. Melén, A. Bergström and I. Kull, *Sci. Rep.*, 2023, **13**, 11300, DOI: 10.1038/s41598-023-38315-2.

14 O. M. Chemych, K. Nehreba, A. Yemchura, Y. Kubrak, A. M. Loboda, N. V. Klymenko, O. K. Melekhovets, O. H. Vasylieva and K. O. Smiian, *Likars'ka Sprava*, 2025, **1**, 20–31, DOI: 10.31640/LS-2025-1-03.

15 M. Bajo-Fernández, É. A. Souza-Silva, C. Barbas, M. F. Rey-Stolle and A. García, *Front. Mol. Biosci.*, 2024, **10**, 1295955, DOI: 10.3389/fmolb.2023.1295955.

16 L. Díaz de León-Martínez, G. Flores-Rangel, L. E. Alcántara-Quintana and B. Mizaikoff, *ACS Sens.*, 2024, **10**, 1564–1578, DOI: 10.1021/acssensors.4c02280.

17 J. Li, A. Hannon, G. Yu, L. A. Idziak, A. Sahasrabhojanee, P. Govindarajan, Y. A. Maldonado, K. Ngo, J. P. Abdou and N. Mai, *ACS Sens.*, 2023, **8**, 2309–2318, DOI: 10.1021/acssensors.3c00367.

18 E. Borras, M. M. McCartney, C. H. Thompson, R. J. Meagher, N. J. Kenyon, M. Schivo and C. E. Davis, *J. Breath Res.*, 2021, **15**, 046004, DOI: 10.1088/1752-7163/ac1a61.

19 A. M. I. Saktiawati, K. Triyana, S. D. Wahyuningtias, B. Dwihardiani, T. Julian, S. N. Hidayat, R. A. Ahmad, A. Probandari and Y. Mahendradhata, *PLoS One*, 2021, **16**, e0249689, DOI: 10.1371/journal.pone.0249689.

20 X. Liu, Q. Chen, S. Xu, J. Wu, J. Zhao, Z. He, A. Pan and J. Wu, *Adv. Sci.*, 2024, **11**, 2401695, DOI: 10.1002/advs.202401695.

21 V. Binson, M. Subramoniam and L. Mathew, *Clin. Chim. Acta*, 2021, **523**, 231–238, DOI: 10.1016/j.cca.2021.10.005.

22 J. Glöckler, J. Mitrovics, S. Beeken, M. Leja, T. Weleareugay, L. Österlund, H. Haick, G. Shani, C. Di Natale and R. Murillo, *ACS Sens.*, 2025, **10**, 427–438, DOI: 10.1021/acssensors.4c02725.

23 L. Díaz de León-Martínez, M. Rodríguez-Aguilar, P. Gorocica-Rosete, C. A. Domínguez-Reyes, V. Martínez-Bustos, J. A. Tenorio-Torres, O. Ornelas-Rebolledo, J. A. Cruz-Ramos, B. Balderas-Segura and R. Flores-Ramírez, *J. Breath Res.*, 2020, **14**, 046009, DOI: 10.1088/1752-7163/aba83f.

24 V. A. Binson, M. Subramomiam and L. Mathew, *Expert Rev. Mol. Diagn.*, 2021, **21**, 1223–1233, DOI: 10.1080/14737159.2021.1971079.

25 L. Savito, S. Scarlata, A. Bikov, P. Carratù, G. E. Carpagnano and S. Dragonieri, *World J. Clin. Cases*, 2023, **11**, 4996, DOI: 10.12998/wjcc.v11.i21.4996.

26 W. G. Cochran, *Sampling techniques*, John Wiley & Sons, 1977.

27 C. Jaeschke, J. Glöckler, O. El Azizi, O. Gonzalez, M. Padilla, J. Mitrovics and B. Mizaikoff, *ACS Sens.*, 2019, **4**, 2277–2281, DOI: 10.1021/acssensors.9b01244.

28 J. Glöckler, C. Jaeschke, M. Padilla, J. Mitrovics and B. Mizaikoff, *ACS Meas. Sci. Au*, 2024, **4**, 184–187, DOI: 10.1021/acsmeasuresciau.3c00053.

29 Y. L. Jimenez, M. E. C. González and K. I. Olivé, *Medicina e Investigación Universidad Autónoma del Estado de México*, 2024, **12**, 23–27, DOI: 10.36677/medicinainvestigacion.v12i1.21509.

30 F. Jaber, M.-A. Hoang, D. E. Feldman, S. Saunders and B. Mazer, *WORK*, 2025, **80**, 1854–1860, DOI: 10.1177/10519815241300409.

31 C. Peters, M. Dulon, C. Westermann, A. Kozak and A. Nienhaus, *Int. J. Environ. Res. Public Health*, 2022, **19**, 6983, DOI: 10.3390/ijerph19126983.

32 V. R. Nidheesh, A. K. Mohapatra, V. K. Unnikrishnan, J. Lukose, V. B. Kartha and S. Chidangil, *Anal. Bioanal. Chem.*, 2022, **414**, 3617–3624, DOI: 10.1007/s00216-022-03990-z.

33 B. N. Zamora-Mendoza, H. Sandoval-Flores, M. Rodríguez-Aguilar, C. Jiménez-González, L. E. Alcántara-Quintana, A. A. Berumen- Rodríguez and R. Flores-Ramírez, *Talanta*, 2023, **256**, 124299, DOI: 10.1016/j.talanta.2023.124299.