Journal of Breath Research

PAPER

Detection of *Clostridioides difficile* infection by assessment of exhaled breath volatile organic compounds

RECEIVED
11 October 2023

REVISED
22 February 2024

ACCEPTED FOR PUBLICATION
19 March 2024

PUBLISHED
28 March 2024

Teny M John[1,2,*], Nabin K Shrestha[2], Leen Hasan[3], Kirk Pappan[4], Owen Birch[4], David Grove[5], Billy Boyle[4], Max Allsworth[4], Priyanka Shrestha[6], Gary W Procop[7] and Raed A Dweik[5]

[1] Department of Infectious Diseases, The University of Texas MD Anderson Cancer Center, Houston, TX, United States of America
[2] Department of Infectious Diseases, Respiratory Institute, Cleveland Clinic, Cleveland, OH, United States of America
[3] Department of Internal Medicine, University of Connecticut, Farmington, CT, United States of America
[4] Owlstone Medical Ltd, Cambridge, United Kingdom
[5] Department of Pulmonary Medicine and Critical Care, Respiratory Institute, Cleveland Clinic, Cleveland, OH, United States of America
[6] Department of Computer Science, Stanford University, Stanford, CA, United States of America
[7] American Board of Pathology, Farmington, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: tmjohn1@mdanderson.org

Keywords: *Clostridioides difficile* infection, exhaled breath, volatile organic compounds

Supplementary material for this article is available online

Abstract

Clostridioides difficile infection (CDI) is the leading cause of hospital-acquired infective diarrhea. Current methods for diagnosing CDI have limitations; enzyme immunoassays for toxin have low sensitivity and *Clostridioides difficile* polymerase chain reaction cannot differentiate infection from colonization. An ideal diagnostic test that incorporates microbial factors, host factors, and host-microbe interaction might characterize true infection. Assessing volatile organic compounds (VOCs) in exhaled breath may be a useful test for identifying CDI. To identify a wide selection of VOCs in exhaled breath, we used thermal desorption-gas chromatography-mass spectrometry to study breath samples from 17 patients with CDI. Age- and sex-matched patients with diarrhea and negative *C.difficile* testing (no CDI) were used as controls. Of the 65 VOCs tested, 9 were used to build a quadratic discriminant model that showed a final cross-validated accuracy of 74%, a sensitivity of 71%, a specificity of 76%, and a receiver operating characteristic area under the curve of 0.72. If these findings are proven by larger studies, breath VOC analysis may be a helpful adjunctive diagnostic test for CDI.

1. Introduction

Clostridioides difficile infection (CDI) is the leading cause of hospital-acquired infective diarrhea. In the United States, the national burden of infection was estimated to be around 462 000 cases in 2017, with patients older than 65 having an in-hospital mortality rate of 8.4% [1]. A major risk factor for CDI is current or recent treatment with antibiotics, as one of their off-target effects is the disruption of the healthy gut microbiota, which allows *C. difficile* to dominate vacant ecological niches and proliferate. This population overgrowth results in toxin production that induces CDI-associated symptoms. Although diarrhea is its main symptom, CDI can also result in life-threatening complications in certain high-risk patients [2]. The key to preventing severe health outcomes in patients with CDI is early diagnosis. However, the 2 most common methods used to diagnose CDI, enzyme immunoassay and polymerase chain reaction (PCR), are suboptimal by themselves; enzyme immunoassay has variable sensitivity with sensitivities as low as 45% reported [1, 2], and PCR cannot differentiate between infection and colonization, and both require stool samples for testing [3]. An ideal diagnostic test that incorporates microbial factors, host factors, and host-microbe interaction might characterize true infection. There is an unmet need for a quick, on-demand, bedside diagnostic test to diagnose CDI.

Compelling evidence suggests that patients with CDI have distinct volatile organic compound (VOC) profiles. Many clinicians who care for patients with CDI recognize a distinct odor in these patients' stool

samples, and recent in vitro studies have identified specific VOCs associated with *C. difficile* [4]. However, whether these VOCs are present in breath is unknown. VOCs produced in the gut can appear in exhaled breath, having been carried via the blood to the lungs. As such, there is a compelling case for identifying a breath VOC profile that could be used to diagnose CDI. One small study that employed selected ion flow tube mass spectrometry has already demonstrated that a selected group of breath VOCs can be used to differentiate healthy and CDI patient breath samples [5].

In the present study, we aimed to identify a novel, more comprehensive selection of exhaled breath VOCs that can be used to discriminate between patients with and without CDI and to obtain biological insight into risk factors associated with CDI.

2. Materials and methods

2.1. Study setting and design

This prospective, case-control study enrolled patients with diarrhea who were hospitalized at a multispecialty academic medical center in the USA. The medical center's Institutional Review Board approved the study.

2.2. Screening, inclusion, and exclusion criteria

Patients aged 18 years or older, who were admitted with or who developed diarrhea during hospitalization and were tested for *C. difficile* by PCR were considered for inclusion in the study. Patients with a positive *C. difficile* PCR test and symptoms compatible with CDI who provided written consent were included in the study. For each enrolled patient with CDI, the single best age- and sex-matched participant with a negative *C. difficile* PCR test the same day was enrolled as a control. We excluded patients without a clinical illness compatible with CDI, those who refused or were unable to give informed consent (e.g. due to intubation, encephalopathy, delirium, or pharmacologic sedation), those requiring supplemental oxygen, and those with CDI in the previous four weeks.

2.3. Clinical variables

Demographic and clinical information, including age, sex, selected comorbid conditions (diabetes mellitus, chronic kidney disease, chronic liver disease, inflammatory bowel disease, malignancy, and transplantation) were collected from participants' medical records.

2.4. Breath sample collection

One breath sample was collected from each participant using a breath collection device (ReCIVA Breath Sampler, Owlstone Medical Ltd, Cambridge, UK). Exhaled breath was collected onto a Breath Biopsy Cartridge, which consisted of Tenax TA/Carbograph 5TD sorbent tubes (Markes International, UK). The breath collection device monitored participants' breathing patterns in real-time using CO_2 and pressure sensors and the system dynamically determined gates using the real-time pressure levels. The ReCIVA breath collection device was configured as outlined in supplementary table 1, such that it collected a broad fraction of exhaled breath from both the upper and lower airways during normal tidal breathing but did not sample while subjects were inhaling. Each pump pulled pressure-gated exhaled breath through 2 sorbent tubes, with 1473 ml collected in each tube. The breath in each pair of tubes was later combined into a single sample to increase the mass of VOCs injected into the thermal desorption-gas chromatography-mass spectrometer (TD-GC-MS) for analysis (and, therefore, to increase signal at the detector) by desorbing both tubes into the thermal desorber cold trap.

2.5. Breath sample analysis

Samples were received by the Breath Biopsy Laboratory (Owlstone Medical Ltd, Cambridge, UK) and manually inspected to identify potential issues, such as loose sorbent tube end caps. Samples were dry purged with helium on a TD100 thermal desorber (Markes International) to remove excess moisture and subsequently analyzed by TD-GC-MS. Tube desorption was performed using a TD100-xr thermal desorption autosampler (Markes International). Each sample consisted of two sorbent tubes, both of which were desorbed into the thermal desorber cold trap for a single analysis. Samples were then transferred onto a Quadrex 007–624 column (30 m $\times$ 0.32 mm $\times$ 3.00 μm) using splitless injection. Chromatographic separation was achieved via a programmed method (40 °C–250 °C in 86.5 min, helium carrier gas flow 3.0 ml min^{-1}) on a 7890B GC oven (Agilent Technologies) and mass spectral data acquired using an electron ionization time-of-flight BenchTOF high-definition (HD) mass spectrometer (Markes International). A cleaning method was run in between each sample to prevent carry-over.

A quality control sample (sorbent tube spiked with a known mixture of chemicals) was run in between every four patient breath samples to monitor the stability of instrumentation. A blank tube was run every four samples and after every quality control sample to monitor background. Patient samples were scaled to the quality control samples run in the same sequence.

2.6. Extraction of molecular features from breath samples

2.6.1. Sample curation

Breath samples were curated using an in-house automated filtering system to quantify the likelihood of a sample being of sufficient quality. Briefly, the filtering system uses a combination of check-in sample

tags (loose sorbent tube end caps, samples collected using the wrong breath collection method, etc) and ReCIVA breath sampler leak metrics (e.g. volume with less than 90% of the expected volume) to identify samples with potential collection issues for further review. Post-processing, column resolution checks were included to ensure that the ability to extract high-quality molecular feature (MFs) was maintained. A sample failing any of these checks was excluded from the analysis. In this study, 6 samples were excluded due to collection leaks or low chromatographic resolution.

2.6.2. Retention time alignment

Retention time shifts due to column events can lead to chromatograms not being aligned. To correct for this, we used the retention times of known quality control (QC) compounds from QC samples. For each QC sample, a piece-wise linear function was constructed by comparing QC retention times in the sample to the compound-specific medians across all QC samples. This piece-wise linear function was then applied to the retention time axis of breath samples that were analyzed immediately after a given specific QC sample.

2.6.3. Feature extraction

Untargeted feature extraction was performed for samples that passed all curation checks. TD-GC-MS chromatograms were converted into MF lists for statistical analysis. The process of extracting features involved identifying a set of characteristics indicative of a compound (including chromatographic retention time, spectral peaks defined by the mass to charge (m/z) ratios of the ions present, and their intensities) and aligning them across all samples to ensure that the same feature was consistently identified and extracted when present in any sample from the dataset. Prior to feature extraction and alignment, a retention time correction step and file conversion step were necessary. Data were acquired in the Markes proprietary format (.lsc) using the BenchTOF-HD platform. Data were converted from the Markes proprietary format to the Agilent Chemstation proprietary format (.d) using TOF-DS version 3.1 (Markes International) and then converted to the Agilent MassHunter proprietary format (.d) using GC/MS Translator version B.07oo (Agilent Technologies). All data were imported into the ProFinder version B.10.00 (Agilent Technologies) to perform feature extraction. The batch recursive molecular feature extraction (RFE) method was used to perform deconvolution. Deconvolution resulted in a list of MFs, each of which consisted of mass spectral ions with similar chromatographic characteristics. Key RFE method parameters were retention time of 5–60 min, retention time extraction window of $\pm$0.3 min, extraction window of m/z $\pm$ 100 ppm (low-resolution mass spectral data), and minimum absolute mass spectral intensity of 300 counts for the largest intensity ion for each feature.

The MFs produced by deconvolution were manually inspected to check extraction/integration consistency across the entire dataset. MFs that demonstrated extraction/integration consistency in at least 30% of all samples during the manual feature inspection by a chemoinformatic scientist were passed for feature processing, curation review, and then data analysis. The list of features was exported from ProFinder as a .csv file. When a feature was not detected or identified with satisfactory confidence in a given sample, the corresponding entry in the feature table was marked as 'data missing.' For compiling the final features table, features were discarded if they did not appear in more than 80% of the samples. For the remaining features, missing values were imputed as described previously [6].

For each MF of interest, deconvolved mass spectra from samples with the greatest peak areas were matched to the NIST 17 library mass spectral library, and the match was inspected manually to ensure consistency across samples using MassHunter Quantitative Analysis software vB.09.00 (Agilent Technologies). A tentative identity was assigned by comparing the spectral characteristics of the MF to those in the NIST library. In addition to a tentative identity, this analysis also returns a score that reflects the percentage match with the library spectrum. All tentative identities had a match score of $>$80%. Scores in this range typically represent a good match, but even MFs with scores of $>$90% can be identified incorrectly owing to the non-specificity of a given spectral pattern. Therefore, the identities are best thought of as guides to the general molecular formula and presence of chemical functional groups (alcohols, aldehydes, unsaturation, etc). These tentative identities can be confirmed by comparisons with true standards. MFs that did not achieve a NIST match score of $>$80% were not assigned a tentative identity and were labeled as MFx, where x is a non-zero whole number.

2.6.4. Feature preprocessing

Small deviations in peak areas were introduced by retention time alignment. These deviations were corrected using the scaling factors derived from the piece-wise linear functions. Because the breath samples were analyzed over a long period of time, it was necessary to correct the peak areas for instrument variation over analytical sequences. The impact of the instrument variation was modeled by the equation $Y_{i,t} = S_{i,t}X_i$, where $Y_{i,t}$ is the peak area of compound i in analytical sequence t, $S_{i,t}$ is the instrument sensitivity to compound i over sequence t, and X_i is the true concentration of compound i in the breath samples. Assuming that the effect of instrument variation on compound sensitivity does not depend on compound identity, $Y_{i,t} = K_t\gamma_i$ and $\gamma_i = \delta_i X_i$, where γ_i is the

theoretical peak area for compound *i* when there is no instrument variation (the target output after scaling), K_t is the sequence-dependent variation in sensitivity, and δ_i is the compound-specific sensitivity constant. Therefore, to remove instrument variation in the peak areas attained, the corrected peak area for compound *m* from sequence *t* was computed as $\gamma_m = Y_{m,t}/K_t$.

2.7. Statistical analysis

2.7.1. Data structure

We used principal component analysis to identify outliers and trends in the data. In checking for patterns or structure in the data, the covariates of interest were sex, age, body mass index (BMI), processing time, smoking status, and patient group.

2.7.2. Univariable analysis

The Mann–Whitney U test [7] was used to determine whether MF abundance differed significantly between participants with and without CDI. *P*-value adjustment using max-T/min-P permutation testing [8] was performed to correct for multiple testing and reduce the likelihood of false positives, however no compounds remained significant after adjustment.. The Wilcoxon signed-rank test [9] was performed to test if MF abundance differed significantly between paired participants. Differences in VOC concentrations between CDI patients and their matched controls were assessed using the paired student t-test.

Spearman rank correlation was used for continuous clinical variables such as age, BMI, and creatinine level. The Mann–Whitney U test or, if the category contained more than 2 groups, 1-way ANOVA, was used for categorical clinical variables. The only comorbidities with enough power and homogeneity for statistical testing were diabetes mellitus and chronic kidney disease, which, along with clinical variables (age, sex, BMI, smoking status, and creatinine level) were tested for associations with VOCs of interest.

2.7.3. Predictive modeling

After identifying VOCs with a univariable association with group (case or control), we performed classification modeling to examine the ability of one or more compounds to discriminate between case and control participants. The model space was reduced by considering only MFs with an uncorrected Mann–Whitney U test *P*-value of <0.2, which yielded 21 candidate MFs for use in a predictive model.

A non-linear model was trialed to try and achieve a better classification performance. The model selected was quadratic discriminant analysis (QDA), a more general version of Fisher linear discriminant analysis [10]. We used step-forward floating selection to limit the number of MFs used by the model, thereby reducing the chance of the model overfitting.

2.7.4. Identification of molecules and their potential biological significance

MFs were assigned tentative identities by matching their mass spectra with those in a library maintained by the NIST. We reviewed the scientific literature to determine these compounds' routes of biological origin, metabolism, and/or external exposures, correlations with diseases and disease processes, and associations with other biological phenotypes. We performed paired testing to control for any confounding effects that age or sex may have had on MF abundance in the participants' breath samples.

3. Results

We collected 40 breath samples from 40 patients. We excluded 6 samples due to collection leaks or chromatographic resolution checks, which left 34 samples (17 from patients with CDI and 17 from patients without CDI) available for analysis.

3.1. Patient characteristics

The demographic and clinical characteristics of the 34 patients whose collected breath samples passed curation are presented in table 1. Because this population was recruited at a hospital, it had a high rate of comorbidities; in fact, every participant had at least one listed comorbidity. Most comorbidities occurred more frequently among patients with CDI.

3.2. Data structure

A plot of the first 2 principal components colored by the participant group showed no clear trends (figure 1, panel a). Coloring the first 2 principal components by other clinical variables revealed no trends for age, sex, race, BMI, or smoking status (data not shown). Coloring samples by their analytical sequence number revealed a trend in the second principal component (figure 1, panel (b)). The samples were analyzed when they were received, and sequences were ordered chronologically, so this trend could suggest a degree of time-dependent variability in the breath samples. The second principal component could account for about 10.6% of the variation in the data, but the modest pattern of samples colored by their analytical sequence suggests that the chronological sequence cannot account for all the variation.

3.3. VOCs with different abundances between paired participants

In total, 65 VOCs were detected and qualified for inclusion in the final curated dataset. Among these, 4 showed some evidence of association with CDI at a level of significance of 0.1 (table 2; figure 2). A lower *P*-value threshold ($P < 0.1$) was considered significant with only 12 pairs of matched participants available.

Table 1. Patients' demographic and clinical characteristics.

Clinical variable	Patients without CDI, $n = 17$	Patients with CDI, $n = 17$	All patients, $n = 34$
Sex			
Male	8	9	17
Female	9	8	17
Age, mean ± standard deviation, years	55 ± 13.5	59.7 ± 16.7	57.4 ± 14.9
BMI, mean ± standard deviation, kg m^{-2}	29.2 ± 6.9	27.2 ± 6.7	28.2 ± 6.8
Current or ex-Smoker	6	8	14
Diabetes mellitus	4	3	7
Coronary artery disease	0	5	5
Heart failure	0	3	3
Chronic kidney disease	2	7	9
ESRD	0	2	2
Chronic liver disease	1	2	3
Inflammatory bowel disease	4	1	5
COPD	1	0	1
Malignancy	2	5	7
Post-stem cell transplantation	1	3	4
Concurrent infection	4	2	6
Other[a]	7	6	13

Note: Data are no. of patients unless otherwise indicated.

Abbreviations: CDI, *Clostridioides difficile* infection; BMI, body mass index; ESRD, end-stage renal disease; COPD, chronic obstructive pulmonary disease.

[a] Includes A1AT deficiency and cystic fibrosis.

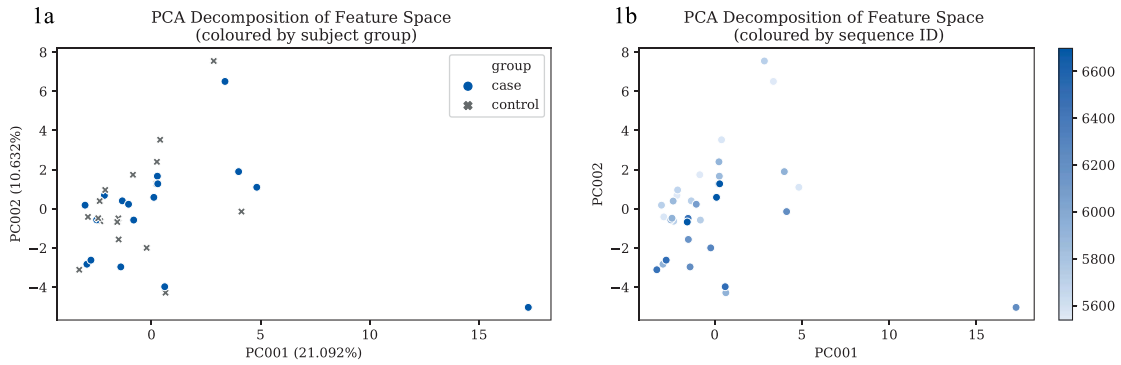

Figure 1. Principal component analysis showing the first 2 principal components. (Panel (a)) The first 2 principal components colored by subject group. The sample from patient CC1026 on the right side of the plot was flagged as an outlier. (Panel (b)) The first 2 principal components colored by analytical sequence.

Table 2. Statistically significant breath volatile organic compounds in 12 pairs of patients with and without *Clostridioides difficile* infection.

MF	NIST tentative identity	Match %	*P*-value[a]
MF41	2-Ethyl-1-hexanol	95	0.050
MF23	p-Xylene	93.8	0.084
MF56	Isophorone	75.8	0.099
MF18	Tetrachloroethylene	97.6	0.099

Abbreviations: MF, molecular feature; NIST, National Institute of Standards and Technology.

[a] Significance was defined using a *P*-value of <0.1.

3.4. VOCs with significantly different abundances between groups

The VOCs whose difference of abundance neared significance ($P < 0.2$) between CDI patients and controls are shown in table 3, and boxplots for the most significant MFs ($P < 0.05$) are shown in figure 3. Of the 4 most significant MFs, 3 had higher abundances on average in the CDI group (figure 4). n-Hexane and 3-methylundecane, which are products of lipid peroxidation, had significantly higher concentrations in the CDI group than in the control group (table 3; figure 3). 2-Phenyl-2-propanol (also known as cumyl alcohol; MF53) was 2.47-fold higher on average in the breath samples from CDI cases than in those from control participants, but this difference was not statistically significant (table 3). Tetrachloroethylene (also known as perchloroethylene; MF18) was less abundant in the CDI group than in the control group (figure 3). Analysis with the unpaired Mann–Whitney U test showed this difference to be significant (table 3), whereas

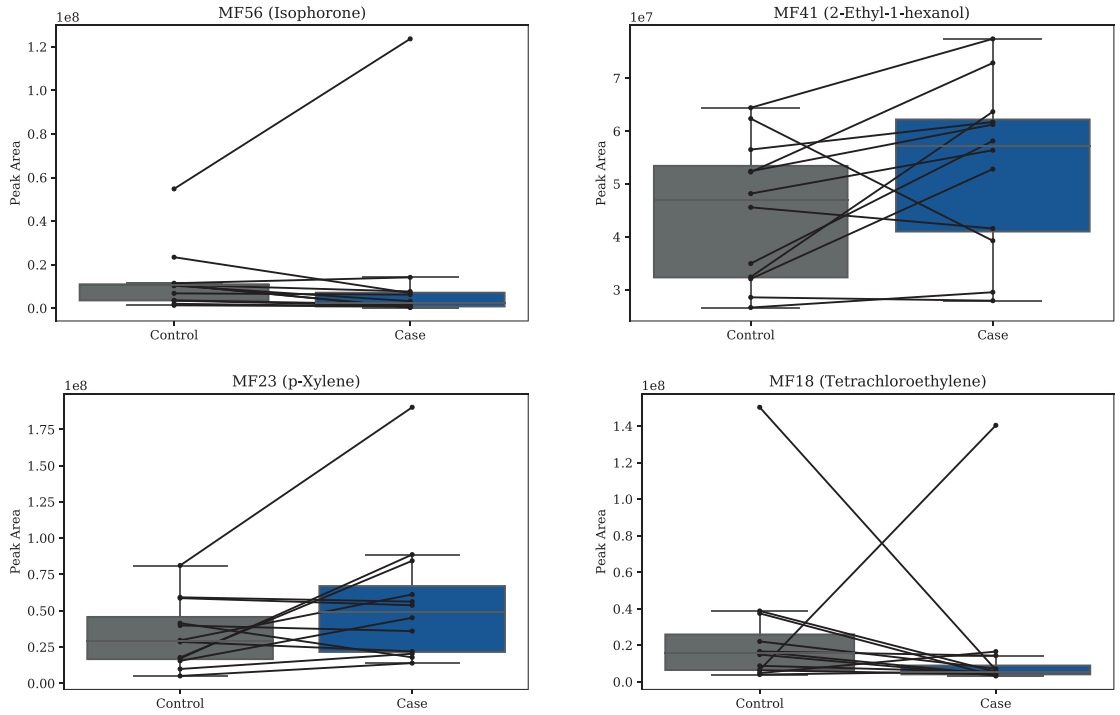

Figure 2. Paired box plots of the 4 molecular features with significant abundance differences between cases and controls after paired testing.

Table 3. Breath volatile organic compounds with near significant ($P < 0.2$), and significantly different abundances between patients with and without *Clostridioides difficile* infection.

MF	NIST tentative identity	Match %	Fold change (case/control)	Uncorrected *P*-value[a]
MF2	n-Hexane	81.93	1.3	0.023
MF18	Tetrachloroethylene	97.59	0.7	0.034
MF23	p-Xylene	93.80	1.6	0.039
MF55	3-Methylundecane	89.80	1.2	0.043
MF53	2-phenyl-2-propanol	80.87	2.5	0.065
MF25	4-Heptanone	93.79	2.4	0.069
MF26	o-Xylene	84.98	1.5	0.069
MF16	Toluene	87.05	1.4	0.074
MF36	6-methyl-5-hepten-2-one	95.58	0.7	0.114
MF41	2-Ethyl-1-hexanol	94.98	1.1	0.128
MF56	Isophorone	75.75	1.1	0.143

Abbreviations: MF, molecular feature; NIST, National Institute of Standards and Technology.
[a] Significance was defined using a *P*-value of <0.2).

analysis with the paired Wilcoxon signed-rank test did not (table 2). 6-Methyl-5-hepten-2-one (MF36) was detected in breath samples and, like tetrachloroethylene, its abundance in case samples was lower than that in control samples, but this difference was not statistically significant (table 3). Compared with those from controls, breath samples from CDI patients had higher levels of the aromatic hydrocarbons p-xylene, o-xylene, and toluene. Analysis with the unpaired Mann–Whitney U-test revealed the difference in the p-xylene level between the groups to be significant (table 3), whereas analysis with the paired Wilcoxon signed rank test did not (table 2). The unpaired test also showed that the between-group differences in the levels of both o-xylene and toluene were not significant (table 3).

3.5. Classification model performance

QDA with step-forward floating selection identified 9 MFs, yielding a model with a final cross-validated accuracy of 0.74. The cross-validated accuracy of each step of the MF selection process is given in supplementary figure 1. A summary of the MFs selected in order of their selection is given in table 4. This model had a sensitivity of 0.71, specificity of 0.76, and mean area under the receiver operating characteristic curve (AUC) of 0.72. The confusion matrix for this model

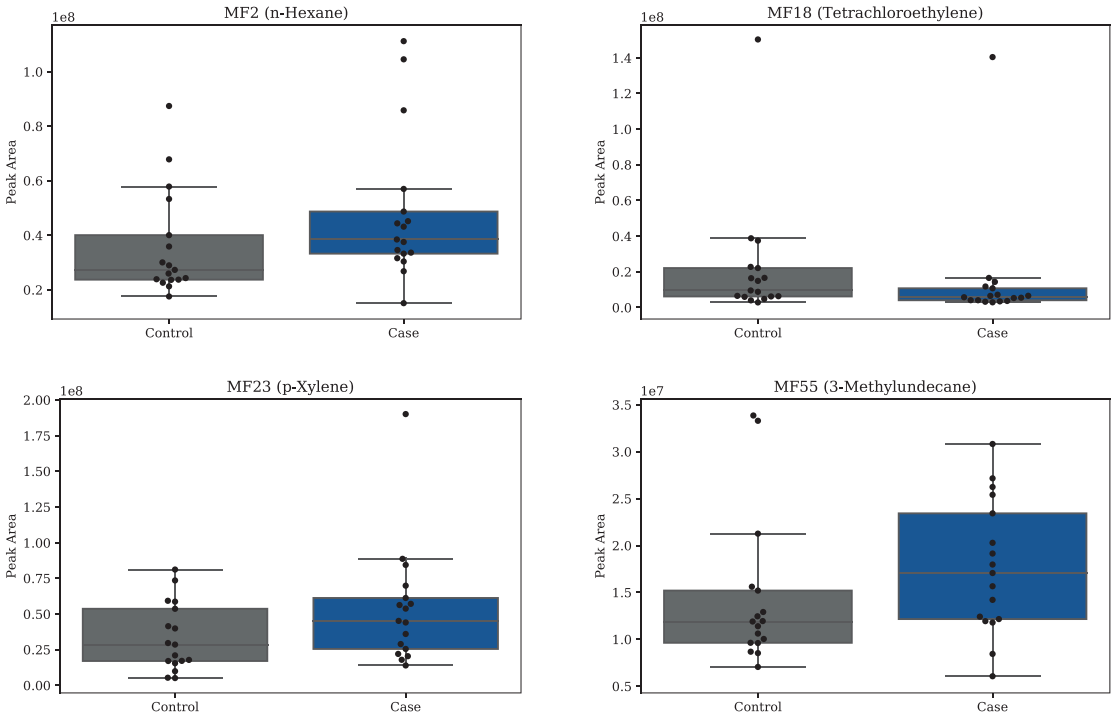

Figure 3. Box plots of molecular features showing significant differences between groups.

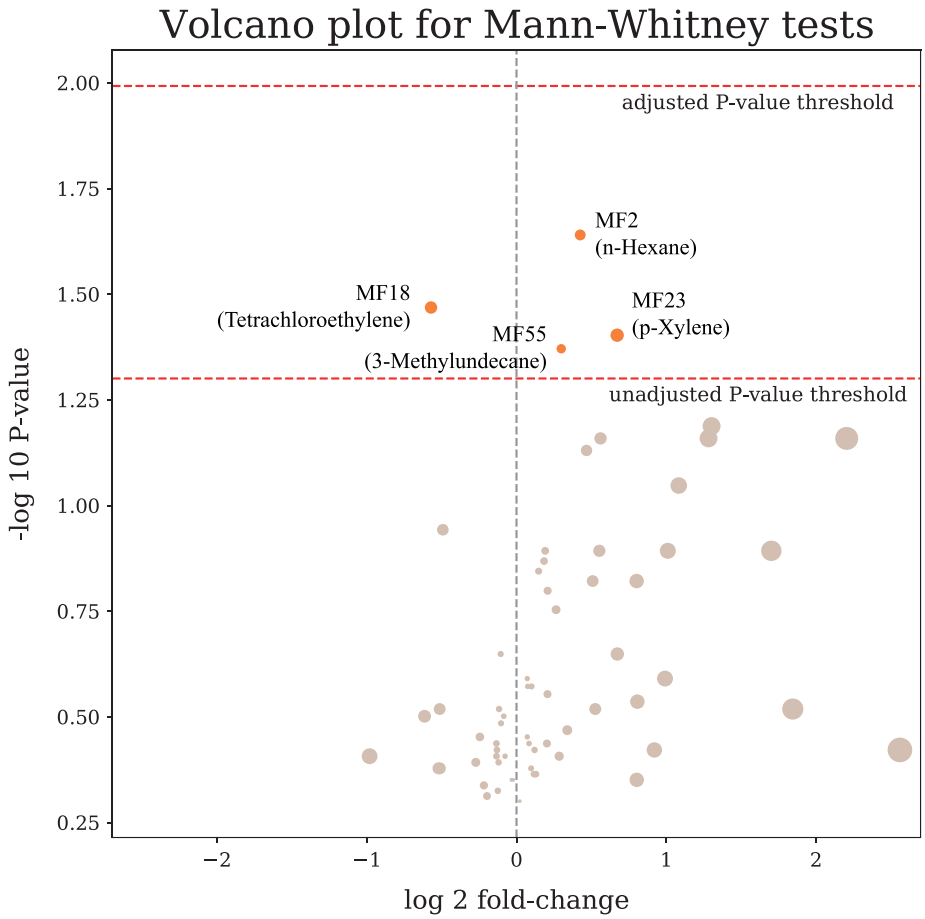

Figure 4. A volcano plot of all molecular features showing a higher abundance of 3 molecular features, MF23 (p-xylene), MF2 (n-hexane), and MF55 (3-methylundecane), in the *Clostridioides difficile* infection group.

Table 4. Breath volatile organic compounds selected by the quadratic discriminant analysis classifier model.

MF	NIST tentative identity	Match %	Mann-Whitney *P*-value	Wilcoxon signed-rank *P*-value	Effect size
MF2	n-Hexane	81.93	0.023	0.480	0.522
MF18	Tetrachloroethylene	97.59	0.034	0.099	−0.219
MF53	2-Phenyl-2-propanol	80.87	0.065	0.308	0.682
MF26	o-Xylene	84.98	0.069	0.136	0.549
MF25	4-Heptanone	93.79	0.069	0.136	0.720
MF41	2-Ethyl-1-hexanol	94.98	0.128	0.050	0.132
MF46	2,6-Dimethyl-7-octen-2-ol	78.71	0.128	0.272	0.451
MF31	1,2,4-Trimethyl-benzene	71.41	0.151	0.480	0.266
MF50	3-7-Dimethyl-3-octanol	91.29	0.176	0.136	0.251

Note: Features are sorted in order of selection, and *P*-values are given for both paired and unpaired testing. Effect size was calculated as Cohen's d, the difference of 2 groups' means (control—case), divided by their pooled standard deviation.

Abbreviations: MF, molecular feature; NIST, National Institute of Standards and Technology.

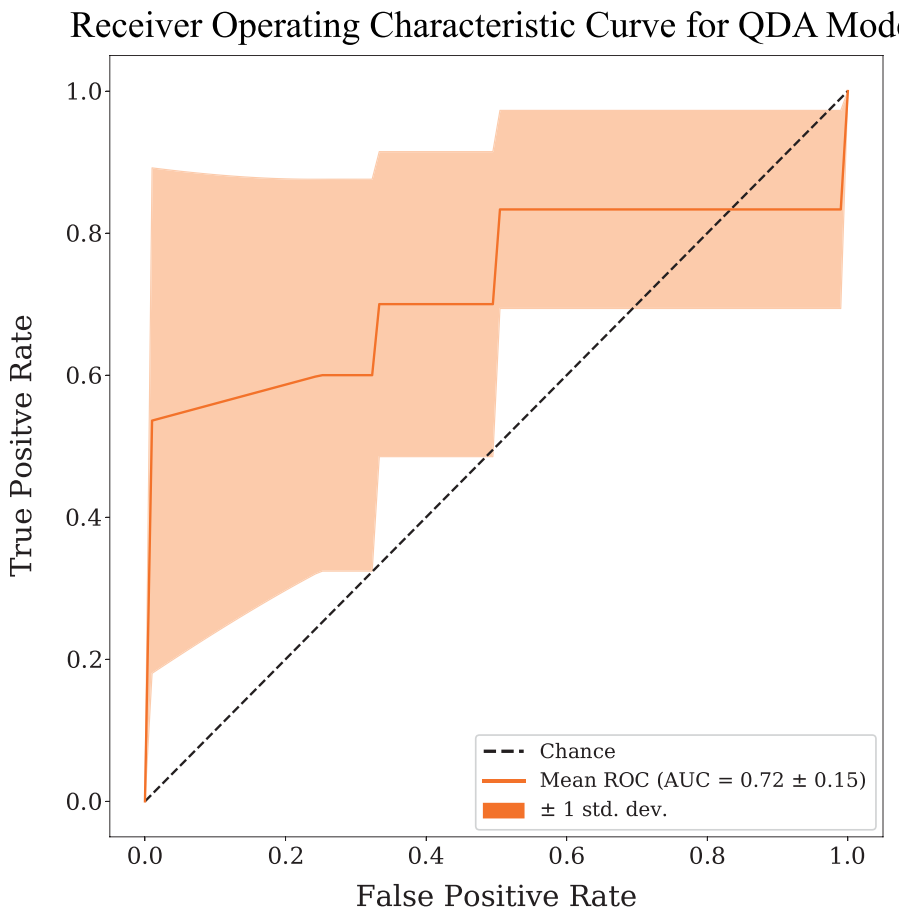

Figure 5. The receiver operating characteristic curve for the quadratic discriminant analysis model with selected features. The mean area under the curve was 0.72.

is shown in supplementary figure 2 and the receiver operating characteristic curve is shown in figure 5.

A QDA model with just 2 MFs—n-hexane (MF2) and tetrachloroethylene (MF18)—yielded a cross-validated accuracy of 0.67.

4. Discussion

This study provides additional evidence supporting the development of a viable breath analysis test for the diagnosis of CDI. We found that a model of 9 breath VOCs could be used to diagnose CDI with moderate sensitivity and specificity.

In a previous study in which selected ion flow tube mass spectrometry was used to analyze VOCs in samples from 31 patients with CDI and 31 controls, John *et al* [5] found evidence that CDI patients have a characteristic VOC profile. A *K*-nearest neighbors classifier model ($k = 7$) showed excellent accuracy for identifying CDI, with an AUC of 93%. Because the

study could not identify molecules of interest, its findings cannot be compared with those of other studies, including the current study. In the present study, using an untargeted approach to identify VOCs, a QDA model comprising 9 features yielded a sensitivity of 71% and specificity of 76%. The low diagnostic accuracy in the current study may have been due to the small sample size and the CDI patients' higher rates of comorbidities. Nonetheless, both studies suggest that patients with CDI have a characteristic breath VOC profile.

The sensitivity of CDI testing by breath analysis such as employed in this pilot study is substantially better than CDI testing by EIA, although not quite as good as that of PCR. The specificity of CDI testing by the breath analysis employed in this study is lower than that of EIA or PCR. With these operating characteristics, breath testing has potential to be developed into a rapid, convenient, and inexpensive screening test. For some patients results of breath testing alone may be enough to make treatment decisions. Others will need a confirmatory test with PCR. If the accuracy of breath testing can be improved, a smaller proportion of tested patients will need a confirmatory test.

In our study, the presence of certain VOCs in the breath of patients with CDI suggested ongoing environmental exposures, including to chemicals and pollutants such as di-(2-ethylhexyl)-phthalate plasticizers, tobacco smoke, and vehicle exhaust. Compared with controls, patients with CDI had elevated levels of 4-heptanone and 2-ethyl-1-hexanol, both of which are products in the di-(2-ethylhexyl) phthalate (DEHP) degradation pathway [11]. 4-Heptanone is also associated with lipid oxidation [12], fungal metabolism [13], and the metabolism of phthalate plasticizers [11, 14, 15]. In neonates, exposure to these compounds result in changes in the gut microbiome and altered immune responses [16]. Previous studies employing headspace analyses of VOCs showed that 4-heptanone levels in feces from CDI patients are higher than those in feces from non-CDI patients with diarrhea of unknown origin [17]. Future studies are needed to determine the role of phthalate-containing plastics in the pathogenesis of CDI.

Our findings suggest that, among the VOCs identified, 4-heptanone is a potential biomarker of CDI. If 4-heptanone is confirmed as such, exposure to DEHP, which is a metabolic precursor to 4-heptanone, could be a novel risk factor for CDI. Interestingly, in this study, patients with chronic kidney disease had significantly elevated levels of 4-heptanone, mirroring previous reports of its association with changes in renal function [18, 19]. Compared with controls, patients with CDI had elevated breath levels of the aromatic hydrocarbon VOCs p-xylene, o-xylene, and toluene, which may reflect exposure to tobacco smoke or vehicle exhaust, a possibility strengthened by the fact that smoking is a risk factor for CDI. In the current study, however, these VOCs did not stand out as markers that differentiated patients who were current ($n = 2$) or former ($n = 12$) smokers from those who were never-smokers.

Compared with controls, CDI patients appeared to have higher levels of VOCs related to lipid peroxidation in their breath samples, as evidenced by the significantly higher levels of n-hexane and 3-methylundecane. Lipid peroxidation is a critical step in *C. difficile*–mediated colonic damage [20]. The biological origin of n-hexane involves the non-enzymatic degradation of unsaturated and polyunsaturated fatty acids, mostly esterified membrane phospholipids and stored triglycerides, in response to lipid peroxidation [21]. n-Hexane is also commonly encountered in the environment as a solvent used in painting and industry. Lipid aldehydes, such as heptanal, octanal, nonanal, and decanal, which are frequently detected in lipid peroxidation studies [22, 23], were not among the tentatively identified compounds in the present study. However, free aldehydes are reactive and tend to form covalent adducts with proteins and other macromolecules, so their absence may indicate that these molecules do not get absorbed well from the gut, or the distance between the intestinal tract and the lungs is too great for aldehydes to traverse so that they undergo further metabolism in the liver.

As some *Clostridioides* species have been reported to metabolize tetrachloroethylene [24], the lower levels of tetrachloroethylene in the breath samples from CDI patients may have been related to the degradation of the VOC by *C. difficile*.

Additional analyses that would strengthen our observations and help further elucidate the role of breath VOCs in the diagnosis of CDI include the following. First, approaches to quantify DEHP exposure through epidemiological surveys or by measuring DEHP and mono-(2-ethylhexyl)- phthalate (MEHP) in blood from CDI and control patients should be considered. Given the possibility of contamination from plastics during sample collection, the characterization of the hepatic metabolite mono-(2-ethylhexyl)-phthalate-β-D-glucuronide [25] may allow a more specific estimation of exposure. Second, the extent to which *C. difficile* can metabolize tetrachloroethylene should be assessed in vitro [24] to see if *C. difficile*, like other *Clostridioides* species, can degrade this compound, which would explain its relatively lower levels in the breath of the CDI patients in the present study. Accomplishing this using headspace analyses of pure cultures could enable the identification of additional metabolic products as CDI biomarkers. In addition, in vitro studies with gene ablation could determine whether tetrachloroethylene pathway is essential for *C. difficile* pathogenesis and potentially identify a novel therapeutic target. Third, these findings should be validated in

an independent cohort, and the compound identities should be confirmed in reference to genuine standards, to develop diagnostic, prognostic, and/or monitoring applications that use metabolomic biomarkers.

Our study had several limitations. First, its small sample size made it difficult to confirm the validity of its findings. Second, frequencies of medical comorbidities like chronic kidney disease and coronary artery were more common, and that of inflammatory bowel disease was less common, in patients with CDI than in controls. This is a limitation of this small study, and it will be important for future larger studies to control for these important comorbidities. Third, there is a potential misidentification of VOCs while matching with the NIST database, whose data are collected under different analytical methods, that may have led to inaccurate biological associations. Last, the case group had more comorbidities than the control group did, which might have affected the results.

In conclusion, this study identified several VOCs that are associated with CDI and demonstrates that a classification algorithm based on the quantitation of such VOCs can differentiate between patients with and without CDI.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

We thank Joseph A Munch, Senior Scientific Editor in the Research Medical Library at The University of Texas MD Anderson Cancer Center, for editing this article.

Funding

This research was supported by the Cleveland Clinic Foundation Research Program Committee grant (Grant Number: 290, 13 March 2018)

Ethical statement

The Cleveland Clinic Institutional Review Board approved the protocol (IRB #18-1324). All participants provided written informed consent for participation. The study was performed in accordance with the Declaration of Helsinki and the Health Insurance Portability and Accountability Act.

ORCID iDs

Teny M John https://orcid.org/0000-0002-2675-1529
Nabin K Shrestha https://orcid.org/0000-0001-6766-9874
David Grove https://orcid.org/0000-0002-0496-9746
Raed A Dweik https://orcid.org/0000-0002-4425-1288

References

[1] Lessa F C *et al* 2015 Burden of *Clostridium difficile* infection in the United States *New Engl. J. Med.* **372** 825–34
[2] Guh A Y *et al* 2020 Trends in U.S. Burden of *Clostridioides difficile* infection and outcomes *New Engl. J. Med.* **382** 1320–30
[3] Surawicz C M, Brandt L J, Binion D G, Ananthakrishnan A N, Curry S R, Gilligan P H, McFarland L V, Mellow M and Zuckerbraun B S 2013 Guidelines for diagnosis, treatment, and prevention of clostridium difficile infections *Am. J. Gastroenterol.* **108** 478–98
[4] Rees C A, Shen A and Hill J E 2016 Characterization of the *Clostridium difficile* volatile metabolome using comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry *J. Chromatogr.* B **1039** 8–16
[5] John T M, Shrestha N K, Procop G W, Grove D, Leal S M Jr, Jacob C N, Butler R and Dweik R 2021 Diagnosis of *Clostridioides difficile* infection by analysis of volatile organic compounds in breath, plasma, and stool: a cross-sectional proof-of-concept study *PLoS One* **16** e0256259
[6] Wei R, Wang J, Su M, Jia E, Chen S, Chen T and Ni Y 2018 Missing value imputation approach for mass spectrometry-based metabolomics data *Sci. Rep.* **8** 1–10
[7] Mann H B and Whitney D R 1947 On a test of whether one of two random variables is stochastically larger than the other *Inst. Stat. Math.* **18** 50–60
[8] Westfall P H and Young S S 1993 *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment* (Wiley)
[9] Wilcoxon F 1992 Individual comparisons by ranking methods *Breakthroughs in Statistics* (Springer) pp 196–202
[10] Fisher R A 1936 The use of multiple measurements in taxonomic problems *Ann. Eugen.* **7** 179–88
[11] Wahl H G, Hong Q, Hildenbrand S, Risler T, Luft D and Liebich H 2004 4-Heptanone is a metabolite of the plasticizer di (2-ethylhexyl) phthalate (DEHP) in haemodialysis patients *Nephrol. Dial. Transplant.* **19** 2576–83
[12] Alnoumani H, Ataman Z A and Were L 2017 Lipid and protein antioxidant capacity of dried Agaricus bisporus in salted cooked ground beef *Meat Sci.* **129** 9–19
[13] Moularat S, Robine E, Ramalho O and Oturan M A 2008 Detection of fungal development in closed spaces through the determination of specific chemical targets *Chemosphere* **72** 224–32
[14] Stingel D, Feldmeier P, Richling E, Kempf M, Elss S, Labib S and Schreier P 2007 Urinary 2-ethyl-3-oxohexanoic acid as major metabolite of orally administered 2-ethylhexanoic acid in human *Mol. Nutr. Food Res.* **51** 301–6
[15] Walker V and Mills G A 2001 Urine 4-heptanone: a β-oxidation product of 2-ethylhexanoic acid from plasticisers *Clin. Chim. Acta* **306** 51–61

[16] Yang Y-N, Yang Y-C S, Lin I-H, Chen Y-Y, Lin H-Y, Wu C-Y, Su Y-T, Yang Y-J, Yang S-N and Suen J-L 2019 Phthalate exposure alters gut microbiota composition and IgM vaccine response in human newborns *Food Chem. Toxicol.* **132** 110700

[17] Patel M, Fowler D, Sizer J and Walton C 2019 Faecal volatile biomarkers of *Clostridium difficile* infection *PLoS One* **14** e0215256

[18] Mochalski P, King J, Haas M, Unterkofler K, Amann A and Mayer G 2014 Blood and breath profiles of volatile organic compounds in patients with end-stage renal disease *BMC Nephrol.* **15** 43

[19] Pagonas N, Vautz W, Seifert L, Slodzinski R, Jankowski J, Zidek W and Westhoff T H 2012 Volatile organic compounds in uremia *PLoS One* **7** e46258

[20] Chandrasekaran R and Lacy D B 2017 The role of toxins in *Clostridium difficile* infection *FEMS Microbiol. Rev.* **41** 723–50

[21] Kivits G A, Ganguli-Swarttouw M A and Christ E J 1981 The composition of alkanes in exhaled air of rats as a result of lipid peroxidation in vivo. Effects of dietary fatty acids, vitamin E and selenium *Biochim. Biophys. Acta* **665** 559–70

[22] Tanaka M *et al* 2020 Identification of characteristic compounds of moderate volatility in breast cancer cell lines *PLoS One* **15** e0235442

[23] Zanella D, Henket M, Schleich F, Dejong T, Louis R, Focant J-F and Stefanuto P-H 2020 Comparison of the effect of chemically and biologically induced inflammation on the volatile metabolite production of lung epithelial cells by GC× GC-TOFMS *Analyst* **145** 5148–57

[24] Bowman K S, Rainey F A and Moe W M 2009 Production of hydrogen by Clostridium species in the presence of chlorinated solvents *FEMS Microbiol. Lett.* **290** 188–94

[25] Silva M J, Barr D B, Reidy J A, Kato K, Malek N A, Hodge C C, Hurtz D, Calafat A M, Needham L L and Brock J W 2003 Glucuronidation patterns of common urinary and serum monoester phthalate metabolites *Arch. Toxicol.* **77** 561–7