Identification of profiles of volatile organic compounds in exhaled breath by means of an electronic nose as a proposal for a screening method for breast cancer: a case-control study

Lorena Díaz de León-Martínez®, Maribel Rodríguez-Aguilar®, Patricia Gorocica-Rosete®, Carlos Alberto Domínguez-Reyes®, Verónica Martínez-Bustos®, Juan Alberto Tenorio-Torres®, Omar Ornellas-Rebolledo®, José Alfonso Cruz-Ramos®, Berenice Balderas-Segura® and Rogelio Flores-Ramírez®

1 Center for Applied Research in Environment and Health, CIACYT, Medicine Faculty, Autonomous University of San Luis Potosí, Av. Venustiano Carranza 2405, CP 78210, San Luis Potosí, SLP, Mexico
2 Biochemistry research department of National Institute Respiratory Diseases, Ismael Cosío Villegas, Mexico city, Mexico
3 Fundación del Cáncer de Mama A.C., Mexico city, Mexico
4 Labinnova Center of Investigation in Breath for early detection diseases, Guadalajara, Mexico
5 Instituto Jalisciense de Cáncerología, Guadalajara, Mexico
6 CONACyT Research Fellow, Coordinación para la Innovación y Aplicación de la Ciencia y la Tecnología (CIACYT), Universidad Autónoma de San Luis Potosí, Avenida Sierra Leona No. 550, Colonia Lomas Segunda Sección, CP 78210, San Luis Potosí, SLP, México
E-mail: rflowesra@conacyt.mx

Keywords: breast cancer, screening, breath analysis, electronic nose, multivariate analysis

Abstract

The objective of the present study was to identify volatile prints from exhaled breath, termed breath-print, from breast cancer (BC) patients and healthy women by means of an electronic nose and to evaluate its potential use as a screening method. A cross-sectional study was performed on 443 exhaled breath samples from women, of whom 262 had been diagnosed with BC by biopsy and 181 were healthy women (control group). Breath-print analysis was performed utilizing the Cyranose 320 electronic nose. Group data were evaluated by principal component analysis (PCA), canonical discriminant analysis (CDA), and support vector machine (SVM), and the test’s diagnostic power was evaluated by means of receiver operating characteristic (ROC) curves. The results obtained using the model generated from the CDA, which best describes the behavior of the assessed groups, indicated that the breath-print of BC patients was different from that of healthy women and that they presented with a variability of up to 98.8% and a correct classification of 98%. The sensitivity, specificity, negative predictive value, and positive predictive value reached 100% according to the ROC curve. The present study demonstrates the capability of the electronic nose to separate between healthy subjects and BC patients. This research could have a beneficial impact on clinical practice as we consider that this test could probably be used at the first point before the application of established gold tests (mammography, ultrasound, and biopsy) and substantially improve screening tests in the general population.

1. Introduction

Breast cancer (BC) is an oncological process in which healthy breast gland cells degenerate into tumor cells, proliferating and multiplying to form a tumor that can spread to other organs through the lymph nodes. This cancer is the most common type in women and is the second leading cause of death in the world. In 2018, the World Health Organization (WHO) estimated that there were 2.09 million cases of BC worldwide and more than 627,000 women died from this disease [1].

Several studies have shown that the risk of suffering from BC may be due to a combination of different factors, among the most important are: biological factors, such as age over 40 years;
environmental factors, such as exposure to ionizing radiation and environmental pollutants; lifestyle factors, such as obesity, sedentary life, and smoking; reproductive history (e.g. no experience of a first pregnancy after 30 years of age); and genetic factors, such as being a carrier of mutations in the BCRA1 or BCRA2 genes [2–4]. Although the direct causative agents of BC remain largely unknown, so far early detection has been the most important aspect of efforts to fight this disease. Strategies such as awareness, early detection, accurate diagnosis, timely treatment, and supportive care are critical to reducing the burden of BC [5–7].

Mammography is generally used for screening programs, and the WHO states that the use of this technology can reduce mortality by up to 20%. However, as with other screening tests, mammography leads to false positive/negative results in over 20% of cases [8]. When used in settings with sufficient resources available, mammography screening programs are recommended for women aged 50–69 years, with testing every two years; for women aged 40–49 years, this test has been shown to have less sensitivity, with one of the main reasons being the high breast density in this age group, where the sensitivity of the test is reduced from 37.6% to 67.6% in early stages of the disease [9].

The WHO states that in resource-constrained settings and deficient health systems, population-based mammography screening programs may be neither cost-effective nor feasible. Therefore, early diagnosis and treatment should be the priority in these settings, where clinical breast examination appears to be a promising screening method [1].

To address these challenges, effort has been invested into the discovery of new diagnostic biomarkers and screening through genomic analysis, leading to improvements in clinical outcomes and survival rates [10–13]. Tests such as the identification of BRCA1 and BRCA2 mutations decrease the cost of timely BC identification programs for women with a family history [14, 15]. However, it is costly and applies only to people with identified risk factors and not to the general population. Evidence-based frameworks for prevention and cost-effective management of BC are essential for achieving equitable outcomes and addressing the growing burden of disease [16].

The analysis of cancer biomarkers for screening based on the volatiolome represents a promising development and has the characteristics of being minimally or non-invasive, fast, low-cost, and with potential for a greater coverage of the population. This technology analyzes and identifies the volatile organic compounds (VOCs) that are produced from the metabolism of normal and cancer cells and interact with their microenvironment to be detectable in biological fluids, depending on their tissue-blood/urine/breath partition coefficients [17]. This approach is based on the production of VOCs by metabolism in pathophysiological processes, including hypoxia, increased energy expenditure by hyperproliferation, excessive inflammatory activity, the generation of reactive oxygen species, and many other processes that occur during the evolution of cancer and lead to changes in the types and amounts of VOCs generated [17–19].

Gas Chromatography coupled to Mass Spectrometry (GC-MS) has frequently been used for the identification and quantification of VOCs in exhaled breath. Our study intended to identify the chemical nature of the composition of VOCs in women with BC, allowing the exploration of each compound through multivariate statistics, and comparing this to the composition of VOCs in healthy women in order to statistically discriminate the differences and, thus, determine a chemical print that acts as a biomarker of the disease [20, 21]. In 2014, Wang et al evaluated VOCs in exhaled breath from 85 patients with histologically confirmed BC and 45 healthy subjects using GC-MS, and identified that the VOCs 2,5,6-trimethyltoluene, 1,4-dimethoxy-2,3-butanediol and cyclohexanone differed between the study groups [22]. This study suggested that these metabolites can be used in the identification of specific disease biomarkers; however, the economic cost of the analysis made it unfeasible to propose this approach as a screening method.

Another widely used methodology for the global identification of VOC patterns uses gas sensors such as the electronic nose. This approach measures the set of breath signals produced, without details of the chemical composition of the sample; this is defined as an olfactory print (breath-print), and its association with disease and differentiation between healthy subjects is accomplished through multivariate analysis, including principal component analysis (PCA), canonical discriminant analysis (CDA), Support Vector Machine (SVM), Artificial Neural Networks, and other methods. This approach is not able to identify every VOC, but it can reveal the overall pattern and, therefore, can be used in the future to identify similar cases [23]. In 2011, Shuster et al evaluated the exhaled breath of 7 healthy women, 16 women with benign breast conditions and 13 women with malignant lesions using an artificial nanoscale nose. The authors managed to distinguish between the groups, and this pilot study has guided the use of the electronic nose as a cost effective, fast, and reliable method in BC screening [24].

The present article aimed to analyze the olfactory print of patients with BC and healthy women using an electronic nose, in order to evaluate its potential as a complementary screening method. A secondary objective was to test the methodology in a pilot study in women categorized as BI-RADS-0.
2. Materials and methods

2.1. Patients, healthy subjects, and study design
A prospective, observational, case-control study was conducted with patients from the Breast Cancer Foundation A.C. (FUCAM) in Mexico City (2240 m above sea level). This site serves 7% of the BC population of Mexico. A convenience non-probabilistic sampling method was performed from March to December 2019. The sample size calculation was performed using the statistical package STATISTICA V7 (Power Analysis and Sample Size Calculation in Experimental Design-Calculating Required Sample Size), assuming the comparison of different means using a one-way analysis of variance. Conservative values were used considering a normal distribution, 50% variability, two experimental groups, with a Power (1-β) of 90% and a significance value (α) of 1%. For the calculation of sample size, it was considered that the prevalence of BC in Mexico City was 26,000 cases per year. For the purposes of this work, the sample size required for each group was 166 subjects.

The study inclusion criteria for BC patients were: (i) recently diagnosed with BC by biopsy; (ii) with no previous oncological treatment for at least one year; (iii) stage I to IV ductal carcinoma. The study elimination criteria were: lung diseases such as COPD, lung cancer, recent respiratory infections, pregnancy, insufficient sample, not following the breath sample protocol, and not signing the informed consent.

A group of healthy women was included in the BI-RADS scale of 1 and 2 as healthy controls. Inclusion criteria for healthy subjects were: (i) absence of BC and (ii) no history of any type of cancer. Also, a questionnaire about family health history, comorbidities (hypertension, diabetes, and other diseases), number of BI-RADS, socioeconomic characteristics, risk activities, such as tobacco smoking, number of births, among others, was applied to both groups. Control group elimination criteria included: diagnosed lung diseases, recent infections, failure to take a sample, failure to follow the breath sample protocol, and failure to sign the informed consent. The study was reviewed and approved by the Ethics Committee of INER, with the number CS4-18, and the Scientific Committee of FUCAM. This protocol was approved by COFEPRIS to be carried out in humans.

2.2. Exhaled breath collection
The collection of exhaled breath samples was conducted based on previous studies proposed by our research group and as proposed by the European Respiratory Society guide [23, 25]. The relaxed participants underwent three deep breaths followed by a deep exhalation into breathing collection bags (BCB) consisting of a 1.4-liter metalized plastic bag that had been pre-purged twice with ultra-pure nitrogen [26]. Every participant was asked to repeat the process. The conditions under which the sample was collected, both for healthy patients and patients with BC, were as follows, (i) fasting conditions (minimum 5 h of fasting), (ii) without smoking before the study, (iii) without oral hygiene and, (iv) before taking any medication. The samples were transported at 4 °C and subsequently analyzed. In addition, an environmental control sample was taken to eliminate possible interferences [27].

2.3. Analysis of exhaled breath
The Cyranose 320 (Sensigent®, California, US) was employed to determine the breath-print of the study groups. This technology is equipped with 32 polymer-based sensors, each with different sensitivities to various VOCs causing an increase in the electrical resistance of each sensor.

The sample processing was as follows, first, the BCBs were incubated at 37 °C for 5 min before the reading. The electronic nose setting consisted of a constant flow rate of 120 ml min⁻¹, with 40 s of baseline with ultra-pure nitrogen. The duration of the Cyranose output signal collection was 90 s. Then, the flow rate was increased to 180 ml min⁻¹ of ultra-pure nitrogen for sample line purging and air intake, and the substrate temperature was 32 °C. As an internal quality control, the resistance of the 32 sensors was recorded every day of the experiment to evaluate the quality of the analysis and determine the variation of the sensors’ responses to methanol, and Undecane standards were used at a concentration of 1 part per million. In addition to this measurement, 10 ultrapure-nitrogen samples were randomly recorded to verify the baseline signal. Low coefficients of variation were obtained (Supplementary material, 10 ultrapure-nitrogen samples were randomly recorded to verify the baseline signal. Low coefficients of variation were obtained (Supplementary material, tables 1 and 2 (available online at stacks.iop.org/JBR/14/046009/mmedia)).

2.4. Statistical analysis
The choice of the data pre-processing algorithm has been shown to affect the performance of the pattern recognition stage. Software written in MATLAB 6.1, CDAnalysis (Sensigent®), was used to extract features from the data in terms of the static change in sensor resistance. All data were normalized using a fractional difference model: R/Ro = (Rmax-Ro)/Ro, where R is the response of the system to the sample gas, and Ro is the baseline reading, the reference gas being the ultra-pure nitrogen flow.

In addition, a self-scaling was carried out to eliminate the effects of the magnitude of the sensor responses, by subtracting the average of the samples from the individual response of each sample and dividing it by the standard deviation of the samples.

To capture the greatest amount of variability in the data, PCA was performed using the Chemometric Data Analysis software CDAnalysis (Sensigent®), thus, reducing the data from the 32 sensors to three main components.
Table 1. Clinical characteristics of Breast Cancer patients and healthy subjects.

| Clinical parameters       | Breast Cancer          | Healthy        | p value |
|---------------------------|------------------------|----------------|---------|
| N                         | 262                    | 181            | –       |
| Age (years)               | 53.3 ± 13.0            | 42.7 ± 12.2    | <0.0001 <a> |
| >40 years                 | 221                    | 113            |         |
| <39 years                 | 41                     | 68             |         |
| BMI (kg/m²)               | 27.4 ± 5.1             | 24.1 ± 3.2     | <0.0001 <a> |
| Weight (kg)               | 67.2 ± 12.8            | 64.3 ± 9.8     | >0.05   |
| Height (m)                | 1.55 ± 0.06            | 1.61 ± 0.12    | <0.0001 |
| SAH (n)                   | 85                     | 6              | <0.0001 <a> |
| T2DM (n)                  | 57                     | 4              | <0.0001 <a> |
| **Schooling (n)**         |                        |                |         |
| Non                       | 38                     | 12             |         |
| Elementary                | 67                     | 8              | >0.05   |
| High School               | 94                     | 26             |         |
| University                | 61                     | 97             |         |
| Postgraduate              | 2                      | 38             |         |
| **Smokers (n)**           |                        |                |         |
| Current                   | 13                     | 27             | >0.05   |
| Non-smokers               | 249                    | 154            |         |
| **Tumor location (n)**    |                        |                |         |
| Bilateral tumor           | 5                      | N/A            |         |
| Right Breast              | 95                     | N/A            |         |
| Left Breast               | 162                    | N/A            |         |
| **Type of Cancer (n)**    |                        |                |         |
| IDC                       | 178                    | N/A            |         |
| ISDC                      | 17                     | N/A            |         |
| ILC                       | 16                     | N/A            |         |
| Other                     | 51                     | N/A            |         |
| **Biological and genetic characteristics (n) N = 89** | | | |
| Luminal A (low grade or Ki-67 absent) | 29 | N/A | |
| Luminal B (high grade or Ki-67 positive) | 22 | N/A | |
| Her2 overexpression        | 10                     | N/A            |         |
| Triple Negative            | 28                     | N/A            |         |
| **BIRADS (n)**            |                        |                |         |
| 0                         | 7                      | N/A            |         |
| 1                         | 0                      | N/A            |         |
| 2                         | 2                      | N/A            |         |
| 3                         | 7                      | N/A            |         |
| 4                         | 85                     | N/A            |         |
| 5                         | 82                     | N/A            |         |
| 6                         | 79                     | N/A            |         |
| **Stage (n)**             |                        |                |         |
| 0                         | 16                     | N/A            |         |
| 1                         | 19                     | N/A            |         |
| IA                        | 10                     | N/A            |         |
| IIA                       | 76                     | N/A            |         |
| IIB                       | 42                     | N/A            |         |
| IIIA                      | 44                     | N/A            |         |
| IIIB                      | 23                     | N/A            |         |
| IIIC                      | 15                     | N/A            |         |
| IV                        | 17                     | N/A            |         |

Mean ± standard deviation. <a> P < 0.05. SAH: systemic arterial hypertension. T2DM: Type 2 Diabetes mellitus. IDC: invasive ductal carcinoma; ISDC: in-situ ductal carcinoma; ILC: invasive lobular carcinoma.

CDA and SVMs discrimination models were used to assess clustering within the data sets of healthy women and BC patients. These exploratory techniques are used to investigate how the data cluster in the multi-sensor space.

The sensors with a higher importance index were used to obtain the CDA and SVM discrimination models through a cross-validation value (leaving one out of the procedure and, thus, predicting the group association and obtaining overall classification success rates) and the Mahalanobis distance between the group means in units of standard deviation.

SVMs is a kernel-based (radial Gaussian) supervised learning classification method that determines the optimal boundaries (support vectors) that precisely separate groups [28]. By giving n training pairs
\((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), where \(x_i\) is an input vector, and \(y_i \in [25]\), the SVM solves the following main problem:

\[
\min_{\beta, b} \frac{1}{2} \beta^T \beta + C \sum_{i=1}^{n} \xi_i
\]

\[
s.t. y_i \{\beta^T \phi(x_i) + b\} \geq 1 - \xi_i,
\]

\[
\xi_i \geq 0, i = 1, 2, \ldots, n.
\]

Where \(\beta\) is a unit vector (i.e. ||\(\beta|| = 1\)), \(T\) denotes the transposition of the matrix to Kernel, \(C\) is the adjustment parameter denoting the compensation between the margin width and the training data error and \(\xi_i \geq 0\) are stationary variables. For an unknown input pattern \(x\), we have the decision function:

\[
f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b
\]

Where, \(\alpha_i, i = 1, 2, \ldots, n; \alpha_i \geq 0\) are the Lagrange multipliers, and \(K(x, x_i) \equiv \phi(x_i) T \phi(x_i)\) is the Kernel function. The Gaussian radial base function is used as the kernel function

\[
K(x, x_i) = \exp(-\gamma ||x - x_i||^2)
\]

Where \(\gamma > 0\) are fixed parameters, \(\gamma ||x - x_i||^2 = (x - x_i, x - x_i)\) [28, 29]. The vector limit display was shown in blue for a matrix equal to zero, indicating a low confidence prediction, and green display was shown in blue for a matrix equal to zero, indicating a low confidence prediction.

Cross-validation (CV) is a standard technique for adjusting the hyperparameters of predictive models. In K-fold CV, the available data \(S\) was partitioned into \(K\) subsets \(S_1, \ldots, S_K\). Each data point in \(S\) was randomly assigned to one of the subsets, such that these are of an almost equal size (i.e. \([|S|/K| \leq |S_i| \leq |S|/K]\)). Furthermore, we defined \(S_i = \cup_{j=1, \ldots, K \neq i} S_j\) as the union of all data points, except those in \(S_i\). For each \(i = 1, \ldots, K\), an individual model was built by applying the algorithm to the training data \(S_i\). This model was then evaluated by means of a cost function using the test data in \(S_i\). The average of the \(K\) outcomes of the model evaluations is called cross-validation (test) performance or cross-validation (test) error and was used as a predictor of the performance of the algorithm when applied to \(S\). Typical values for \(K\) are 10 [30].

We evaluated differences in the CDA model between the groups of Healthy vs BC, <39 vs Healthy, >40 vs Healthy, IDC, other types of cancer, BI-RADS 2 3, and BI-RADS 4 5 6, in order to explore differences in healthy and BC groups. The performance of the CDA model was evaluated by receiver operating characteristic (ROC) using the PC1 axis values obtained from the training model, with an analysis with a 95% confidence interval, and the threshold value was selected with the highest specificity/sensitivity ratio. An external validation was conducted by selecting 70% of the population for the definition of the groups and the other 30% were randomly selected to validate the model.

The analysis was performed using XLSTAT version 12.0 (StatSoft®, Tulsa, Oklahoma, USA). The sensitivity = [number of true positives/(number of true positives + number of false negatives)]. Specificity = [number of true negatives/(number of true negatives + number of false positives)]. The predictive positive value (PPV) = [number of true positives/(number of true positives + number of false positives)]. The negative predictive value (NPV) = [number of true negatives/(number of true negatives + number of false negatives)].

3. Results

Table 1 presents the general characteristics of the study participants. A total of 443 women participated, including 262 women with BC diagnosed by biopsy and 181 healthy women, with similar characteristics to those of the BC group who did not present significant differences. The general characteristics of the women with BC were an average age of 53.3 ± 13.0 years, with 84.4% over 40 years, 32.3% presenting with hypertension, and 21.7% with T2DM. Ninety-five percent mentioned not having smoked. In 61.6% of cases the tumor was located in the left breast, and 67.6% presented with invasive ductal carcinoma (IDC), 6.5% had in-situ ductal carcinoma (ISDC), and 6.1% had invasive lobular carcinoma (ILC). The most frequently reported BI-RADS were 4, 5, and 6, covering 93.8% of the population of the study. With respect to stages, 17.4% were reported as stage 0, I and IA, 44.5% as stage II A and IIB, 31.7% as stage IIIA, IIB and IIC, and 6.4% as stage IV. 1(a–c) shows the PCA, CDA and SVM Model of the general model comparing groups of healthy women against women with BC. In figure 1(a), corresponding to the PCA, an evident separation between the study groups is shown, with the percentage of variability between the groups being 98.87%. In figure 1(b), it can be seen that 28 sensors presented with a higher index and were used to construct the CDA model. The SVM model in figure 1(c) also shows a high discrimination between the two study groups (100%), indicating that there are differences between healthy female VOCs and VOCs from BC patients.

With the information from the CDA model (Supplementary material- figure 1), the percentages of correct prediction between groups of healthy women vs women with BC were projected to reach 98%. On
Figure 1. Breath-print models of control group (Healthy) (green square) vs Breast-cancer patients (yellow circle). (a) Principal Component Analysis (PCA); (b) Canonical Discriminant Analysis (CDA), and (c) Support Vector Machine (SVM).

Table 2. Percentage of correct prediction obtained in canonical discriminant analysis with mean centering.

| Canonical discriminant analysis: evaluated subgroups | Number of PCs | Percentage of Correct Prediction | p-value |
|-----------------------------------------------------|---------------|---------------------------------|---------|
| a) BC vs Healthy                                    | 4             | 98.0                            | <0.001  |
| b) Age < 39 vs Healthy                              | 4             | 97.8                            | <0.001  |
| c) Age > 40 vs Healthy                              | 4             | 98.4                            | <0.001  |
| d) IDC vs Healthy                                   | 4             | 98.3                            | <0.001  |
| e) Other types of cancer vs Healthy                  | 5             | 97.4                            | <0.001  |
| f) BI-RADS 2 3 vs Healthy                           | 4             | 97.6                            | <0.001  |
| g) BI-RADS 4 5 6 vs Healthy                         | 4             | 97.8                            | <0.001  |

PC: Principal components; IDC: Infiltrating ductal cancer

the other hand, the same analysis was performed for healthy women vs women with BC under 39 years old, women with BC over 40 years old, women with IDC and women with other types of BC, reaching 97.8%, 98.4%, 98.3%, and 97.4% of correct classification, respectively (table 2). This indicates that the methodology presents a high level of correct classification and can also be used in people under 40 years of age and other types of BC, regardless of their location. Furthermore, in table 2 (Supplementary material- figure 1), the results indicate that the subgroups within the BC group do not have an influence on the model's ability to distinguish the groups or the percentage of correct prediction.
Table 3 presents the Cross-Validation Correct Rate performed by the Support Vector Machine model. A correct classification with 100% was obtained in the general model of the BC group vs healthy women. The same analysis was applied to establish the classifications between healthy women vs women with BC under 39 years, women with BC over 40 years, women with IDC and women with other types of BC, reaching 97.8%, 98.4%, 98.3% and 97.4% of correct classification, respectively. This can be observed clearly in figure 2 of the supplementary material.

Furthermore, with the values created in the CDA score, a cut-off point of 0.048 was established, which provided 100% sensitivity (Confidence Interval: 0.982%–100%) and specificity (Confidence Interval: 0.974%–100%), with a negative predictive value and a positive predictive value of 100%, and 100% accuracy over the entire ROC curve (figure 2). The values of the external validation of the CDA model obtained a percentage of correct prediction of 98.7%.

Finally, the methodology developed was applied to a double-blind study of seven women with BI-RADS 0, and the Mahalanobis distance and the probability of belonging to the group of healthy women or the group with BC were calculated (table 4). 100% of the women showed similar breath print characteristics to the BC group (Supplementary material-figure 3).

4. Discussion

This study demonstrates the great potential of electronic noses for discrimination between the breath-print of women with BC and healthy women. The main contribution of this work is that of having well-characterized groups of BC, and the discrimination between healthy women with respect to the types of BC (Supplementary material- figure 1). Moreover, in the pilot study, we managed to determine the similarity of the breath-print in women with BI-RADS 0 and women with BC diagnosed with a biopsy. It is important to highlight that, in this group, the corresponding gold tests were performed, and six women were ultimately diagnosed with infiltrating ductal carcinoma and one with a malignant phyllodes tumor (five women in grade III and two in grade II).

Similar studies using electronic noses have been shown to discriminate between healthy women, those with benign tumors and women with malignant stages of cancer. In women from China, an investigation was carried out where 30 healthy women were included, 52 with benign conditions, 25 with ductal carcinoma in-situ and 169 women with malignant lesions. This model reached 83% accuracy [31] by using an artificially intelligent nanoarray based on Gold Nanoparticles (GNPs) and Single Wall Carbon Nanotubes (SWCNTs) coated with different organic layers. In Israel, a model utilizing the Cyranose 320 was built using an artificial neural network (ANN) algorithm with 46 BC women and 39 healthy women, achieving up to 86.1% accuracy, 87.5% sensitivity, and 85% specificity [32]. Although the Cyranose 320 was used in the study proposed by Or Herman-Saffar et al, 2018 [32], the main difference to our study was the analytical method; it is worth highlighting the following differences: (i) the sample analysis time proposed was 330 s, (ii) the sample was taken by connecting a mask directly to the input of the electronic nose, and (iii) the patient exhaled for 40 s. The advantage of the sampling proposed in our study is the reduction of the effect of the ambient air via our hermetic system, in addition to obtaining a greater amount of alveolar air, in which a greater amount of sample is concentrated. Our method also ensures a continuous flow that reduces the variability in the response of the sensors. On the other hand, the method of Herman-Saffar and colleagues obtained good sensitivity and specificity using 85 samples (46 sick and 39 healthy). However, the limitations suggested by the study indicated that the data extraction proposed for the elimination of sensor noise did not improve the percentages of sensitivity, specificity, and accuracy; therefore, data processing did not improve performance in the Artificial Neural Network model.

Other studies used to discriminate between exhaled breath VOCs of healthy subjects and BC patients have applied gas chromatography coupled to mass spectrometry, and these have developed models with a number from 10 to 80 patients with BC compared to healthy women. The results of these studies reached a sensitivity from 68.2% to 93.8%, a specificity of 61% to 91.7% and an accuracy from 79%
to 90.2% [26, 31, 33–35]. This shows that there are differences between the profiles of VOCs in these groups, indicating their ability to be used as biomarkers of disease.

The studies using GC-MS concluded that increased oxidative stress in breast tissue, caused by cancer, is linked to the alteration of biomarkers in the breath. Saturated hydrocarbons (ethane, pentane, nonane, tridecane), methylated derivatives (5-methyl undecane, 3-methyl pentadecane), and alcohols (2-ethyl-1-hexanol) have been identified as candidates for possible biomarkers [26, 31, 36–39].

Remarkably, the electronic nose provides pattern recognition methods that allow great discrimination between healthy and patient groups; however, in this study it does not manage to identify between stages or genotypes of cancer (table 5, supplementary material- figure 4(e)). Though the specificity is clinically relevant since it defines the guideline for the treatment of the disease, we consider that due to the obtained data, the analysis of the breathprint will be able to substantially improve the screening capacity of the BC. Moreover, the mathematical models developed are able to discriminate patients in BI-RADS 0 (table 4, supplementary material- figure 3) and even manage to position the samples from early stages of the disease.

The limitations of our study include that it was conducted at only one center, and we believe that other types of cancer should be included to establish the differences between the diseases. As for the strengths of our work, the high number of samples used in our model allowed us to assess factors that can be considered confounding, such as age, smoking, diabetes, and hypertension. Several authors point out that age is an independent variable of the breathprint and that it does not change the accuracy of the test [39], in agreement with our results presented among the groups under 39 and over 40 years old (table 2). Furthermore, when evaluating the influence of comorbidities (diabetes and hypertension) and habits (smoking) on the model (Supplementary material- table 3 and figures 5 to 8), it was demonstrated that the separation between the groups is not due to these factors. This result is consistent with other studies evaluating VOCs in exhaled breath to

---

**Figure 2.** ROC curve for the diagnosis of BC when using the PC1 axis. An AUC of 1.00 was obtained, and a PPV and NPV of 100% when using a cut-off point of 0.048.

---

**Table 4.** BI-RADS 0 samples in the model.

| Sample   | BIRADS | CProb  | MDist  |
|----------|--------|--------|--------|
| FCP0005  | 0      | 1.0000 | 0.2832 |
| FCP0008  | 0      | 1.0000 | 0.6400 |
| FCP0168  | 0      | 1.0000 | 0.6975 |
| FCP0186  | 0      | 1.0000 | 0.0052 |
| FCP0191  | 0      | 1.0000 | 0.0839 |
| FCP0240  | 0      | 1.0000 | 0.1364 |
| FCP0246  | 0      | 1.0000 | 0.9788 |

CProb: Correct probability percentage; MDist: Mahalanobis distance
Table 5. Percentage of correct prediction obtained in canonical discriminant analysis by genetic and biological cells characteristics.

| Canonical discriminant analysis model                          | Number of PCs | Percentage of Correct Prediction | p-value |
|---------------------------------------------------------------|---------------|----------------------------------|---------|
| a) Luminal A vs Healthy                                       | 4             | 97.7                             | <0.001  |
| b) Luminal B vs Healthy                                       | 4             | 94.6                             | <0.001  |
| c) Her 2 vs Healthy                                           | 3             | 95.3                             | <0.001  |
| d) Triple Negative vs Healthy                                 | 4             | 97.0                             | <0.001  |
| e) Luminal A vs Luminal B vs Her 2 vs Triple negative        | 6             | 0                                | >0.05   |

PC: Principal components

identify chemical prints of lung diseases using mass spectrometry and the electronic nose [23, 40–42]. We consider that further evaluations of this technique should include the evaluation of other types of cancer, as well as including a multi-center study to allow the development of this technology so that it can be used as a screening method, taking into account that our external validation results indicate a high percentage of correct classification (97.8%). Besides, although different authors show similar results to those obtained in this work, with a lower number of samples, there is a lack of standardization of sample collection among the different studies to implement this methodology in clinical practice.

This research could have a beneficial impact on clinical practice [43], and we consider that this test could probably be used as a first point of contact before the application of established gold tests (mammography, ultrasound, and biopsy) and substantially improve screening tests in the general population. Future development will involve innovation or a combination of advanced techniques for sampling and detection but, in addition, the analysis of the breath-print needs to be validated and standardized for real-world clinical use.

Acknowledgments

The authors acknowledge the research contribution to M.S. Masiel Vásquez Tinoco and Engineer Dalia Barrera de la Rosa of the National Institute of Respiratory Diseases (INER) and to M.S. María Fernanda Negrete FUCAM.

Funding

The study was financed by Labinnova Center of Investigation in Breath for early detection diseases.

Conflict of interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The author Omar Ornelas-Rebolledo who is CEO of the company LABINNOVA which financed the study and the Cyranose 320 equipment.

ORCID iDs

Lorena Díaz de León-Martínez https://orcid.org/0000-0002-8969-9924
Rogelio Flores-Ramírez https://orcid.org/0000-0003-2263-6280

References

[1] WHO 2018 Cancer. Key Facts World Health Organization (Geneva: World Health Organization)
[2] Akdeniz D, Klaver M M, Smith C Z A, Koppert L B and Hooning M J 2020 The impact of lifestyle and reproductive factors on the risk of a second new primary cancer in the contralateral breast: a systematic review and meta-analysis Cancer Causes Control 31 403–16
[3] Sisti J S et al Group WSC 2015 Reproductive factors, tumor estrogen receptor status and contralateral breast cancer risk: results from the WECARE study Springerplus 4 825
[4] King M-C, Marks J H and Mandell J B 2003 Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2 Science 302 643–6
[5] Pharoah P D, Antoniou A C, Easton D F and Ponder B A 2008 Polygenes, risk prediction, and targeted prevention of breast cancer N. Engl. J. Med. 358 2796–803
[6] Kyu H H et al 2016 Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013 BMJ 354 i3857
[7] Nothacker M, Duda V, Hahn M, Warm M, Degenhardt F, Madjar H, Weinbrenner S and Albert U-S 2009 Early detection of breast cancer: benefits and risks of supplemental breast ultrasound in asymptomatic women with mammographically dense breast tissue. A systematic review BMC Cancer 9 335
[8] WHO 2014 WHO Position Paper on Mammography Screening (Geneva: World Health Organization)
[9] Buist D S M, Porter P L, Lehman C, Taplin S H and White E 2004 Factors contributing to mammography failure in women aged 40–49 years JNCI: J. Natl Cancer Inst. 96 1432–40
[10] Dent R, Trudeau M, Pritchard K I, Hanna W M, Kahn H K, Sawka C A, Lickley I A, Rawlinson E, Sun P and Narod S A 2007 Triple-negative breast cancer: clinical features and patterns of recurrence Clin. Cancer Res. 13 4429–34
Sharma P, Bhattacharyya D K and Kalita J 2017 Disease biomarker identification from gene network modules for metastasized breast cancer Sci. Rep. 7 1072
Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Horvath I, Marie M, Bukovskis M, Strazda G, Jurka N and Taivans I 2015 Current challenges in systemic therapy Breast Cancer: Basic Clin. Res. 1 39–52
Balažna J, Sant J, Bonfill X, Casado A, Rué M, Gich I, Díez O, Sabate J M, Baiget M and Alonso M C 2004 Genetic counseling program in familial breast cancer: analysis of its effectiveness, cost and cost-effectiveness ratio Int. J. Cancer 112 647–52
Sarvazyan A, Egovor V, Son J S and Kaufman C S 2008 Article commentary; cost-effective screening for breast cancer Worldwide: current state and future directions Breast Cancer: Basic Clin. Res. 1 BCBCR.S774
Broza Y Y, Zuri L and Haick H 2014 Combined volatolomics for monitoring of human body chemistry Sci. Rep. 4 4611
Hakim M, Broza Y Y, Barash O, Peled N, Phillips M, Amann A and Haick H 2012 Volatile organic compounds of lung cancer and possible biochemical pathways Chem. Rev. 112 5949–66
Schmidt K and Podmore I 2015 Current challenges in volatile organic compounds analysis as potential biomarkers of cancer J. Biomark 10 319–33
Hanna G B, Boshier P R, Markar S R and Romano A 2019 Accuracy and methodologic challenges of volatile organic compound–based exhaled breath tests for cancer diagnosis: a systematic review and meta-analysis JAMA Oncol. 5 e182815-c
Phillips M, Catanese R, Lebauer C, Mundada M and Saunders C 2017, Breast mass ion biomarkers of breast cancer J. Breath Res. 11 016004
Wang C et al. 2014 Volatile organic metabolites identify patients with breast cancer, cyclomastopathy and mammary gland fibroma Sci. Rep. 4 5383
Rodriguez-Aguilar M, Díaz de Leon–martinez L, Gorocica-Rosete P, Padilla R P, Thirion-Romero I, Ornelas-Rebolledo O and Flores-Ramírez R 2020 Identification of breath–prints for the COPD detection associated with smoking and household air pollution by electronic nose Respir. Med. 163 105901
Schumer E M, Trivedi J R, van Berkel V, Black M C, Li M, Fu X A and Bousamra M 2015 High sensitivity for lung cancer detection using analysis of exhaled carbonyl compounds J. Thorac. Cardiovasc. Surg. 150 1517
Horvath I et al. 2017 A European Respiratory Society technical standard: exhaled biomarkers in lung disease Eur. Respir. J. 49 00956–2016
Phillips M, Catanese R N, Cruz-Ramos J A, Huston J, Ornelas O, Pappas N and Pathak S 2018 Prediction of breast cancer risk with volatile biomarkers in breath Breast Cancer Res. Treat. 170 543–50
Koo I, Shi X, Kim S and Zhang X 2014 iMatch2: compound identification using retention index for analysis of gas chromatography-mass spectrometry data J. Chromatogr. A 1337 202–10
Tsujitani M and Tanaka Y 2011 , Cross-validation, bootstrap, and support vector machines Adv. Artif. Neural Syst. 2011 302572
Tirzite M, Bukovskis M, Strazda G, Jurka N and Taivans I 2017 , Detection of lung cancer in exhaled breath with an electronic nose using support vector machine analysis J. Breath Res. 11 036009
Hastie T, Tibshirani R and Friedman J 2009 The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Berlin: Springer)
Barash O, Zhang W, Halpern J M, Hua Q-L, Pan –Y-Y, Kayal H, Khoury K, Liu H, Davies M P A and Haick H 2015 Differentiation between genetic mutations of breast cancer by breath volatolomics Onatarget 6 48864–76
Hereman-Saffar O, Roger Z, Libson S, Lieberman D, Gonen R and Zeiri Y 2018 Early non-invasive detection of breast cancer using exhaled breath and urine analysis Comput. Biol. Med. 96 227–32
Li J, Peng Y, Liu Y, Li W, Jin Y, Tang Z and Duan Y 2014 Investigation of potential breath biomarkers for the early diagnosis of breast cancer using gas chromatography-mass spectrometry Clin. Chim. Acta 436 59–67
Peng G, Hakim M, Broza Y Y, Billan S, Abdah-Bortnyak R, Kuten A, Tisch U and Haick H 2010 Detection of lung, colorectal, and prostate cancers from exhaled breath using a single array of nanosensors Br. J. Cancer 103 542–51
Phillips M, Catanese R N, Ditkoff B A, Fisher P, Greenberg J, Gunawardena R, Kwon C S, Tietje O and Wong C 2006 Prediction of breast cancer using volatile biomarkers in the breath Breast Cancer Res. Treat. 99 19–21
Phillips M, Catanese R N, Greenberg I, Goodman R, Gunawardena R and Naidu A 2003 Effect of oxygen on breath markers of oxidative stress Eur. Respir. J. 21 48–51
Jeziorska-Drutel A, Rosenzweig S A and Neumann C A 2013 Role of oxidative stress and the microenvironment in breast cancer development and progression Adv. Cancer Res. 119 107–25
Lavra L et al. 2015 Investigation of VOCs associated with different characteristics of breast cancer cells Sci. Rep. 5 13246
Phillips M, Catanese R N, Lebauer C, Mundada M and Saunders C 2017, Breast mass ion biomarkers of breast cancer J. Breath Res. 11 016004
Fens N, Zwinderman A H, Pappas N and Pathak S 2018 Prediction of breast cancer risk with volatile biomarkers in breath Breast Cancer Res. Treat. 170 543–50
Schwarz K 2015 Methodological challenges of beating the odds: a perspective Breast Cancer: Basic Clin. Res. 1 3–5
Barash O, Zhang W, Halpern J M, Hua Q-L, Pan –Y-Y, Kayal H, Khoury K, Liu H, Davies M P A and Haick H 2015 Differentiation between genetic mutations of breast cancer by breath volatolomics Onatarget 6 48864–76
Hereman-Saffar O, Roger Z, Libson S, Lieberman D, Gonen R and Zeiri Y 2018 Early non-invasive detection of breast cancer using exhaled breath and urine analysis Comput. Biol. Med. 96 227–32
Li J, Peng Y, Liu Y, Li W, Jin Y, Tang Z and Duan Y 2014 Investigation of potential breath biomarkers for the early diagnosis of breast cancer using gas chromatography-mass spectrometry Clin. Chim. Acta 436 59–67
Peng G, Hakim M, Broza Y Y, Billan S, Abdah-Bortnyak R, Kuten A, Tisch U and Haick H 2010 Detection of lung, colorectal, and prostate cancers from exhaled breath using a single array of nanosensors Br. J. Cancer 103 542–51
Phillips M, Catanese R N, Ditkoff B A, Fisher P, Greenberg J, Gunawardena R, Kwon C S, Tietje O and Wong C 2006 Prediction of breast cancer using volatile biomarkers in the breath Breast Cancer Res. Treat. 99 19–21
Phillips M, Catanese R N, Greenberg I, Goodman R, Gunawardena R and Naidu A 2003 Effect of oxygen on breath markers of oxidative stress Eur. Respir. J. 21 48–51
Jeziorska-Drutel A, Rosenzweig S A and Neumann C A 2013 Role of oxidative stress and the microenvironment in breast cancer development and progression Adv. Cancer Res. 119 107–25
Lavra L et al. 2015 Investigation of VOCs associated with different characteristics of breast cancer cells Sci. Rep. 5 13246
Phillips M, Catanese R N, Lebauer C, Mundada M and Saunders C 2017, Breast mass ion biomarkers of breast cancer J. Breath Res. 11 016004
Schwarz K 2015 Methodological challenges of beating the odds: a perspective Breast Cancer: Basic Clin. Res. 1 3–5
Barash O, Zhang W, Halpern J M, Hua Q-L, Pan –Y-Y, Kayal H, Khoury K, Liu H, Davies M P A and Haick H 2015 Differentiation between genetic mutations of breast cancer by breath volatolomics Onatarget 6 48864–76
Hereman-Saffar O, Roger Z, Libson S, Lieberman D, Gonen R and Zeiri Y 2018 Early non-invasive detection of breast cancer using exhaled breath and urine analysis Comput. Biol. Med. 96 227–32
Li J, Peng Y, Liu Y, Li W, Jin Y, Tang Z and Duan Y 2014 Investigation of potential breath biomarkers for the early diagnosis of breast cancer using gas chromatography-mass spectrometry Clin. Chim. Acta 436 59–67
Peng G, Hakim M, Broza Y Y, Billan S, Abdah-Bortnyak R, Kuten A, Tisch U and Haick H 2010 Detection of lung, colorectal, and prostate cancers from exhaled breath using a single array of nanosensors Br. J. Cancer 103 542–51