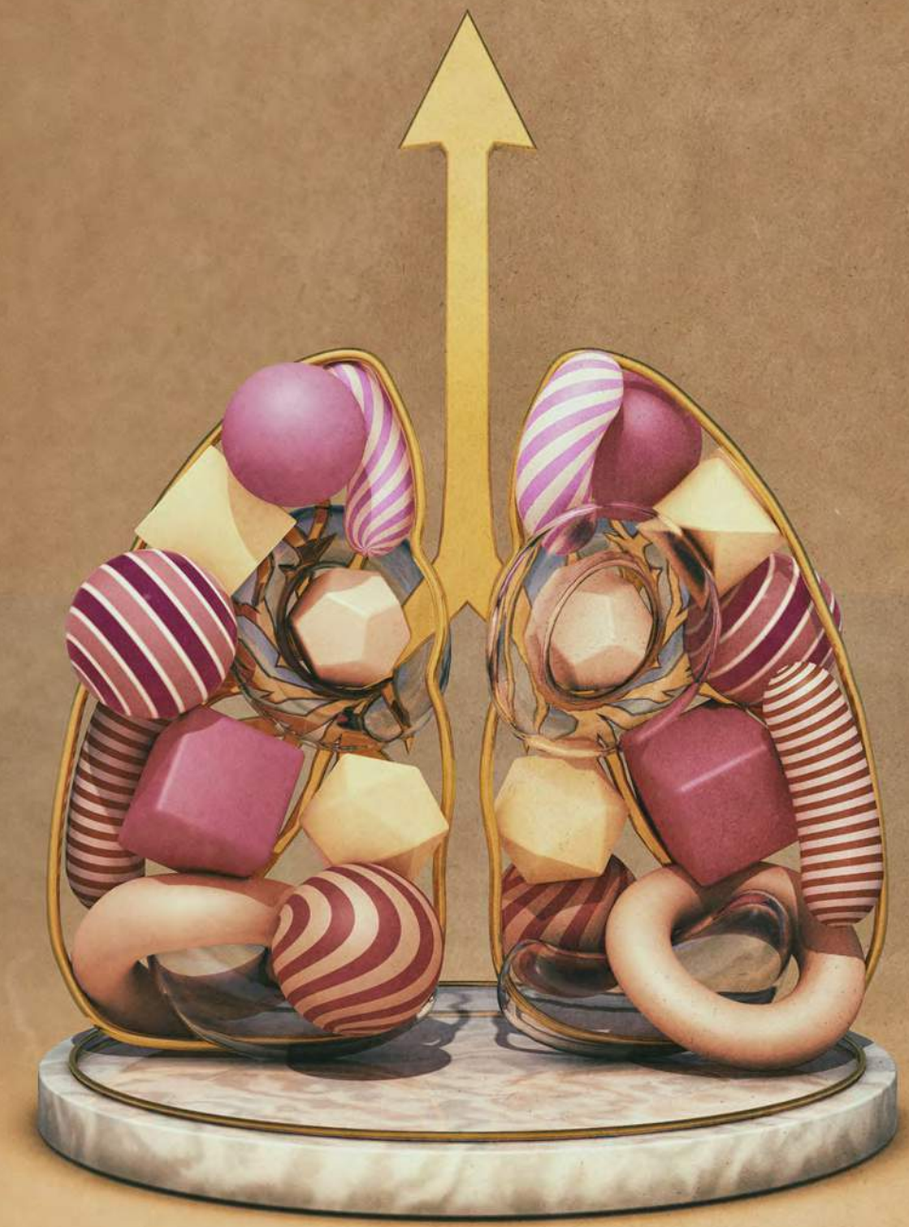
DETECTION OF LUNG CANCER IN EXHALED BREATH WITH ELECTRONIC NOSE TECHNOLOGY

SHARINA KORT

# DETECTION OF LUNG CANCER IN EXHALED BREATH WITH ELECTRONIC NOSE TECHNOLOGY



SHARINA KORT

This dissertation has been approved by:

| | |
|---|---|
| Supervisor: | Prof. Dr. J. van der Palen |
| Co-supervisors: | Dr. M.G.J. Brusse-Keizer |
| | Dr. J.H. Schouwink |

Detection of lung cancer in exhaled breath with electronic nose technology.

Department of Respiratory Medicine, Medisch Spectrum Twente, Enschede, the Netherlands

Department of Research Methodology, Measurement, and Data-analysis, Faculty of Behavioural Sciences, University of Twente, Enschede, the Netherlands

Thesis, University of Twente, 2022

ISBN: 978-90-365-5452-7
DOI: 10.3990/1.9789036554527
Artwork: Jody Venema – Digital Donut
Printed by: Gildeprint, Enschede

The publication of this thesis was financially supported by Medical School Twente, Medisch Spectrum Twente and The eNose Company, Zutphen.

# DETECTION OF LUNG CANCER IN EXHALED BREATH WITH ELECTRONIC NOSE TECHNOLOGY

# DISSERTATION

to obtain

the degree of doctor at the University of Twente,

on the authority of the rector magnificus,

prof.dr.ir. A. Veldkamp,

on account of the decision of the doctorate board,

to be publicly defended

on Friday the 7th of October 2022 at 16.45 hours

by

**Sharina Kort**

born on the 27th of December 1991

in Enschede, the Netherlands

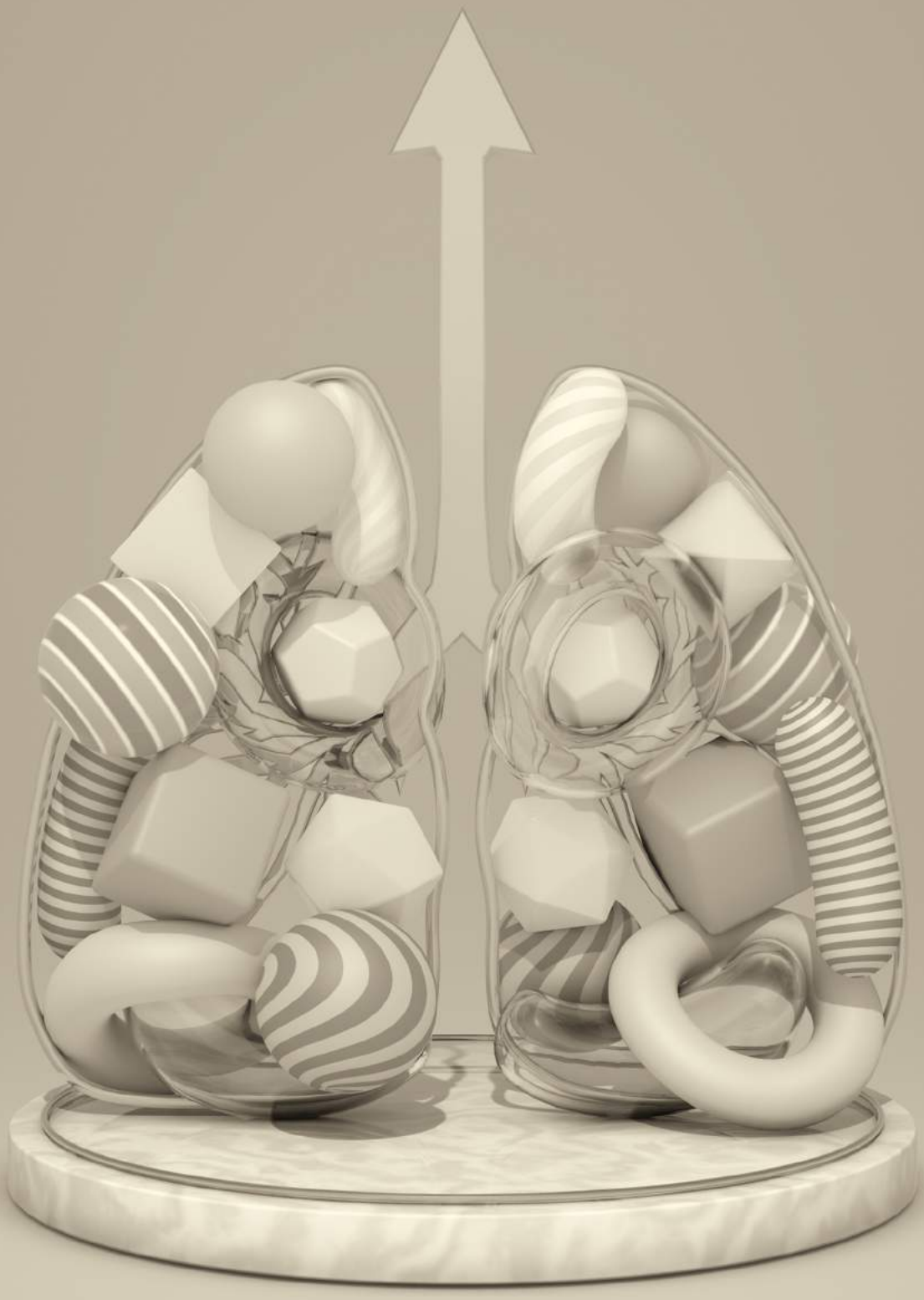
## Promotiecommissie

| | | |
|---|---|---|
| **Voorzitter/secretaris** | Prof. Dr. T. Bondarouk | Faculty of Behavioural Sciences, University of Twente |
| **Promotor** | Prof. Dr. J. van der Palen | Medisch Spectrum Twente/ Universiteit van Twente |
| **Co-promotores** | Dr. J.H. Schouwink | Medisch Spectrum Twente |
| | Dr. M.G.J. Brusse-Keizer | Medisch Spectrum Twente / Universiteit van Twente |
| **Overige leden** | Prof. Dr. P. Baas | Leiden UMC / NKI-AVL |
| | Prof. Dr. Ir. M.J.A.M. Van Putten | Medisch Spectrum Twente / Universiteit van Twente |
| | Prof. Dr. G.C.M. Kusters | Universiteit van Twente |
| | Prof. Dr. G. Sonke | Amsterdam UMC / NKI-AVL |
| | Prof. Dr. M.M. van den Heuvel | Radboud UMC |
| | Dr. P. Brinkman | Amsterdam UMC |

## Contents

# 01

# GENERAL INTRODUCTION AND OUTLINE OF THE THESIS



S. KORT

## Lung cancer – Impact and diagnosis

Lung cancer is the leading cause of cancer-related death worldwide, accounting for approximately 5% of total mortality in many countries (1, 2). The main types of lung cancer are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), accounting for 15% and 85% of the established cases respectively. NSCLC can be subdivided into two major subtypes: squamous cell carcinoma (SCC) and adenocarcinoma (AC), which differ in clinical, radiological, and histological characteristics (3). Despite substantial improvements in treatment options, such as rapidly evolving developments in targeted therapies, immune therapy, and surgical procedures, its high mortality rate reflects the large proportion of patients that present with advanced-staged disease (>50%), which is not curable. The 5-year survival rate for localized stage non-small cell lung cancer approximates 60%, whilst the 5-year survival rate for metastatic disease equals 5%. In case of SCLC, 5-year survival rate for localized disease approximates 30%, whilst the 5-year survival rate for metastatic disease conforms 3%. Therefore, an essential step to reduce lung cancer mortality is early detection through non-invasive, point-of-care diagnostic strategies (1). Small cell lung cancer (SCLC) substantially differs from non-small cell lung cancer, i.e. SCLC is characterized by a rapid growth rate, early regional and distant dissemination, and high sensitivity to chemo- and radiotherapy, albeit temporary (4). Despite significant improvements in the treatment options for NSCLC patients, unfortunately this is not yet the case for the small cell lung cancer field. The most important, internationally accepted staging system to characterize the extent of the lung cancer is the Tumor, Node, Metastasis (TNM) system as issued by the IASLC, where staging depends on the size of the tumor (T), involvement of lymph nodes (N), and presence of distant metastases (M) (5, 6). These features, combined as one disease stage at time of initial diagnosis, correlate with survival, and determine treatment recommendations (5).

In case a subject is suspected of lung cancer, based on symptoms or abnormal imaging, a diagnostic path is initiated to agree upon a final diagnosis and to determine the most optimal treatment for this individual patient. Due to the improvements in directed treatment options, current guidelines emphasize the necessity of proper tissue sampling, i.e. invasive diagnostics, in order to provide optimal personalized treatment for a patient (7). Tissue sampling can either be performed by bronchoscopy in case of a central endobronchial lesion, by endobronchial ultrasound for central lesions or lymph node sampling, or by transthoracic biopsy for peripheral intrapulmonary lesions. Besides, in case of a suspected distant metastasis, an ultrasound-guided or CT-guided biopsy from a liver, brain, or bone lesion can be acquired. In case these diagnostic techniques do not allow for a definite diagnosis, more invasive techniques, such as mediastinoscopy or video-assisted thoracoscopic surgery (VATS) are performed. However, all invasive diagnostics involve a certain risk of complications including pneumothorax, local bleeding, infection, and death (8-10). Therefore,

there is an increasing demand for innovative, non-invasive, point-of-care diagnostic tools to detect lung cancer at an early stage.

## Exhaled breath analysis

In the past decades, various non-invasive strategies have been investigated as a potential tool to diagnose lung cancer early (11-14). Over the recent years, there has been growing interest using exhaled breath in the diagnosis of different diseases. Exhaled breath contains a gas mixture mainly composed of inorganic compounds, such as carbon dioxide, water vapour, nitrogen, and inert gases. However, it also contains thousands of volatile organic compounds (VOCs) in very low concentrations that reflect metabolic processes in the body at tissue level (15-17). Exhaled breath analysis focuses on shifts in the composition of these VOCs, indicating biochemical changes at tissue level in different (patho)physiological processes, such as infection, inflammation, and malignancy. Breath sampling and VOC detection can generally be performed in two ways. It can be performed through pattern recognition techniques, using machine learning and artificial intelligence for classification of VOC mixtures through cross-reactive, non-specific sensors in electronic noses. Alternatively, it can be performed by separation methods for specific identification of individual VOCs (e.g. gas chromatography and mass spectrometry). The first mimics human olfaction in which an odorant triggers a biochemical cascade to eventually interpret the odorant as a familiar, previously recognized smell in a non-invasive way (Figure 1). However, the use of smell as a diagnostic aid is not completely innovative. It has been known since ancient times when Hippocrates already mentioned the additional value of smell in his work 'Aphorisms', written in 400 BC (18). Much later, it was Pauling in 1971 who described the presence of VOCs in exhaled-breath (19). Ever since, various electronic nose devices with innovative sensor technologies, and improved classification techniques have been developed. In the past years, both types of breath sampling techniques have shown promising results in pilot studies to diagnose lung cancer and other conditions (20-28).

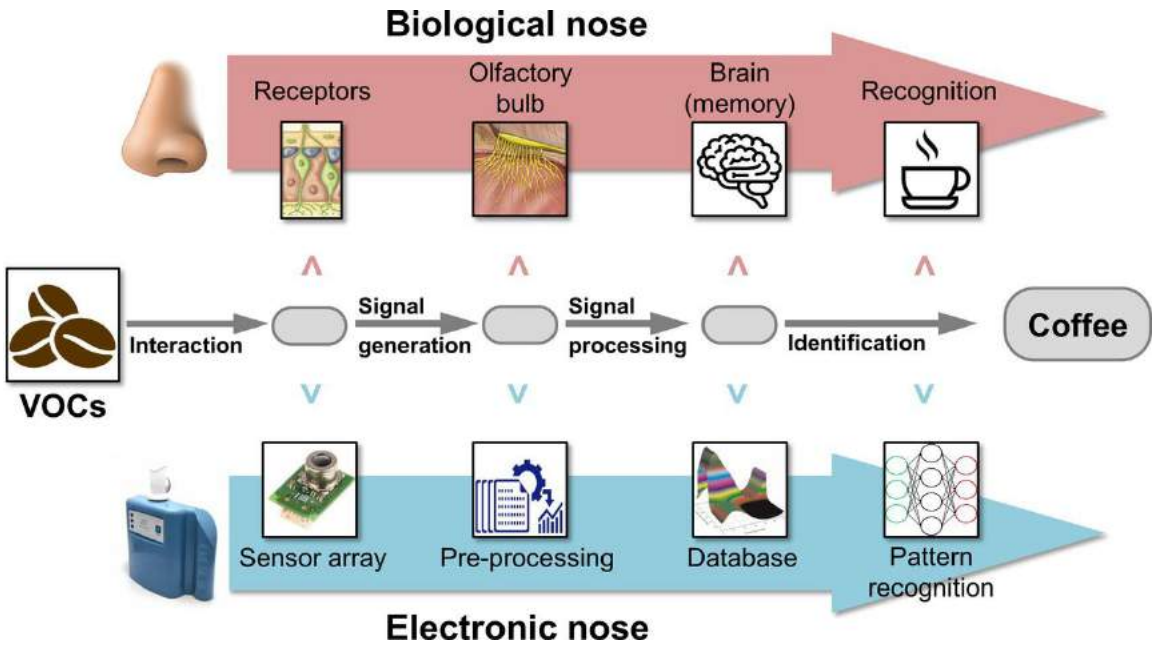



**Figure 1**. Schematic overview of the working mechanism of the Aeonose™ mimicking human olfaction. Reference: Santos et al. Electronic noses Applications in Beer Technology. Chapter 9, July 2017.

## Aeonose™ technology

The Aeonose™ (the eNose Company, Zutphen, the Netherlands) is a handheld electronic nose device featuring an array of three metal oxide sensors that offer the opportunity for real-time breath analysis (Figure 2). The metal-oxide sensors are mass producible, vary in terms of metal-oxide type and catalyst, and, when temperature-controlled properly, enable simple transfer of calibration models between Aeonose™ devices (29). When exposed to VOCs at higher temperatures, redox reactions can occur leading to consecutive conductivity changes that result into a digital exhaled breath profile consisting of conductivity values.

**Figure 2.** Aeonose™ device.

## Breath sampling procedure

A measurement with the Aeonose™ comprises 5 minutes of tidal breathing through the non-rebreathing Aeonose™ device (29). The Aeonose™ is equipped with a disposable mouthpiece containing an active carbon filter to filter inhaled air and a HEPA-filter to prevent contaminating the interior of the device. Besides, the subject's nose is clipped to prevent nose breathing and subsequently entrance of unfiltered, environmental air. During the first 2 minutes of use only rinsing of the lungs takes place with air guided through the active carbon filter. No measurements are recorded then. In the next 3 minutes, the metal-oxide sensors are exposed to exhaled breath and conductivity values are recorded. After 5 minutes, the device is put aside, after which the sensors are regenerated by guiding clean air through a second carbon active filter. Then, an internal Tenax-tube that collected VOCs during the measurement is heated, whilst VOCs are released and recorded. Redox reactions of VOCs at the sensor surfaces are recorded as conductivity changes and subsequently quantified and displayed as a unique breath signal. A complete breath test cycle encompasses 15 minutes in total.

The Aeonose™ device has been investigated in various studies involving a broad spectrum of diseases. Not only lung cancer has shown the ability to be detected by exhaled breath, among others: Barrett's oesophagus, colorectal cancer, head-and-neck cancer, prostate cancer, pancreatic adenocarcinoma, respiratory infections in COPD, multiple sclerosis, and tuberculosis have shown the potential in training studies to be diagnosed by exhaled-breath analysis based on pattern recognition techniques (27, 30-36).

## Development and validity of prediction models

Data analysis of the multidimensional breath data obtained, is executed by Aethena, a proprietary software package from The eNose Company, incorporating data pre-processing, data compression, machine learning algorithms for data classification, internal validation techniques (leave-10%-out cross validation and bootstrapping), and model selection. Initially, only artificial neural network (ANN) was used as a classification technique to analyse breath data. Neural networks use multiple layers of calculations to imitate how the human brain interprets information (37). However, in the course of time, the repertoire of machine learning techniques has been enhanced with additional machine learning techniques such as Support Vector Machine (SVM), Random Forest (RF), XG Boost, and logistic regression, all based on supervised learning of a large amount of data where the subject's outcome is pre-identified to construct a prediction model.

Evaluation of classification performance (also called accuracy) of a diagnostic tool or prediction model is important to determine how well this tool or model distinguishes between subjects with and without the condition of concern, compared to the reference test or gold standard. By providing an evidence-based evaluation of the accuracy of the diagnostic test, clinicians can carefully consider which diagnostic test to subject a patient to, and how to interpret the results of the test.

**Table 1. 2x2 cross table for a diagnostic test**

| | Reference standard test result positive | Reference standard test result negative | |
|---|---|---|---|
| Observed result index test positive | True positive (TP) | False positive (FP) | Total positive test results index test |
| Observed result index test negative | False negative (FN) | True negative (TN) | Total negative test results index test |
| | Total positive results reference standard test | Total negative results reference standard test | Total number of observations |

Important terms to assess diagnostic performance (Table 1):

| | |
|---|---|
| ***Sensitivity (true positive rate):*** | TP / (TP + FN) |
| ***Specificity (true negative rate):*** | TN / (FP + TN) |
| ***Positive predictive value:*** | TP / (TP + FP) |
| ***Negative predictive value:*** | TN / (FN + TN) |
| ***Accuracy:*** | (TP + TN) / (TN + FP + FN + TN) |
| ***Receiver operating characteristic curve:*** | Summary statistic to numerically and graphically represent the performance of an algorithm for a |

binary classification (see Figure 3).

***Area under the receiver operating characteristic curve:*** Measure of the ability of a classifier to distinguish between classes, i.e. summary of the ROC-curve.

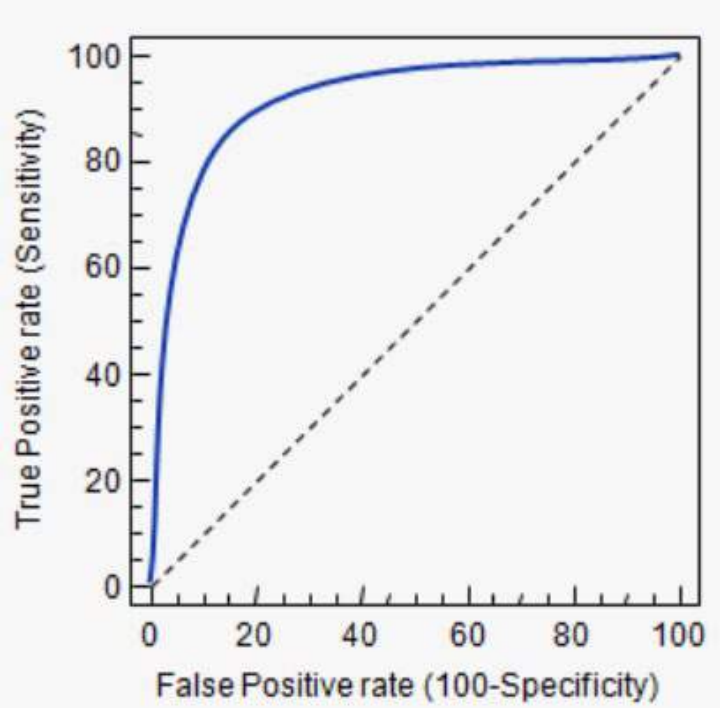



**Figure 3.** Receiver Operating Characteristic curve (Reference: www.medcalc.org)

An ideal diagnostic test is characterized by a high sensitivity and high specificity in which none of the subjects having the condition are missed, and none of the "healthy" subjects are incorrectly classified as having the condition. Considering lung cancer being a condition with a high mortality rate when not promptly recognized, a tool to diagnose lung cancer primarily has to be characterized by a high sensitivity and high negative predictive value in order to miss as few cases as possible, and to exclude negatively tested subjects safely from further invasive diagnostics. Besides, subjects without the condition should, whenever possible, be prevented from undergoing unnecessary interventions. Therefore, a minimum of specificity should also be considered in evaluating the diagnostic test. The ROC-curve plots the sensitivity and specificity, whereupon a threshold can be determined that is relevant for clinical practice.

The various machine learning techniques develop thousands of prediction models, eventually one being considered the most optimal model for further evaluation based on relevant characteristics. Diagnostic performance of this prediction model is assessed on collected breath data as a training or pilot model. However, since the proof of the pudding is in the eating, validation of the prediction model, i.e. determining the accuracy of a prediction model in a completely new, independent cohort with a varying case-mix, that has not yet been exposed to the trained model, is fundamental. This

validation tackles the risk of overfitting data and assures reliability of previous study findings and generalisability to the overall population. Nevertheless, this ideal design for external validation is not always feasible due to the fact that it concerns a time-consuming process or rapidly evolving techniques in which the diagnostic tool investigated has already been improved or innovated.

### Clinical parameters

Various clinical characteristics of subjects are known to be related to an increased risk of developing lung cancer. These clinical parameters have extensively been studied in the past in epidemiological, multivariate studies to: 1) construct clinical tools for lung cancer risk prediction; 2) determine adequate screening criteria for lung cancer screening programmes; and 3) incorporate in guidelines on follow up of solitary pulmonary nodules, such as the Brock and Herder model (38, 39). Important clinical parameters indicating a higher risk of developing lung cancer are male sex, an active or former smoking status and a higher number of pack-years smoked, higher age, presence of emphysema, and a positive family history. Addition of biomarker assays, such as exhaled-breath data to these easily available clinical parameters might improve the diagnostic performance to diagnose lung cancer.

The aim of the research presented in this thesis is to analyse and validate exhaled breath analysis, based on pattern recognition techniques with electronic nose technology, as a non-invasive diagnostic method to diagnose lung cancer.

## Outline of the thesis

In **Chapter 2** the proposed study design to train the Aeonose™ as a diagnostic tool to diagnose lung cancer is described in detail, mainly focusing on the technical working mechanism of the device, as well as the statistical analyses in the 'black box' of the Aeonose™ measurements to classify subjects as having lung cancer or not based on multidimensional data.

**Chapter 3** outlines a proposed stepwise design to simultaneously develop and validate prediction models based on machine learning techniques, involving datasets with a large number of data. In this chapter the relevance of the study design as proposed in specific situations is discussed, where a study design regarding true external validation is possibly inefficient. The proposed study design is demonstrated with our previous performed study outlined in **Chapter 4** as an example.

**Chapter 4** shows the results of the exploratory multicentre training study performed in four hospitals in which exhaled-breath analysis based on pattern recognition techniques with the Aeonose™

distinguishes between subjects with and without lung cancer. Sub-analyses are performed where subjects without lung cancer were divided into a group that was suspected of lung cancer, but was proven negative, and a group of healthy controls matched on sex and age. Also, sub-analyses are performed on histology subtypes of non-small cell lung cancer, such as adenocarcinoma and squamous cell carcinoma, and a small subset with small-cell lung cancer (SCLC) patients.

In **Chapter 5** the original prediction model as obtained in the multicentre training study from **Chapter 4** is extended to improve the diagnostic performance to diagnose non-small cell lung cancer by adding readily available clinical parameters in two ways. In a multivariate logistic regression analysis, the classification value of the Aeonose™, as well as significant clinical parameters for the presence of lung cancer, are combined. Furthermore, clinical parameters are *a priori* added to the artificial neural network in the training process of the Aeonose™ as described in **Chapter 2**.

**Chapter 6** shows the results of the multicentre, multinational validation study in which a prediction model is trained and subsequently validated to distinguish non-small cell lung cancer patients from subjects without lung cancer based on exhaled-breath patterns. Also, similar to **Chapter 4**, clinical parameters are added to the final prediction model based on exhaled breath data only, to improve to accuracy of the diagnosis of lung cancer.

Finally, in **Chapter 7** the general discussion is outlined in which we place the main results of the performed studies in a broader context and discuss the relevance of the findings and future implications.

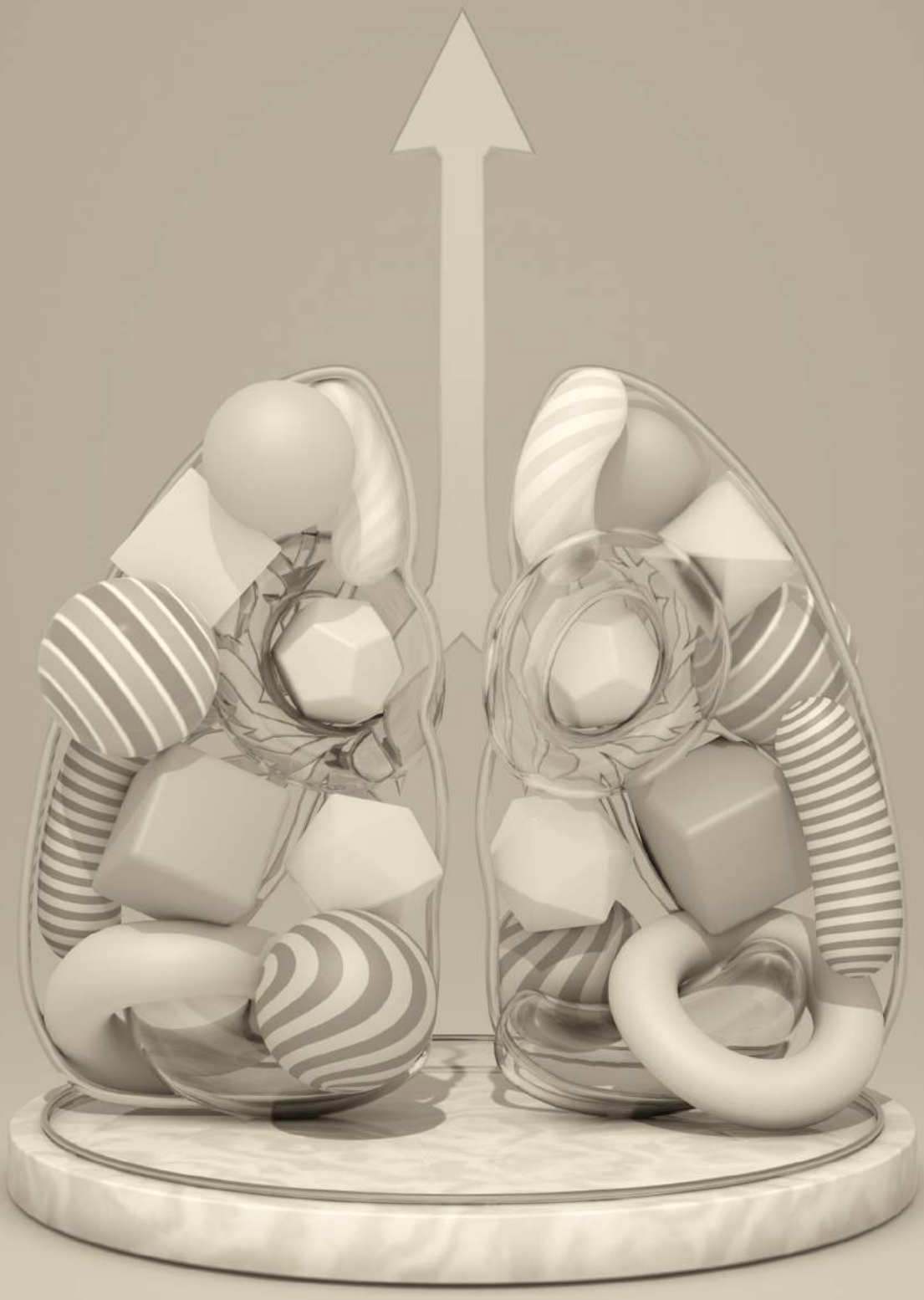
## References


1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72(1):7-33.
2. Dagenais GR, Leong DP, Rangarajan S, Lanas F, Lopez-Jaramillo P, Gupta R, et al. Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study. Lancet. 2020;395(10226):785-94.
3. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. J Thorac Oncol. 2015;10(9):1243-60.
4. Ganti AKP, Loo BW, Bassetti M, Blakely C, Chiang A, D'Amico TA, et al. Small Cell Lung Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2021;19(12):1441-64.
5. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol. 2016;11(1):39-51.
6. Rami-Porta R, Bolejack V, Giroux DJ, Chansky K, Crowley J, Asamura H, et al. The IASLC lung cancer staging project: the new database to inform the eighth edition of the TNM classification of lung cancer. J Thorac Oncol. 2014;9(11):1618-24.
7. Planchard D, Popat S, Kerr K, Novello S, Smit EF, Faivre-Finn C, et al. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2018;29(Suppl 4):iv192-iv237.
8. Khan MF, Straub R, Moghaddam SR, Maataoui A, Gurung J, Wagner TO, et al. Variables affecting the risk of pneumothorax and intrapulmonal hemorrhage in CT-guided transthoracic biopsy. Eur Radiol. 2008;18(7):1356-63.
9. Lee SM, Park CM, Lee KH, Bahn YE, Kim JI, Goo JM. C-arm cone-beam CT-guided percutaneous transthoracic needle biopsy of lung nodules: clinical experience in 1108 patients. Radiology. 2014;271(1):291-300.
10. Kano H, Kubo T, Ninomiya K, Ichihara E, Ohashi K, Rai K, et al. Comparison of bronchoscopy and computed tomography-guided needle biopsy for re-biopsy in non-small cell lung cancer patients. Respir Investig. 2021;59(2):240-6.
11. Broodman I, Lindemans J, van Sten J, Bischoff R, Luider T. Serum Protein Markers for the Early Detection of Lung Cancer: A Focus on Autoantibodies. J Proteome Res. 2017;16(1):3-13.
12. Jiang R, Dong X, Zhu W, Duan Q, Xue Y, Shen Y, et al. Combining PET/CT with serum tumor markers to improve the evaluation of histological type of suspicious lung cancers. PLoS One. 2017;12(9):e0184338.
13. Rolfo C, Russo A. Liquid biopsy for early stage lung cancer moves ever closer. Nat Rev Clin Oncol. 2020;17(9):523-4.
14. Seijo LM, Peled N, Ajona D, Boeri M, Field JK, Sozzi G, et al. Biomarkers in Lung Cancer Screening: Achievements, Promises, and Challenges. J Thorac Oncol. 2019;14(3):343-57.
15. Boots AW, van Berkel JJ, Dallinga JW, Smolinska A, Wouters EF, van Schooten FJ. The versatile use of exhaled volatile organic compounds in human health and disease. J Breath Res. 2012;6(2):027108.
16. Mansurova M, Ebert BE, Blank LM, Ibanez AJ. A breath of information: the volatilome. Curr Genet. 2018;64(4):959-64.
17. Buszewski B, Kesy M, Ligor T, Amann A. Human exhaled air analytics: biomarkers of diseases. Biomed Chromatogr. 2007;21(6):553-66.
18. Scholtz M. Hippocrates' Aphorisms. Cal West Med. 1940;52(3):125-6.
19. Pauling L, Robinson AB, Teranishi R, Cary P. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. Proc Natl Acad Sci U S A. 1971;68(10):2374-6.

20. Amal H, Leja M, Funka K, Lasina I, Skapars R, Sivins A, et al. Breath testing as potential colorectal cancer screening tool. Int J Cancer. 2016;138(1):229-36.
21. Bajtarevic A, Ager C, Pienz M, Klieber M, Schwarz K, Ligor M, et al. Noninvasive detection of lung cancer by analysis of exhaled breath. BMC Cancer. 2009;9:348.
22. Bos LD, Schultz MJ, Sterk PJ. Exhaled breath profiling for diagnosing acute respiratory distress syndrome. BMC Pulm Med. 2014;14:72.
23. D'Amico A, Pennazza G, Santonico M, Martinelli E, Roscioni C, Galluccio G, et al. An investigation on electronic nose diagnosis of lung cancer. Lung Cancer. 2010;68(2):170-6.
24. Dragonieri S, Annema JT, Schot R, van der Schee MP, Spanevello A, Carratu P, et al. An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD. Lung Cancer. 2009;64(2):166-70.
25. Gasparri R, Santonico M, Valentini C, Sedda G, Borri A, Petrella F, et al. Volatile signature for the early diagnosis of lung cancer. J Breath Res. 2016;10(1):016007.
26. Phillips M, Altorki N, Austin JH, Cameron RB, Cataneo RN, Greenberg J, et al. Prediction of lung cancer using volatile biomarkers in breath. Cancer Biomark. 2007;3(2):95-109.
27. van de Goor R, Hardy JCA, van Hooren MRA, Kremer B, Kross KW. Detecting recurrent head and neck cancer using electronic nose technology: A feasibility study. Head Neck. 2019;41(9):2983-90.
28. Lamote K, Brinkman P, Vandermeersch L, Vynck M, Sterk PJ, Van Langenhove H, et al. Breath analysis by gas chromatography-mass spectrometry and electronic nose to screen for pleural mesothelioma: a cross-sectional case-control study. Oncotarget. 2017;8(53):91593-602.
29. Bruins MGJ, Gerritsen JW, Van de Sande W, Van Belkum A, Bos A. Enabling a transferable calibration model for metal-oxide type electronic noses. Sensors and Actuators B: Chemical. 2013;188:1187-95.
30. Peters Y, Schrauwen RWM, Tan AC, Bogers SK, de Jong B, Siersema PD. Detection of Barrett's oesophagus through exhaled breath using an electronic nose device. Gut. 2020;69(7):1169-72.
31. Saktiawati AMI, Stienstra Y, Subronto YW, Rintiswati N, Sumardi, Gerritsen JW, et al. Sensitivity and specificity of an electronic nose in diagnosing pulmonary tuberculosis among patients with suspected tuberculosis. PLoS One. 2019;14(6):e0217963.
32. Steenhuis EGM, Schoenaker IJH, de Groot JWB, Fiebrich HB, de Graaf JC, Brohet RM, et al. Feasibility of volatile organic compound in breath analysis in the follow-up of colorectal cancer: A pilot study. Eur J Surg Oncol. 2020;46(11):2068-73.
33. Uslu HI, Dolle AR, Dullemen HM, Aktas H, Kolkman JJ, Venneman NG. Pancreatic ductal adenocarcinoma and chronic pancreatitis may be diagnosed by exhaled-breath profiles: a multicenter pilot study. Clin Exp Gastroenterol. 2019;12:385-90.
34. van Geffen WH, Bruins M, Kerstjens HA. Diagnosing viral and bacterial respiratory infections in acute COPD exacerbations by an electronic nose: a pilot study. J Breath Res. 2016;10(3):036001.
35. Waltman CG, Marcelissen TAT, van Roermund JGH. Exhaled-breath Testing for Prostate Cancer Based on Volatile Organic Compound Profiling Using an Electronic Nose Device (Aeonose): A Preliminary Report. Eur Urol Focus. 2020;6(6):1220-5.
36. Ettema AR, Lenders M, Vliegen J, Slettenaar A, Tjepkema-Cloostermans MC, de Vos CC. Detecting multiple sclerosis via breath analysis using an eNose, a pilot study. J Breath Res. 2021;15(2).
37. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. 2018;284(6):603-19.
38. Baldwin DR, Callister ME, Guideline Development G. The British Thoracic Society guidelines on the investigation and management of pulmonary nodules. Thorax. 2015;70(8):794-8.
39. Herder GJ, van Tinteren H, Golding RP, Kostense PJ, Comans EF, Smit EF, et al. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. Chest. 2005;128(4):2490-6.

# 02

# DATA ANALYSIS OF ELECTRONIC NOSE TECHNOLOGY IN LUNG CANCER: GENERATING PREDICTION MODELS BY MEANS OF AETHENA



S. KORT - M. BRUSSE-KEIZER - J.W. GERRITSEN - J. VAN DER PALEN

J BREATH RES. 2017;11(2):026006

## Data analysis of electronic nose technology in lung cancer: Generating prediction models by means of Aethena.

Sharina Kort [(1)]; Marjolein Brusse-Keizer [(2)]; Jan-Willem Gerritsen [(3)]; Job van der Palen [(2,4)].

1: Department of Pulmonary Medicine, Medisch Spectrum Twente, Enschede, the Netherlands

2: Medical School Twente, Medisch Spectrum Twente, Enschede, the Netherlands

3: The eNose Company, Zutphen, The Netherlands

4: Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, the Netherlands

## Abstract

**Introduction:** Only 15% of lung cancer cases present with potentially curable disease. Therefore, there is large interest in a fast, non-invasive tool to detect lung cancer earlier. Exhaled breath analysis by electronic nose technology measures volatile organic compounds (VOC's) in exhaled breath which are associated with lung cancer.

**Methods:** The diagnostic accuracy of the Aeonose™ is currently being studied in a multi-centre, prospective study in 210 subjects suspected for lung cancer, where approximately half will have a confirmed diagnosis and the other half will have a rejected diagnosis of lung cancer. We will also include 100-150 healthy control subjects. The eNose Company (provider of the Aeonose™) uses a software program, called Aethena, comprising pre-processing, data compression and neural networks to handle big data analysis. Each individual exhaled-breath measurement comprises a data matrix with thousands of conductivity values. This is followed by data compression using a Tucker-3-like algorithm, resulting in a vector. Subsequently, model selection takes place after entering vectors with different presets in an Artificial Neural Network to train and evaluate the results. Next, a "judge model" is formed which is a combination of models for optimizing performance. Finally, two types of cross-validation, being 'leave-10%-out' cross-validation and 'bagging', are used when recalculating the judge models. These judge models are subsequently used to classify new, blind measurements.

**Discussion:** Data analysis in eNose technology is principally based on generating prediction models which need to be validated internally and externally for eventual use in clinical practice. This paper describes the analysis of big data, captured by eNose technology in lung cancer. This is done by means of generating prediction models with Aethena, a data analysis program especially developed for analyzing VOC data.

## Introduction

Lung cancer is the leading cause of cancer death among males and females worldwide, accounting for approximately 5% of total mortality in many countries (1). Lung cancer is not a well-defined single entity. It is a heterogeneous disease, arising in many different clinical pathological patterns. The World Health Organization classification recognizes 20 different types of malignant lung neoplasms (2). The main types of lung cancer are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) where the latter can be subdivided into three major histological types: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Chest radiography and computed tomography (CT), which are considered non-invasive diagnostic techniques, are the first steps in the diagnostic work-up to detect and stage lung cancer. Histopathological diagnosis following an invasive bronchoscopic intervention still remains the gold standard to prove or rule out the diagnosis of lung cancer. However, this investigation is accompanied by associated risks and substantial costs, which makes it not suitable for population-based screening.

The diagnosis of early stage lung cancer is essential for curative therapy by means of surgery and substantially determines life expectancy (3). Five-year survival for those with pathological stage IA non-small cell lung cancer is 73%, whereas metastatic disease has a miserable prognosis with a five-year survival of merely 13% (4;5). Unfortunately, only 15% of the lung cancer cases present with localised, potentially curable disease, which means that the majority of the cases is diagnosed in an advanced stage with consequently poor survival rates.

There has been a lot of interest in secondary prevention involving screening tests for the detection of early-stage lung cancer. Screening tests using sputum cytology and chest radiography have been attempted with unfortunately limited success (6). Although low-dose computed tomography (LDCT) is able to detect early-stage lung cancers (3), in practice it does not sufficiently demonstrate a survival benefit, reduce the incidence of advanced stage cancers or reduce lung cancer mortality (7;8). The observed increased survival time with screening can be overestimated due to lead time bias, when survival time is measured from the time of diagnosis. But also, length bias can give an overestimation of survival duration among screening detected cases by the relative excess of slowly progressing cases. These cases are disproportionately identified by screening because the probability of detection is directly proportional to the length of time during which they are detectable (and thereby inversely proportional to the rate of progression). Furthermore, maybe less important in lung cancer, overdiagnosis bias can play a role in screening research, which could lead to overestimation of survival duration among screen-detected cases caused by inclusion of pseudo disease—subclinical disease that would not become overt before the patient dies of other causes

(9;10). However, there are several lung cancer screening trials by means of CT-scanning ongoing with some optimistic results (11-15), but these results still are insufficient for screening to be incorporated in clinical practice since high numbers needed to screen and a large number of false positives continues to question cost-effectiveness especially concerning determining the definition of the screening population and the screening frequency (16-19). Hence, there is large interest in a fast, simple, cost-effective and non-invasive tool to detect lung cancer at an early stage, preferably during a visit at the general practitioner.

This has led to the introduction of exhaled breath analysis by means of electronic nose technology. This diagnostic approach seems very promising in the lung cancer field, though it is yet far from being incorporated in clinical practice (20-23).

The concept of an electronic nose is based on the availability of powerful personal computing making it possible to apply pattern recognition techniques to complex measurement data. The desire is to have a general, broadly responsive sensor system that generates complex multidimensional measurement data and uses pattern recognition techniques to match measured response patterns to previously observed response patterns in order to identify specific scents present within complex mixtures. This is analogous to the physiology of the human smell, where the brain combines received signals and determines what characteristic scent pattern is smelled, but doesn't distinguish specific components. Hence, the name 'electronic nose'.

Electronic nose technology is based on the usability of volatile organic compounds (VOC's) in exhaled breath. Exhaled breath is mainly composed of inorganic compounds, such as nitrogen, oxygen, carbon dioxide, water vapour and inert gases. In addition, it contains thousands of VOC's, which are exhaled in very low concentrations, but reflect pathological processes, such as inflammation, oxidation, infection and neoplasms, where they can serve as non-invasive biomarkers for certain diseases (24). The perspective is that metabolic and biochemical processes that occur in different pathological situations cause different endogenous VOCs to arise, which can be detected with different chemical sensors and can therefore be promising disease biomarkers. All these methods are directed at measurable changes in physical properties of the sensors when being exposed to a gas mixture.

However, the use of VOCs in electronic nose technology is only one method. There are several other methods utilized for breath sampling, like Multi-Capillary Column-Ion Mobility Spectrometry or Gas Chromatography-Mass Spectrometry that look for specific compounds in exhaled air (25-28). Contrary to determining VOCs in exhaled breath, these techniques do not apply pattern recognition techniques, since they are aimed at identifying individual molecules in exhaled breath instead of a

unique composite breath signal. Recently, Schallschmidt et al. published results of an observational study on the profiles of volatile organic compounds where they showed that the use of solid phase microextraction-Gas Chromatography-Mass Spectrometry is not reliable enough to discriminate between cancer patients and healthy controls (29). An important remark they make relates to the limited capability of current analytical procedures to detect unstable marker candidates.

The use of human breath as a diagnostic tool is not completely innovative. The use of smell as a diagnostic aid has been known since ancient times when Hippocrates mentioned the diagnostic value of smell in his work 'Aphorisms' which was written in 400 BC (30)**.** However, it was Pauling who described in 1971 the presence of VOC's in exhaled breath that this method became of great scientific interest (31). Over the last few decades, several electronic nose devices have been developed, which contain different sensors to detect the VOC's and generate a quantifying measure for these VOC's. A lot of research has been performed with the Cyranose 320, and analyses performed by Machado et al. and Dragonieri et al. provided some promising results in the lung cancer field (26;32). Also promising was the gold particle nanosensor developed by Peng et al (33). Peled et al showed an accuracy of the nanoarray in discriminating between malignant and benign pulmonary disease of 88% with an area under the curve (AUC) of 0.986 (34). However, these results are based on a small study population (n=69) without performed external validation.
In this manuscript, the Aeonose™, developed by The eNose Company (Zutphen, The Netherlands) will be discussed. The Aeonose™ differs from other electronic nose devices that it offers the opportunity for transferring calibration models and therefore enables large-scale application (35).

An important aspect of the electronic nose concept is that a substance, or a mixture of substances (VOC's), can only be recognized after a calibration phase, i.e. the pattern must be known beforehand ('seen' before). This is why the electronic nose must be trained and a database of patterns, called breath prints, must be developed. This searchable, digital database systematically stores previous measurements with characteristic scent patterns. In this way, new scent patterns can be matched with an existing scent profile through comparative pattern recognition analysis.

When comparing breath patterns between subjects diagnosed with and without a certain disease, the eNose can be trained to distinguish between these two groups. In this way, a new diagnostic device can be developed for screening or diagnosing diseases based on people's exhaled breath.

The aim of this manuscript is to describe our study concerning the detection of lung cancer with the Aeonose™, where we will focus on the statistical analysis in the 'black box' of the Aeonose™ measurements for classifying whether lung cancer is present or not.

Objectives

The main objective of this study is training the Aeonose™ to build a database for recognition for the detection of lung cancer. This study aims to investigate the diagnostic accuracy of exhaled breath analysis with the Aeonose™ to distinguish breath of subjects suspected for lung cancer who are truly diagnosed with lung cancer from subjects suspected for lung cancer in which this diagnosis is rejected after histopathological diagnosis following a bronchoscopic intervention. The obtained patterns will also be compared with breath patterns of healthy subjects who are not suspected for lung cancer. Additionally, we will investigate whether the Aeonose™ recognizes patterns between different types of lung cancer (NSCLC vs. SCLC) and between different lung cancer stages.

## Material and Methods

Design

It concerns a multi-centre, prospective, non-invasive study in subjects suspected for lung cancer, who are referred for a histological biopsy through bronchoscopy. Subjects who are suspected for lung cancer will be compared in a cross-sectional design, where breath patterns from those who are truly diagnosed with lung cancer are compared to those where this diagnosis is rejected. Also, breath patterns of healthy subjects will be compared with confirmed and rejected lung cancer cases. It concerns a single measurement in the pulmonology departments of Medisch Spectrum Twente Enschede, Ziekenhuis Bernhoven Uden, Medisch Centrum Leeuwarden, and Deventer Ziekenhuis, all in the Netherlands.

Study population

Adult subjects who have a scheduled visit at the outpatient clinic of the pulmonology departments of the participating hospitals due to suspicion of lung cancer will be asked to participate. Suspected subjects will be divided in a group with a confirmed diagnosis of lung cancer and a group with a rejected diagnosis of lung cancer based on histopathology following a bronchoscopic intervention**.** Healthy subjects will be recruited from partners, relatives or friends of eligible subjects. They will be frequency matched on age and gender distribution to the subjects suspected for lung cancer. When we calculated a sample size to ensure a study with a reasonable power, we took into account a desired sensitivity of 90% with a two-sided confidence interval of 82.5% - 95%. In this way, we approximately need 105 subjects diagnosed with lung cancer. When we presume a realistic 1:1 ratio of a confirmed versus a rejected diagnosis of lung cancer in suspected subjects, we also approximately need 105 subjects with a rejected diagnosis, which gives a total of 210 suspected subjects. Given the possibility to observe a bigger contrast between suspected subjects with a

confirmed diagnosis of lung cancer and subjects not suspected for lung cancer at all, we also include 100-150 'healthy' subjects without any suspicion for lung cancer.

Inclusion criteria

Recruitment of these subjects has started in June 2015 and is expected to conclude in the winter of 2016. We aim to include a total of 210 patients where the number of patients per hospital depends on the catchment population of each hospital. From these 210 patients, based on hospital data, approximately 105 patients will have a confirmed diagnosis of lung cancer and 105 of the suspected cases will have a rejected diagnosis of lung cancer. Additionally, we aim to include 100-150 healthy subjects. This should be sufficient for training the Aeonose™ and determining whether it can reliably detect differences in breathing substances.

Suspected subjects need to meet the following criteria to be eligible:

1) Referred for a histological biopsy due to suspicion for lung cancer;

2) Age ≥ 18 years.

Eligible healthy subjects need to meet the following criterion:

1) Age ≥ 18 years.

The only exclusion criterion for all subjects is:

1) Known with an active malignancy.

In setting-up the study protocol, we tried to exclude correlated features between cases and controls as much as possible. In an exploratory analysis, however, we have noticed an (unexpected) decrease of AUC when we used supposedly healthy partner controls. This might be due to correlated features, such as similar diet and smoking behavior, or at least residing in the same indoor atmosphere. In case of a suspicion for correlated features, cluster analysis could be helpful using e.g. a software package like Carotta (36).

Aeonose™ technology

The Aeonose™ consists of three micro hotplate metal-oxide sensors (MOS) that are rigid, mass producible, and offer the opportunity for transferring calibration models. This means that once a calibration model has been developed, it can easily be transferred to other Aeonose™ devices. Several metal-oxides behave as semi-conductors at higher temperatures. The sensors vary in terms of metal oxide type and catalysing agent. Redox reactions occurring at the sensor surface result in changes in conductivity that can be measured and quantified resulting in a unique breath signal. These redox reactions depend on the type of metal oxide and catalyst, the reacting gas(es), and the temperature. A broad range of VOCs in exhaled breath will give a redox reaction.

Thermal cycling

Redox reactions are temperature dependent, and by using thermal cycling this temperature dependency can be determined as a function of time. Different VOC's show different responses at varying temperatures for the same chemical sensor type (figure 1A). The breath patterns are obtained by taking the response of a complete cycle and can be presented as a function of the temperature (figure 1B). In this way the temperature dependency of the redox reactions is acquired on a single sensor. The patterns obtained by thermal cycling do not only depend on the applied temperatures, but also on the dynamics of the temperature, because intermediate products created at the sensor surface have limited life times.

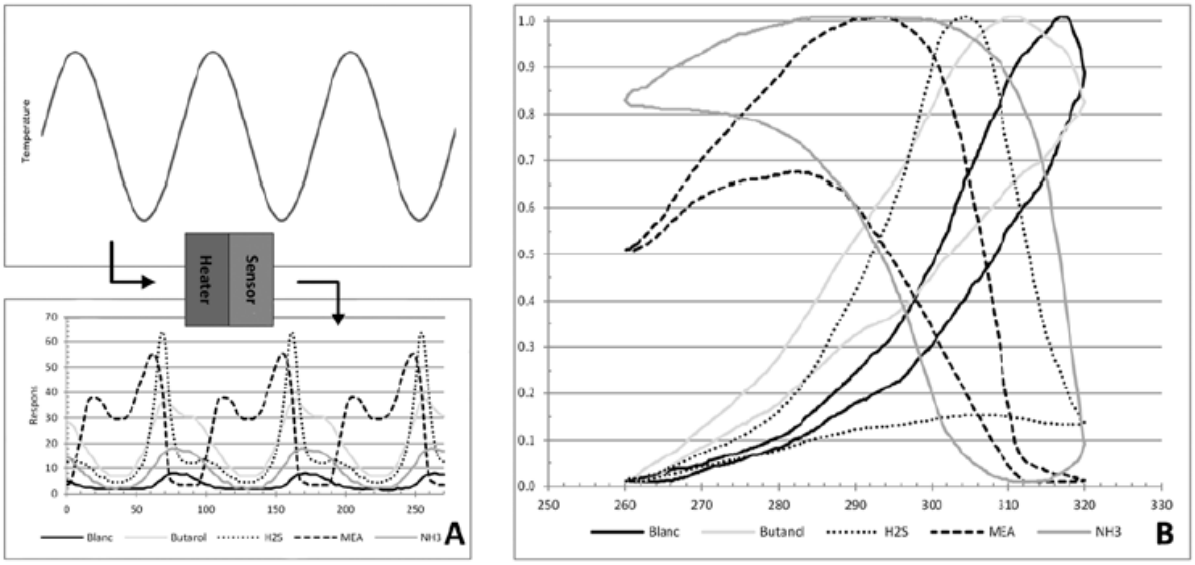



**Figure 1. A:** Measuring principle of the Aeonose™: A continuous sinusoidal temperature cycle in which the heater is thermally cycled (top) and the conductivity of the sensor (bottom) is recorded as a function of momentary temperature. The temperature profile is applied to the heater while the response is recorded at the sensor. Substances can show temperature-dependent responses for the same chemical sensor type. (Blanc = clean air, H2S = hydrogen sulphide, MEA = methylamine, NH3 = ammonia). **B**: Thermal response loops resulting from the normalized conductivity of the sensor plotted as a function of the heater temperature during a full period. (Blanc = clean air, H2S = hydrogen sulphide, MEA = methylamine, NH3 = ammonia). (Taken from: Bruins et al. Transferable odour differentiation models for infectious disease diagnostics, 2014 (48))

**Statistical analysis**

Predictive models are important tools to provide estimations of diagnostic outcomes. There are various resampling methods to estimate the performance of a model in a new sample of independent subjects (the test set) after a training set of observations has been created, i.e. these methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model. The resampling methods provide estimates of the test-set prediction error (test error) and error of the parameter estimates for future observations

(prediction error).

First, it is important to know whether a data set is either low-dimensional or high-dimensional. Low-dimensional implies that there are more subjects present than parameters in a data set (n>p). In contrast, high-dimensional implies having more parameters than subjects in a dataset (p>n). A high dimensional data set, as obtained with the Aeonose™, poses statistical challenges where too many predictors will overfit the data and result in a model that looks appropriate on the training data used to develop it, but will poorly perform on future observations from the test data. This problem of overfitting can be avoided by using a combination of analytical techniques such as data compression, cross-validation and bootstrapping. In statistics, cross-validation is a model validation technique for assessing how the results of a prediction model will generalise to a new independent data set (37;38). Bootstrapping is a useful technique to get an idea of the variability or standard deviation of an estimate and its bias (39;40).

The eNose Company uses a proprietary software for data analysis, called 'Aethena'. This package retrieves raw data from a database and takes care of data compression, data analysis and data reporting. In this section we will illustrate the methods used to obtain the best prediction models. During an exhaled breath measurement, for each sensor, 64 * 36 data points are being recorded. In this way, each individual patient measurement comprises of a data matrix with thousands of records. In the course of the data analysis and pattern recognition, the following steps can be distinguished:

Pre-processing

As mentioned before, the sensor's temperature control enables accurate reproducibility of the results. However, slight variations between sensors among Aeonoses™ can be seen. In order to cope with these variations, the data are being standardized in several ways, creating multiple representations of the same dataset.

1. Data of a measurement are scaled between 0 and 1 per measurement cycle.
2. Data of the full measurement are scaled between 0 and 1.

Data compression

As the matrix sizes are too large for classification, the data are compressed using a Tucker3-solution (41). This needs to be done to avoid the so-called spurious correlations. Spurious correlations become of greater importance since modern eNoses collect increasing amount of data. The compression results into a vector for one of the seven sensor combinations of the three metal-oxide sensors (A, B, C, AB, AC, BC and ABC). In the case of lung cancer this results in 11 components per patient in which redundant information and noise is removed, but in which information concerning the distinction between healthy and sick subjects is maintained.

We start with all subjects, called dataset A. When classifying subjects, we set aside 20% of the data in a blinded fashion in order to create a test set for external validation (dataset C, also called test set). Of the remaining 80% (dataset B, also called training set) the true lung cancer status based on pathology is known. The vectors generated in the study will be entered into an artificial neural network (ANN). Figure 2 describes the principle of an artificial neural network. There is one input layer consisting of the obtained vector in the compression phase. By means of algorithms based on trial and error the components of the input layer and hidden layer will be given different weights to determine the best output.

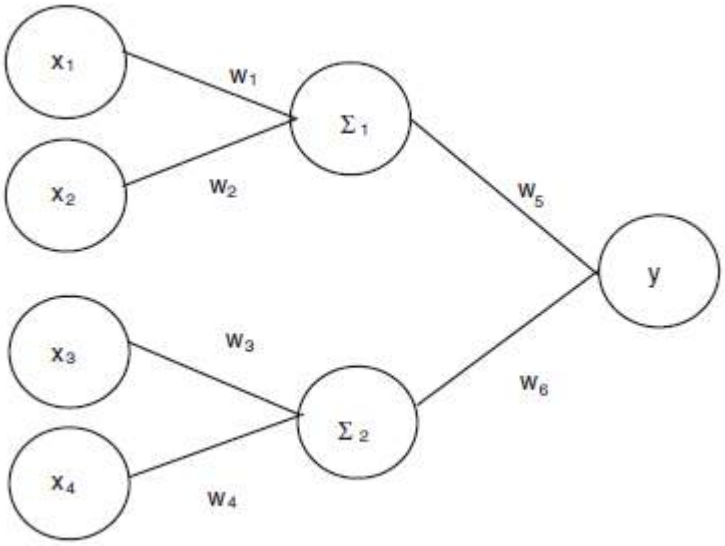



**Figure 2.** General structure of an artificial neural network (ANN). A feed-forward neural network with one input layer consisting of four nodes (*x*), one hidden layer with two nodes (*Σ*), and one output layer (*y*), The connections between the layers have associated connection strengths or weights ($w_i$),which can be varied. (Taken from: Genet. Epidemiol. 2008 May;32(4):325-40. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD) (49).

Several statistical learning methods could be applied for data classification. For optimal results all of them require fine-tuning. Up till now we have been focusing on applying neural networks. However, also other methods like Random Forest and Support Vector Machine could be applied. Actually, in another study (submitted for publication), the neural network results were compared to results obtained from Random Forest, Support Vector Machine, and Gaussian Process showing comparable AUC values. Hauschild et al. have described different classification methods as well (42). Up till now we have no compelling evidence that other classification techniques will show better results than neural networks. However, for specific diseases, it could be favorable to use other classification techniques (e.g. Random Forest). Therefore, we intend extending our software package with other classification techniques in the near future.

10-Fold cross-validation or Leave-10%-Out cross validation

Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used when one wants to estimate how accurately a predictive model will perform in practice.

10-fold cross-validation comprises 10 rounds of validation (figure 3). One round of cross-validation involves partitioning the training set (dataset B) into complementary subsets (80% (dataset D), 10% (dataset E), 10% (dataset F)), performing the analysis on dataset D, and validating the analysis on the 10% in subset E. Dataset F is used as a stop criterion in order to decide how long the model needs to be trained. To reduce variability, 10 rounds of cross-validation are performed using different partitions in such a way that after 10 rounds all data have been used once in dataset D, E, and F, and all patients are predicted once. The validation results are averaged over the 10 rounds, resulting in one combined AUC.

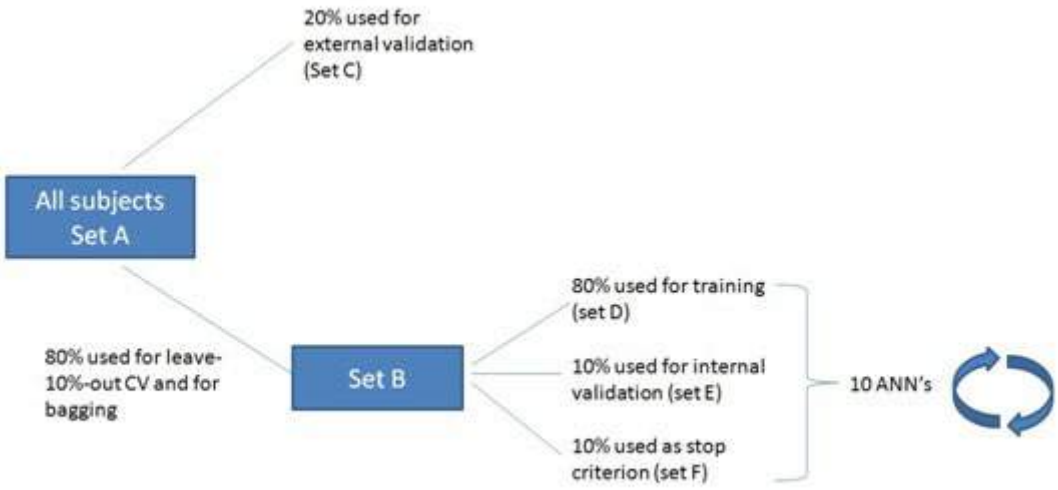



**Figure 3**. First round cross-validation.

Model selection

The process described is executed for all 7 sensor combinations and for different pre-processing techniques. In this way, a large number of possible ANN-models are being generated, each with its specific performance measures. The output consists of a list including ranked ROCs with performance calculated by means of the Area Under the Curve (AUC), sensitivity and specificity. Higher AUCs usually indicate better performance.

Subsequently, based on ranked AUC's, various models will be selected for optimizing diagnostic performance. First, a model is selected that is able to properly separate positive and negative

subjects (figure 4). Subsequently, for both negative and positive subjects, two different complementary *combined* models are constructed, based on single models, which minimize the number of false positives and false negatives. For positive subjects we use models that accurately predict positives and for negative subjects models are used that accurately predict negatives.

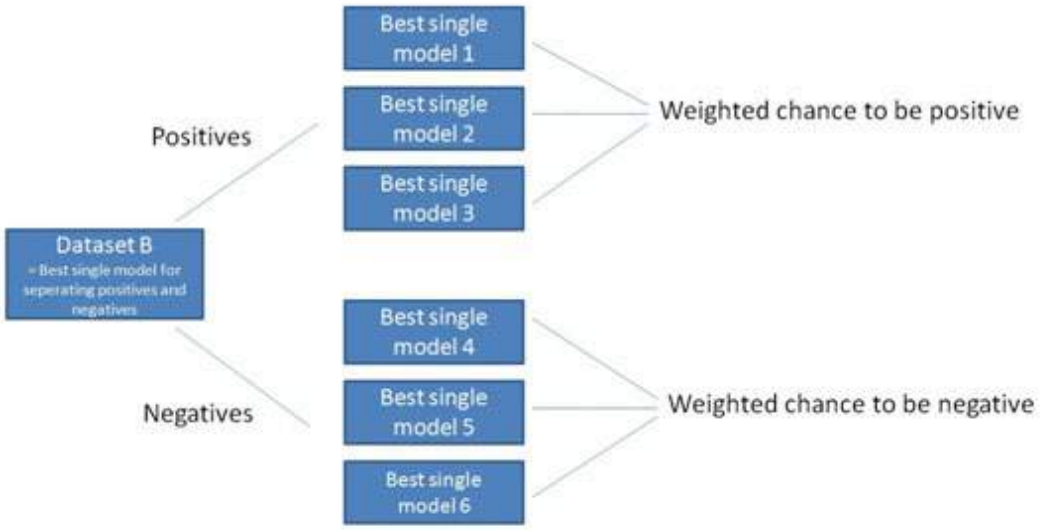



**Figure 4**. Mechanism of judge models.

A *combination* of the best models showing the smallest error is called a "judge" model. Smallest error is defined as the sum of false positives and false negatives. Every judge model gives one AUC and all models are independent from each other. The next step will be to fine-tune and validate these selected models.

Two sets of ROC plots can be constructed: at first, the neural network is being trained using samples with known classifications, and applying leave-10%-out cross validation. The results can be represented in a ROC plot that should be representative for blind samples as well because of the cross validation process. Secondly, blind samples are classified using the trained neural network. When these classifications are compared with the gold standard results, threshold dependent confusion matrices can be constructed followed by a corresponding ROC plot. If the blind samples have similar characteristics as the training set, the ROC curves of training set and blind samples can be expected being almost identical.

Subsequent cross-validation

In our analysis, two types of cross-validation techniques are used: 'Leave-10%-Out' and 'Bagging' (bootstrap aggregation) (43;44). When using the 'Leave-10%-Out' method, the selected single models and obtained judge models are recalculated as previously described. However, fine-tuning of

the ANN's is being applied for optimal performance where new weights for every model are calculated, which means that the ANN's are generated a few more times from scratch to determine whether the ANN's are stable; i.e. whether comparable ROC-curves are derived. However, the input is not the full dataset. Only the positives from the first separation are entered in the upper models and only the negatives from the first separation are entered in de lower models, which eventually lead to one AUC.

Bagging is an alternative cross-validation technique to provide stable networks. From dataset B, a random sample of measurements is chosen, used for training an ANN, and this sample is subsequently replaced, contrary to 10-fold cross-validation. This procedure is repeated many times (i.e. >1000). Per person a large number of calculated risks for lung cancer are derived and are averaged to one chance which is used to calculate the AUC. This obtained AUC will be compared with the obtained AUC from the 10-fold cross-validation. Finally, the best ANN's generated by bagging will be used to classify the blind measurements from dataset C, the test set.

The bagging technique is mainly used to check whether the leave-10%-out procedure succeeded and gives a smoother model fit with a better balance between potential bias and variance. An important difference compared with 10-fold cross-validation is that in bagging models are not further adapted and no judge models are constructed. The calculated weight remains constant.

**Example**

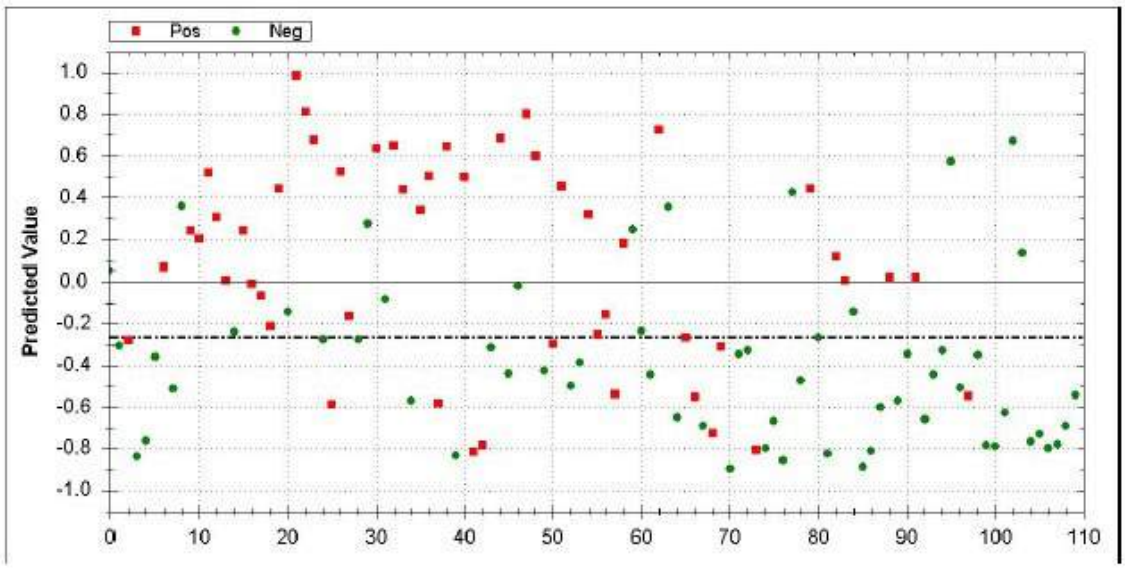



**Figure 5A.** Separation plot

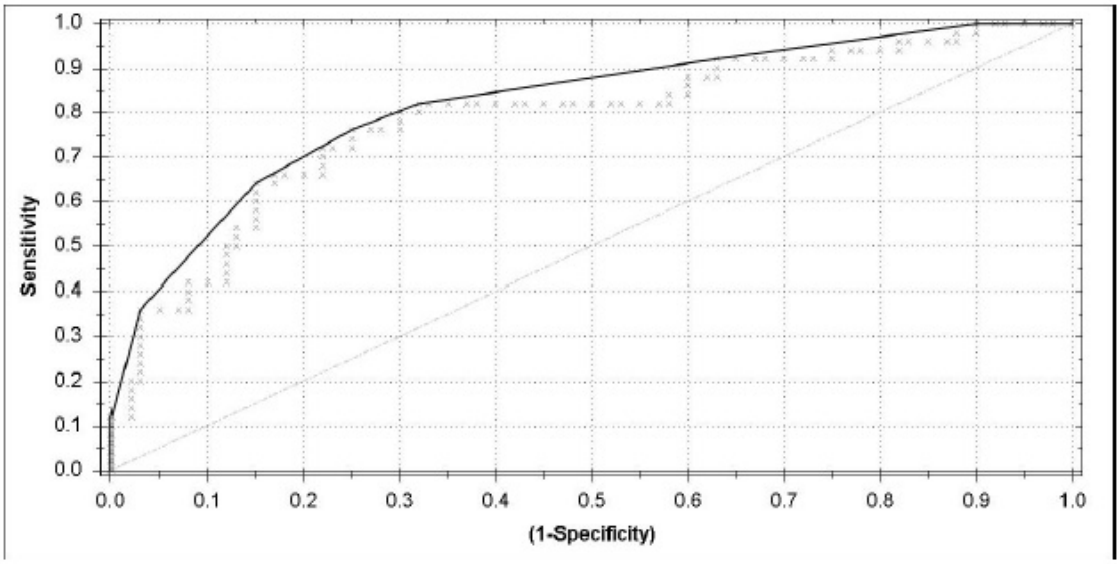



**Figure 5B.** ROC curve

Figures 5A shows a separation plot, based on training data, showing predicted values for 50 patients with lung cancer (pos) and 60 healthy controls (neg), according to the statistical procedures as described in this manuscript, representing a demonstration of the principle. Figure 5B shows the corresponding ROC curve, again based on preliminary training data.

## Discussion

Despite modest advances in the treatment of lung cancer, it remains a fatal disease with overall 5 year survival rates having not increased over a few decades (45;46). Therefore, it is of great matter to detect lung cancer at an early, potentially curable state. Screening programs concerning lung cancer have proven evidence of reducing lung cancer-specific mortality, but results must be implemented carefully. There should be a clear balance between maximizing benefits and minimizing harms with acceptable costs. As seen with lung cancer screening, the high amount of false positives involves substantial costs and therefore drives the cost-effectiveness of lung cancer screening downward. A positive CT-scan triggers additional diagnostics ranging from rather easily repeating the CT-scan to invasive diagnostics like biopsy and surgical resection. These interventions however also involve associated risks, such as morbidity and mortality from complications and high emotional stress. Therefore, the lung cancer screening field can be extended with alternative forms of diagnostics instead of just focusing on imaging techniques.

Exhaled breath analysis by means of electronic nose technology is a young field of research, but has been of great scientific interest the last few years and is a rapid emerging field of medical diagnostics. However, it has not yet been implemented in clinical practice. Several electronic noses with varying underlying technologies have been tried with some promising results, but the limited

amount external validation studies have not yet given sufficient trust in these methods. Recently, Leopold et al. have published an article concerning external validation in studies using various methods of electronic nose technology in lung cancer (47). They evaluated 46 studies regarding different approaches to dimension reduction, classification and validation in electronic nose technology. Only 7 studies performed external validation on an independent dataset with rather 4 datasets available for re-analysis. External validation resulted in a lower area under the receiver operating characteristics curve (ROC-AUC) compared to the internal validation in 2 out of 4 datasets. The other 2 datasets did not show decreased ROC-AUC's when applying external validation. However, no single combination of dimension reduction and classification methods gave consistent results between internal and external validation sets in these 4 datasets. Therefore, to show accurate diagnostic performance, it is important to estimate diagnostic performance on an independent dataset (external validation). Robustness of the models is important indeed, especially when one plans on classifying blind samples. Next to high overall AUC, we therefore also require models to show a small AUC standard deviation between the 10 consecutive steps during the 10-fold Leave-10%-Out cross validation.

The ideal diagnostic test should be both sensitive (a high percentage of sick subjects who are correctly identified as having the condition) and specific (a high percentage of healthy subjects who are correctly identified as not having the condition). This overall percentage of correctly diagnosed subjects determines the test accuracy. The results of the new diagnostic test are compared to the results of the reference test called the gold standard.

The Aeonose™ used in our study is a hand-held electronic nose device, which is convenient to use, includes non-invasiveness and gives fast results with consistent copy-and-paste between different Aeonoses™. However, possible disadvantages which need to be taken into account are the inability to differentiate between endogenous and exogenous compounds and the influence of many exogenous factors, such as smoking, diet and other scents. In this study, we investigate whether exhaled breath patterns from patients with lung cancer can be distinguished from healthy subjects. After completing the training phase with approximately 350 subjects, we should have an idea whether the Aeonose™ can reliably detect differences in breath patterns of patients with lung cancer and subjects without lung cancer. After the training phase, an external validation phase must follow with an independent group of sick and healthy subjects in a different setting.

## Conflicts of interest and Source of Funding

Prof. Dr. J van der Palen and Dr. M. Brusse report no conflict of interest. Miss S. Kort was partly financed by an unrestricted research grant from The eNose Company. Dr. J.W. Gerritsen is employed by the company producing the e-nose devices used.

## References


(1) Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin 2015 Mar;65(2):87-108.

(2) Beasley MB, Brambilla E, Travis WD. The 2004 World Health Organization classification of lung tumors. Semin Roentgenol 2005 Apr;40(2):90-7.

(3) Henschke CI, Yankelevitz DF, Libby DM, Pasmantier MW, Smith JP, Miettinen OS. Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med 2006 Oct 26;355(17):1763-71.

(4) Goldstraw P, Crowley J, Chansky K, Giroux DJ, Groome PA, Rami-Porta R, et al. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. J Thorac Oncol 2007 Aug;2(8):706-14.

(5) Quadrelli S, Lyons G, Colt H, Chimondeguy D, Buero A. Clinical characteristics and prognosis of incidentally detected lung cancers. Int J Surg Oncol 2015;2015:287604.

(6) Manser RL, Irving LB, Byrnes G, Abramson MJ, Stone CA, Campbell DA. Screening for lung cancer: a systematic review and meta-analysis of controlled trials. Thorax 2003 Sep;58(9):784-9.

(7) Bach PB, Jett JR, Pastorino U, Tockman MS, Swensen SJ, Begg CB. Computed tomography screening and lung cancer outcomes. JAMA 2007 Mar 7;297(9):953-61.

(8) Infante M, Cavuto S, Lutman FR, Brambilla G, Chiesa G, Ceresoli G, et al. A randomized study of lung cancer screening with spiral computed tomography: three-year results from the DANTE trial. Am J Respir Crit Care Med 2009 Sep 1;180(5):445-53.

(9) Gates TJ. Screening for cancer: concepts and controversies. Am Fam Physician 2014 Nov 1;90(9):625-31.

(10) Gill RR, Jaklitsch MT, Jacobson FL. Controversies in Lung Cancer Screening. J Am Coll Radiol 2016 Feb;13(2 Suppl):R2-R7.

(11) Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011 Aug 4;365(5):395-409.

(12) Baldwin DR, Duffy SW, Wald NJ, Page R, Hansell DM, Field JK. UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer. Thorax 2011 Apr;66(4):308-13.

(13) Becker N, Motsch E, Gross ML, Eigentopf A, Heussel CP, Dienemann H, et al. Randomized study on early detection of lung cancer with MSCT in Germany: study design and results of the first screening round. J Cancer Res Clin Oncol 2012 Sep;138(9):1475-86.

(14) Lopes PA, Picozzi G, Mascalchi M, Maria CF, Carrozzi L, Comin C, et al. Design, recruitment and baseline results of the ITALUNG trial for lung cancer screening with low-dose CT. Lung Cancer 2009 Apr;64(1):34-40.

(15) van Klaveren RJ, Oudkerk M, Prokop M, Scholten ET, Nackaerts K, Vernhout R, et al. Management of lung nodules detected by volume CT scanning. N Engl J Med 2009 Dec 3;361(23):2221-9.

(16) Chirikos TN, Hazelton T, Tockman M, Clark R. Cost-effectiveness of screening for lung cancer. JAMA 2003 May 14;289(18):2358-9.

(17) Mahadevia PJ, Fleisher LA, Frick KD, Eng J, Goodman SN, Powe NR. Lung cancer screening with helical computed tomography in older adult smokers: a decision and cost-effectiveness analysis. JAMA 2003 Jan 15;289(3):313-22.

(18) Marshall D, Simpson KN, Earle CC, Chu CW. Economic decision analysis model of screening for lung cancer. Eur J Cancer 2001 Sep;37(14):1759-67.

(19) Marshall D, Simpson KN, Earle CC, Chu C. Potential cost-effectiveness of one-time screening for lung cancer (LC) in a high risk cohort. Lung Cancer 2001 Jun;32(3):227-36.

(20) Bajtarevic A, Ager C, Pienz M, Klieber M, Schwarz K, Ligor M, et al. Noninvasive detection of lung cancer by analysis of exhaled breath. BMC Cancer 2009;9:348.

(21) Chen X, Xu F, Wang Y, Pan Y, Lu D, Wang P, et al. A study of the volatile organic compounds exhaled by lung cancer cells in vitro for breath diagnosis. Cancer 2007 Aug 15;110(4):835-44.

(22) D'Amico A, Pennazza G, Santonico M, Martinelli E, Roscioni C, Galluccio G, et al. An investigation on electronic nose diagnosis of lung cancer. Lung Cancer 2010 May;68(2):170-6.

(23) Hubers AJ, Brinkman P, Boksem RJ, Rhodius RJ, Witte BI, Zwinderman AH, et al. Combined sputum hypermethylation and eNose analysis for lung cancer diagnosis. J Clin Pathol 2014 Aug;67(8):707-11.

(24) Boots AW, Bos LD, van der Schee MP, van Schooten FJ, Sterk PJ. Exhaled Molecular Fingerprinting in Diagnosis and Monitoring: Validating Volatile Promises. Trends Mol Med 2015 Oct;21(10):633-44.

(25) Gordon SM, Szidon JP, Krotoszynski BK, Gibbons RD, O'Neill HJ. Volatile organic compounds in exhaled air from patients with lung cancer. Clin Chem 1985 Aug;31(8):1278-82.

(26) Machado RF, Laskowski D, Deffenderfer O, Burch T, Zheng S, Mazzone PJ, et al. Detection of lung cancer by sensor array analyses of exhaled breath. Am J Respir Crit Care Med 2005 Jun 1;171(11):1286-91.

(27) Phillips M, Altorki N, Austin JH, Cameron RB, Cataneo RN, Greenberg J, et al. Prediction of lung cancer using volatile biomarkers in breath. Cancer Biomark 2007;3(2):95-109.

(28) Phillips M, Altorki N, Austin JH, Cameron RB, Cataneo RN, Kloss R, et al. Detection of lung cancer using weighted digital analysis of breath biomarkers. Clin Chim Acta 2008 Jul 17;393(2):76-84.

(29) Schallschmidt K, Becker R, Jung C, Bremser W, Walles T, Neudecker J, et al. Comparison of volatile organic compounds from lung cancer patients and healthy controls-challenges and limitations of an observational study. J Breath Res 2016 Oct 12;10(4):046007.

(30) Soled M. Aphorisms of Hippocrates. N J Med 1991 Jan;88(1):33.

(31) Pauling L, Robinson AB, Teranishi R, Cary P. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. Proc Natl Acad Sci U S A 1971 Oct;68(10):2374-6.

(32) Dragonieri S, Annema JT, Schot R, van der Schee MP, Spanevello A, Carratu P, et al. An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD. Lung Cancer 2009 May;64(2):166-70.

(33) Peng G, Tisch U, Adams O, Hakim M, Shehada N, Broza YY, et al. Diagnosing lung cancer in exhaled breath using gold nanoparticles. Nat Nanotechnol 2009 Oct;4(10):669-73.

(34) Peled N, Hakim M, Bunn PA, Jr., Miller YE, Kennedy TC, Mattei J, et al. Non-invasive breath analysis of pulmonary nodules. J Thorac Oncol 2012 Oct;7(10):1528-33.

(35) Bruins M, Gerritsen JW, van de Sande WWJ, van Belkum A, Bos A. Enabling a transferable calibration model for metal-oxide type electronic noses. Sensors and Actuators B: Chemical 2013 Nov;188:1187-95.

(36) Hauschild AC, Frisch T, Baumbach JI, Baumbach J. Carotta: Revealing Hidden Confounder Markers in Metabolic Breath Profiles. Metabolites 2015 Jun 10;5(2):344-63.

(37) Schumacher M, Hollander N, Sauerbrei W. Resampling and cross-validation techniques: a tool to reduce bias caused by model building? Stat Med 1997 Dec 30;16(24):2813-27.

(38) Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001 Aug;54(8):774-81.

(39) de la Cruz R, Fuentes C, Meza C, Nunez-Anton V. Error-rate estimation in discriminant analysis of non-linear longitudinal data: A comparison of resampling methods. Stat Methods Med Res 2016 Jul 8.

(40) Hesterberg TC. What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. Am Stat 2015 Oct 2;69(4):371-86.

(41) Sidiropoulos ND, Kyrillidis A. Multi-way compressed sensing for sparse low-rank tensors. IEEE Signal Processing Letters 2012;19(11):757-60.

(42) Hauschild AC, Baumbach JI, Baumbach J. Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification. Genet Mol Res 2012 Aug 16;11(3):2733-44.

(43) Hughey JJ, Hastie T, Butte AJ. ZeitZeiger: supervised learning for high-dimensional data from an oscillatory system. Nucleic Acids Res 2016 May 5;44(8):e80.

(44) Jaffe AE, Storey JD, Ji H, Leek JT. Gene set bagging for estimating the probability a statistically significant result will replicate. BMC Bioinformatics 2013;14:360.

(45) Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011 Mar;61(2):69-90.

(46) Rafiemanesh H, Mehtarpour M, Khani F, Hesami SM, Shamlou R, Towhidi F, et al. Epidemiology, incidence and mortality of lung cancer and their relationship with the development index in the world. J Thorac Dis 2016 Jun;8(6):1094-102.

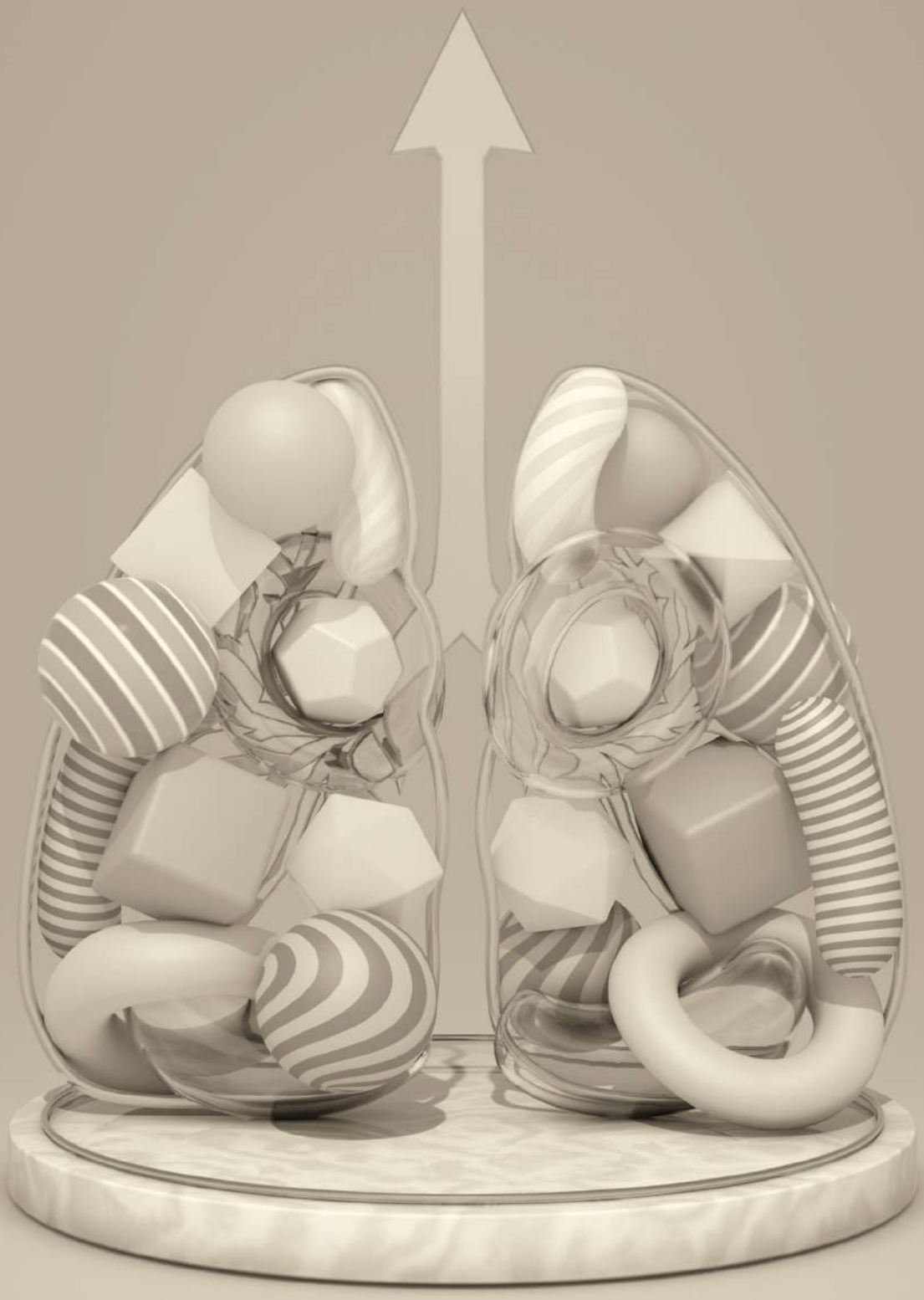
(47) Leopold JH, Bos LD, Sterk PJ, Schultz MJ, Fens N, Horvath I, et al. Comparison of classification methods in breath analysis by electronic nose. J Breath Res 2015 Dec;9(4):046002.

(48) Bruins MG. Transferable odor differentiation models for infectious disease diagnostics.2014. Thesis.

(49) Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiololgy. Genet. Epidemiol 2008 May;32(4):325-40.

# 03

# USING A STEPWISE APPROACH TO SIMULTANEOUSLY DEVELOP AND VALIDATE MACHINE LEARNING BASED PREDICTION MODELS



M. HAALBOOM - S. KORT - J. VAN DER PALEN

J CLIN EPIDEMIOL. 2022 FEB;142:305-310

# Using a stepwise approach to simultaneously develop and validate Machine Learning based prediction models

Haalboom M[1*], Kort S[2], van der Palen J[1]

1. Department of Epidemiology, Medisch Spectrum Twente, Enschede, the Netherlands

2. Department of Pulmonary Medicine, Medisch Spectrum Twente, Enschede, the Netherlands

## Abstract

Accurate diagnosis of a disease is essential in healthcare. Prediction models, based on classical regression techniques, are widely used in clinical practice. Machine Learning (ML) techniques might be preferred in case of a large amount of data per patient and relatively limited numbers of subjects. However, this increases the risk of overfitting, and external validation is imperative. However, in the field of ML, new and more efficient techniques are developed rapidly, and if recruiting patients for a validation study is time consuming, the ML technique used to develop the first model might have been surpassed by more efficient ML techniques, rendering this original model no longer relevant. We demonstrate a stepwise design for simultaneous development and validation of prediction models based on ML techniques. The design enables – in one study - evaluation of the stability and robustness of a prediction model over increasing sample size as well as assessment of the stability of sensitivity/specificity at a chosen cut-off. This will shorten the time to introduction of a new test in health care. We finally describe how to use regular clinical parameters in conjunction with ML based predictions, to further enhance differentiation between subjects with and without a disease.

## WHAT IS NEW?

- Validation of clinically relevant prediction models is an essential step.
- Traditional external validation can be too time-consuming for diagnostic tests that evolve over time.
- Stepwise development and validation is useful for highly relevant and rapidly changing diagnostic tests.
- Our design facilitates using promising emerging less-traditional diagnostic techniques in clinical practice.
- Evaluation of final model stability should prevent using suboptimal and overfitted models.

## Introduction

Accurate and preferably fast diagnosis of a disease is essential in healthcare. Without an accurate diagnosis, health care providers are not able to provide the appropriate treatment to cure, or prevent progression of the disease and further complications.

During the diagnostic process, a medical doctor often combines a variety of data to determine which disease or condition is present, e.g. anamnesis, physical examination, laboratory measurements, imaging techniques. It can, however, be difficult to combine all available data and provide a clear and rational diagnosis. The increase in available tests and measurements only further complicates this. As a tool, statistical models can be useful to combine test results and predict whether a disease is present or not. These so-called prediction models are already widely used in clinical practice, such as PREDICT (www.predictcancer.org). Most prediction models in healthcare are based on classical statistical regression techniques to model data, such as logistic regression, resulting in a probability of the disease. Based on a chosen cut-off of the probability of the presence or absence of the disease under investigation, sensitivity, specificity, positive predicted value, and negative predicted value can be calculated. However, if the amount of available data per patient increases drastically, modern and more flexible modelling techniques such as machine learning techniques might be preferred [1]. An example of a situation where the amount of available data becomes too large for classical regression techniques is for example in the diagnosis of sleep apnoea, where not only the number of apnoea's or hypopnoea's is used for the diagnosis, but where the raw oxygen saturation signal during the entire night is used.

The most often used machine learning approach in diagnostic research is so-called *supervised* machine learning. In supervised machine learning, the diagnosis of patients is known, based on the regular diagnostic process, and this is used as the "gold standard". This is similar to logistic regression, where the outcome as a binary variable (the disease is present or not) also needs to be known to construct a prediction model. However, in contrast to logistic regression, machine learning builds very many (many thousands is not uncommon) models, using all the available information in a huge variety of combinations, and trying different weights for all these parameters, to find models that best predict the diagnosis [2,3]. This also results in a probability of the disease being present, but what parameters are used in this model and what weight is assigned to them, is not known, as it works like a "black box". Also, it is data-driven and not hypothesis-driven and this data-driven construction of prediction models hugely increases the risk of overfitting [4]. Therefore, machine learning approaches always include statistical methods to reduce the chances of ending up with spurious models, that are based on chance only. Some of these statistical methods are cross-validations and bootstrapping techniques [5]. Nevertheless, due to the nature of machine learning,

overfitting is still likely to be present, just as in logistic regression, because it is fitted on a specific set of data. A statistically overfitted model can provide accurate diagnoses for the studied patient population, based on the data used to develop the model but will not perform as well in new patients. Therefore, external validation, in which the accuracy of a model is determined in a completely new, independent study population with a slightly different case-mix, is an essential step to determine the usability in clinical practice [6]. It is therefore striking that many prediction models based on classical regression techniques have not been (appropriately) externally validated [7,8]. For prediction models based on machine learning, this is not different.

A major challenge in external validation studies is that it is often a time-consuming process to properly design and conduct such a study, while it is known that the study result, i.e. the performance of the model, will probably be lower than the results from the original study [8]. Moreover, in machine learning-based prediction models, the original machine learning technique used to develop the model might have been improved in the intervening time. In the relatively new field of machine learning, new and more efficient techniques are developed rapidly, and if recruiting patients for an external validation study is time consuming, the machine learning technique used to develop the first model might have been surpassed by more efficient machine learning techniques, rendering this original model no longer relevant. This might motivate researchers that use machine learning to rather develop a new model based on improved modelling techniques than to externally validate the original model.

Another challenge in external validation studies is the difficulty to determine the appropriate sample size to accurately determine the performance of the model in a slightly different setting [7]. Although the importance of an appropriate sample size in external validation studies has been advocated over the past years, clear guidance on how to determine the sample size is missing. Vergouwe et al. (2005) and Collins, Ogundimu and Altman (2015) suggested to include at least 100 events and 100 non-events in an external validation study, but they also point out that specific hypotheses may require substantially larger sample sizes [9,10]. Especially in the case of screening, where the prevalence of a disease is relatively low, the total sample size will be very large to obtain sufficient cases with a positive diagnosis.

Both in classical regression techniques and in machine learning techniques, predictions become more accurate when the number of available patients increases. For logistic regression it is relatively simple to see what the effect of adding more patients is. In machine learning this is not the case, because many machine learning techniques, such as artificial neural networks, are like black boxes.

When data of additional patients is added, new predictions can be based on completely different underlying models, which are not visible.

In this article, we demonstrate a design in which one can simultaneously develop and validate prediction models based on machine learning techniques. Furthermore, the design enables the evaluation of the stability and robustness of the models over increasing sample size.
We will use the external validation of the prediction model developed by Kort et al. (2018) [11] as a clinical example to explain our proposed design. Kort et al. (2018) developed a prediction model using machine learning techniques (artificial neural network) to discriminate between subjects with and without non-small cell lung cancer. Of all patients in the study breath data was collected through an electronic nose: the Aeonose™ (The Enose Company, Zutphen, the Netherlands). Human breath contains thousands of volatile organic compounds (VOCs). The idea is that the mixture of VOCs changes when a particular disease (here: lung cancer) is present. This VOCs mixture is measured by the Aeonose™ and used as input for the machine learning algorithms [12]. Including many hundreds of subjects, suspected of lung cancer, in an external validation study, is time consuming. In the meantime more efficient machine learning techniques were adopted by the Enose Company. Next to artificial neural networks, analysis techniques now also included Support Vector Machine, XGBoost, Random Forest, but also classical logistic regression and linear discriminant analysis. Furthermore, in the process of CE certification, some small changes were introduced in the hardware of the Aeonose™. This led to the assumption that the data that was collected by Kort et al. in 2018 for building the original model with the non-CE certified Aeonose™ was not compatible to new Aeonose™ data collected during the external validation study. This is not only relevant for changes in hardware, but also for changes in software. Think of machine learning prediction models based on MRI imaging data or continuous EEG signals. If the method to obtain MRI or EEG data is slightly changed due to e.g. a software update, this might also influence the collected data, and thus the machine learning-based prediction models.

## Methods

### 2.1 Artificial Intelligence approach

Data of exhaled breath are analysed by Aethena, a big-data software package, which includes data pre-processing, data compression, and building models based on the aforementioned analysis techniques, including artificial neural networks, Support Vector Machine, XGBoost, Random Forest, logistic regression, and linear discriminant analysis. Many thousands prediction models are obtained that show varying degrees of separation between subjects with and without lung cancer. Statistical validation techniques are employed to prevent overfitting of models, such as leave-10%-out cross

validation and bootstrapping techniques. This results in a “best” model, that predicts the presence of lung cancer.

2.2 Stepwise validation design

We have created a stepwise design (Figure 1 and Table 1) to enable 1) the development of new prediction models based on improved machine learning modelling techniques; 2) the validation of these new models; 3) evaluation of the increase in predictive power and stability of the new modelling technique over an increasing sample size; 4) a (split-sample) validation of the final new model; 5) the assessment of the stability of the sensitivity and specificity at a chosen cut-off value of the probability of having the disease.

Our stepwise design starts with a large study population containing both subjects with and without the disease. This population will be split into a *training cohort* for the development of new machine leaning based models and a *test cohort* for the external validation of these models. For clarity: the *test cohort* is kept blinded, while the data of the *training cohort* are unblinded for supervised machine learning. The sizes of the *test* and *training cohort* depend on the prevalence of the disease, estimated sensitivity/specificity and desired accuracy. In our example, we estimated a sample size of 350 subjects for the external validation of the prediction model developed by Kort et al. (2018) [11] (i.e. the *test cohort*) based on an estimated sensitivity of 90%, specificity of 75%, a prevalence of 40% and desired accuracy of 5% of sensitivity. Based on experience, it was estimated that 400 subjects in our *training cohort* should be adequate for developing new machine learning based models.

During the first step, the first many thousands models will be developed based on data from 100 subjects (*training set*) from the *training cohort*. Statistical cross-validation is employed as a statistical validation technique at this point and this will already eliminate the vast majority of these models. The models that pass the internal cross-validation phase are validated by using blinded data from the next 50 patients from the *training cohort.* Subsequently, these 50 patients are unblinded and are added to the *training set,* the ‘best’ model can be based on a different analysis technique (e.g. artificial neural network, XGBoost or Random Forest). These steps are repeated until all 400 patients from the *training cohort* have been added, in steps of 50, to the *training set*. Note that the step size can vary depending on disease prevalence; approximately 20-25 subjects need to be present for each disease state to obtain meaningful improvements in the models. After each step of adding another 50 subjects, the new models are again first internally validated using cross-validations, and subsequently validated on the next 50 blinded subjects from the training cohort, and the Area Under the Curve (AUC) is calculated to observe the improvement in diagnostic performance of the “best”

model. By inspecting the change in AUC after each step, an impression of the stability of the models over increasing sample size can be obtained.

During the last step (Table 1, step 7), the final model will be validated (split-sample validation) on the *test cohort* (n = 350). Moreover, the final model can be "validated" on the smaller *training sets* from previous steps to see how stable this final model is (Table 1). If the final model is still an overfitted model, it will not perform consistently on these smaller datasets from which it had been developed. If the final model performs similarly well in all these steps, this will enhance the believability of the machine learning based model for use in clinical practice.

| **Table 1. Stepwise development and validation of prediction models** |
|---|
| **Step 1a.** Develop new models based on the data of the first 100 subjects (training set) from the *training cohort*. |
| **Step 1b.** Use the 'best' model (1) from step 1a to predict the next 50 subjects from the *training cohort* in a blinded fashion and determine model performance (e.g. the AUC) |
| **Step 2a.** Unblind these 50 subjects and add them to the training set (n=150) to develop new models. |
| **Step 2b.** Use the 'best' model (2) from step 2a to predict the next 50 subjects from the *training cohort* in a blinded fashion and determine model performance (e.g. the AUC) |
| **Step 3a.** Unblind these 50 subjects and add them to the training set (n=200) to develop new models. |
| **Step 3b.** Use the 'best' model (3) from step 3a to predict the next 50 subjects from the *training cohort* in a blinded fashion and determine model performance (e.g. the AUC) |
| **Step 4a.** Unblind these 50 subjects and add them to the training set (n=250) to develop new models. |
| **Step 4b.** Use the 'best' model (4) from step 4a to predict the next 50 subjects from the *training cohort* in a blinded fashion and determine model performance (e.g. the AUC) |
| **Step 5a.** Unblind these 50 subjects and add them to the training set (n=300) to develop new models. |
| **Step 5b.** Use the 'best' model (5) from step 5a to predict the next 50 subjects from the *training cohort* in a blinded fashion and determine model performance (e.g. the AUC) |
| **Step 6a.** Unblind these 50 subjects and add them to the training set (n=350) to develop new models. |
| **Step 6b.** Use the 'best' model (6) from step 6a to predict the next 50 subjects from the *training cohort* in a blinded fashion and determine model performance (e.g. the AUC) |
| **Step 7a.** Unblind these 50 subjects and add them to the training set (n=400) to develop new models. |
| **Step 7b.** Validate the 'best' **final model** (7) from step 7a to predict the 350 subjects from the ***test cohort*** in a blinded fashion and determine model performance (e.g. the AUC) |
| **Step 8.** Evaluate the stability of the final model over increasing sample size by predicting the subjects in the *training sets* used in steps "a" (1a, 2a, etc) in a blinded fashion and determine relevant performance measures )e.g. AUC, sensitivity/specificity) of the final model for each training set (n=100, n=150, …., n=400). |

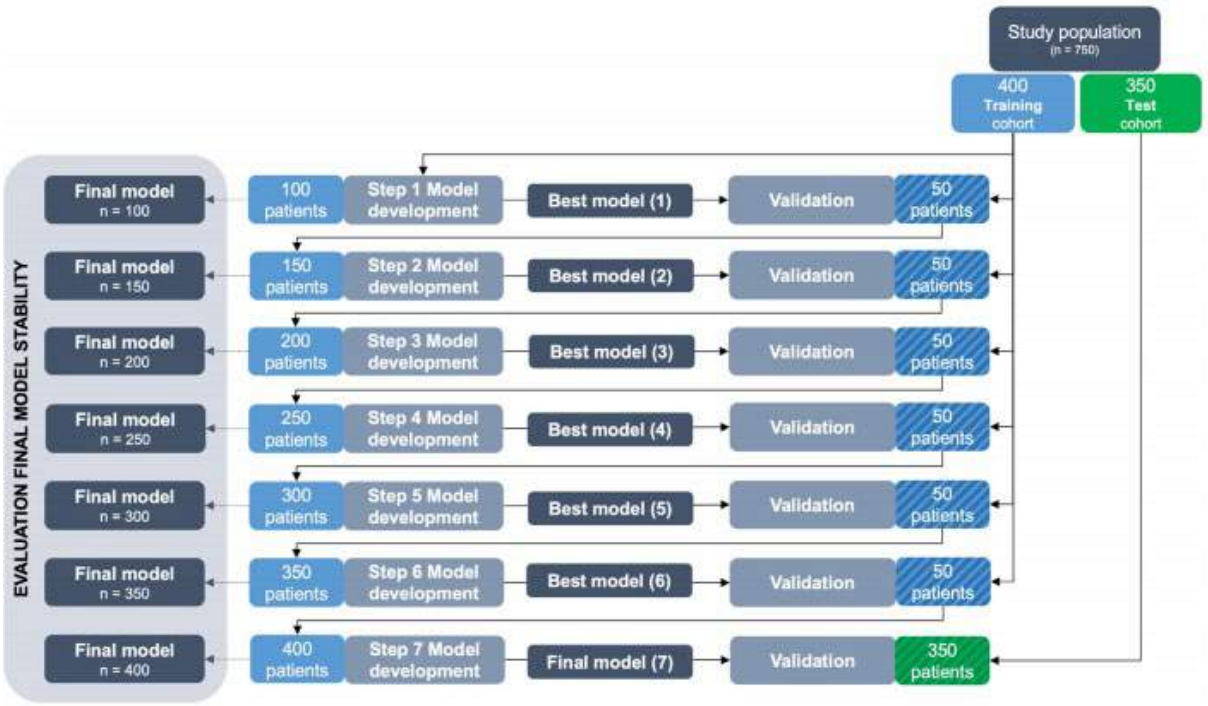



**Figure 1.** Stepwise model development and validation design

## Discussion

The strategy we propose has several advantages, but also requirements. The traditional external validation strategy is time-consuming, as a first study is needed for model development and a separate, completely new, second study is needed for external validation. Our strategy allows for simultaneously developing and validating prediction models, based on data collected in one study, which will shorten the time to the introduction of a new test in health care. Especially when patient recruitment is slow and therefore time-consuming, the proposed strategy will aid in new tests to become available much sooner.

By using a stepwise approach with ever increasing sample size, the trend in the change in AUC after each step can be seen as an indication of whether the performance of the “best” model is plateauing or not. If this is the case, this will lead to a larger confidence in the obtained final model. It should be noted that this is only a qualitative impression.

Ideally one pre-specifies the number of steps to be taken in advance. However, it would be wise to also pre-specify an interim decision moment to adjust the number of steps, based on a minimal or maximal AUC. E.g. if after already a few steps a very high AUC is achieved, larger than a prespecified threshold, one can decide to reduce the number of steps, resulting in a larger *test set* to validate the final model on. This will lead to smaller 95% confidence intervals around the test parameters:

sensitivity, specificity, positive predicted values (PPV) and negative predicted values (NPV) and AUC. Conversely, if after the prespecified number of steps the minimal requested AUC has not yet been reached, and a plateau has also not yet been reached, one can decide to increase the number of steps, at the cost of the size of the *test cohort*, with resulting wider 95% confidence intervals around the test parameters. To be able to make a valid comparison of the AUC over the subsequent (larger number of) steps, one should estimate how many steps will be necessary, and reduce the size of *test cohort* accordingly. Then, one should restart at step 1.

The focus should not only be on the AUC, but also on sensitivity and specificity. In practice, one chooses a cut-off value of the probability to have the disease or not, based on the consequences of a false positive or false negative diagnosis. By choosing a specific cut-off value, one can choose to focus on a higher sensitivity or specificity, and therewith, given a certain prevalence of the disease, also the PPV and NPV. It should be noted, however, that at each step from step 1 through 7, new models are built, based on increasing sample sizes and that the best models can - and in all likelihood will - result from different analysis techniques. The best model after step one can e.g. be based on a Random Forest analysis, while after step 2 XGBoost might produce the best model. If the researchers a-priori wish to obtain a model that has at least e.g. 90% sensitivity, a cut-off can be chosen to result in this high sensitivity. However, the chosen cut-off based on the best model in step 1 will invariably be different from the cut-off after step 2, if a different analysis technique has resulted in a new “best” model. This can be remedied following the final model developing step 7 (and as described in step 8 in our example). Based on the final “best" model resulting from step 7 (n=400), a preferred cut-off value should be determined that results in the desired sensitivity or NPV (or alternatively specificity and PPV). Then, this final “best” model should be “retrofitted” to the training sets from steps 1 through 7 to evaluate the stability of this final “best” model over increasing sample sizes by predicting the subjects in the training sets used in steps “a” (1a, 2a, etc) in a blinded fashion and determine the AUC of the final “best” model for each training set (n=100, n=150, …., n=400). One can now see whether the chosen cut-off results in a similar sensitivity and NPV at each “retrofitted” step. The variation in sensitivity and NPV at each step will provide an indication whether the cut-off value chosen after step 7 is indeed a valid and stable cut-off value, which can be used in clinical practice.

Because PPV and NPV depend on the prevalence of the disease, one should ensure that during each step where new data is added to build new models, the prevalence of the disease is constant by sampling a fixed ratio of subjects with and without the disease.

Next to the data on which the prediction models are based (e.g. breath data), relatively easily obtainable clinical parameters are often also available, such as age, gender, smoking behaviour and

the presence of e.g. relevant comorbidities. These variables can be used to further enhance the differentiation between subjects with and without lung cancer, as already has been show by Kort et al [13]. A straightforward multivariate logistic regression analysis, including these clinical parameters plus the probability of disease resulting from the final "best" model, will produce a probability of having lung cancer or not for each subject. Again a cut-off can be chosen that will result in the desired sensitivity and NPV or specificity with corresponding PPV. This logistic regression model can also be retrofitted again to the smaller dataset from steps 1a, 2a, 3a, 4a and so on, to assess whether the chosen cut-off results in a sensitivity or specificity that is stable enough to be used in clinical practice.

Of course, our proposed validation design is still a form of "split-file" validation and not a true substitute for a classical external validation study in which one recruits a complete new study population. However, when the validation steps as described above have led to the implementation of the test in clinical practice, routine data collection will result in an ever increasing number of subjects with a positive or negative diagnosis of the disease of interest. This offers the opportunity to do a classical external validation study or to further improve the diagnostic potential of the test with multi-centre or even multi-national data with a variety of case-mixes. The choice to redo the analyses as described in figure 1 and table 1, but with larger numbers and/or with more or fewer steps depends on how well the test already performs after the original validation. If the results are already excellent, not much is to be gained. If there is still room for improvement, one can go through the same validation steps again and see whether much larger numbers provide better and more robust models. Whether the same or new analysis techniques are used is moot, as long as a good and stable "best" model with a stable cut-off point results from the analyses.

In summary, validation of any new diagnostic test is imperative, and given the unique circumstances surrounding prediction models based on machine learning techniques, a clear testing and validation strategy needs to be described upfront, prior to actually performing the analyses. The steps described in this manuscript might provide guidance as how to do this.

## References


1. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak.* 2019;19(1):281.

2. Deo RC. Machine Learning in Medicine. *Circulation.* 2015;132(20):1920-1930.

3. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. *J Intern Med.* 2018;284(6):603-619.

4. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine.* 2016;375(13):1216-1219.

5. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning.* New York: Springer-Verlag; 2013.

6. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol.* 2003;56(9):826-832.

7. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 2014;14:40.

8. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol.* 2015;68(1):25-34.

9. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med.* 2016;35(2):214-226.

10. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology.* 2005;58(5):475-483.

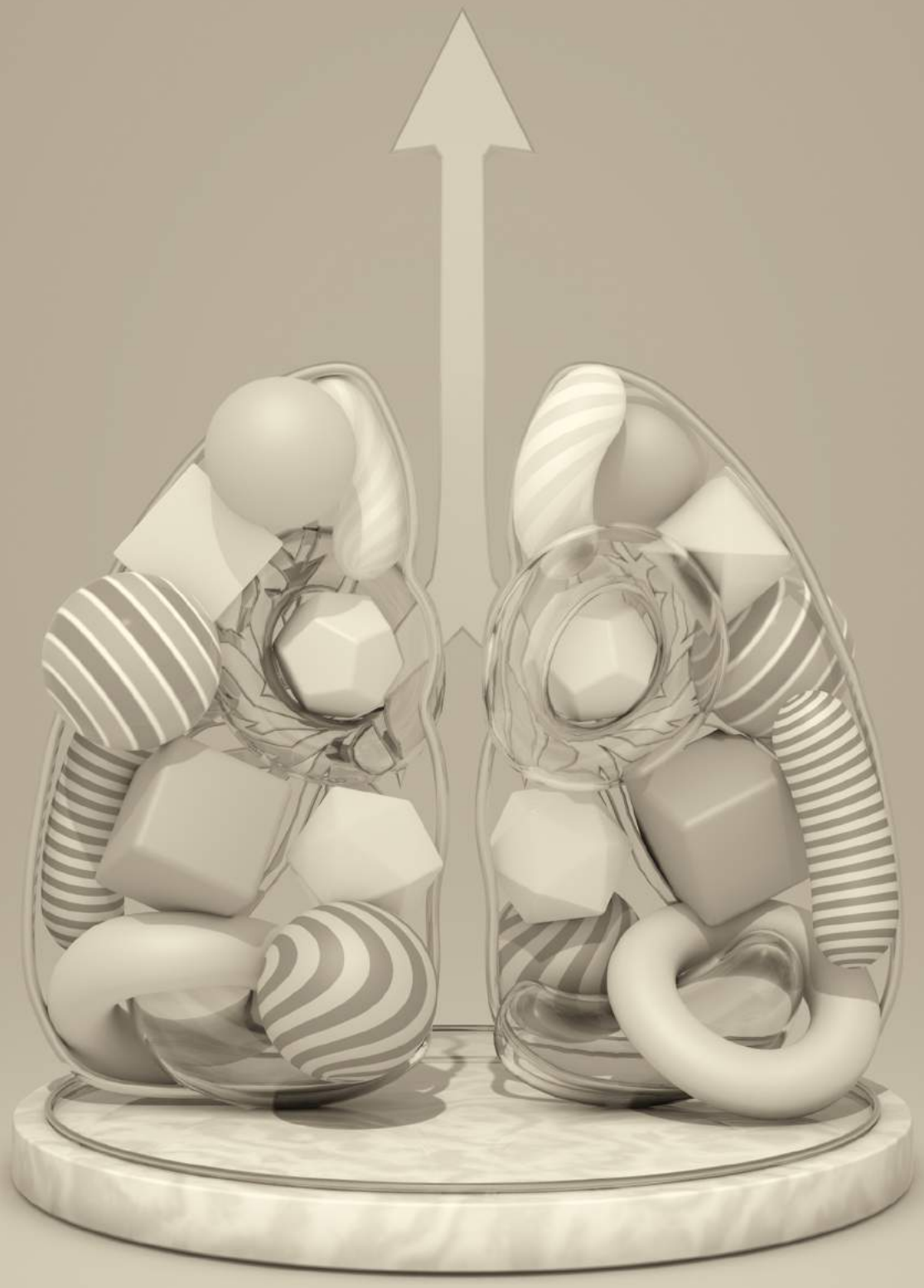
11. Kort S, Tiggeloven MM, Brusse-Keizer M, et al. Multi-centre prospective study on diagnosing subtypes of lung cancer by exhaled-breath analysis. *Lung Cancer.* 2018;125:223-229.

12. Bruins M, Gerritsen JW, van de Sande WWJ, van Belkum A, Bos A. Enabling a transferable calibration model for metal-oxide type electronic noses. *Sensors and Actuators B: Chemical.* 2013;188:1187-1195.

13. Kort S, Brusse-Keizer M, Gerritsen JW, et al. Improving lung cancer diagnosis by combining exhaled-breath data and clinical parameters. *ERJ Open Res.* 2020;6(1):00221-02019.

# 04

# MULTI-CENTRE PROSPECTIVE STUDY ON DIAGNOSING SUBTYPES OF LUNG CANCER BY EXHALED-BREATH ANALYSIS



S. KORT - M.M. TIGGELOVEN - M. BRUSSE-KEIZER - J.W. GERRITSEN - J.H. SCHOUWINK - E. CITGEZ - F.H.C. DE JONGH S. SAMII - J. VAN DER MATEN - M. VAN DEN BOGART J. VAN DER PALEN

LUNG CANCER. 2018 NOV;125:223-229

## Multi-centre prospective study on diagnosing subtypes of lung cancer by exhaled-breath analysis

S. Kort [(1)]; M.M. Tiggeloven [(1)]; M. Brusse-Keizer [(2)]; J.W. Gerritsen [(3)]; J.H. Schouwink [(1)]; E. Citgez [(1)]; F.H.C. de Jongh [(1)]; S. Samii [(4)]; J. van der Maten [(5)]; M. van den Bogart [(6)]; J. van der Palen [(2,7)].

1: Department of Pulmonary Medicine, Medisch Spectrum Twente, Enschede, the Netherlands
2: Medical School Twente, Medisch Spectrum Twente, Enschede, the Netherlands
3: The eNose Company, Zutphen, the Netherlands
4: Department of Pulmonary Medicine, Deventer Ziekenhuis, Deventer, the Netherlands
5: Department of Pulmonary Medicine, Medisch Centrum Leeuwarden, Leeuwarden, the Netherlands
6: Department of Pulmonary Medicine, Bernhoven Uden, Uden, the Netherlands
7: Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, the Netherlands

## List of abbreviations

| | |
|---|---|
| AC | adenocarcinoma |
| AUC | area under the curve |
| LDCT | low dose computed tomography |
| NLST | national lung screening trial |
| NPV | negative predictive value |
| NSCLC | non-small cell lung cancer |
| PPV | positive predictive value |
| ROC | receiver operating curve |
| SCC | squamous cell carcinoma |
| SCLC | small cell lung carcinoma |
| VOC | volatile organic compound |

## Abstract

**Objectives:** Lung cancer is a leading cause of mortality. Exhaled-breath analysis of volatile organic compounds (VOC's) might detect lung cancer early in the course of the disease, which may improve outcomes. Subtyping lung cancers could be helpful in further clinical decisions.

**Materials and Methods:** In a prospective, multi-centre study, using 10 electronic nose devices, 144 subjects diagnosed with NSCLC and 146 healthy subjects, including subjects considered negative for NSCLC after investigation, breathed into the Aeonose™ (The eNose Company, Zutphen, Netherlands). Also, analyses into subtypes of NSCLC, such as adenocarcinoma (AC) and squamous cell carcinoma (SCC), and analyses of patients with small cell lung cancer (SCLC) were performed.

**Results:** Choosing a cut-off point to predominantly rule out cancer resulted for NSCLC in a sensitivity of 94.4%, a specificity of 32.9%, a positive predictive value of 58.1%, a negative predictive value (NPV) of 85.7%, and an area under the curve (AUC) of 0.76. For AC sensitivity, PPV, NPV, and AUC were 81.5%, 56.4%, 79.5%, and 0.74, respectively, while for SCC these numbers were 80.8%, 45.7%, 93.0%, and 0.77, respectively. SCLC could be ruled out with a sensitivity of 88.9% and an NPV of 96.8% with an AUC of 0.86.

**Conclusion:** Electronic nose technology with the Aeonose™ can play an important role in rapidly excluding lung cancer due to the high negative predictive value for various, but not all types of lung cancer. Patients showing positive breath tests should still be subjected to further diagnostic testing.

## Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide (1). The main types of lung cancer are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), accounting for 15% and 85% of the established cases respectively. NSCLC can be subdivided into two major subtypes: squamous cell carcinoma (SCC) and adenocarcinoma (AC), which differ in clinical, radiological, and histological characteristics (2). The diagnosis of early-stage lung cancer is crucial for successful curative therapy, because treatment options and prognosis directly depend on tumour size and metastatic spread at the time of diagnosis (3). Five-year survival rates for those with pathological stage IA NSCLC is 73%, whereas metastatic disease has a five-year overall survival rate of only 29% with local lymphatic spread and 4.5% for patients with distant metastases (4-6). SCLC is associated with even worse survival rates where limited disease has a five-year survival of 10-20%, and metastatic disease <1%. Unfortunately, only 16% of lung cancer cases present with localised, potentially curable disease, which explains the poor survival rates (6).

The current gold standard for diagnosing lung cancer is a histological or cytological proof, either from the primary or metastatic lesion.

There have been many attempts to develop screening tests in order to detect early-stage lung cancer. Currently, the only screening method implying reduced lung cancer mortality in high-risk groups is annual low-dose computed tomography (LDCT) (7;8). However, several issues still need to be addressed, such as the high rate of false positives cases (up to 96.4%) in the National Lung Screening Trial (NLST), leading to unnecessary invasive procedures, radiation risk, and unnecessary anxiety. In most countries in Europe, results of the Dutch-Belgian lung cancer screening trial (NELSON) are awaited before a decision on implementation of screening programs will be made (7;9). One approach could be adding a simple, non-invasive and reliable test to reduce the number of false positives and consequently unnecessary invasive interventions.

Lately, sensor technologies based on pattern recognition in exhaled breath have been developed. These so-called electronic noses allow fast, low-cost, and non-invasive analysis of exhaled breath. Although this diagnostic approach seems very promising in the lung cancer field, it has not been incorporated in clinical practice so far (10-16). This can partly be explained by the fact that in most cases, calibration models for electronic noses aren't transferrable among different devices. On the other hand, the negative predictive value (NPV) of electronic noses is still too low to allow clinical implementation.

The concept of the electronic nose as described in this manuscript, the Aeonose™ (the Enose Company, Zutphen, the Netherlands), is based on the availability of powerful IT solutions, allowing the application of pattern recognition techniques to complex measurement data without the need of

specific identification of individual molecules. An electronic nose can measure low concentrations of volatile organic compounds (VOC's) in exhaled breath, that represent a breath print and reflect pathological processes in the body on tissue level, such as inflammation, infection, and neoplasms (17). In this way, a combination of VOC's can serve as a non-invasive, diagnostic biomarker for metabolic changes associated with different pathological conditions. These VOC's can be detected with multiple, highly-sensitive electro-chemical sensors. This detection method is directed at changes in physical properties of the sensors, such as surface conductivity when being exposed to VOC's (18).

Recently, a pilot study on detecting lung cancer using the Aeonose™ was reported by van de Goor et al (19). In this study, the Aeonose™ was used to distinguish between patients with lung cancer and healthy controls. A total of 167 subjects were included of whom 107 were diagnosed with lung cancer. They found a promising sensitivity of 83%, a specificity of 84% with an area under the curve (AUC) of 0.83. However, this study was single center, and the researchers did not distinguish between various types of lung cancer.

**Goals and objectives**

The aims of this multi-centre study were: 1) to rapidly prove or reject the diagnosis of lung cancer in a cohort of patients suspected of lung cancer and healthy controls, 2) to discriminate between the subtypes of NSCLC: adenocarcinoma and squamous cell carcinoma, and 3) to distinguish SCLC patients from non-SCLC subjects in patients suspected of lung cancer and healthy controls.

## Material and methods

It concerns a multi-centre, prospective diagnostic study in subjects suspected for lung cancer who were referred for a histological biopsy, as well as in healthy volunteers. The four secondary teaching hospitals participating in this study were Medisch Spectrum Twente Enschede, Ziekenhuis Bernhoven Uden, Medisch Centrum Leeuwarden, and Deventer Ziekenhuis, all in the Netherlands. Each hospital weekly diagnoses approximately 2-3 patients with lung cancer. For patients who turned out to have lung cancer, staging was established according to the 7$^{th}$ edition of the American Joint Committee on Cancer TNM staging system (5). For all subjects, demographic parameters (e.g age), smoking status, amount of packyears, and comorbidities were recorded.

Participants with suspected lung cancer visiting the outpatient clinic of the pulmonology departments of the participating hospitals were included between June 2015 and December 2017. Suspected subjects were divided into a group with confirmed lung cancer and a group with a rejected diagnosis of lung cancer, based on imaging and/or derived histopathology. Subjects with a suspicion

of lung cancer were not biopsied when the CT-scan showed no evidence of lung cancer, even when the chest X-ray did. Also, some subjects showed a spontaneous decrease in nodule size without any treatment, which does not fit with the suspicion of lung cancer. A few patients with a high suspicion of lung cancer did not undergo biopsy because of their weak condition, but these subjects were excluded from the analyses. Finally, subjects who had a negative biopsy, but still a very high clinical suspicion of lung cancer were directed for a re-biopsy that eventually led to a confirmed diagnosis lung cancer. Healthy volunteers with a minimum age of 50 were recruited through an advertisement at the hospitals' website. The only exclusion criterion for all subjects was being diagnosed with an other active malignancy. We compared breath patterns from patients with a proven diagnosis of lung cancer prior to initiation of treatment with subjects without lung cancer, i.e. healthy volunteers and suspected subjects with a rejected diagnosis of lung cancer. The study protocol was approved by the medical ethics committee of Medisch Spectrum Twente, and the board of directors at each participating centre. All patients provided written informed consent.

Aeonose™ technology and procedure of breath sampling

The Aeonose™ is a hand-held electronic nose, containing three micro-hotplate metal-oxide sensors (MOS) that are mass producible, and offer the opportunity for transferring calibration models. This means that once a calibration model for a specific indication has been developed, it can easily be transferred to other Aeonose™ devices (20). In this study we used 10 Aeonose™ devices which were randomly applied to subjects to avoid specific device dependent variations. Patients were instructed to perform tidal breathing through the non-rebreathing Aeonose™ device for 5 minutes during a single visit. A disposable mouthpiece with a carbon active filter was used (filtering inhaled air) and the patient's nose was clipped to prevent nose breathing. A washout period during the first 2 minutes was used for clearing the lungs from ambient, possibly polluted air with a carbon filter and the nose clip, without recording any measurements. During the next 3 minutes, metal-oxide sensors were exposed to exhaled breath and conductivity values of the sensors were recorded.

Redox reactions of VOCs at the sensor surfaces were recorded in terms of conductivity changes. After these 5 minutes, the Aeonose™ was put aside, and the sensors were regenerated by guiding clean air to them through another active carbon filter. Then, a build-in Tenax™-tube that collected VOC's during the measurement was heated, and these VOCs released were guided over the sensors and recorded, providing additional information on the breath profile. Finally, another regeneration step with clean air was enforced. Using this protocol, the total breath-test cycle took approximately 15 minutes.

Sample size

We calculated a sample size taking into account a required sensitivity of 90% with a confidence interval of 82.5% - 95%. Therefore, approximately 105 subjects diagnosed with lung cancer must be included. Presuming a 1:1 ratio of a confirmed versus a rejected diagnosis of lung cancer in suspected subjects, we also needed 105 subjects with a rejected diagnosis. We also planned to include approximately 75 'healthy' subjects without any suspicion for lung cancer.

Statistical analysis

Clinical characteristics are reported as means with standard deviations when normally distributed or as medians with interquartile range (IQR). Nominal variables are reported as numbers with corresponding percentages. To assess differences between the different groups, either the ANOVA test for normally distributed continuous variables, Kruskall Wallis non-parametric test for skewed distributed continuous or ordinal variables, or chi-squared test ($x^2$) for nominal and categorical variables were applied. We used the Bonferroni Holm correction to adjust for multiple testing.
Data of exhaled breath were analysed by Aethena, a proprietary big-data software package from The eNose Company (21). In the course of the big data analysis and pattern recognition (using artificial neural networks), several steps can be distinguished such as pre-processing of data, data compression, leave-10%-out cross-validation, model selection, and combining prediction models with promising AUC's.
Sensitivity, specificity, positive predictive value (PPV) and (NPV) were calculated for the diagnosis of lung cancer and its subtypes. Receiver operating characteristics (ROC) curves were composed and AUCs were calculated with 95% confidence intervals. A scatter plot showing values between -1 and +1 was calculated for each subject indicating the degree to which the subject was classified as positive (maximum value +1) or negative (minimum value -1) for lung cancer. During the analysis, a cut-off value was chosen, which showed best separation between the two groups in terms of optimal sensitivity and NPV to exclude lung cancer early, together with an acceptable number of false positives. All analyses were based on the complete dataset after including all participating subjects.

In order to rule out any influence of device characteristics on results during the training phase, it was required for every Aeonose™ to measure at least four positive and four negative samples. If this condition was not met, some measurements from that specific device were excluded from the analysis. No Aeonose™ device was excluded during the study.
All statistical tests were two-sided with a significance level at 0.05. SPSS V.22.0 was used.

## Results

Of the 308 subjects included, 144 had confirmed NSCLC, 18 had confirmed SCLC, 61 were suspected for lung cancer due to complaints or an abnormal chest X-ray, but were considered negative after investigation, and 85 subjects were healthy volunteers (Figure 1). No adverse effects were found when performing the breath measurements. Clinical characteristics of the subjects are described in Table 1. Healthy volunteers were significantly younger, more likely to be female and non-smoker, had smoked less packyears, and did not have COPD (all p<0.001). Suspected patients without lung cancer were more often never-smokers and had smoked less pack-years than confirmed NSCLC patients (p<0.001). Out of the 144 NSCLC patients, 93 had AC, 42 had SCC, 4 had large cell carcinoma and 5 were NSCLC not otherwise specified (Figure 1). Approximately 75% of the lung cancer patients were classified as stage III or IV disease.

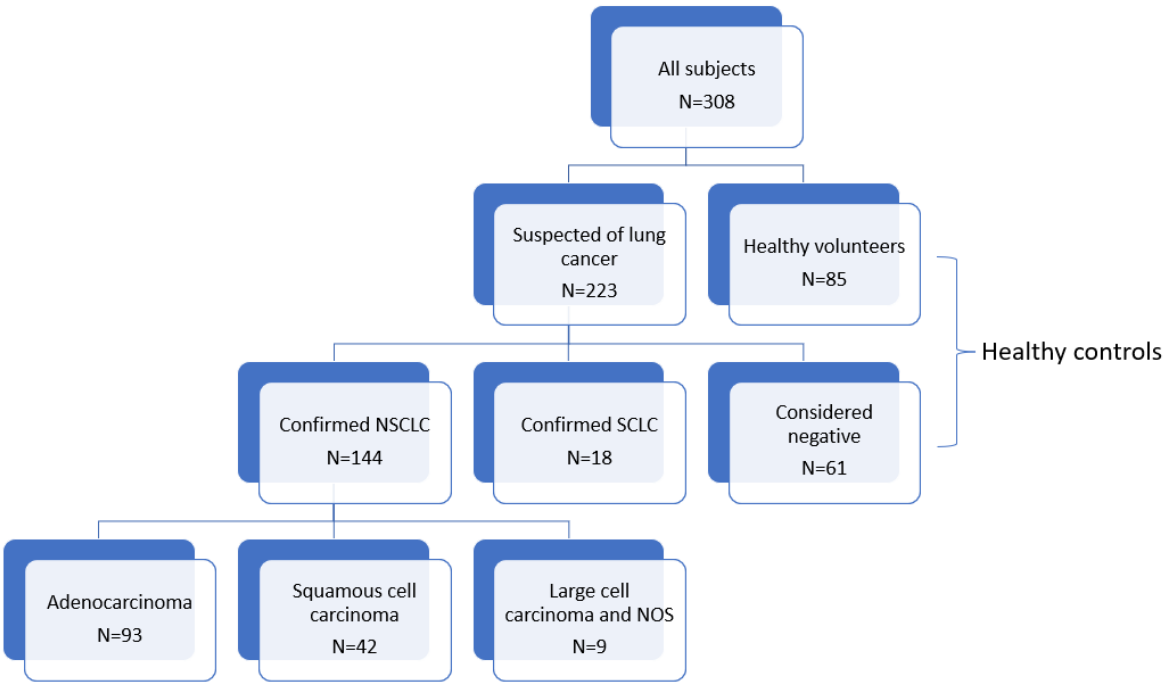



**Figure 1.** Flow chart showing the different groups. NOS: not otherwise specified.

| Table 1. Clinical characteristics of subjects. | | | | | | |
|---|---|---|---|---|---|---|
| | All subjects<br><br>N=308 | Confirmed NSCLC<br>N=144 | Confirmed non-NSCLC<br>N=61 | Confirmed SCLC<br>N=18 | Healthy<br><br>N=85 | P-value |
| **Age in years, mean (SD)** | 64.6 (8.5) | 67.1 (9.0) | 65.1 (8.8) | 63.2 (8.2) | 60.0 (4.4) | <0.001[a] |
| **Sex, number of males (%)** | 142 (49%) | 83 (57.6) | 32 (52.5) | 10 (55.6) | 27 (31.8) | 0.001[b] |
| **BMI, mean (SD)** | 25.6 (5.2) | 25.3 (5.5) | 27.0 (5.9) | 28.0 (4.8) | 25.2 (3.8) | 0.056 |
| **Smoking status, N (%)**<br>**Current smoker**<br>**Ex-smoker**<br>**Never smoked** | <br>71 (24.5)<br>164 (56.6)<br>55 (19) | <br>51 (35.4)<br>86 (59.7)<br>7 (4.9) | <br>13 (21.3)<br>33 (54.1)<br>15 (24.6) | <br>7 (38.9)<br>10 (55.6)<br>1 (5.6) | <br>7 (8.2)<br>45 (52.9)<br>33 (38.8) | <br><0.001[c] |
| **Pack years[d], median (IQR)** | 21.5 (3.25-40.0) | 35.0 (20.0-46.75) | 20.0 (1.25-32.75) | 45.0 (27.75-52-75) | 2.0 (0.0-14.5) | <0.001[c] |
| **COPD, N (%)** | 89 (37) | 66 (46.5) | 21 (34.4) | 8 (44.4) | 1 (1.2) | <0.001[b] |

**[a]After Games-Howell correction, there was a significant difference between healthy volunteers and confirmed NSCLC and healthy volunteers and confirmed non-NSCLC. [b] After Holm-Bonferroni correction there was a significant difference between healthy volunteers and confirmed NSCLC, confirmed non-NSCLC and confirmed SCLC. [c] Between all 4 groups. [d] 5 subjects missing pack years. Abbreviations: BMI, body mass index; COPD: chronic obstructive pulmonary disease**

Table 2 summarizes the diagnostic performance of the Aeonose™ for the different groups in terms of sensitivity, specificity, PPV, NPV, and AUC with corresponding 95% confidence intervals. Limited case sample sizes resulted in different group sizes for healthy subjects.

When focusing on a high sensitivity and NPV, a high-sensitivity point was chosen, based on the ROC-curve to distinguish between NSCLC and all negatives, which led to a sensitivity of 94.4%, an NPV of 85.7%, at an AUC of 0.76.

**Table 2. Diagnostic performance of the Aeonose™.**

| Groups | N | Cut-off chosen | TP | TN | FP | FN | Sensitivity (%) | Specificity (%) | PPV (%) | NPV (%) | AUC (95% CI) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NSCLC vs. all negatives[a] | 144 vs. 146 | -0.265 | 136 | 48 | 98 | 8 | 94.4 | 32.9 | 58.1 | 85.7 | 0.76 (0.71-0.82) |
| NSCLC vs considered negative after investigation | 105 vs. 43* | -0.350 | 95 | 11 | 32 | 10 | 90.5 | 25.6 | 74.8 | 52.4 | 0.73 (0.64-0.82) |
| NSCLC vs healthy volunteers | 103 vs. 84[b] | -0.295 | 95 | 43 | 41 | 8 | 92.2 | 51.2 | 69.9 | 84.3 | 0.85 (0.79-0.90) |
| Adenocarcinoma vs all negatives[a] | 81 vs 109* | -0.365 | 66 | 58 | 51 | 15 | 81.5 | 53.2 | 56.4 | 79.5 | 0.74 (0.67-0.82) |
| Squamous cell carcinoma vs all negatives[a] | 26 vs 91* | -0.015 | 21 | 66 | 25 | 5 | 80.8 | 72.5 | 45.7 | 93.0 | 0.78 (0.67-0.88) |
| SCLC vs. all negatives[a] | 18 vs. 75 | -0.575 | 16 | 60 | 15 | 2 | 88.9 | 80.0 | 51.6 | 96.8 | 0.86 (0.78-0.95) |

TP, true positive; TN, true negative; FP, false positive; FN, false negative; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; NSCLC, non-small cell lung cancer; SCLC, small cell lung cancer.
[a] All negatives include suspected subjects considered negative after investigation and healthy volunteers.
[b]Limited case sample sizes resulted in different group sizes

When only suspected subjects with a rejected diagnosis of NSCLC were distinguished from NSCLC patients, we observed a relatively decreased performance when compared to all negatives and when compared to healthy volunteers only. This analysis revealed a sensitivity of 90.5%, an NPV of 52.4%, and an AUC of 0.73. At the same time, diagnostic performance improved when discriminating breath prints of NSCLC patients from healthy volunteers, resulting in a sensitivity of 92.2%, an NPV of 84.3%. an AUC of 0.85. The corresponding scatterplots are presented in figure 2A-C.

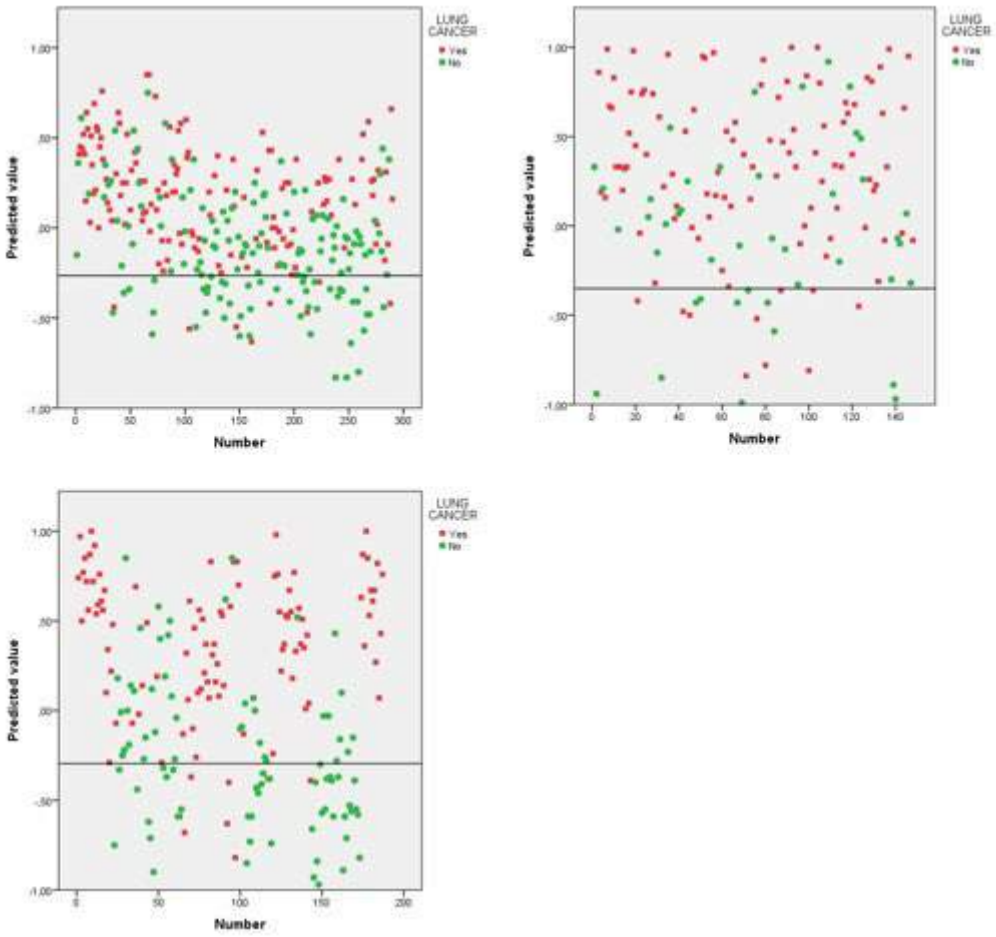



**Figure 2**. Scatterplots with chosen optimal cut-off values. 2A. NSCLC vs all negatives (AUC 0.76). 2B. NSCLC vs proven negatives (AUC 0.73). 2C. NSCLC vs healthy volunteers (AUC 0.85).

We investigated whether the most prevalent subtypes of NSCLC, being AC and SCC could be discriminated more accurately compared to the combined group of NSCLC patients. The results are presented in Table 2. The diagnostic accuracy in diagnosing AC from healthy subjects resulted into a sensitivity of 81.5% with an NPV of 79.5% and a corresponding AUC of 0.74.

When discriminating SCC patients from healthy subjects, we found an interesting performance of the Aeonose™ to rule out SCC with a sensitivity of 80.8%, an NPV of 93.0% and a corresponding AUC of 0.77.

The corresponding scatterplots of the analyses of the subtypes are presented in figure 3.

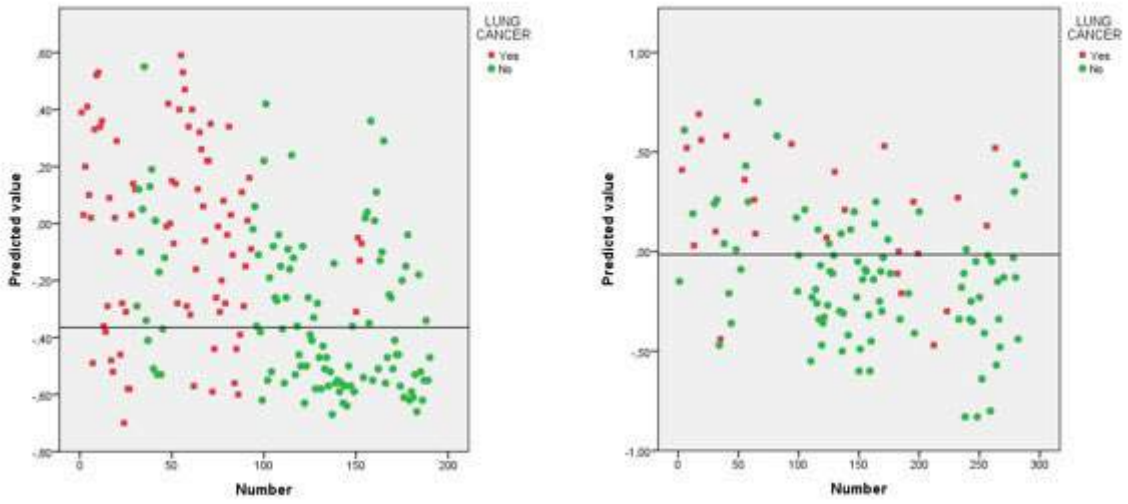



**Figure 3.** Scatterplots with chosen optimal cut-off values. 3A. Adenocarcinoma vs. all negatives (AUC 0.74). 3B. Squamous cell carcinoma vs all negatives (AUC 0.77).

Due to the lower prevalence of SCLC, analyses could only be performed in a limited number of patients to distinguish SCLC-patients from healthy controls (Table 2). Ninety-three subjects of whom 18 had pathologically confirmed SCLC were included (Figure 4). The diagnostic accuracy in diagnosing SCLC resulted in a sensitivity of 88.9%, a specificity of 80.0% with a PPV of 51.6%, an NPV of 96.8%, and an AUC of 0.86.

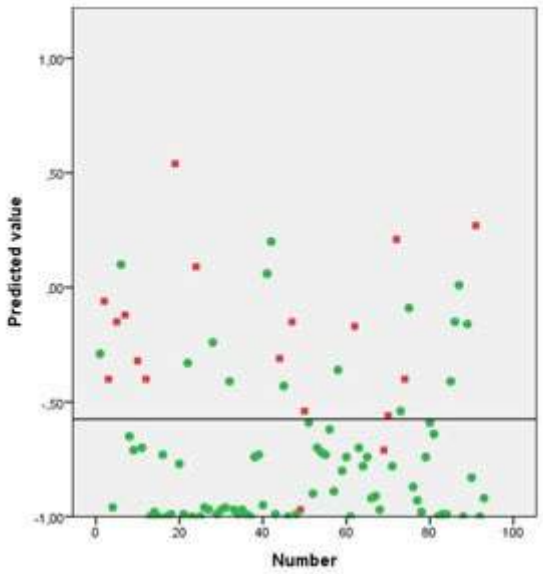



**Figure 4.** Scatterplot with a chosen optimal cut-off value. SCLC vs. all negatives

## Discussion

This exploratory study showed that exhaled-breath analysis with the Aeonose™ can differentiate between patients with lung cancer and healthy subjects, including suspected subjects that are considered negative after investigation. The Aeonose™ could discriminate between breath prints of NSCLC patients and all negatives with a high sensitivity and high NPV for various, but not all types of lung cancer, implying that many subjects could be prevented from undergoing unnecessary invasive diagnostic procedures. These results are in agreement with results published previously (13;19;22-24).

The Aeonose™ was also able to distinguish NSCLC patients from patients who were suspected for lung cancer. However, the diagnostic performance of the Aeonose™ differed when negative subjects were split in suspected, but considered negative after investigation, and healthy volunteers with an AUC of 0.73 and 0.85 respectively, showing a remarkable decline in performance for the suspected subjects. This could be explained by the fact that not lung cancer caused complaints, leading to referral, but other diseases such as COPD or pneumonia. These other diseases could lead to different breath prints, possibly more resembling lung cancer patterns, and could therefore not properly be distinguished by the pattern recognition software. The Aeonose™ could likely be trained to distinguish these other diseases as well, when the number of participants in these groups are sufficiently large. Another explanation could be the overlap in smoking behaviour between the suspected patients without lung cancer and patients with lung cancer, which could lead to a considerable resemblance in metabolism and breath pattern.

From Table 1 it can be seen that the healthy volunteers are more often female and never smokers. What effect this might have on the diagnostic parameters of the Aeonose™ is unknown and needs to be investigated in a larger study.

We found better sensitivity (94.4%), at a noticeable lower specificity (32.9%) in our analyses than reported in other eNose studies (13;25). This might be explained by the fact that in clinical use high NPV and sensitivity are essential when using Enose technology in an early diagnostic stage, on which we based our position at the ROC-curve. As a consequence, this leads to a lower specificity in our study.

The Aeonose™ was able to exclude SCC with a NPV of 93%, which accounts for a clinical relevant diagnostic power. This could be explained by the often central origin of this type of tumour (2). However, the incidence of SCC was lower than in the other groups. Therefore, including more subjects should prove the validity of this high NPV. AC itself could also be distinguished significantly from non-adenocarcinoma, but with lesser performance than SCC. Since AC's are known for their histological heterogeneity, these tumours could probably be subclassified further into tumours with

similar characteristics and consequently improved performance in Aeonose™ diagnostics (26). This hypothesis is supported further by findings of Shlomi et al. who showed a diagnostic accuracy of 83% to discriminate between AC patients with and without an EGFR mutation (27). Findings of improved performance of exhaled-breath analysis in lung cancer when subdividing tumours on histological or molecular biological grounds have also been presented by Barash et al (28). They reported an accuracy of 96% when discriminating adenocarcinoma from squamous cell carcinoma when using gold nanoparticle sensors, albeit with fewer number of patients and the need to use multiple different sensors.

Next to this, we found promising results in excluding SCLC from healthy controls with a high NPV of 96.8%, taking into consideration that this analysis was performed with a relatively low number of subjects due to the lower prevalence of SCLC.

Pattern recognition of a large amount of VOC's leading to a breath signature is only one of the methods used in electronic nose technology. Other methods used for breath sampling in lung cancer, such as gas chromatography-mass spectrometry (GCMS) or multicapillary column-ion mobility spectrometry aim for the detection, identification and quantification of specific, individual chemical compounds in exhaled breath (29-31). In principle, these complex methods are sensitive, but more expensive and time-consuming, and require a specialized operator for the system. When looking for a convenient and low-cost tool to detect lung cancer, point-of-care VOC pattern recognition techniques are favourable.

We performed a study with a relative large study population in a multi-centre setting where we observed an acceptable difference in breath prints of lung cancer patients versus subjects without lung cancer, despite different environments. Next to this, we showed that subdivision of NSCLC types can improve performance and requires further investigation, as earlier shown by Peled et al (32). This was however analysed with GCMS. Our findings further support the transferability of calibration models between different Aeonose™ devices, which supports the results of a smaller, single-centre study of van de Goor et al (22).

Results from this multi-centre study are promising. The technique seems especially valuable in addition to a screening trial based on periodical low dose CT scanning. Electronic nose technology could be able to diminish the number of false positive cases by choosing a cut-off point resulting in an NPV of nearly 100%. Subjects with a false positive diagnosis according to LDCT can subsequently be excluded without having to undergo an invasive bronchoscopy. However, it must be noted that our study population with subjects suspected of lung cancer differs from the high-risk subjects included in LDCT-screening.

This point of view can be seen as a limitation of our study since the majority of the included NSCLC patients was classified as stage III and stage IV disease. These are not the patients that would benefit most from screening programs. However, screening is mostly aimed at patients without symptoms, so when introducing a screening program, probably more cases of stage I and stage II disease can be detected. In this study, the prevalence of stage I and II lung cancer was too low to draw firm conclusions about the detection rates in early stage lung cancer. In future, larger studies, it should become clear if stage I and II tumours could be detected by exhaled-breath analysis as well. In such studies, really large numbers of participants will be required, including supposingly healthy persons.

In this study we showed the training phase of the Aeonose™ to detect or exclude NSCLC with promising results. Including more subjects for training the artificial neural network will likely lead to improved stability, and a better prediction model. Especially differentiation of lung cancer from other lung diseases is expected to improve when data of more patients are analysed, and breath profiles relating to other lung diseases can be taken into account.

Next to further training of the predictive performance, external validation of the obtained results needs to take place in a new study population, preferably in a multi-centre setting as well. It should be noted, however, that all results presented in this study were obtained using leave-10%-out cross validation. This implies in fact that -in 10 consecutive steps- all data were predicted as if they were blind data, based on a training model built from the remaining 90% of data. So, it is to be expected, true blind data will be predicted with similar results as in the cross validation, provided the cohorts are similar.

## Conclusion

Exhaled breath analysis is a rapidly developing field. Electronic nose technology with the Aeonose™ is a non-invasive diagnostic tool that can discriminate between patients with lung cancer and healthy subjects, including subjects suspected of lung cancer with a rejected diagnosis and healthy volunteers. The Aeonose™ is also able to discriminate between lung cancer patients with different subtypes of NSCLC, such as adenocarcinoma and squamous cell carcinoma from healthy subjects, and SCLC from healthy subjects. The data suggest that the Aeonose™ can contribute to the early diagnostic workup of lung cancer where it could provide added value in screening for lung cancer. However, the results must first be validated externally in a new multi-center study with a larger study population.

## Conflicts of interest and Source of Funding

Miss S. Kort was partly financed by an unrestricted research grant from The eNose Company. Dr. J.W. Gerritsen is employed by the company producing the e-nose devices used. The other authors report no conflict of interest.

## References


(1) Torre LA, Siegel RL, Jemal A. Lung Cancer Statistics. Adv Exp Med Biol 2016;893:1-19.

(2) Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. J Thorac Oncol 2015 Sep;10(9):1243-60.

(3) Henschke CI. Survival of patients with clinical stage I lung cancer diagnosed by computed tomography screening for lung cancer. Clin Cancer Res 2007 Sep 1;13(17):4949-50.

(4) Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol 2016 Jan;11(1):39-51.

(5) Quadrelli S, Lyons G, Colt H, Chimondeguy D, Buero A. Clinical characteristics and prognosis of incidentally detected lung cancers. Int J Surg Oncol 2015;2015:287604.

(6) Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C, et al. Cancer statistics, 2006. CA Cancer J Clin 2006 Mar;56(2):106-30.

(7) Patz EF, Jr., Greco E, Gatsonis C, Pinsky P, Kramer BS, Aberle DR. Lung cancer incidence and mortality in National Lung Screening Trial participants who underwent low-dose CT prevalence screening: a retrospective cohort analysis of a randomised, multicentre, diagnostic screening trial. Lancet Oncol 2016 May;17(5):590-9.

(8) Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011 Aug 4;365(5):395-409.

(9) Oudkerk M, Devaraj A, Vliegenthart R, Henzler T, Prosch H, Heussel CP, et al. European position statement on lung cancer screening. Lancet Oncol 2017 Dec;18(12):e754-e766.

(10) Machado RF, Laskowski D, Deffenderfer O, Burch T, Zheng S, Mazzone PJ, et al. Detection of lung cancer by sensor array analyses of exhaled breath. Am J Respir Crit Care Med 2005 Jun 1;171(11):1286-91.

(11) Chen X, Xu F, Wang Y, Pan Y, Lu D, Wang P, et al. A study of the volatile organic compounds exhaled by lung cancer cells in vitro for breath diagnosis. Cancer 2007 Aug 15;110(4):835-44.

(12) Bajtarevic A, Ager C, Pienz M, Klieber M, Schwarz K, Ligor M, et al. Noninvasive detection of lung cancer by analysis of exhaled breath. BMC Cancer 2009 Sep 29;9:348.

(13) Dragonieri S, Annema JT, Schot R, van der Schee MP, Spanevello A, Carratu P, et al. An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD. Lung Cancer 2009 May;64(2):166-70.

(14) D'Amico A, Di NC, Falconi C, Martinelli E, Paolesse R, Pennazza G, et al. Detection and identification of cancers by the electronic nose. Expert Opin Med Diagn 2012 May;6(3):175-85.

(15) Peled N, Hakim M, Bunn PA, Jr., Miller YE, Kennedy TC, Mattei J, et al. Non-invasive breath analysis of pulmonary nodules. J Thorac Oncol 2012 Oct;7(10):1528-33.

(16) Hubers AJ, Brinkman P, Boksem RJ, Rhodius RJ, Witte BI, Zwinderman AH, et al. Combined sputum hypermethylation and eNose analysis for lung cancer diagnosis. J Clin Pathol 2014 Aug;67(8):707-11.

(17) Boots AW, van Berkel JJ, Dallinga JW, Smolinska A, Wouters EF, van Schooten FJ. The versatile use of exhaled volatile organic compounds in human health and disease. J Breath Res 2012 Jun;6(2):027108.

(18) van der Schee MP, Paff T, Brinkman P, van Aalderen WMC, Haarman EG, Sterk PJ. Breathomics in lung disease. Chest 2015 Jan;147(1):224-31.

(19) van de Goor R, van Hooren M, Dingemans AM, Kremer B, Kross K. Training and Validating a Portable Electronic Nose for Lung Cancer Screening. J Thorac Oncol 2018 May; 13(5):676-681.

(20) Bruins MGJ, van de Sande W, van Belkum A, Bos A. Enabling a transferable calibration model for metal-oxide type electronic noses. Sensors and actuators B 188, 1187-1195. 2013.

(21) Kort S, Brusse-Keizer M, Gerritsen JW, van der Palen J. Data analysis of electronic nose technology in lung cancer: generating prediction models by means of Aethena. J Breath Res 2017 Jun 1;11(2):026006.

(22) Mazzone PJ, Hammel J, Dweik R, Na J, Czich C, Laskowski D, et al. Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array. Thorax 2007 Jul;62(7):565-8.

(23) Mazzone PJ. Analysis of volatile organic compounds in the exhaled breath for the diagnosis of lung cancer. J Thorac Oncol 2008 Jul;3(7):774-80.

(24) Schumer EM, Trivedi JR, van B, V, Black MC, Li M, Fu XA, et al. High sensitivity for lung cancer detection using analysis of exhaled carbonyl compounds. J Thorac Cardiovasc Surg 2015 Dec;150(6):1517-22.

(25) D'Amico A, Pennazza G, Santonico M, Martinelli E, Roscioni C, Galluccio G, et al. An investigation on electronic nose diagnosis of lung cancer. Lung Cancer 2010 May;68(2):170-6.

(26) Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 2008 Oct 23;455(7216):1069-75.

(27) Shlomi D, Abud M, Liran O, Bar J, Gai-Mor N, Ilouze M, et al. Detection of Lung Cancer and EGFR Mutation by Electronic Nose System. J Thorac Oncol 2017 Oct;12(10):1544-51.

(28) Barash O, Peled N, Tisch U, Bunn PA, Jr., Hirsch FR, Haick H. Classification of lung cancer histology by gold nanoparticle sensors. Nanomedicine 2012 Jul;8(5):580-9.

(29) Machado RF, Laskowski D, Deffenderfer O, Burch T, Zheng S, Mazzone PJ, et al. Detection of lung cancer by sensor array analyses of exhaled breath. Am J Respir Crit Care Med 2005 Jun 1;171(11):1286-91.

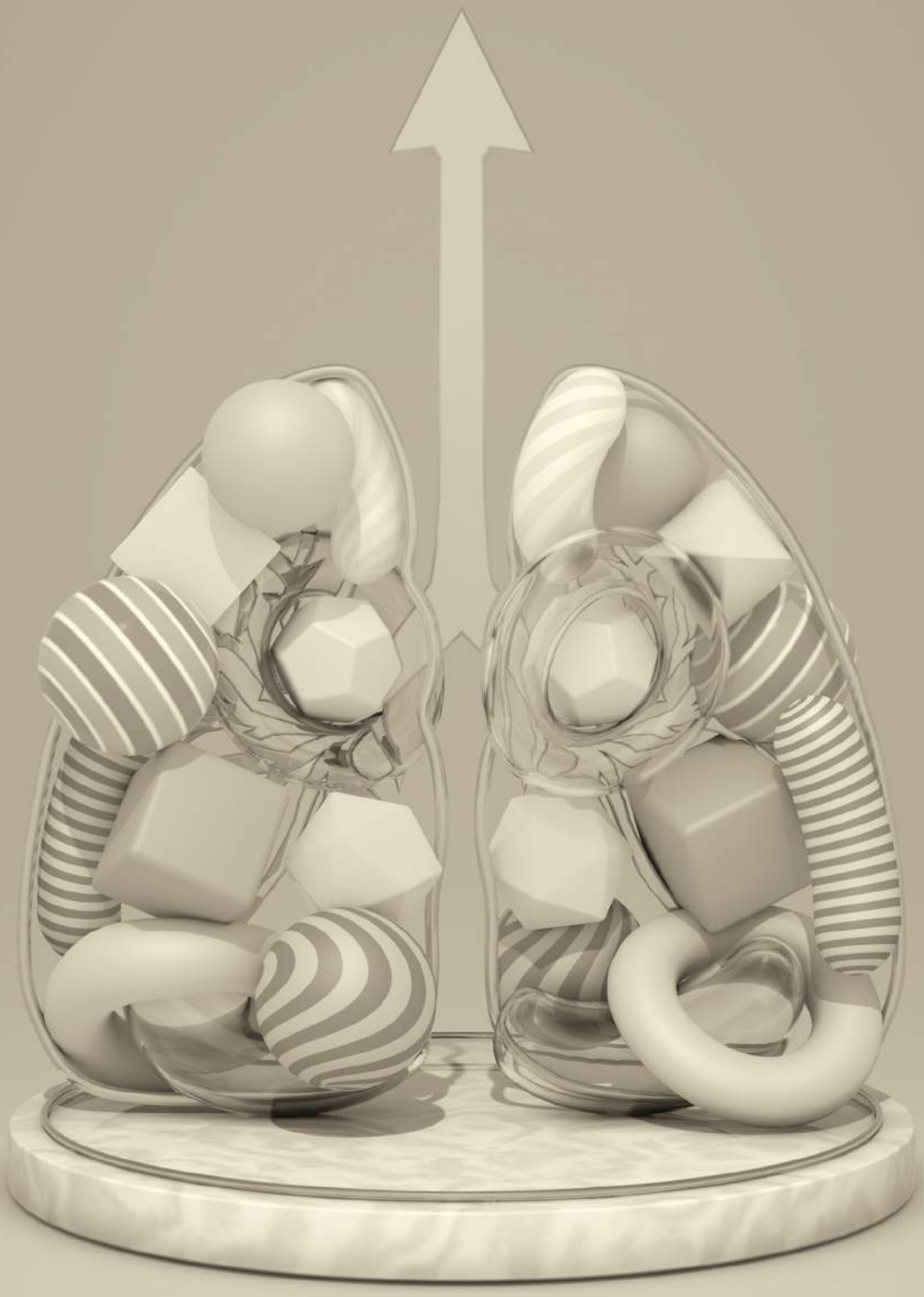
(30) Gordon SM, Szidon JP, Krotoszynski BK, Gibbons RD, O'Neill HJ. Volatile organic compounds in exhaled air from patients with lung cancer. Clin Chem 1985 Aug;31(8):1278-82.

(31) Phillips M, Altorki N, Austin JH, Cameron RB, Cataneo RN, Kloss R, et al. Detection of lung cancer using weighted digital analysis of breath biomarkers. Clin Chim Acta 2008 Jul 17;393(2):76-84.

(32) Peled N, Hakim M, Bunn PA, Jr., Miller YE, Kennedy TC, Mattei J, et al. Non-invasive breath analysis of pulmonary nodules. J Thorac Oncol 2012 Oct;7(10):1528-33.

# 05

# IMPROVING LUNG CANCER DIAGNOSIS BY COMBINING EXHALED-BREATH DATA AND CLINICAL PARAMETERS



S. KORT - M. BRUSSE-KEIZER - J.W. GERRITSEN - J.H. SCHOUWINK - E. CITGEZ
F.H.C. DE JONGH - S. SAMII - J. VAN DER MATEN - M. VAN DEN BOGART
J. VAN DER PALEN

ERJ OPEN RES. 2020 MAR 16;6(1):00221-2019

## Improving lung cancer diagnosis by combining exhaled-breath data and clinical parameters

S. Kort [1]; M. Brusse-Keizer [2]; J.W. Gerritsen [3]; J.H. Schouwink [1]; E. Citgez [1]; F.H.C. de Jongh [1]; J. van der Maten [4]; S. Samii [5]; M. van den Bogart [6]; J. van der Palen [2,7].

1: Department of Pulmonary Medicine, Medisch Spectrum Twente, Enschede, the Netherlands

2: Medical School Twente, Medisch Spectrum Twente, Enschede, the Netherlands

3: The eNose Company, Zutphen, the Netherlands

4: Department of Pulmonary Medicine, Medisch Centrum Leeuwarden, Leeuwarden, the Netherlands

5: Department of Pulmonary Medicine, Deventer Ziekenhuis, Deventer, the Netherlands

6: Department of Pulmonary Medicine, Bernhoven Uden, Uden, the Netherlands

7: Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, the Netherlands

## Abstract

**Introduction:** Exhaled-breath analysis of volatile organic compounds (VOC's) could detect lung cancer earlier, possibly leading to improved outcomes. Combining exhaled-breath data with clinical parameters may improve lung cancer diagnosis.

**Methods**: Based on data of a previous performed multi-centre study, the current manuscript reports on additional analyses. 138 subjects with non-small cell lung cancer (NSCLC) and 143 controls without NSCLC breathed into the Aeonose™. The diagnostic accuracy, presented as area under the receiver operating characteristic curve (AUC-ROC), of the Aeonose™ itself was compared with a) performing a multivariate logistic regression analysis of the distinct clinical parameters obtained, and b) by using this clinical information beforehand in the training process of the artificial neural network (ANN) for the breath analysis.

**Results:** NSCLC patients (67.1 (9.1) years; 58% male) were compared with controls (62.1 (7.0) years; 40.6% male). The AUC-ROC of the classification value of the Aeonose™ itself was 0.75 (95% CI: 0.69-0.81). Adding age, number of pack years, and presence of COPD to this value in a multivariate regression analysis resulted in an improved performance with an AUC-ROC of 0.86 (95% CI: 0.81-0.90). Adding these clinical variables beforehand to the ANN for classifying the breath print also led to an improved performance with an AUC-ROC of 0.84 (95% CI 0.79-0.89).

**Conclusions:** Adding readily available clinical information to the classification value of exhaled-breath analysis with the Aeonose™, either post-hoc in a multivariate regression analysis or a-priori to the ANN, significantly improves the diagnostic accuracy to detect the presence or absence of lung cancer.

## List of abbreviations

| | |
|---|---|
| ANN | artificial neural network |
| AUC-ROC | area under the receiver operating characteristic curve |
| GC/MS | gas chromatography / mass spectrometry |
| LDCT | low dose computed tomography |
| NPV | negative predictive value |
| NSCLC | non-small cell lung cancer |
| PPV | positive predictive value |
| ROC | receiver operating curve |
| SCC | squamous cell carcinoma |
| SCLC | small cell lung carcinoma |
| VOC | volatile organic compound |

## Introduction

Lung cancer remains the leading cause of cancer-related death worldwide, accounting for approximately 5% of total mortality in many countries (1). Unfortunately, the majority of patients presents with late-stage disease (stage III and IV) accompanied by limited effective treatment options and consequently high mortality rates with a five-year survival rate of less than 10% (2;3).
Currently, the only screening method leading to reduced lung cancer mortality in high-risk groups is annual low-dose computed tomography (LDCT) (4;5). However, LDCT-screening for lung cancer has also resulted in a notable rate of false positive cases leading to unnecessary invasive procedures, risks due to radiation exposure, and unnecessary anxiety (4-8). In Europe results of the Dutch-Belgian lung cancer screening trial (NELSON) are awaited before a decision on implementation of screening programs in Europe will be made (9). Hence, there is an increasing demand for innovative, non-invasive, point-of-care diagnostic tools to detect lung cancer at an early stage.

Exhaled-breath analysis with electronic nose technology is a technique based on detecting combinations of volatile organic compounds (VOC's) that are exhaled in very low concentrations. These VOC's reflect pathophysiological processes, such as infection, inflammation and neoplasms (10-12). VOCs are of interest since they might be directly related to the presence of diseases and can test non-invasively by pattern recognition techniques serving as classifiers for diseases. Several studies on exhaled-breath analysis have supported the hypothesis that VOC-patterns alter when lung cancer is present (13-20).

Recently, we reported results of a study (including 290 subjects) differentiating subjects with lung cancer, including classification into subtypes of lung cancer, and healthy individuals by means of exhaled-breath analysis with the Aeonose™ (13). In this study, the artificial neural network (ANN) was trained using exhaled-breath data only.
The Aeonose™ was able to diagnose patients with non-small cell lung cancer (NSCLC) with a sensitivity of 94%, a negative predictive value (NPV) of 85%, and an area under the receiver operating characteristic curve (AUC-ROC) of 0.76. Also subtyping NSCLC into adenocarcinoma and squamous cell carcinoma showed promising results. These diagnostic parameters were based on the analysis of exhaled VOC's only and did not take into account any subjects' risk factors, such as age, gender, smoking status (number of pack years), and presence of chronic obstructive pulmonary disease (COPD). This paper describes the potential of adding specific clinical information to the classification value obtained from the Aeonose™ on the diagnostic accuracy to diagnose lung cancer. The hypothesis is that adding clinical information would improve the diagnostic performance. This was assessed in two ways: First, the clinical information was added afterwards to the classification value

of the Aeonose™ as obtained from the ANN by applying multivariate logistic regression analysis, and second, by using this clinical information a-priori in the training process of the ANN.

## Methods

Data were obtained from a previous performed prospective, multi-centre study where subjects suspected for lung cancer, as well as healthy volunteers were asked to participate (13). The originally collected breath samples were currently used for additional analyses. The four secondary teaching hospitals participating were Medisch Spectrum Twente Enschede, Bernhoven Uden, Medisch Centrum Leeuwarden, and Deventer Ziekenhuis. For patients with confirmed lung cancer based on histopathology, staging was established according to the 7$^{th}$ edition of the American Joint Committee on Cancer TNM staging system (21). The control group consisted of suspected subjects with a rejected diagnosis based on imaging and/or derived histopathology and healthy volunteers. Healthy volunteers with a minimum age of 50 were recruited through an advertisement at the hospitals' website. There were two exclusion criteria for all subjects: another active malignancy in the past five years or the inability to perform a complete Aeonose™ measurement. Demographic data were collected including age, gender, body mass index (BMI), smoking status, number of pack years, and presence of COPD, hypertension and diabetes mellitus.

The Aeonose™ is a hand-held electronic-nose device containing three metal-oxide sensors (13;22;23). This device is a non-invasive, easy-to-use, low-cost tool that is, once trained and validated, able to perform real-time analysis to detect lung cancer. Temporary storage of the breath sample is not required. Subjects were instructed to breathe through the Aeonose™ for five minutes with their nose clipped to prevent nose breathing.
The study protocol was approved by the medical ethics committee of Medisch Spectrum Twente, and by the board of directors at each participating centre. All subjects signed an informed consent.

Statistical analysis

Continuous variables are reported as mean with corresponding standard deviation (SD) or as median with interquartile range (IQR). Nominal variables are reported as numbers with corresponding percentages. To assess differences between the groups, either the T-test for normally distributed continuous variables, Mann-Whitney U test for skewed distributed continuous or ordinal variables, or chi-squared test ($x^2$) for nominal and categorical variables was applied. Number of pack years was categorized as none, up to 20 pack years, between 21 and 40 pack years, and more than 40 pack years. Based on clinical reasoning we assumed a strong relationship between smoking status and number of pack years, which was confirmed ($p<0.001$). Number of pack years contained most

relevant information. Therefore, we excluded smoking status as clinical variable from the multivariate analysis.

Data of exhaled breath were analysed by Aethena, a proprietary, dedicated software package from The eNose Company. The software package comprises techniques for data pre-treatment, data compression methods, ANN training and classification to assess the probability of lung cancer, ranging a single value between -1 and +1 (23). ANN's have been developed as an alternative statistical technique to perform multifactorial analyses by interconnecting nodes by weighted connection lines to predict outcomes or classifying values on an individualized basis (24).

Sensitivity, specificity, positive predictive value (PPV) and NPV were calculated for the diagnosis of lung cancer based on the classification by the Aeonose™, and receiver operating characteristics (ROC) curves were composed with a corresponding AUC-ROC with 95% confidence interval.

Clinical variables that were univariately associated with the presence of lung cancer ($p<0.15$) were entered in a multivariate logistic regression analysis where variables with the highest p-values were eliminated step-by-step (backward method), until the fit of the model decreased significantly, based on the -2 log likelihood. This analysis was based on clinical variables only.

Subsequently, two types of multivariate analysis were performed where breath data were included: First, another multivariate logistic regression analysis, consistent with the abovementioned method, was performed together with the classification value of the Aeonose™ as obtained from the ANN.

Second, clinical variables that were univariately associated with the presence of lung cancer ($p<0.15$) were added to the vector containing breath profile information once data compression had been realised. These extended vectors (one per subject) were used for training the ANN.

Sensitivity, specificity, PPV, NPV and AUC-ROC were then calculated for the diagnosis of lung cancer according to the selected multivariate logistic regression model and the extended ANN. These outcomes were compared with the diagnostic accuracy obtained by the classification result of the exhaled-breath analysis only.

The multivariate regression model was internally validated by 1000 iterations of bootstrap.

All statistical tests were two-sided with a significance level at 0.05. SPSS V.24.0 was used to perform statistical mathematics.

## Results

A total of 281 subjects were included, of whom 138 had confirmed NSCLC. The control group consisted of 143 subjects without lung cancer of whom 59 were suspected for lung cancer, but were considered negative after investigation, and 84 subjects were healthy volunteers (Figure 1). Table 1 provides a description of the study participants including clinical characteristics for both groups.

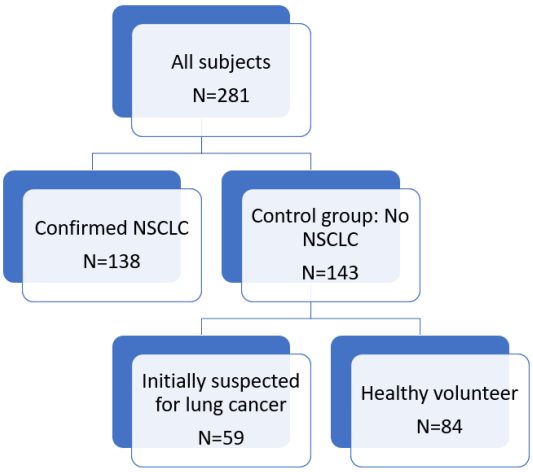



**Figure 1.** Flow chart showing the different groups. NSCLC: non-small cell lung cancer.

| Table 1. Clinical characteristics of subjects. | | | | | |
|---|---|---|---|---|---|
| | Confirmed NSCLC N=138 | Total control group N=143 | Suspected, proven negative N=59 | Healthy volunteer N=84 | P-value |
| **Age in years, mean (SD)** | 67.1 (9.1) | 62.1 (7.0) | 65.2 (8.8) | 59.8 (4.3) | <0.001[a] |
| **Sex, number of males (%)** | 80 (58.0) | 58 (40.6) | 31 (52.5) | 27 (32.1) | <0.001[b] |
| **Smoking status, N (%)** | | | | | |
| **Current smoker** | 49 (35.5) | 19 (13.3) | 13 (22.0) | 6 (7.1) | <0.001[b] |
| **Ex-smoker** | 82 (59.4) | 76 (53.1) | 32 (54.2) | 44 (52.4) | |
| **Never smoked** | 7 (5.1) | 48 (33.6) | 14 (23.7) | 34 (40.5) | |
| **Pack years, N (%)** | | | | | |
| **0** | 7 (5.1) | 48 (33.6) | 14 (23.7) | 34 (40.5) | |
| **1-20** | 30 (21.7) | 53 (37.1) | 18 (30.5) | 35 (41.7) | <0.001[b] |
| **21-40** | 53 (38.4) | 25 (17.5) | 17 (28.8) | 8 (9.5) | |
| **>40** | 48 (34.8) | 17 (11.9) | 10 (16.9) | 7 (8.3) | |
| **COPD, N (%)** | 66 (47.8) | 22 (15.4) | 21 (35.6) | 1 (1.2) | <0.001[b] |
| **BMI, mean (SD)** | 25.6 (4.6) | 25.9 (4.8) | 26.9 (5.9) | 25.2 (3.8) | 0.104 |
| **Type of NSCLC, N (%)** | | | | | |
| **Adenocarcinoma** | 88 (63.8) | | | | |
| **Squamous cell carcinoma** | 41 (29.7) | | | | |
| **Large cell carcinoma** | 4 (2.9) | | | | |
| **NOS** | 5 (3.6) | | | | |
| **NSCLC stage, N (%)[#]** | | | | | |
| **I** | 25 (14.5) | | | | |
| **II** | 15 (10.8) | | | | |
| **III** | 39 (28.3) | | | | |
| **IV** | 64 (46.4) | | | | |

**[a] After Games-Howell correction, there was a significant difference between healthy volunteers and confirmed NSCLC and healthy volunteers and suspected, proven negative subjects.**
**[b] After Holm-Bonferroni correction there was a significant difference between healthy volunteers and confirmed NSCLC and suspected proven negative subjects.**
**[#] Staging established according to the 7th edition of the American Joint Committee on Cancer TNM staging system**
**Abbreviations: COPD: chronic obstructive pulmonary disease; BMI, body mass index; NSCLC, non-small cell lung cancer; NOS, not otherwise specified**

Lung cancer patients were significantly older (mean age 67.1 ± 9.1 years), more likely to be male and current or ex-smoker, had smoked more pack years and were more often diagnosed with COPD than subjects in the control group. Almost 75% of the NSCLC patients were classified as stage III or IV disease.

Table 2 shows that sex, age, smoking status, number of pack years, presence of COPD and the classification value obtained by the Aeonose™ were univariately associated with the presence of lung cancer. Subsequently, we added these candidate variables to a multivariate regression analysis which showed that age, number of pack years, presence of COPD and the value of the Aeonose™ remained

significantly predictive for the presence of lung cancer.

**Table 2. Results of the univariate and multivariate logistic regression analyses for diagnosing lung cancer.**

| Variable | Univariate analysis Odds ratio (95% CI) | Multivariate analysis Odds ratio (95% CI) | Regression coefficient (B) |
|---|---|---|---|
| | | | Constant: -5.54 |
| **Sex** | 2.01 (1.26-3.20) | 1.42 (0.76-2.58) | 0.34 |
| **Age** | 1.08 (1.05-1.11) | 1.05 (1.02-1.09) | 0.05 |
| **BMI** | 0.99 (0.94-1.04) | - | |
| **Smoking status**<br>**Current smoker**<br>**Ex-smoker**<br>**Never smoked** | <br>17.49 (6.79-45.06)<br>7.56 (3.23-17.69)<br>Reference | - | |
| **Pack years**<br>**0**<br>**1-20**<br>**21-40**<br>**>40** | <br>Reference<br>3.88 (1.56-9.65)<br>14.77 (5.89-37.04)<br>19.36 (7.36-50.91) | <br>Reference<br>3.48 (1.25-9.66)<br>10.20 (3.66-28.46)<br>11.69 (4.04-33.87) | <br><br>1.25<br>2.32<br>2.46 |
| **COPD** | 4.90 (2.80-8.58) | 2.29 (1.18-4.43) | 0.83 |
| **Diabetes mellitus** | 0.70 (0.30-1.64) | - | |
| **Classification value Aeonose™(13)** | 24.20 (9.71-60.33) | 12.67 (4.48-35.83) | 2.54 |

**Abbreviations: CI: confidence interval; BMI: body mass index; COPD: chronic obstructive pulmonary disease; -: not added to the multivariate model.**

Each additional year of age was associated with a 5% higher chance of having lung cancer (OR 1.05, 95% CI: 1.02-1.09). Subjects who have smoked up to 20 pack years have a 3.5-fold higher chance of developing lung cancer (OR 3.5, 95% CI 1.25-9.66), whereas those who have smoked more than 40 pack years have a 11.7-fold higher chance (OR 11.7, 95% CI: 4.04-33.87). Patients with COPD had a 2.3-fold increased risk of having lung cancer (OR 2.27, 95% CI: 1.18-4.43). The classification value of the Aeonose™ was also strongly associated with the presence of lung cancer (OR 12.7, 95% CI: 4.48-35.83).

The multivariate logistic regression analysis based on clinical variables only showed a sensitivity of 93.5%, a specificity of 50%, a PPV of 64.5% and an NPV of 88.8%. This corresponded with an AUC-ROC of 0.80 (95% CI: 0.75-0.85).

When the ANN was trained with exhaled-breath data from the Aeonose™ only, we found a sensitivity of 94.2%, a specificity of 44.1%, and a PPV and NPV of respectively 61.9% and 88.7% with an AUC-ROC of 0.75 (95% CI 0.69-0.81) (Table 3). When applying the multivariate logistic regression model including the resulting value (-1 to +1) of the exhaled breath data from the ANN in the exact same study population, we found an improved performance to distinguish NSCLC patients from controls with an AUC-ROC of 0.86 (95% CI: 0.81-0.90). By choosing a relevant threshold value in the ROC-

curve focusing on high sensitivity and high NPV, the analysis showed a sensitivity of 95.7%, a specificity of 59.7%, and a PPV and NPV of 69.5% and 92.5%, respectively. The bootstrap analysis for internal validation showed similar regression coefficients compared to our original model showing robustness of the model.

When training the ANN with exhaled breath data combined with the clinical variables that were univariately associated with the presence of lung cancer, we found an improved diagnostic performance as well to distinguish NSCLC patients from controls, showing an AUC-ROC of 0.84 (95% CI: 0.79-0.89). By choosing an appropriate threshold value in the ROC-diagram, we observed a sensitivity of 94.2%, a specificity of 49.0%, and a PPV and NPV of 64.0% and 89.7% respectively. Figure 2 shows the combined ROC-curve showing the improved performance of both multivariate models.

**Table 3. Diagnostic performance of the three investigated prediction models.**

| | N (positive vs negative | Optimal cut-off | Sensitivity (%) | Specificity (%) | PPV (%) | NPV (%) | AUC (95% CI) |
|---|---|---|---|---|---|---|---|
| **Clinical variables only** | 138 vs 143 | 0.32 | 93.5 | 50.0 | 64.5 | 88.8 | 0.80 (0.75-0.85) |
| **Aeonose™ result only** | 138 vs 143 | -0.38 | 94.2 | 44.1 | 61.9 | 88.7 | 0.75 (0.69-0.81) |
| **Multivariate logistic regression model** | 138 vs 143 | 0.27 | 95.7 | 59.7 | 69.5 | 92.5 | 0.86 (0.81-0.90) |
| **Extended ANN** | 138 vs 143 | -0.65 | 94.2 | 49.0 | 64.0 | 89.7 | 0.84 (0.79-0.89) |

**Abbreviations: ANN, artificial neural network; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; CI, confidence interval.**

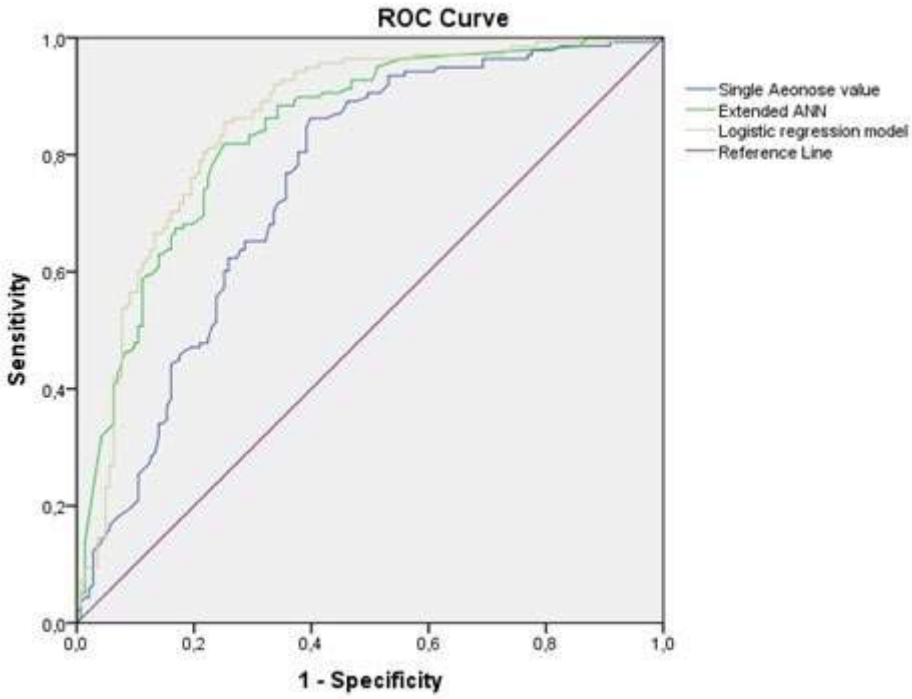



**Figure 2.** Combined ROC-curve showing four predictive algorithms: Logistic regression clinical variables, Single Aeonose™ value, extended ANN and logistic regression including Aeonose™ value. ANN: artificial neural network.

## Discussion

This study assessed the impact of combining exhaled-breath analysis and clinical parameters in diagnosing lung cancer. We showed that adding readily available clinical information to the classification value of exhaled-breath analysis by the Aeonose™ in a relatively easy-to-perform multivariate regression model improved the diagnostic accuracy, expressed as AUC-ROC, from 0.75 to 0.86 to diagnose lung cancer in a non-invasive way. When extending the vector containing compressed breath data with clinical variables, and in this way training the ANN to distinguish between sick and healthy individuals, the diagnostic accuracy, expressed as AUC-ROC, increased from 0.75 to 0.84. Minor differences were observed compared to our previous performed analysis due to the somewhat smaller sample size because of missing information on pack years (previous study n=290, AUC-ROC=0.76) and the fact that outcomes of frequently training an ANN can slightly fluctuate.

It turns out that logistic regression analysis and ANN are equally capable of increasing classification quality of lung cancer diagnosis. We expected to see rather improved accuracy when entering a combination of clinical and exhaled-breath data directly to the yet untrained ANN, because it considers possible interactions. Independency of parameters, like breath profile, COPD and pack years cannot be fully assumed so far. As extending exhaled-breath data with clinical parameters

followed by classification by an ANN is more complex than building a multivariate regression model out of single clinical parameters, the latter is to be recommended for practical use.
Besides ANN and logistic regression analysis, other statistical learning methods for classifying exhaled breath data are available, such as Random Forest and Support Vector Machines (25). In a previous study, results from neural network analysis were compared to results obtained from Random Forest and Support Vector Machine showing comparable diagnostic performance (26). In this study we only focused on the two statistical learning methods described: logistic regression analysis and neural network analysis.
Previous studies have shown that electronic nose technology based on pattern recognition of VOC's, or identifying VOC's with gas chromatography/mass spectrometry (GC/MS) can differentiate between subjects with and without lung cancer (13;14;17;20;27-29). Several studies using techniques for VOC-identification have used logistic regression analysis to identify lung cancer-specific VOC's (30;31). However, logistic regression analysis including clinical parameters in studies using pattern recognition techniques has not been shown often yet. Tirzïte et al. used logistic regression analysis to predict the presence of lung cancer with the Cyranose 320 electronic nose mainly using segments of exhaled breath as input variables for the logistic regression analysis, but also including a few clinical parameters, such as age, smoking status, smoking history and ambient temperature (19). They were able to distinguish subjects with lung cancer from controls with a sensitivity of 96% in both smokers and non-smokers and a specificity >90% in both groups. To our knowledge, no studies have been performed using clinical variables and exhaled-breath data based on pattern recognition combined in an ANN to diagnose lung cancer. However, several studies have used ANN's to detect lung cancer, but without performing exhaled-breath analysis. These studies mainly focused on clinical parameters and biomarkers based on blood and genetic abnormalities (32;33).
As described in our training study, the optimal cut-off point chosen determines the amount of false positive cases contrary to the amount of missed cases concerning lung cancer (13)**.** We focused again on a high sensitivity and a high NPV, since lung cancer has an extremely high mortality when not detected early. By adding the clinical variables to the exhaled breath data, we saw in both models that all diagnostic parameters improved, and thereby reaching higher sensitivity and NPV compared to the training study, but we also observed fewer false-positive cases by achieving higher specificity.

In the near future, the results obtained could be proposed as added value in several ways. First, in case of implementation of LDCT in Europe, the Aeonose™ may be deployed after suspicion of lung cancer has been raised with LDCT. Due to the high NPV with the Aeonose™, subjects could be prevented from undergoing unnecessary invasive interventions, and be monitored with prolonged intervals (8). Also, there is current debate about identifying at-risk groups relevant for LDCT-

screening (34-36). Combining clinical parameters and exhaled-breath data in an ANN could indicate the degree of suspicion of lung cancer and therefore serve as an adjunct for risk stratification in lung cancer screening supporting clinical decision making.

Limitations of our study should also be mentioned. We did not analyse subjects with SCLC, nor did we analyse differences between the histological subtypes of NSCLC and lung cancer stages, since these subgroups were too small to include all relevant variables in the multivariate model. Moreover, we do not know the influence of food intake by subjects. Eating and drinking was not restricted before the exhaled-breath measurement. However, the neural network is being trained comparing breath profiles of positive and negative subjects regardless their food intake. When numbers of subjects are sufficiently large, it can be assumed that food intake is not relevant as it averages out. It should also be noted that the majority of subjects consisted of stage III and IV lung cancer (75%). This population differs from the high-risk asymptomatic subjects suitable for screening where focus lies on early-stage lung cancer. However, the risk factors included in the multivariate analysis are applicable for both early and late-stage lung cancer, so when exhaled-breath analysis is able to detect early-stage lung cancer, readily available clinical information should be incorporated in the analysis. Future analysis, including sufficient stage I and II NSCLC should indicate whether breath patterns already change early in the course of the disease.
We should also note that in high dimensional data set as obtained with the Aeonose™, the problem of overfitting can occur where a prediction model that looks appropriate on training data used to develop it, will perform poorly on future observations. Combining analytical techniques, such as data compression and cross-validation partly overcomes this issue. Currently, an external validation study is performed where a complete new cohort of subjects is included to totally overcome the issue of overfitting.

## Conclusion

Due to the aggressive nature of lung cancer, diagnostic accuracy should be as high as possible. This diagnostic accuracy to detect the presence or absence of lung cancer by exhaled-breath analysis with the Aeonose™ can be improved by adding readily available clinical information either post-hoc in a multivariate logistic regression model, or a-priori in the training process to the ANN compared to the single classification value based on exhaled-breath data only. As both approaches yield similar results, the multivariate logistic regression model should be preferred as its application is more convenient.

## Conflicts of interest

Miss S. Kort was partly financed by an unrestricted research grant from The eNose Company. Dr. J.W. Gerritsen is employed by the company producing the e-nose devices used. The other authors report no conflict of interest.

## References


(1) Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424.

(2) Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol 2016 Jan;11(1):39-51.

(3) Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018 Jan;68(1):7-30.

(4) Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011 Aug 4;365(5):395-409.

(5) Patz EF, Jr., Greco E, Gatsonis C, Pinsky P, Kramer BS, Aberle DR. Lung cancer incidence and mortality in National Lung Screening Trial participants who underwent low-dose CT prevalence screening: a retrospective cohort analysis of a randomised, multicentre, diagnostic screening trial. Lancet Oncol 2016 May;17(5):590-9.

(6) Black WC, Keeler EB, Soneji SS. Cost-effectiveness of CT screening in the National Lung Screening Trial. N Engl J Med 2015 Jan 22;372(4):388.

(7) de Koning HJ, Meza R, Plevritis SK, ten HK, Munshi VN, Jeon J, et al. Benefits and harms of computed tomography lung cancer screening strategies: a comparative modeling study for the U.S. Preventive Services Task Force. Ann Intern Med 2014 Mar 4;160(5):311-20.

(8) Huo J, Xu Y, Sheu T, Volk RJ, Shih YT. Complication Rates and Downstream Medical Costs Associated With Invasive Diagnostic Procedures for Lung Abnormalities in the Community Setting. JAMA Intern Med 2019 Jan 14.

(9) Oudkerk M, Devaraj A, Vliegenthart R, Henzler T, Prosch H, Heussel CP, et al. European position statement on lung cancer screening. Lancet Oncol 2017 Dec;18(12):e754-e766.

(10) Boots AW, van Berkel JJ, Dallinga JW, Smolinska A, Wouters EF, van Schooten FJ. The versatile use of exhaled volatile organic compounds in human health and disease. J Breath Res 2012 Jun;6(2):027108.

(11) Buszewski B, Kesy M, Ligor T, Amann A. Human exhaled air analytics: biomarkers of diseases. Biomed Chromatogr 2007 Jun;21(6):553-66.

(12) Nakhleh MK, Amal H, Jeries R, Broza YY, Aboud M, Gharra A, et al. Diagnosis and Classification of 17 Diseases from 1404 Subjects via Pattern Analysis of Exhaled Molecules. ACS Nano 2017 Jan 24;11(1):112-25.

(13) Kort S, Tiggeloven MM, Brusse-Keizer M, Gerritsen JW, Schouwink JH, Citgez E, et al. Multi-centre prospective study on diagnosing subtypes of lung cancer by exhaled-breath analysis. Lung Cancer 2018 Nov;125:223-9.

(14) van de Goor R, van HM, Dingemans AM, Kremer B, Kross K. Training and Validating a Portable Electronic Nose for Lung Cancer Screening. J Thorac Oncol 2018 May;13(5):676-81.

(15) Phillips M, Altorki N, Austin JH, Cameron RB, Cataneo RN, Greenberg J, et al. Prediction of lung cancer using volatile biomarkers in breath. Cancer Biomark 2007;3(2):95-109.

(16) Machado RF, Laskowski D, Deffenderfer O, Burch T, Zheng S, Mazzone PJ, et al. Detection of lung cancer by sensor array analyses of exhaled breath. Am J Respir Crit Care Med 2005 Jun 1;171(11):1286-91.

(17) Bajtarevic A, Ager C, Pienz M, Klieber M, Schwarz K, Ligor M, et al. Noninvasive detection of lung cancer by analysis of exhaled breath. BMC Cancer 2009 Sep 29;9:348.

(18) Mazzone PJ, Wang XF, Lim S, Jett J, Choi H, Zhang Q, et al. Progress in the development of volatile exhaled breath signatures of lung cancer. Ann Am Thorac Soc 2015 May;12(5):752-7.

(19) Tirzite M, Bukovskis M, Strazda G, Jurka N, Taivans I. Detection of lung cancer with electronic nose and logistic regression analysis. J Breath Res 2018 Nov 20;13(1):016006.

(20) Gasparri R, Santonico M, Valentini C, Sedda G, Borri A, Petrella F, et al. Volatile signature for the early diagnosis of lung cancer. J Breath Res 2016 Feb 9;10(1):016007.

(21) Goldstraw P, Crowley J, Chansky K, Giroux DJ, Groome PA, Rami-Porta R, et al. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. J Thorac Oncol 2007 Aug;2(8):706-14.

(22) Bruins MGJ, van de Sande W, van Belkum A. Enabling a transferable calibration model for metal-oxide type electronic noses. Sens Actuators B Chem 2013;188, 1187-1195.

(23) Kort S, Brusse-Keizer M, Gerritsen JW, van der Palen J. Data analysis of electronic nose technology in lung cancer: generating prediction models by means of Aethena. J Breath Res 2017 Jun 1;11(2):026006.

(24) Hanai T, Yatabe Y, Nakayama Y, Takahashi T, Honda H, Mitsudomi T, et al. Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. Cancer Sci 2003 May;94(5):473-7.

(25) Leopold JH, Bos LD, Sterk PJ, Schultz MJ, Fens N, Horvath I, et al. Comparison of classification methods in breath analysis by electronic nose. J Breath Res 2015 Dec 15;9(4):046002.

(26) Hauschild AC, Baumbach JI, Baumbach J. Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification. Genet Mol Res 2012 Aug 16;11(3):2733-44.

(27) Antoniou SX, Gaude E, Ruparel M, van der Schee MP, Janes SM, Rintoul RC. The potential of breath analysis to improve outcome for patients with lung cancer. J Breath Res 2019 Apr 24;13(3):034002.

(28) Gasparri R, Romano R, Sedda G, Borri A, Petrella F, Galetta D, et al. Diagnostic biomarkers for lung cancer prevention. J Breath Res 2018 Feb 6;12(2):027111.

(29) Peng G, Tisch U, Adams O, Hakim M, Shehada N, Broza YY, et al. Diagnosing lung cancer in exhaled breath using gold nanoparticles. Nat Nanotechnol 2009 Oct;4(10):669-73.

(30) Corradi M, Poli D, Banda I, Bonini S, Mozzoni P, Pinelli S, et al. Exhaled breath analysis in suspected cases of non-small-cell lung cancer: a cross-sectional study. J Breath Res 2015 Jan 29;9(2):027101.

(31) Mazzone PJ, Hammel J, Dweik R, Na J, Czich C, Laskowski D, et al. Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array. Thorax 2007 Jul;62(7):565-8.

(32) Duan X, Yang Y, Tan S, Wang S, Feng X, Cui L, et al. Application of artificial neural network model combined with four biomarkers in auxiliary diagnosis of lung cancer. Med Biol Eng Comput 2017 Aug;55(8):1239-48.

(33) Hart GR, Roffman DA, Decker R, Deng J. A multi-parameterized artificial neural network for lung cancer risk prediction. PLoS One 2018;13(10):e0205264.

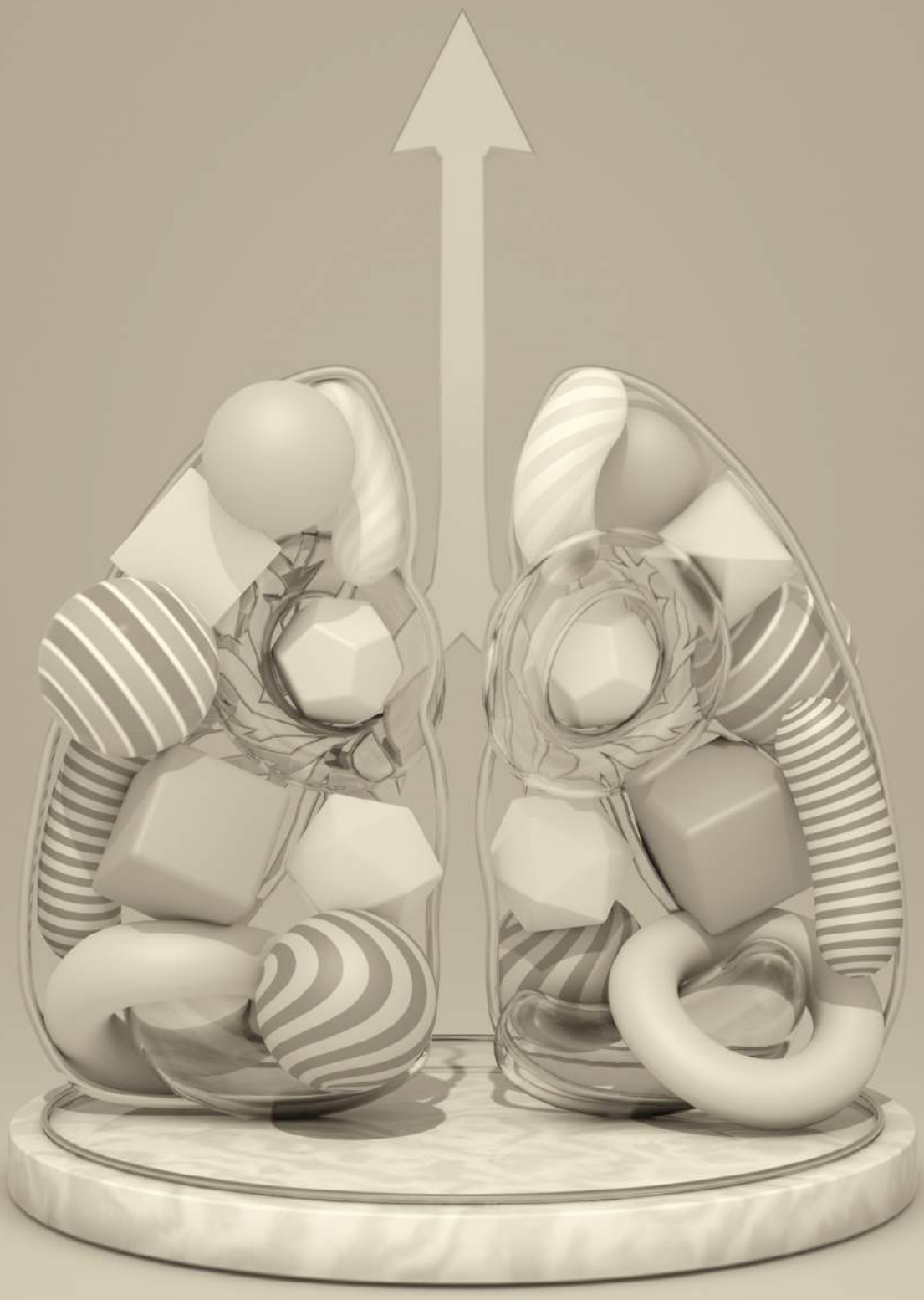
(34) Katki HA, Kovalchik SA, Berg CD, Cheung LC, Chaturvedi AK. Development and Validation of Risk Models to Select Ever-Smokers for CT Lung Cancer Screening. JAMA 2016 Jun 7;315(21):2300-11.

(35) Kovalchik SA, Tammemagi M, Berg CD, Caporaso NE, Riley TL, Korch M, et al. Targeting of low-dose CT screening according to the risk of lung-cancer death. N Engl J Med 2013 Jul 18;369(3):245-54.

(36) Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. N Engl J Med 2013 Feb 21;368(8):728-36.

# 06

# DIAGNOSING NON-SMALL CELL LUNG CANCER BY EXHALED-BREATH PROFILING USING AN ELECTRONIC NOSE: A MULTICENTRE VALIDATION STUDY



S. KORT - M. BRUSSE-KEIZER - J.H. SCHOUWINK - E. CITGEZ - F.H.C. DE JONGH - D. STOLZ
J.W.G. VAN PUTTEN - B.E. VAN DEN BORNE - E.A. KASTELIJN - M.M.F. SCHUURBIERS
M.M. VAN DEN HEUVEL - W.H. VAN GEFFEN - J. VAN DER PALEN

MANUSCRIPT SUBMITTED

**ORIGINAL ARTICLE**

# Diagnosing non-small cell lung cancer by exhaled-breath profiling using an electronic nose: a multicentre validation study.

*Exhaled-breath analysis for the diagnosis of lung cancer.*

S. Kort, MD[1*], Dr. M. Brusse-Keizer[2,10], Dr. J.H. Schouwink[1], E. Citgez, MD[1], Dr. F.H.C. de Jongh[1,10], Dr. J.W.G. van Putten[3], Dr. B.E. van den Borne[4], Dr. E.A. Kastelijn[5], Prof. Dr. D. Stolz[6,7], M.M.F. Schuurbiers, MD[8], Prof. Dr. M.M van den Heuvel[8], Dr. W.H van Geffen[9], Prof. Dr. J. van der Palen[2,10]

Affiliations:

[1]Department of Respiratory Medicine, Medisch Spectrum Twente Enschede, the Netherlands

[2]Medical School Twente, the Netherlands

[3]Department of Respiratory Medicine, Martini Ziekenhuis Groningen, the Netherlands

[4]Department of Respiratory Medicine, Catharina Ziekenhuis Eindhoven, the Netherlands

[5]Department of Respiratory Medicine, Sint Antonius Ziekenhuis Utrecht, the Netherlands

[6]Clinic for Pulmonary Medicine and Respiratory Cell Research, Universitätspital Basel, Switzerland

[7]Clinic for Respiratory Medicine, Medical Center, University of Freiburg, Faculty of Medicine, University of Freiburg, Germany

[8]Department of Respiratory Medicine, Radboud UMC Nijmegen, the Netherlands

[9] Department of Respiratory Medicine, Medisch Centrum Leeuwarden, the Netherlands

[10]Universiteit Twente, the Netherlands

## Abbreviations list

AI = Artificial Intelligence

ANN = Artificial Neural Network

AUC-ROC = Area Under the Receiver Operating Characteristic Curve

BMI = Body mass index

ERS = European Respiratory Society

IQR = Interquartile range

LDCT = Low dose computed tomography

NLST = National Lung Screening Trial

NPV = Negative predictive value

NSCLC = Non-small cell lung cancer

RF = Random Forest

SD = Standard deviation

SVM = Support Vector Machine

TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

VOC = Volatile Organic Compound

## Abstract

**Background:** Despite the potential of exhaled-breath analysis of volatile organic compounds to diagnose lung cancer, clinical implementation has not been realized partly due to the lack of validation studies.

**Research question:** This study addressed two questions: 1) Can we simultaneously train and validate a prediction model to distinguish non-small cell lung cancer (NSCLC) patients from non-lung cancer subjects based on exhaled-breath patterns? 2) Does addition of clinical variables to exhaled-breath data improve the diagnosis of lung cancer?

**Study design and methods:** In this multicentre study, subjects with NSCLC and control subjects, performed a measurement of 5 minutes of breathing in the Aeonose™. A training cohort was used for developing a prediction model based on breath data, whereas a blinded cohort was used for validation. Multivariate logistic regression analysis was performed including breath data and clinical variables, where the formula and cut-off value for the probability of lung cancer were applied on the validation data.

**Results:** 376 Subjects formed the training, and 199 subjects formed the validation set. The full training model, including clinical parameters and breath data showed, at a cut-off probability of 16% for lung cancer, a sensitivity of 95%, specificity of 51%, a negative predictive value (NPV) of 94% with an area under the receiver operating characteristic curve (AUC) of 0.87. Performance of the prediction model on the validation cohort showed corresponding results with a sensitivity of 95%, specificity of 49%, NPV of 94%, and an AUC of 0.86.

**Interpretation:** Combining exhaled-breath data and clinical variables in a multicentre, multi-device validation study can adequately distinguish lung cancer patients from subjects without lung cancer in a non-invasive manner. This study paves the way to implement exhaled-breath analysis in the daily practice of diagnosing lung cancer.

**Trial Registration Number:** The Netherlands Trial Register, NL7025

## Background

Lung cancer is the leading cause of cancer mortality worldwide [1,2]. Its high mortality rate is generally a consequence of advanced-stage disease at the time of initial diagnosis. Despite striking progress in treatment options in advanced-stage lung cancer, such as molecular-targeted therapies and immunotherapy, an essential step to reduce lung cancer mortality is early detection through non-invasive, point-of-care strategies[3-6].

Exhaled-breath contains a gas mixture of thousands of volatile organic compounds (VOCs) in low concentrations that reflect metabolic processes at tissue level [7,8]. Exhaled-breath analysis is based on shifts of this VOCs composition due to biochemical changes in different (patho)physiologic processes. This method has extensively been investigated in clinical research as a non-invasive tool to diagnose a variety of conditions [9,10]. Studies on pattern recognition for classification of VOC mixtures through non-specific cross-reactive sensors mimicking human and animal olfaction (e.g. electronic noses) as well as identifying individual VOCs by separation methods (e.g. gas chromatography mass spectrometry) have shown promising results in pilot studies to diagnose lung cancer [11-17].

Besides, studies based on imaging techniques have shown to be effective in screening purposes to diagnose lung cancer in high-risk asymptomatic subjects. Significant mortality reduction in high-risk subjects was observed by the National Lung Screening Trial (NLST), and the Dutch-Belgian lung cancer screening trial (NELSON) [18,19]. However, screening of high-risk subjects has not yet been implemented in Europe. Furthermore, determination of accurate screening criteria remains debatable since only subjects at the highest risk for lung cancer are targeted in current screening programmes.

The Aeonose™ (the eNose Company, Zutphen, the Netherlands) is a handheld electronic nose device featuring an array of three metal-oxide sensors that enables real-time breath analysis. The technology and breath sampling method have previously been described in detail [20,21]. After exposure to VOCs, consecutive conductivity changes at the sensors are recorded resulting in a digital exhaled-breath profile consisting of conductivity values. Exhaled-breath profiles of lung cancer patients can then be distinguished from profiles of non-lung cancer subjects using Artificial Intelligence (AI) techniques. Once a model has been developed for separating the groups, a new breath profile can be classified using this model. In previous studies, several malignant and non-malignant conditions have been investigated using the Aeonose™ [12,22-24].
We have previously reported the results of a proof-of-concept multicentre study performed with the Aeonose™, in which a prediction model, based on exhaled-breath profiles, was developed using supervised machine-learning techniques to discriminate subjects with and without non-small cell lung cancer (NSCLC) in a hospital setting [12]. An artificial neural network (ANN) trained with 290

subjects was able to classify breath samples with a sensitivity of 94%, a specificity of 33%, and an area under the receiver operating characteristic curve (AUC-ROC) of 0.76. Resampling techniques including leave-10%-out cross validation and bootstrapping were incorporated to reduce the risk of overfitting of the diagnostic model. Adding readily available clinical information, i.e. sex, age, number of pack-years, smoking status, and COPD-status to the exhaled-breath data resulted in a relevant improvement in diagnosing lung cancer patients [25].

To date, no single breath test has yet been approved for clinical practice to diagnose lung cancer. For this, validation studies are required, preferably involving multiple devices in multiple centres, where part of the data is used for developing a diagnostic model, and the remainder remains blinded to validate this model. Several studies on external validation of breath biomarkers in lung cancer have been performed, however these studies are aimed at identification of specific VOCs rather than exhaled-breath patterns [13,26,27]. Regarding pattern recognition techniques, Fens et al. and Bos et al. assessed validation of exhaled-breath molecular patterns in pulmonary diseases other than lung cancer, based on previous created training sets, showing moderate to high accuracy [28,29].

The objective of this prospective multicentre study using multiple devices is to train and subsequently validate a prediction model to distinguish NSCLC patients from subjects, initially suspected of lung cancer, but considered negative, and healthy control subjects, based on their exhaled-breath patterns.

## Study design and methods

Study design and participants

Participants suspected of lung cancer were recruited from 7 outpatient pulmonary departments between May 2018 and April 2020. The participating hospitals included Medisch Spectrum Twente Enschede, Radboud UMC Nijmegen, Medisch Centrum Leeuwarden, Martini Ziekenhuis Groningen, Catharina Ziekenhuis Eindhoven, Sint Antonius Ziekenhuis Utrecht (all in the Netherlands), and Universitätsspital Basel (Switzerland). Each centre used one Aeonose™ device, except for Basel using two devices. Since a single Aeonose™ device needs, as a rule of thumb, a minimum number of 30 observations in the smallest group (in this case positive measurements) to calibrate the device, and hence form reliable conclusions considering the training data, data from devices with an insufficient number of measurements were not used for further analyses.

Subjects suspected of lung cancer were divided into a group with confirmed NSCLC based on pathology and a group with a rejected diagnosis of lung cancer (control subjects), based on imaging and/or pathology. Types of lung cancer other than NSCLC were excluded. Additional healthy control

subjects with a minimum age of 55 years were recruited through an alert at the hospitals' websites. In case of pathologically confirmed lung cancer, staging was established according to the 8th edition of the American Joint Committee on Cancer TNM staging system [30]. Patients suspected of lung cancer in whom pathology (gold standard) was not performed due to insufficient clinical performance were excluded from the analyses. Demographic data and data on comorbidities were collected for all subjects. All participants were asked to complete a short questionnaire on recent smoking, eating, and alcohol intake, and were instructed to perform tidal breathing through the non-rebreathing Aeonose™ device for 5 minutes with their nose clipped.
The study protocol was approved by the institutional review board of Medisch Spectrum Twente and the board of directors of all participating institutions (eAppendix). All eligible patients provided written informed consent.

In this study we made use of the second generation, CE-certified Aeonose™ device. Since the training study was performed with the first generation, CE-uncertified device, these previously collected data were deemed not compatible, and therefore not used [12]. Instead, we decided to create a split-sample study design in which we enabled development and subsequent validation of new prediction models, which conforms to the European Respiratory Society (ERS) criteria for exhaled biomarkers [31].
Collected breath data were split into a training cohort for supervised learning and internal cross-validation, and a validation cohort, which was kept blinded, for model validation. Subjects were randomly assigned to the training and validation set, taking into consideration an equal prevalence of lung cancer patients in both sets.

Statistical analysis

Clinical characteristics are reported as means with standard deviations (SD) in case of a normal distribution, or as medians with interquartile ranges (IQR). Nominal variables are reported as numbers with corresponding percentages. To assess differences between the groups, T-tests, Mann-Whitney U tests, or Chi-squared-tests (X2) were applied, as appropriate.
Analysis of exhaled-breath data was executed by Aethena™, a proprietary software package, incorporating data pre-processing, data compression, machine learning algorithms for classification (e.g. ANN, Support Vector Machine (SVM), Random Forest (RF), XGBoost, logistic regression), internal validation techniques (leave-10%-out cross validation and bootstrapping), and model selection. Analyses yielded values between -1 and +1 per subject, indicating the degree to which the subject was classified as having lung cancer (maximum value +1) or not having lung cancer (minimum value -1). Details on the software package Aethena™ have been published previously [21].
We selected and trained five different models (each using a different classifier: ANN, Logistic Regression, RF, RF Extreme, and XGBoost, respectively) each showing proper discriminative

performance. As the different classifying techniques could interpret the data differently, we envisioned that averaging results over these five models would increase classification robustness. A cut-off value for the probability of lung cancer was determined for the training set to obtain a high sensitivity and negative predictive value (NPV), together with an acceptable number of false positive cases, as deemed relevant for clinical practice. Receiver operating characteristics (ROC) curves were composed and AUCs were calculated with 95% confidence intervals.

Subsequently, clinical variables, i.e. sex, age, number of pack-years, COPD, diabetes, hypertension, BMI, and the absolute value obtained from the Aeonose™ (between -1 and +1) were entered in a multivariate logistic regression analysis. Non-significant variables were eliminated according to the backward method until the fit of the model decreased significantly, based on the -2-log likelihood. Age and sex were included regardless of their significance. A cut-off value for the probability of lung cancer based on this multivariate model was again chosen to obtain a high sensitivity and NPV together with an acceptable number of false positive cases.

The diagnostic performance of this final logistic regression model, based on the training data, was validated on the blinded data set, where the β-coefficients were fixed. The same cut-off value, chosen for the training data to determine the presence of lung cancer, was applied to the logistic regression analysis in the validation set. Results are expressed as sensitivity, specificity, predictive values, and AUC.

A calibration plot was constructed to demonstrate how well the predicted probability of lung cancer matches the observed probability of lung cancer.

Stratification for variables to evaluate possible influences on exhaled-breath outcomes was performed in explorative analyses for sex, age, presence of COPD, lung cancer stage, and type of histology. Early-stage lung cancer was classified as either stage I or II, whereas late-stage lung cancer was classified as stage III or IV.

SPSS version 24.0 was used. All statistical tests were two-sided with a significance level at 0.05.

## Results

A total of 575 subjects were enrolled in the analyses (Figure 1). Approximately two-thirds formed the training set (376 subjects; 160 lung cancer patients, 51 suspected, but negative, and 165 healthy control subjects), whereas the remaining one-third comprised the validation set (199 subjects; 79 lung cancer patients, 32 suspected, but negative, and 88 healthy control subjects). Subject characteristics are described in Table 1. Data were obtained using 5 Aeonose™ devices.

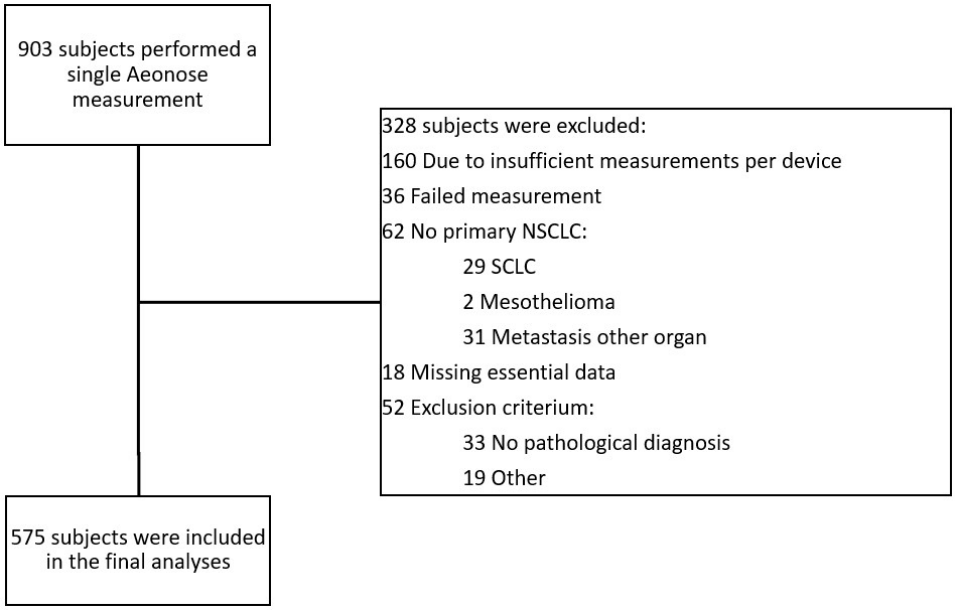



**Figure 1.** Flow chart study cohort.

| **Table 1. Clinical characteristics of all enrolled subjects.** | | | | | | |
|---|---|---|---|---|---|---|
| | Training set (N = 376) | | | Validation set (N = 199) | | |
| | Lung cancer (N = 160) | Control subjects (N = 216) | *P*-value | Lung cancer (N = 79) | Control subjects (N = 120) | *P*-value |
| **Age in years, mean (SD)** | 68.4 (8.6) | 64.6 (8.2) | <0.001 | 69.0 (7.9) | 63.4 (9.4) | <0.001 |
| **Sex (males), N (%)** | 97 (60.6) | 131 (60.6) | 0.996 | 49 (62.0) | 59 (49.2) | 0.075 |
| **Smoking status, N (%)** | | | <0.001 | | | 0.001 |
| **Current smoker** | 48 (30.0) | 41 (19.0) | | 30 (38.0) | 27 (22.5) | |
| **Ex-smoker** | 103 (64.4) | 130 (60.2) | | 45 (57.0) | 65 (54.2) | |
| **Never smoker** | 9 (5.6) | 45 (20.8) | | 4 (5.1) | 28 (23.3) | |
| **Pack-years, N (%)[#]** | | | <0.001 | | | <0.001 |
| **0** | 8 (5.1) | 45 (20.8) | | 2 (2.6) | 28 (23.3) | |
| **1-20** | 37 (23.4) | 56 (25.9) | | 14 (18.4) | 38 (31.7) | |
| **21-40** | 52 (32.9) | 55 (25.5) | | 28 (36.8) | 19 (15.8) | |
| **>40** | 61 (38.6) | 60 (27.8) | | 32 (42.1) | 35 (29.2) | |
| **COPD, N (%)** | 71 (44.4) | 94 (43.5) | 0.869 | 37 (46.8) | 52 (43.3) | 0.627 |
| **Hypertension, N (%)[$]** | 66 (41.3) | 74 (34.3) | 0.166 | 27 (34.6) | 38 (31.9) | 0.695 |
| **Diabetes, N (%)[$]** | 15 (9.4) | 22 (10.2) | 0.794 | 11 (13.9) | 10 (8.4) | 0.217 |
| **BMI, mean (SD)** | 26.4 (4.4) | 25.8 (4.7) | 0.210 | 26.2 (5.0) | 25.7 (4.4) | 0.402 |
| **Type of NSCLC, N (%)** | | | | | | |
| **Adenocarcinoma** | 101 (63.1) | | | 39 (50.0) | | |
| **Squamous cell carcinoma** | 43 (26.9) | | | 32 (41.0) | | |
| **Large cell carcinoma** | 6 (3.8) | | | 4 (5.1) | | |
| **NOS** | 10 (6.3) | | | 3 (3.8) | | |
| **Stage*, N (%)** | | | | | | |
| **I** | 54 (33.8) | | | 21 (26.6) | | |
| **II** | 23 (14.4) | | | 15 (19.0) | | |
| **III** | 38 (23.8) | | | 19 (24.1) | | |
| **IV** | 45 (28.2) | | | 24 (30.4) | | |
| **Hospital, N (%)** | | | | | | |
| **MST** | 66 (41.3) | 69 (31.9) | | 30 (38.0) | 30 (25.0) | |
| **Radboud UMC** | 31 (19.4) | 29 (13.4) | | 20 (25.3) | 12 (10.0) | |
| **MCL Leeuwarden** | 29 (18.1) | 34 (15.7) | | 17 (21.5) | 33 (27.5) | |
| **US Basel** | 34 (21.3) | 84 (38.9) | | 12 (15.2) | 45 (37.5) | |

**BMI: Body mass index, NSCLC: Non-small cell lung cancer, NOS: Not otherwise specified, [#] 5 missing subjects, [$] 1 missing subject, * according to the eighth edition of the American Joint Committee on Cancer TNM staging system**

The training model, exclusively based on breath data from the Aeonose™, showed, at a cut-off value of -0.36, an AUC of 0.83 (95% CI 0.79-0.87), a sensitivity of 91%, a specificity of 54%, and an NPV of 89%. The diagnostic performance of the Aeonose™, maintaining the same cut-off value in the validation set, reached an AUC of 0.79 (95% CI 0.72-0.85), with a sensitivity of 88%, a specificity of 52%, and an NPV of 87%, which conforms to the training model.

Due to the multicollinearity of smoking status and number of pack-years, we chose to include number of pack-years in our analyses, because this parameter contained the most detailed information. The multivariate analysis based on solely clinical data from the training set, including sex, age, and number of pack-years, showed an AUC of 0.67 (95% CI 0.61-0.72), while the validation set showed an AUC of 0.75 (95% CI 0.68-0.82).

Exhaled-breath data and clinical parameters from the training set were combined in a multivariate logistic regression analysis, maintaining a cut-off of 16% probability of lung cancer, resulting in a sensitivity of 95%, a specificity of 51%, and an NPV of 94%, which was based on clinical relevance (Tables 2 and 3). This corresponded to an AUC of 0.87 (95% CI 0.83-0.90). When applying the identical multivariate logistic regression model on the validation set, maintaining the selected cut-off probability of 16%, we observed a sensitivity of 95%, a specificity of 49%, a PPV of 54%, and an NPV of 94%, with a corresponding AUC of 0.86 (0.81-0.91) (Table 3 and Figure 2). In case of this cut-off probability of 16%, 63 of the 196 subjects (32%) were classified as “no lung cancer” (Table 4). Corresponding performance of breath data only, with an equal cut-off probability of lung cancer in the training and validation set is also displayed in Table 3.

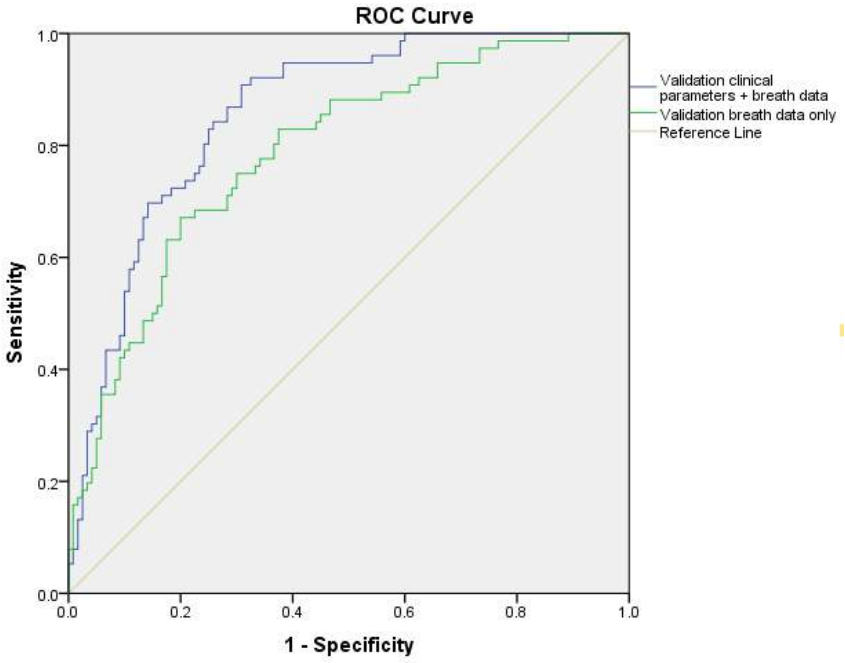



**Figure 2.** Combined ROC curve of the validation models based on either exhaled-breath data only and a prediction model including clinical parameters.

| **Table 2. Results of the univariate and multivariate logistic regression analysis for diagnosing lung cancer.** | | | |
|---|---|---|---|
| **Variable** | Univariate analysis Odds ratio (95% CI) | Multivariate analysis Odds ratio (95% CI) | Regression coefficient (B) |
| | | | -4.949 (intercept) |
| **Sex (ref female)** | 1.01 (0.66-1.53) | 0.68 (0.38-1.19) | -0.393 |
| **Age** | 1.06 (1.03-1.08) | 1.05 (1.02-1.09) | 0.049 |
| **BMI** | 1.03 (0.98-1.08) | - | |
| **Smoking status** | | | |
| **Current smoker** | 7.53 (3.06-18.49) | - | |
| **Ex-smoker** | 5.09 (2.21-11.77) | | |
| **Never smoked** | Ref | | |
| **Pack years** | | | |
| **0** | Ref | Ref | |
| **1-20** | 3.72 (1.57-8.77) | 5.19 (1.91-14.1) | 1.647 |
| **21-40** | 5.32 (2.29-12.35) | 8.11 (3.02-21.76) | 2.092 |
| **>40** | 5.72 (2.49-13.14) | 8.69 (3.22-23.50) | 2.162 |
| **COPD** | 1.06 (0.70-1.60) | - | |
| **Hypertension** | 1.35 (0.88-2.06) | - | |
| **Diabetes mellitus** | 0.86 (0.42-1.73) | - | |
| **Classification value Aeonose™** | 22.8 (12.0-43.3) | 27.9 (14.0-55.5) | 3.328 |
| **CI: confidence interval; BMI: body mass index; COPD: chronic obstructive pulmonary disease; -: not added to the multivariate model.** | | | |

| Table 3. Diagnostic accuracy of exhaled breath analysis in the training and validation set | | | | | | |
|---|---|---|---|---|---|---|
| | Cut-off probability* | Sensitivity | Specificity | PPV | NPV | AUC (95% CI) |
| **Training breath data only[$]** | 20% | 93.0 | 54.2 | 59.8 | 91.4 | 0.83 (0.79-0.87) |
| **Validation breath data only*[$]** | 20% | 88.2 | 48.3 | 51.9 | 86.6 | 0.79 (0.72-0.85) |
| **Training clinical parameters + breath data[$]** | 16% | 94.9 | 50.5 | 58.4 | 93.2 | 0.87 (0.83-0.90) |
| **Validation clinical parameters + breath data*[$]** | 16% | 94.7 | 49.2 | 54.1 | 93.7 | 0.86 (0.81-0.91) |

***Corresponding cut-off values and fixed β-coefficients based on logistic regression analyses in the training set.**
**[$]All analyses are performed in subjects without missing data (Training data: N=374, validation data: N=196).**
**PPV: positive predictive value; NPV: negative predictive value; AUC: area under the curve**

| Table 4. 2x2 table of the final multivariate prediction model including clinical parameters + breath data, and a cut-off probability of 16%. | | | |
|---|---|---|---|
| | Lung cancer (gold standard) | No lung cancer (gold standard) | |
| **Lung cancer (final model)** | 72 | 61 | 133 |
| **No lung cancer (final model)** | 4 | 59 | 63 |
| | 76 | 120 | 196 |

A calibration plot with the predicted probability of lung cancer in deciles of the validation cohort is shown in *e-Figure 1*. The figure shows good concordance between the predicted probability of lung cancer in each decile, and the observed prevalence of lung cancer the same decile.

Explorative subgroup analyses show equal performance of the Aeonose™ in early and late-stage lung cancer, in both sexes, different age groups, and different types of histology (*e-Tables 1-12*). In stage I and II lung cancer, sensitivity and NPV were 94% and 97%, respectively, while in stage III and IV lung cancer, sensitivity and NPV were 84% and 90%, respectively.

## Discussion

In this study, we trained and subsequently validated exhaled-breath data to distinguish between patients with NSCLC and clinically relevant control subjects in a multicentre setting using multiple devices. Our findings show that patients with NSCLC can successfully be discriminated from subjects without NSCLC using exhaled-breath patterns based on a training set concentrating on a high NPV to exclude the diagnosis of lung cancer in a non-invasive manner. Discrimination between both groups improves significantly when readily available clinical variables, i.e. age, sex, and number of pack-years are added to the prediction model. Classifying new subjects, not used for training of the Aeonose™, shows excellent performance.

Our previously performed training study indicated that exhaled-breath patterns differ between patients with lung cancer and subjects without lung cancer [12]. The current study provided the necessary essential step where a prediction model based on a training set was validated on "blind" subjects in a multicentre and multinational setting, using multiple devices.

To our knowledge, this is the first NSCLC study to validate blinded exhaled-breath profiles based on pattern recognition techniques in a multicentre split-sample design, including readily available clinical variables, whilst using multiple electronic nose devices.
In the past, Machado et al. performed a similar study in which they used a split-sample design to validate a prediction model to distinguish lung cancer patients from control subjects [32]. However, as they showed promising results, the study was performed in a single centre setting and had a very small study population (14 individuals with bronchogenic carcinoma in both the training phase and validation phase, respectively). Also, Mazzone et al. performed a split-sample study design using pattern recognition techniques based on exhaled-breath to distinguish lung cancer patients from control subjects [15]. This concerned a two-centre study with the application of only one electronic nose device, and diagnostic performance in the validation set could be considered moderate. A recent study of Long et al. showed interesting results in an external validation study of exhaled-breath biomarkers to diagnose lung cancer [26]. Although they made use of the GC-MS technique, with several Tedlar bags and one GC-MS station, to identify molecules in exhaled-breath, they also focused on the possible origin of breath biomarkers by explaining specific metabolic processes in lung cancer pathogenesis. This strict study protocol may, however, be not easily implemented in daily clinical practice and contrary to the Aeonose™, it does not offer a point-of-care solution.

The reported AUC of 0.86 in our study as obtained by the multivariate validation model provides very good accuracy, but is lower than some of the reported accuracies by other studies using pattern recognition techniques in exhaled-breath analysis to diagnose lung cancer [11,17,33-35]. Possible

explanations for these discrepancies are incomparable study designs and control groups, a single centre versus multicentre setting, small datasets with the inherent risk of overfitting of models, the use different sensor technologies, use of a single device, and reporting results based on training data that are not validated.

Validation of a prediction model, as performed in this study, is a pivotal step for clinical integration of exhaled-breath analysis in the diagnostic path of lung cancer. To assess the feasibility and acceptability of the electronic nose in clinical practice, we envision using the Aeonose™ in parallel with current practice in a hospital setting. Although based on an exploratory analysis, the validated model seems to be able to distinguish early-stage lung cancer from non-lung cancer with relatively high accuracy. In case of doubt, e.g. based on CT-scans, and a low probability of lung cancer, based on the validated model, a wait and see strategy could be employed.

Exhaled-breath analysis might have promise as an element in an integral lung cancer screening program, most likely combined with other non-invasive tests such as low dose CT (LDCT) screening. However, the Aeonose™ should then be trained on a sample of subjects with an increased probability of lung cancer, such as heavy smokers. Despite the fair number of early-stage lung cancer cases in our cohort, we did not specifically analyse pulmonary nodules, which has been inherently the focus of LDCT screening. Future studies should focus on solitary pulmonary nodules and assess whether exhaled-breath analysis can fulfil a substantial role in lung cancer screening, possibly serving a synergistic role combined with LDCT and guide risk assessment prior to LDCT screening as a pre-selection tool or after LDCT screening to determine surveillance intervals (6). However, in such a setting new prediction models must be built with data based on current screening criteria. Besides assessment of lung cancer risk, exhaled-breath analysis could also serve as a prognostic biomarker to predict response on therapies and possible recurrence risk [36,37].

A notable strength of this study is the addition of clinical variables to the prediction model. This easily available information has previously shown to be informative, including development of clinical prediction scores in lung cancer screening based on imaging [38,39]. Our results show significant improvement of the prediction model when adding clinical variables, which was confirmed in the validation cohort.
Another strength is the excellent match between training and validation results. This is not straightforward as AI techniques are usually applied with far larger datasets.

The Aeonose™ device not only features the possibility to perform real-time analysis of breath data without the necessity of breath sample storage; it has also incorporated a wash-out period of 2 minutes where the lungs are fully cleared of dead space ventilation and analysis is solely performed

on VOC's originating from metabolic processes in peripheral tissues.
Other strong points worth mentioning are the multicentre and multinational design. Multiple devices were used for gathering training data, leading to a prediction model capable of classifying blinded samples, also collected with multiple devices.

The study follows the recommendations of the TRIPOD statement (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) [40]. Unfortunately, in our case, due to the use of the second generation, CE-certified Aeonose™ device, we could not use previously collected data. Given the long timeframe of collecting the necessary data, we decided to use a split-sample study design in which we simultaneously trained and validated a prediction model. We intended to use breath data from all 8 Aeonoses™ (7 hospitals) to create a training cohort for supervised learning and cross-validation, and a validation cohort, which was kept blinded, for validation. However, it turned out that in some of the participating hospitals the amount of breath data, due to limited positive and negative lung cancer diagnoses, was not sufficient for adequate data analysis.

## Interpretation

In summary, combining exhaled-breath data and clinical parameters in a multicentre, multi-device validation study can adequately distinguish lung cancer patients from subjects without lung cancer in a non-invasive manner. This study paves the way to implement exhaled-breath analysis in the daily practice of diagnosing lung cancer.

## Acknowledgements

### Author contributions

SK is the guarantor of the content of the manuscript, data, and analysis, and was responsible for study conception and design, data analysis and interpretation, and manuscript writing
SK, MBK, MMvdH, WHvG, and JvdP were involved in the conception and design of the study.
SK, JHS, EC, JWGvP, BEvdB, EAK, DS, MMFS, MMvdH, and WHvG acquired the data.
SK, MBK, and JvdP analysed and interpreted the data.
SK, MBK, JHS, EC, FHCdJ, JWGvP, BEvdB, EAK, DS, MMFS, MMvdH, WHvG, and JvP wrote the article or were substantially involved in its revision before submission.

### Conflict of interest

S. Kort reports an unrestricted research grant paid to her institution by The eNose Company, Zutphen, during the conduct of the study. All remaining authors have declared no conflict of interest.

### Funding/support information

This work was supported by the eNose Company, Zutphen, the Netherlands. The funder of the study, the eNose Company, had no role in the study design, data collection, data interpretation, or writing of the report. However, they performed part of the data analysis, but for model validation they had explicitly no access to blinded data and the classification of these study subjects. S. Kort and J. van der Palen had access to raw data. S. Kort, the corresponding author, had full access to all the data in the study and the final responsibility for the decision to submit for publication.

## Take home points

**Study Question:** Can exhaled-breath patterns of patients with non-small cell lung cancer (NSCLC) and without NSCLC adequately be discriminated with an electronic nose in a multicentre, multi-device validation study?

**Results:** Exhaled-breath data can adequately distinguish lung cancer patients from subjects without lung cancer in a non-invasive manner in this multicentre, multi-device study including 575 subjects. Adding clinical variables relevantly improves the diagnostic performance to diagnose lung cancer.

**Interpretation:** Validation of a prediction model, as performed in this study, is a pivotal step for clinical integration of exhaled-breath analysis in the diagnostic path of lung cancer.

## References

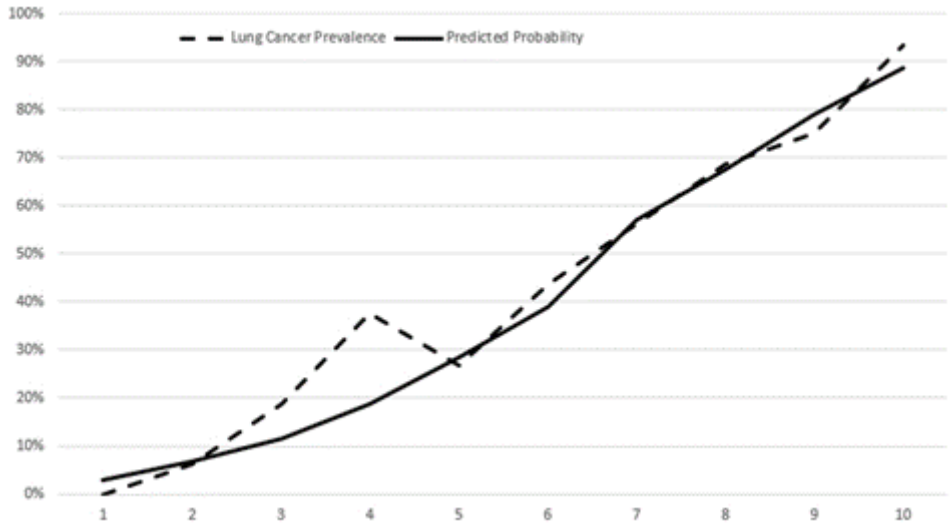

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* 2020;70(1):7-30.
2. Dagenais GR, Leong DP, Rangarajan S, et al. Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study. *Lancet.* 2020;395(10226):785-794.
3. Broodman I, Lindemans J, van Sten J, Bischoff R, Luider T. Serum Protein Markers for the Early Detection of Lung Cancer: A Focus on Autoantibodies. *J Proteome Res.* 2017;16(1):3-13.
4. Jiang R, Dong X, Zhu W, et al. Combining PET/CT with serum tumor markers to improve the evaluation of histological type of suspicious lung cancers. *PLoS One.* 2017;12(9):e0184338.
5. Rolfo C, Russo A. Liquid biopsy for early stage lung cancer moves ever closer. *Nat Rev Clin Oncol.* 2020;17(9):523-524.
6. Seijo LM, Peled N, Ajona D, et al. Biomarkers in Lung Cancer Screening: Achievements, Promises, and Challenges. *J Thorac Oncol.* 2019;14(3):343-357.
7. Mansurova M, Ebert BE, Blank LM, Ibanez AJ. A breath of information: the volatilome. *Curr Genet.* 2018;64(4):959-964.
8. Boots AW, van Berkel JJ, Dallinga JW, Smolinska A, Wouters EF, van Schooten FJ. The versatile use of exhaled volatile organic compounds in human health and disease. *J Breath Res.* 2012;6(2):027108.
9. Nakhleh MK, Amal H, Jeries R, et al. Diagnosis and Classification of 17 Diseases from 1404 Subjects via Pattern Analysis of Exhaled Molecules. *ACS Nano.* 2017;11(1):112-125.
10. Wilson AD. Advances in electronic-nose technologies for the detection of volatile biomarker metabolites in the human breath. *Metabolites.* 2015;5(1):140-163.
11. Gasparri R, Santonico M, Valentini C, et al. Volatile signature for the early diagnosis of lung cancer. *J Breath Res.* 2016;10(1):016007.
12. Kort S, Tiggeloven MM, Brusse-Keizer M, et al. Multi-centre prospective study on diagnosing subtypes of lung cancer by exhaled-breath analysis. *Lung Cancer.* 2018;125:223-229.
13. Phillips M, Altorki N, Austin JH, et al. Prediction of lung cancer using volatile biomarkers in breath. *Cancer Biomark.* 2007;3(2):95-109.
14. van de Goor R, van Hooren M, Dingemans AM, Kremer B, Kross K. Training and Validating a Portable Electronic Nose for Lung Cancer Screening. *J Thorac Oncol.* 2018;13(5):676-681.
15. Mazzone PJ, Wang XF, Lim S, et al. Progress in the development of volatile exhaled breath signatures of lung cancer. *Ann Am Thorac Soc.* 2015;12(5):752-757.
16. Rocco G, Pennazza G, Santonico M, et al. Breathprinting and Early Diagnosis of Lung Cancer. *J Thorac Oncol.* 2018;13(7):883-894.
17. Tirzite M, Bukovskis M, Strazda G, Jurka N, Taivans I. Detection of lung cancer with electronic nose and logistic regression analysis. *J Breath Res.* 2018;13(1):016006.
18. National Lung Screening Trial Research T, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med.* 2011;365(5):395-409.
19. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med.* 2020;382(6):503-513.
20. Bruins MGJ, Gerritsen JW, Van de Sande W, Van Belkum A, Bos A. Enabling a transferable calibration model for metal-oxide type electronic noses. *Sensors and Actuators B: Chemical.* 2013;188:1187-1195.
21. Kort S, Brusse-Keizer M, Gerritsen JW, van der Palen J. Data analysis of electronic nose technology in lung cancer: generating prediction models by means of Aethena. *J Breath Res.* 2017;11(2):026006.
22. Peters Y, Schrauwen RWM, Tan AC, Bogers SK, de Jong B, Siersema PD. Detection of Barrett's oesophagus through exhaled breath using an electronic nose device. *Gut.* 2020;69(7):1169-1172.

23. van de Goor R, Hardy JCA, van Hooren MRA, Kremer B, Kross KW. Detecting recurrent head and neck cancer using electronic nose technology: A feasibility study. *Head Neck.* 2019;41(9):2983-2990.
24. Uslu HI, Dolle AR, Dullemen HM, Aktas H, Kolkman JJ, Venneman NG. Pancreatic ductal adenocarcinoma and chronic pancreatitis may be diagnosed by exhaled-breath profiles: a multicenter pilot study. *Clin Exp Gastroenterol.* 2019;12:385-390.
25. Kort S, Brusse-Keizer M, Gerritsen JW, et al. Improving lung cancer diagnosis by combining exhaled-breath data and clinical parameters. *ERJ Open Res.* 2020;6(1).
26. Long Y, Wang C, Wang T, et al. High performance exhaled breath biomarkers for diagnosis of lung cancer and potential biomarkers for classification of lung cancer. *J Breath Res.* 2021;15(1):016017.
27. Phillips M, Bauer TL, Cataneo RN, et al. Blinded Validation of Breath Biomarkers of Lung Cancer, a Potential Ancillary to Chest CT Screening. *PLoS One.* 2015;10(12):e0142484.
28. Fens N, Roldaan AC, van der Schee MP, et al. External validation of exhaled breath profiling using an electronic nose in the discrimination of asthma with fixed airways obstruction and chronic obstructive pulmonary disease. *Clin Exp Allergy.* 2011;41(10):1371-1378.
29. Bos LD, Schultz MJ, Sterk PJ. Exhaled breath profiling for diagnosing acute respiratory distress syndrome. *BMC Pulm Med.* 2014;14:72.
30. Goldstraw P, Chansky K, Crowley J, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol.* 2016;11(1):39-51.
31. Horvath I, Barnes PJ, Loukides S, et al. A European Respiratory Society technical standard: exhaled biomarkers in lung disease. *Eur Respir J.* 2017;49(4).
32. Machado RF, Laskowski D, Deffenderfer O, et al. Detection of lung cancer by sensor array analyses of exhaled breath. *Am J Respir Crit Care Med.* 2005;171(11):1286-1291.
33. Li W, Liu H, Xie D, He Z, Pi X. Lung Cancer Screening Based on Type-different Sensor Arrays. *Sci Rep.* 2017;7(1):1969.
34. Mazzone PJ, Wang XF, Xu Y, et al. Exhaled breath analysis with a colorimetric sensor array for the identification and characterization of lung cancer. *J Thorac Oncol.* 2012;7(1):137-142.
35. D'Amico A, Pennazza G, Santonico M, et al. An investigation on electronic nose diagnosis of lung cancer. *Lung Cancer.* 2010;68(2):170-176.
36. de Vries R, Muller M, van der Noort V, et al. Prediction of response to anti-PD-1 therapy in patients with non-small-cell lung cancer by electronic nose analysis of exhaled breath. *Ann Oncol.* 2019;30(10):1660-1666.
37. Buma AIG, Muller M, de Vries R, et al. eNose analysis for early immunotherapy response monitoring in non-small cell lung cancer. *Lung Cancer.* 2021;160:36-43.
38. Baldwin DR, Callister ME, Guideline Development G. The British Thoracic Society guidelines on the investigation and management of pulmonary nodules. *Thorax.* 2015;70(8):794-798.
39. Herder GJ, van Tinteren H, Golding RP, et al. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. *Chest.* 2005;128(4):2490-2496.
40. Patzer RE, Kaji AH, Fong Y. TRIPOD Reporting Guidelines for Diagnostic and Prognostic Studies. *JAMA Surg.* 2021;156(7):675-676.

## Supplementary Data

**Diagnosing non-small cell lung cancer by exhaled-breath profiling using an electronic nose: a multicentre validation study.**

Sharina Kort, Marjolein Brusse-Keizer, Hugo Schouwink, Emanuel Citgez, Frans de Jongh, John van Putten, Ben van den Borne, Lisanne Kastelijn, Daiana Stolz, Milou Schuurbiers, Michel van den Heuvel, Wouter van Geffen, Job van der Palen




**e-Figure 1.** Calibration plot with predicted probability of lung cancer, in deciles, in the validation cohort, and the corresponding observed lung cancer prevalence (in %) for the same decile.

**e-Tables 1-12.** Explorative subgroup analyses, displayed in 2x2 tables. They each show the diagnostic performance in terms of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Area under the Receiving Operator Characteristic Curve (AUC).
Sensitivity, specificity, PPV, and NPV are expressed as percentages.

## e-Table 1. Early-stage non-small cell lung cancer (NSCLC) (Stage I and II)

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 31 | 58 | 89 |
| Aeonose - | 2 | 62 | 64 |
| | 33 | 120 | 153 |

Sensitivity: 93.9
Specificity: 51.2
PPV: 34.8
NPV: 96.9
AUC: 0.83 (0.75-0.91)

## e-Table 2. Late-stage non-small cell lung cancer (NSCLC) (Stage III and IV)

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 36 | 58 | 94 |
| Aeonose - | 7 | 62 | 69 |
| | 43 | 120 | 163 |

Sensitivity: 83.7
Specificity: 51.2
PPV: 38.3
NPV: 89.9
AUC: 0.75 (0.67-0.84)

## e-Table 3. Males

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 45 | 33 | 78 |
| Aeonose - | 4 | 26 | 30 |
| | 49 | 59 | 108 |

Sensitivity: 91.8
Specificity: 44.1
PPV: 57.7
NPV: 86.7
AUC: 0.83 (0.75-0.91)

**e-Table 4. Females**

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 27 | 28 | 55 |
| Aeonose - | 0 | 33 | 33 |
| | 27 | 61 | 88 |

Sensitivity: 100
Specificity: 54.1
PPV: 49.1
NPV: 100
AUC: 0.90 (0.83-0.96)

**e-Table 5. Age ≥65.3 (median age)**

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 47 | 33 | 80 |
| Aeonose - | 3 | 14 | 17 |
| | 50 | 47 | 97 |

Sensitivity: 94.0
Specificity: 29.8
PPV: 58.8
NPV: 82.4
AUC: 0.80 (0.71-0.89)

**e-Table 6. Age <65.3 (median age)**

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 25 | 28 | 53 |
| Aeonose - | 1 | 45 | 46 |
| | 26 | 73 | 99 |

Sensitivity: 96.2
Specificity: 61.6
PPV: 47.2
NPV: 97.8
AUC: 0.89 (0.83-0.96)

**e-Table 7. COPD +**

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 37 | 36 | 73 |
| Aeonose - | 0 | 16 | 16 |
| | 37 | 52 | 89 |

Sensitivity: 100
Specificity: 30.8
PPV: 50.7
NPV: 100
AUC: 0.81 (0.72-0.90)

**e-Table 8. COPD-**

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 35 | 25 | 60 |
| Aeonose - | 4 | 43 | 47 |
| | 39 | 68 | 107 |

Sensitivity: 89.7
Specificity: 63.2
PPV: 58.3
NPV: 91.4
AUC: 0.90 (0.84-0.96)

**e-Table 9. Active smoking**

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 28 | 21 | 49 |
| Aeonose - | 2 | 6 | 9 |
| | 30 | 27 | 57 |

Sensitivity: 93.3
Specificity: 22.2
PPV: 57.1
NPV: 66.9
AUC: 0.79 (0.67-0.91)

**e-Table 10. Ex + never smoking**

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 44 | 40 | 84 |
| Aeonose - | 2 | 53 | 55 |
| | 46 | 93 | 139 |

Sensitivity: 95.7
Specificity: 57.0
PPV: 52.4
NPV: 96.4
AUC: 0.88 (0.83-0.94)

**e-Table 11. Adenocarcinoma**

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 34 | 58 | 92 |
| Aeonose - | 5 | 62 | 67 |
| | 39 | 120 | 159 |

Sensitivity: 87.2
Specificity: 51.2
PPV: 37.0
NPV: 92.5
AUC: 0.79 (0.71-0.87)

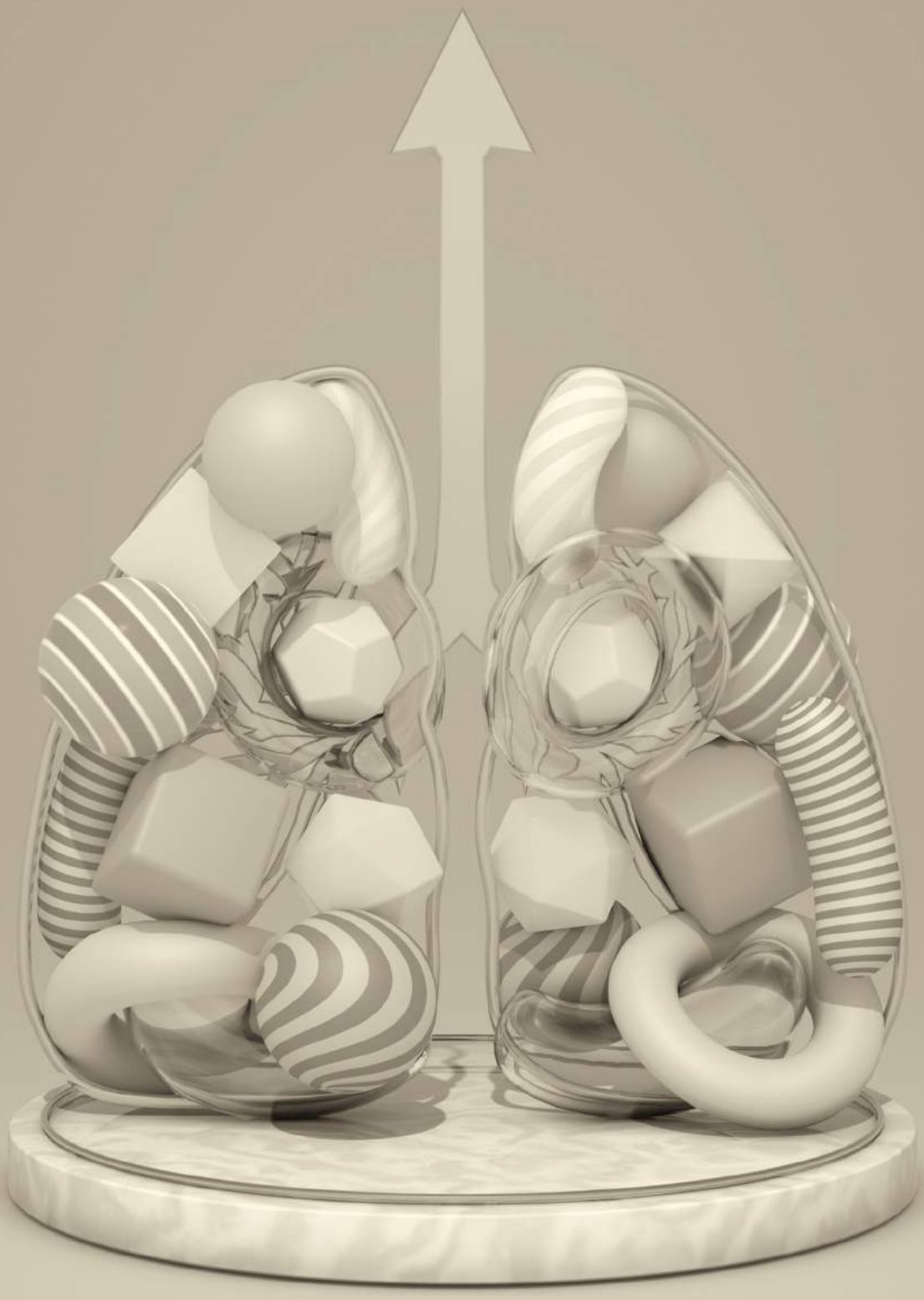
**e-Table 12. Squamous cell carcinoma**

| | NSCLC + | NSCLC - | |
|---|---|---|---|
| Aeonose + | 28 | 58 | 86 |
| Aeonose - | 4 | 62 | 66 |
| | 32 | 120 | 152 |

Sensitivity: 87.5
Specificity: 51.2
PPV: 32.6
NPV: 93.9
AUC: 0.78 (0.69-0.86)

# 07

# GENERAL DISCUSSION



S. KORT

The main objective of this thesis was to investigate and validate exhaled breath analysis based on pattern recognition techniques with the Aeonose™, as a non-invasive diagnostic tool to discriminate lung cancer patients from non-lung cancer subjects. Since lung cancer is a common type of cancer with a high incidence, and an extremely high mortality rate due to late detection, innovative diagnostic tools to detect lung cancer are desired, preferably non-invasive, low-cost, and with the ability to provide real-time analysis to realise a point-of care diagnosis.

Various non-invasive diagnostic tools have been investigated in the past decades with a promising role for exhaled breath analysis. Exhaled breath analysis is based on the detection of volatile organic compounds (VOCs) that reflect metabolic processes in the body, and moreover, changes in metabolic states in case of disease. These VOCs can be used as non-invasive biomarkers to identify disease activity, but also as prognostic measures in the follow up of a disease after treatment or as an indication of treatment efficacy prior to the start of a treatment (1-3). Analysis of these exhaled VOCs is generally based on either pattern recognition with electronic nose devices (eNose), such as the Aeonose™, or exact identification and quantification of individual VOCs by analytical separation methods, such as gas-chromatography-mass spectrometry (GC-MS), or ion-mobility spectrometry (IMS). Although both analytical methods are aimed at diagnosing a certain condition in exhaled breath, and both make use of highly sensitive sensors that are able to detect VOCs at part per billion (ppb) concentration levels, they differ in terms of breath sampling methods, selectivity of VOCs, complexity to perform a measurement, costs, and the ability to perform real-time analyses. Identification methods involving mass spectrometry are characterized by the ability to identify specific VOCs, but are associated with high costs, a time-consuming process, requirement of expertise to handle the technique, necessity to store exhaled breath before analyses can be performed, and the impossibility to obtain point-of-care results. In contrast, exhaled breath analysis based on pattern recognition techniques is generally considered low-cost, convenient to perform, and can be performed real-time without the need to store the exhaled breath. However, since pattern recognition techniques lack selectivity of the sensors as they are cross-reactive, it is not possible to identify or quantify individual molecules as a tool to investigate underlying biological and pathophysiological mechanisms. Both types of analytical methods have been extensively investigated in the past decades. However, to date, no unique disease-specific VOC or a combination of VOCs, indicative for lung cancer, has been identified.

One of the first studies that used GC-MS to discriminate lung cancer patients from non-lung cancer subjects was published by Phillips et al. in 2003. Using nine VOCs, 67 primary lung cancer patients could be discriminated from 41 healthy controls with 85% sensitivity and 81% specificity (4). A consecutive study of Phillips et al. in 2007, including over 400 subjects, achieved 85% sensitivity and

80% specificity to distinguish lung cancer patients from non-lung cancer subjects, based on a prediction model including 16 VOCs, with the important remark that the accuracy was not affected by TNM stage or tobacco smoking (5).
In 2009, Bajtarevic et al. used GC-MS to identify VOCs that could differentiate lung cancer patients from controls. They found that an increasing number of VOCs resulted in a higher sensitivity for discrimination, i.e. the sensitivity for the detection of lung cancer, based on 4 different VOCs present in exhaled breath of lung cancer patients, was 52%, whereas this sensitivity improved to 71% when the prediction model was based on 15 different VOCs, with both models showing a specificity of 100% (6). Both studies, however, have not validated their results. A recent study of Long et al. also showed interesting results in an external validation study of exhaled breath biomarkers to diagnose lung cancer (7). They used the GC-MS technique to identify molecules in exhaled-breath showing a very high accuracy to detect lung cancer in the training group, which remained in the validation (over 90% sensitivity, 88% specificity, and an area under the receiver operating characteristic curve (AUC) of 0.93). It must however be noted that the discriminating VOCs, as reported in these three studies of Phillips, Bajtarevic, and Long, minimally corresponded.

Electronic noses mostly vary in terms of sensor type, breath sampling techniques, portability of the device, and classification methods, all accompanied by their own advantages and disadvantages. In 2007, Mazzone et al. investigated the usability of colorimetric sensors to identify lung cancer in exhaled breath (8). Colorimetric sensors are characterized by a change in colour after chemically sensitive compounds, e.g. VOCs, interact with the sensors. After including 143 subjects, they reached a sensitivity of 73% and a specificity of 72% to discriminate lung cancer patients from controls, including subjects with other lung diseases and healthy controls. In a consecutive study with a split-sample design in 2015, they also attempted to distinguish lung cancer patients from controls, including subgroups based on histology type and lung cancer stage, reaching accuracies from 79% to 86% when building prediction models based on exhaled breath only (9).
Conductive polymer gas sensors are also considered a potential sensor type, which have been incorporated in the Cyranose 320, a hand-held electronic nose that has extensively been investigated, not only in the lung cancer field. The technical principle is based on a change in the sensors' electrical resistance after exhaled gases are absorbed on the sensor surface. Machado et al. used the Cyranose 320 already in 2005 to analyse exhaled breath of 135 subjects (28 patients with bronchogenic carcinoma) in which they built a prediction model using 59 subjects and validated this prediction model on the remaining subjects (10). Validation of the model resulted in an accuracy of 85%, a sensitivity of 71%, and a specificity of 92%. Dragonieri et al. used the Cyranose 320 to distinguish COPD patients from NSCLC patients and healthy subjects. Based on 30 subjects, patients

with COPD and lung cancer patients could be distinguished with an accuracy of 85%, whereas lung cancer patients could be distinguished from healthy controls with an accuracy of 90% (11). In 2018, Tirzite et al. included 475 subjects (252 lung cancer patients). Based on exhaled breath patterns detected with the Cyranose 320, they could separate both groups with a sensitivity of 96%, and a specificity ranging from 90.6% to 92.3% depending on smoking status, whereas NSCLC patients and healthy controls were discriminated with an accuracy of 90% (12).

Metal oxide sensors (MOS) are characterized by changes in conductivity when gas molecules interact with the sensor surface. Van de Goor et al. used the Aeonose™ in 2018 to distinguish between patients with lung cancer and healthy controls. 167 subjects were included of whom 107 were diagnosed with lung cancer. They found a sensitivity of 83%, a specificity of 84%, with an AUC of 0.83, which remained in their validation cohort (sensitivity 86%, specificity 86%) (13).

Besides the Aeonose™, the device investigated in this thesis, another electronic nose device based on metal oxide sensors concerns the Spironose. In 2015, a study with the Spironose (n=144) investigated the possibility to discriminate lung cancer patients from patients with COPD, asthma, and healthy controls. Lung cancer patients could be distinguished from patients with COPD with an accuracy of 87%, from patients with asthma with an accuracy of 68%, and from healthy controls with an accuracy of 88% (14).

Finally, Shlomi et al. investigated exhaled breath in the diagnosis of lung cancer involving gold nanoparticle sensors. When exposed to air, electrical properties of the gold nanoparticles change. In a study to differentiate between subjects with benign nodules (n=30) and lung cancer patients (n=89), they found an accuracy to do so of 83% (15). Discrimination of early lung cancer from benign lung nodules showed good performance as well with an accuracy of 87%.

Considering all features of the abovementioned electronic nose devices, the Aeonose™ is characterized by highly sensitive metal-oxide sensors, portability of the device, transferability between devices, and without the necessity of temporary storage of the breath sample. These device-related features offer an interesting device for potential use in regular clinical practice.

**Chapter 2** and **Chapter 3** outlined methodological issues concerning exhaled breath analysis based on the principle of pattern recognition with machine learning techniques, where **Chapter 3** focused more on development and validation of machine learning based prediction models in general.

In **Chapter 2** we focused on the technical working mechanism of the Aeonose™ device, we provided a comprehensive overview of the statistical analyses involved in the pattern recognition techniques to give insight in the black box behind the Aeonose™, and we proposed a study design to train the Aeonose™ in a multicentre setting to discriminate lung cancer patients from non-lung cancer subjects.

Since a single measurement with the Aeonose™ yields a large number of parameters, exceeding the number of cases by far, machine learning techniques for supervised learning of 'big data' are favoured over conventional statistical methods (16). In supervised learning, training data are entered in an algorithm in order to identify links between the entered features and the outcome measures (17). However, as machine learning algorithms take into account all available information – also called data-driven-, whereas traditional statistical methods are hypothesis-driven, these machine-learning techniques involve a high risk of overfitting in which a prediction model fits perfectly on the specific training data, but possibly reflects noise instead of true relationships of the data. As a consequence the prediction model fails to generalize the results on unseen or blind data in a new dataset (18). Important techniques and steps to overcome overfitting are internal validation by using resampling techniques to estimate model accuracy and model robustness, and the application of external validation. It must be noted that in case of low risk of bias, a recent literature review showed equal performance of traditional logistic regression models and machine learning models, where machine learning models do not always lead to improved performance over traditional methods (19).

Performed studies on pattern recognition of breath data show large heterogeneity in the applied techniques to obtain validated classification models from raw sensor data (20). However, all techniques require fundamental steps, including data pre-processing, where baseline correction occurs, data compression to reduce data dimensions and eliminate useless information, classification techniques to build prediction models, and internal validation to estimate the performance and stability of the obtained model (21, 22). There are various machine learning techniques available for analysing and classifying breath data, e.g. artificial neural network (ANN), support vector machine (SVM), linear discriminant analysis (LDA), and random forest (RF). Until now, there is no consensus on which statistical techniques for dimension reduction and classification methods should be used and combined when analysing breath data (20). This is partly due to the lack of direct comparison of the techniques, the use of explorative study designs, and consequently a lack of validation studies to prove reproducibility of the obtained results. Hanna et al. have proposed a framework for conducting and reporting studies investigating VOCs in cancer diagnosis. However, this framework has been arranged for identification studies of VOCs without including studies on pattern recognition (23). An overall essential step is proper internal validation by resampling methods to estimate performance and robustness of the training set, and estimate the potential of overfitting, preferably followed by external validation in an independent dataset (17, 24, 25). Two commonly used resampling methods concern cross-validation in which the *leave-n-out* cross-validation routine uses a part of the training data to build a prediction model, and the remaining data are used as a test set. This implies in fact

that – e.g. in 10 consecutive steps - all data are predicted as if they were blind data, based on a training model built from the remaining 90% of the data. In a second commonly used validation method, called bootstrapping, all data are used for model development and resampling a single dataset with replacement creates many simulated samples where the performances of all developed models are averaged to indicate the optimism of the initially developed prediction model (24).
Still, not all published eNose studies report on this essential internal validation step to check stability and robustness of the obtained prediction model (20). Furthermore, external validation of prediction models is lacking in the majority of published eNose studies (20).

In **Chapter 3** we proposed a methodological study design to simultaneously develop and validate prediction models based on machine learning techniques in general. We used our training study as published in 2018 **(Chapter 4)** as a demonstration for applying the proposed study design (26). As previously mentioned, in case of a large amount of data with a relatively limited number of included subjects, machine learning techniques for supervised or unsupervised learning of the training data are favoured over traditional statistical methods. We already mentioned the importance of internal validation to manage the potential issue of overfitting, but in order to implement new diagnostic tools in clinical practice, external validation of prediction models remains fundamental. However, diagnostic techniques evolve rapidly nowadays, due to highly innovative technologies, e.g. continuous improvements in soft- and hardware, and including subjects for external validation of a prediction model might take so much time, that the model under investigation has already been surpassed by a more efficient prediction model. The original prediction model would in that case be no longer relevant, and new subjects need to be externally validated. The proposed study design as outlined in **Chapter 3** would be suitable for such study circumstances in which a new diagnostic tool seems highly relevant for clinical practice, but can change over time due to rapid technological innovations or changes in treatment options. An alternative circumstance where such a study design would be applicable is in case of a relatively rare disease where recruitment of subjects, for both the training model and the external validation model, takes a lot of time.
Another important feature of the proposed design is the ability to concurrently evaluate potential improvement, stability and robustness of a prediction model when increasing the sample size, consequently to discover a plateau at which increasing the sample size no longer results in an improvement of the model. Nevertheless, investigators should guard against overoptimistic estimates of model performance in the test set, which might be tackled by calibration of the model (27).

It must be noted that there is no intention to replace traditional external validation designs as outlined by the STARD and TRIPOD guidelines, which incorporate the important elements such as

similar case-mix, temporal validation in a prospective study design, geographical validation to test transportability of the model, different investigators and alternative, but related settings (26-29). Our goal was to provide an alternative, guiding design in situations where true external validation might be inefficient, especially in case of applying machine learning techniques considering a large amount of data.

**Chapter 4, 5, and 6** show the results of clinical studies with the application of exhaled breath analysis by the Aeonose™ to distinguish lung cancer patients from non-lung cancer subjects. **Chapter 4** reports on the diagnostic performance of the Aeonose™ in an exploratory multicentre training study to observe a potential signal in exhaled breath to diagnose lung cancer. **Chapter 5** investigated whether addition of clinical parameters to the exhaled breath data might improve this potential to diagnose lung cancer. **Chapter 6** shows an overall validation of a prediction model to diagnose lung cancer based on exhaled breath data only, and with the addition of relevant clinical variables to the prediction model.

In **Chapter 4** we reported on the diagnostic performance of the Aeonose™ in an exploratory multicentre training study including 4 hospitals in the Netherlands, to discriminate non-small cell lung cancer patients from subjects without lung cancer. Based on 290 subjects, among which 144 non-small cell lung cancer patients, and 146 controls (61 suspected of lung cancer, but proven negative, and 85 healthy controls), analysis of exhaled breath patterns with the Aeonose™ showed the ability to discriminate non-small cell lung cancer patients from non-lung cancer patients with a sensitivity of 94%, a specificity of 33%, and an AUC of 0.76 (95% confidence interval (CI): 0.71-0.82). A high negative predictive value (NPV) of 86% was observed, implying that a large number of suspected subjects can be prevented from undergoing unnecessary interventions. This reported diagnostic performance in our study as obtained by the prediction model based on exhaled breath data provides fair accuracy, but is lower than some of the earlier reported accuracies in other pattern recognition studies to diagnose lung cancer, especially showing a discrepancy in specificity (13, 30-32). Possible explanations for this discrepancy might be variability in study designs and control groups, our use of a multicentre and multidevice setting, small datasets with consequently risk of overfitting, focus on high sensitivity, and the use of different breath sampling techniques, different sensor technologies, and various statistical methods (33). In order to test generalisability of the individual results to the overall lung cancer population, external validation studies of the obtained results should be performed.

In case of differentiating the control subjects into a group that was suspected of lung cancer, due to complaints or abnormal imaging, and a group of healthy volunteers, matched on age and sex, the

diagnostic performance substantially differed. A decline in diagnostic performance was observed in the analysis to discriminate NSCLC patients from suspected subjects that were proven negative, showing an AUC of 0.73 (95% CI: 0.64-0.82), whereas a remarkable improvement was seen when differentiating NSCLC patients from healthy volunteers, showing an AUC of 0.85 (95% CI: 0.79-0.80). A possible explanation for the reduced diagnostic performance in the suspected subjects might be an alternative cause for the complaints, such as presence of a pulmonary infection, obstructive lung disease or interstitial lung disease that also alter the lung- and airway cell metabolism and the biochemical environment, leading to a change in VOC-pattern with a breath print more resembling lung cancer. After a follow up of five years, we did not observe a single case in the control group that eventually turned out to have lung cancer.

As suspected subjects of lung cancer generally share more risk factors with lung cancer patients than healthy controls do, such as current or past smoking, and presence of COPD, one could discuss the influence of intrinsic and extrinsic factors on exhaled-breath patterns. Various studies have investigated the influence of age, gender, smoking status, presence of comorbidities, and type of tumour stage and histology on the outcomes of exhaled-breath analysis (34-37). Important arguments against the influence of environmental factors in our study, including active smoking, are the incorporation of a nose clip during the Aeonose™ measurement to prevent breathing of environmental air, a 2-minute washout period to rinse the lung from air in the anatomical dead space in which no measurements are recorded, and the multicentre study design. However, it must be noted that our studies have been performed in controlled circumstances with the use of the same room for the measurements and strictly prohibited the use of (hand)alcohol in the room. The possible influence of environmental factors in uncontrolled circumstances are still unsure. A possible solution to overcome the possible intrinsic influence of concurrent comorbidities is to train the Aeonose™ to detect these comorbidities as well. However, these groups need to be sufficiently large to build adequate and stable prediction models. One could start with the performance of exploratory sub-analyses to obtain an impression of possible influencing variables and to investigate whether equal performance is achieved in subgroups, e.g. based on age, sex, presence of comorbidities, and smoking status. Another solution might be adequate matching of the lung cancer and control group in the training phase on these potential influencing variables to cancel out these influences. We also decided not to ask subjects to restrict eating or drinking before a measurement, as the neural network is being trained comparing breath profiles of positive and negative subjects for lung cancer, regardless of their food intake. Also, when numbers of subjects are sufficiently large, it can be assumed that food intake is not relevant as it averages out. However, since not all possible intrinsic and extrinsic influences have been investigated, and therefore cannot be excluded with certainty, in future studies and in clinical practice we might consider prohibition of certain foods and drinks for a

predefined period to standardize the breath sampling procedure. Furthermore, as proposed by Hanna et al., in the case of standardization of the breath sampling procedure one could take into account patient-related factors, environmental considerations and breath sampling methods (23). In this multicentre study we also performed sub-analyses on the two most common NSCLC histology types, i.e. adenocarcinoma (AC), and squamous cell carcinoma (SCC). Especially SCC showed a high negative predictive value of 93%, indicating that in case of a Aeonose™ value lower than -0.015, there is a high certainty, with high clinical relevance, that SCC is absent. Possible explanations for this high diagnostic accuracy might be the often-central origin of the tumour, and lower heterogeneity in this type of tumour compared to adenocarcinoma (38, 39). However, we observed a low incidence of SCC in our study population, and validity of this diagnostic performance should be tested in a larger study population. Besides, whole-genome sequencing for SCC is less available compared to AC, where potential driver mutations have possibly not been explored yet. Contrary, adenocarcinomas are known for their histological and molecular heterogeneity and show a lower diagnostic performance of exhaled breath analysis in our sub-analysis (40). These adenocarcinomas should probably be further divided based on their tumour and molecular characteristics for proper discrimination. Shlomi et al. investigated the difference in exhaled-breath patterns of patients with adenocarcinoma with and without an EGFR mutation (15). They found a diagnostic accuracy of 83% to discriminate both groups which supports our hypothesis that in adenocarcinoma further subdivisions in histology and biochemical features might lead to differences in breath patterns.
In a small sub-analysis to evaluate differences in breath patterns between small cell lung cancer (SCLC) patients and non-lung cancer patients, we found promising results to exclude SCLC with a high NPV of 97%. As mentioned before, this also concerns a very small sample size of SCLC patients, where the results should be interpreted with caution.

In **Chapter 5**, the aforementioned original, multicentre training model as discussed in **Chapter 4** was extended with relevant clinical parameters to assess improvement of diagnostic performance of the prediction model. Various clinical parameters are known as independent risk factors for the development of lung cancer, such as age, sex, presence of COPD, smoking status, number of pack-years smoked, family history, and presence of emphysema (41, 42). Including 281 subjects (138 NSCLC patients, and 143 subjects without lung cancer), univariate analyses showed age, sex, smoking status, number of pack-years, presence of COPD, and the absolute classification value of the Aeonose™ to be associated with the presence of lung cancer. These univariately associated clinical variables were entered into two types of multivariate analysis: 1) a traditional multivariate logistic regression analysis where the absolute classification value of the Aeonose™, as obtained by the neural network analysis, was entered as an independent variable together with the univariately

associated clinical variables, and 2) addition of the univariately associated clinical variables *a priori* to the artificial neural network, together with the vector containing breath profile data.
Both types of multivariate analysis showed a remarkable improvement of diagnostic performance compared to analysis based on solely exhaled breath data. In case of a prediction model built on this current study population based on exhaled breath data only, we found a sensitivity of 94.2%, a specificity of 44.1%, and a positive predictive value (PPV) and NPV of 61.9% and 88.7%, respectively, with an AUC of 0.75 (95% CI 0.69–0.81). Minor differences were observed compared to our previous performed analysis in **Chapter 4** due to a somewhat smaller sample size because of missing information on pack-years. The multivariate logistic regression analysis including the classification value of the Aeonose™ and clinical variables showed, at a chosen cut-off for sensitivity relevant for clinical practice, an improved AUC of 0.86 (95% CI: 0.81-0.90), with a sensitivity of 96%, a specificity of 60%, and an NPV of 93%. In case of training the ANN with the vector containing exhaled breath data together with the univariately associated clinical variables, we observed a relevant improvement in diagnostic performance to distinguish NSCLC patients from non-lung cancer subjects as well, with a sensitivity of 94%, a specificity of 49%, an NPV of 90%, with an AUC of 0.84 (95% CI: 0.79-0.89). We concluded that both multivariate methods were equally capable of increasing classification quality in diagnosing lung cancer, especially leading to an increase in specificity with consequently less false positive diagnoses and unnecessary interventions. Since clinical risk factors for developing lung cancer are known, and they are often readily available, combining these variables with non-invasive breath data is recommended. Other biomarker studies, including serum biomarkers and genetic studies have also used machine learning techniques to build prediction models to diagnose lung cancer including clinical variables (43, 44). However, studies on combining exhaled breath data and clinical variables are rare. Tirzite et al. used logistic regression analysis to predict the presence of lung cancer with the Cyranose 320 mainly using segments of exhaled breath as input variables for the logistic regression analysis, but they also included a number of clinical parameters, such as age, smoking status, smoking history and ambient temperature (12). They were able to distinguish subjects with lung cancer from controls with a sensitivity of 96% in both smokers and non-smokers, and a specificity >90% in both groups. Mazzone et al. performed a study with colorimetric sensor arrays to analyse exhaled breath to diagnose lung cancer in a split-sample design where they found a moderate discriminative performance to diagnose lung cancer (C-statistic of 0.79 (95% CI: 0.78-0.82)), but addition of clinical parameters, such as age, smoking history, and presence of COPD did not lead to a relevant improvement of the prediction model to diagnose lung cancer (C-statistic of 0.80 (95% CI: 0.78-0.82)) (9). However, they did not report on the univariate influences of the clinical variables in their study population.

In **Chapter 6** we trained and validated a prediction model in a split-sample design based on exhaled breath data to distinguish between patients with NSCLC and clinically relevant control subjects in a multicentre (7 hospitals) and multi-device setting with and without the addition of clinical variables. Since our previously performed training study indicated a potential signal in the exhaled breath of lung cancer patients, the next essential step was to validate these results in a multicentre setting on an independent group of subjects. Unfortunately, due to the use of the second generation, CE-certified Aeonose™ device, we could not use the prediction model based on the previously collected data as obtained in **Chapter 4**. Given the long timeframe of collecting the necessary data, we decided to use a split-sample study design in which we enabled development and subsequent validation of a new prediction model as outlined in **Chapter 3**.

The training set consisted of 376 subjects (160 lung cancer patients, 216 controls) and the validation set consisted of 199 subjects (79 lung cancer patients, 120 controls). We observed a moderate model performance to discriminate patients with NSCLC and controls based on exhaled breath data only, with similar results in the validation set including 'blind' subjects. The prediction model based on exhaled breath data only, with a cut-off probability of 20% for the diagnosis of lung cancer, showed in the validation set a sensitivity of 88%, a specificity of 48%, a PPV of 52%, an NPV of 87%, with an AUC of 0.79 (95% CI: 0.72-0.85). Discrimination between both groups improved significantly when readily available clinical variables, i.e. age, sex, and number of pack-years were added to the prediction model. Classifying new subjects, not used for training of the Aeonose™, maintaining the same cut-off probability and regression coefficients as in the training set, showed excellent performance with a sensitivity of 95%, a specificity of 49%, a PPV of 54%, an NPV of 94%, and an AUC of 0.86 (95% CI: 0.81-0.91). Similar to the training study in **Chapter 4**, we focused on high sensitivity and high NPV based on clinical relevance in order to miss as few lung cancer cases as possible due to its high mortality. With the chosen cut-off probability of having lung cancer of 16%, about one third of the subjects (63 of the 196 subjects) were classified as not having lung cancer and could be refrained from undergoing unnecessary invasive diagnostics based on our prediction model.

In this validation study, we applied the original developed model, with the same predictors and regression coefficients, to measure the outcome values in the test set with new individuals that had not been used to develop the original model. Not only did we assess discrimination to test this model's performance, we also assessed calibration in a calibration plot to evaluate the model's predictive performance as a quantitative measure as proposed by Moons et al., showing good concordance between the predicted probability of lung cancer in each decile, and the observed prevalence of lung cancer the same decile (26).

To our knowledge, our study as reported above, was the first study to validate "blind" exhaled breath profiles of patients with NSCLC and subjects without lung cancer based on pattern recognition

techniques in a multicentre, multinational split-sample design, including clinical variables with the use of multiple electronic nose devices. As mentioned above, Mazzone et al. performed a split-sample study design using pattern recognition techniques based on exhaled breath to distinguish lung cancer patients from control subjects (9). This concerned a two-centre study with the application of only one electronic nose device, and diagnostic performance in the validation set could be considered moderate. A recent study of Long et al. showed interesting results in an external validation study of exhaled breath biomarkers to diagnose lung cancer. Although they made use of the GC-MS technique, with several Tedlar bags and one GC-MS station, to identify molecules in exhaled breath, they also focused on the possible origin of breath biomarkers by explaining specific metabolic processes in lung cancer pathogenesis. In a validation cohort including 156 lung cancer patients and 100 controls, eight discriminating VOCs reached an AUC of 0.96 to discriminate between the two groups. Sub-analysis on lung cancer stage and histology subtype also showed excellent performances. This strict study protocol, as well as the high costs, necessity to temporarily store breath samples, and requirement of expertise to handle the technique, may not be easily implemented in daily clinical practice, and contrary to the Aeonose™, it does not offer a point-of-care solution.

In our validation study, we also performed explorative subgroup analyses on lung cancer stage, different sexes and age groups, and different types of histology. An important finding of these analyses was the good diagnostic performance in all lung cancer stages, where we found a sensitivity and NPV of 94% and 97%, respectively in early-stage lung cancer, while in stage III and IV lung cancer, sensitivity and NPV were 84% and 90%, respectively.

## Future implications

The results of these clinical studies pave the way for clinical implementation of exhaled breath analysis with electronic nose technology serving as a non-invasive biomarker in the diagnosis of lung cancer. As large studies focusing on low-dose computed tomography (LDCT) screening have shown reduced lung cancer mortality in high-risk asymptomatic subjects, the potential of combining these non-invasive diagnostic tools should be investigated (45, 46). Due to the high sensitivity and high NPV to exclude lung cancer, the Aeonose™ may be deployed after suspicion of lung cancer has been raised by LDCT, where appropriate monitoring steps, including a watch-and-wait strategy, might be applied. Furthermore, determination of accurate screening criteria for LDCT screening remains debatable since only subjects at the highest risk for lung cancer are targeted in current screening programmes. Combining clinical parameters and exhaled breath data in an ANN could indicate the degree of suspicion of lung cancer, and therefore serve as an adjunct for risk stratification in lung

cancer screening, supporting clinical decision making. However, an important limitation of our training study as discussed in **Chapter 4 and 5** was the small population of early-stage lung cancer (+/- 25%), which would be the target population of such a lung cancer screening programme. In our validation study we included a larger subgroup of early-stage lung cancer (45-50%), that better corresponds with the target population in screening programmes. Results of these sub-analyses are promising. However, our clinical studies were performed in a hospital setting in subjects with a clear suspicion of lung cancer, whereas screening programmes target asymptomatic, high-risk subjects. To assess acceptability and feasibility of the exhaled breath analysis with the Aeonose™ in clinical practice, we envision to first use the Aeonose™ parallel to current practice in a hospital setting. In particular the purpose to prevent subjects to undergo unnecessary invasive interventions could be pursued, and more often a wait-and-see strategy could be applied with strict follow-up. In order to assess the synergistic role of exhaled breath analysis in screening purposes, guidance of risk assessment prior to LDCT as a pre-selection tool, and determination of surveillance intervals in case of solitary pulmonary nodules, a new study design should be formatted including asymptomatic, high-risk subjects with a considerably lower prevalence of lung cancer compared to a hospital setting. Exhaled breath analysis could not only play a substantial role in lung cancer screening, it might also serve an important role as a prognostic biomarker to predict effectiveness of proposed treatments, and monitor possible disease recurrence after treatment. Especially since lung cancer treatment increasingly focuses on the concept of personalized medicine with the current knowledge of the importance of driver mutations and PD-L1 status, each individual patient should be evaluated in terms of effectiveness of his personalized treatment, possibly based on a combination of biomarkers, including radiological and pathological markers, and exhaled breath patterns. Besides, the Aeonose™ could be trained to detect recurrent disease after a curative treatment, possibly in a stage where this recurrent disease is not yet radiologically visible, or to detect minimal residual disease after an intended curative treatment.

Specifically for the Aeonose™, further improvements in soft- and hardware should provide a higher specificity to minimize the number of false-positive subjects and prevent more subjects from undergoing unnecessary invasive interventions.

## References


1. Buma AIG, Muller M, de Vries R, Sterk PJ, van der Noort V, Wolf-Lansdorf M, et al. eNose analysis for early immunotherapy response monitoring in non-small cell lung cancer. Lung Cancer. 2021;160:36-43.
2. Nardi-Agmon I, Abud-Hawa M, Liran O, Gai-Mor N, Ilouze M, Onn A, et al. Exhaled Breath Analysis for Monitoring Response to Treatment in Advanced Lung Cancer. J Thorac Oncol. 2016;11(6):827-37.
3. Schumer EM, Black MC, Bousamra M, 2nd, Trivedi JR, Li M, Fu XA, et al. Normalization of Exhaled Carbonyl Compounds After Lung Cancer Resection. Ann Thorac Surg. 2016;102(4):1095-100.
4. Phillips M, Cataneo RN, Cummin AR, Gagliardi AJ, Gleeson K, Greenberg J, et al. Detection of lung cancer with volatile markers in the breath. Chest. 2003;123(6):2115-23.
5. Phillips M, Altorki N, Austin JH, Cameron RB, Cataneo RN, Greenberg J, et al. Prediction of lung cancer using volatile biomarkers in breath. Cancer Biomark. 2007;3(2):95-109.
6. Bajtarevic A, Ager C, Pienz M, Klieber M, Schwarz K, Ligor M, et al. Noninvasive detection of lung cancer by analysis of exhaled breath. BMC Cancer. 2009;9:348.
7. Long Y, Wang C, Wang T, Li W, Dai W, Xie S, et al. High performance exhaled breath biomarkers for diagnosis of lung cancer and potential biomarkers for classification of lung cancer. J Breath Res. 2021;15(1):016017.
8. Mazzone PJ, Hammel J, Dweik R, Na J, Czich C, Laskowski D, et al. Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array. Thorax. 2007;62(7):565-8.
9. Mazzone PJ, Wang XF, Lim S, Jett J, Choi H, Zhang Q, et al. Progress in the development of volatile exhaled breath signatures of lung cancer. Ann Am Thorac Soc. 2015;12(5):752-7.
10. Machado RF, Laskowski D, Deffenderfer O, Burch T, Zheng S, Mazzone PJ, et al. Detection of lung cancer by sensor array analyses of exhaled breath. Am J Respir Crit Care Med. 2005;171(11):1286-91.
11. Dragonieri S, Annema JT, Schot R, van der Schee MP, Spanevello A, Carratu P, et al. An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD. Lung Cancer. 2009;64(2):166-70.
12. Tirzite M, Bukovskis M, Strazda G, Jurka N, Taivans I. Detection of lung cancer with electronic nose and logistic regression analysis. J Breath Res. 2018;13(1):016006.
13. van de Goor R, van Hooren M, Dingemans AM, Kremer B, Kross K. Training and Validating a Portable Electronic Nose for Lung Cancer Screening. J Thorac Oncol. 2018;13(5):676-81.
14. de Vries R, Brinkman P, van der Schee MP, Fens N, Dijkers E, Bootsma SK, et al. Integration of electronic nose technology with spirometry: validation of a new approach for exhaled breath analysis. J Breath Res. 2015;9(4):046001.
15. Shlomi D, Abud M, Liran O, Bar J, Gai-Mor N, Ilouze M, et al. Detection of Lung Cancer and EGFR Mutation by Electronic Nose System. J Thorac Oncol. 2017;12(10):1544-51.
16. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. Medicina (Kaunas). 2020;56(9).
17. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. 2018;284(6):603-19.
18. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774-81.
19. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.
20. Leopold JH, Bos LD, Sterk PJ, Schultz MJ, Fens N, Horvath I, et al. Comparison of classification methods in breath analysis by electronic nose. J Breath Res. 2015;9(4):046002.

21. Herbig J, Beauchamp J. Towards standardization in the analysis of breath gas volatiles. J Breath Res. 2014;8(3):037101.
22. Horvath I, Barnes PJ, Loukides S, Sterk PJ, Hogman M, Olin AC, et al. A European Respiratory Society technical standard: exhaled biomarkers in lung disease. Eur Respir J. 2017;49(4).
23. Hanna GB, Boshier PR, Markar SR, Romano A. Accuracy and Methodologic Challenges of Volatile Organic Compound-Based Exhaled Breath Tests for Cancer Diagnosis: A Systematic Review and Meta-analysis. JAMA Oncol. 2019;5(1):e182815.
24. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683-90.
25. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.
26. Kort S, Tiggeloven MM, Brusse-Keizer M, Gerritsen JW, Schouwink JH, Citgez E, et al. Multi-centre prospective study on diagnosing subtypes of lung cancer by exhaled-breath analysis. Lung Cancer. 2018;125:223-9.
27. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
28. Bossuyt PM, Cohen JF, Gatsonis CA, Korevaar DA, group S. STARD 2015: updated reporting guidelines for all diagnostic accuracy studies. Ann Transl Med. 2016;4(4):85.
29. Patzer RE, Kaji AH, Fong Y. TRIPOD Reporting Guidelines for Diagnostic and Prognostic Studies. JAMA Surg. 2021;156(7):675-6.
30. Gasparri R, Santonico M, Valentini C, Sedda G, Borri A, Petrella F, et al. Volatile signature for the early diagnosis of lung cancer. J Breath Res. 2016;10(1):016007.
31. Huang CH, Zeng C, Wang YC, Peng HY, Lin CS, Chang CJ, et al. A Study of Diagnostic Accuracy Using a Chemical Sensor Array and a Machine Learning Technique to Detect Lung Cancer. Sensors (Basel). 2018;18(9).
32. Rocco G, Pennazza G, Santonico M, Longo F, Rocco R, Crucitti P, et al. Breathprinting and Early Diagnosis of Lung Cancer. J Thorac Oncol. 2018;13(7):883-94.
33. Antoniou SX, Gaude E, Ruparel M, van der Schee MP, Janes SM, Rintoul RC, et al. The potential of breath analysis to improve outcome for patients with lung cancer. J Breath Res. 2019;13(3):034002.
34. Buszewski B, Ulanowska A, Ligor T, Denderz N, Amann A. Analysis of exhaled breath from smokers, passive smokers and non-smokers by solid-phase microextraction gas chromatography/mass spectrometry. Biomed Chromatogr. 2009;23(5):551-6.
35. Dragonieri S, Schot R, Mertens BJ, Le Cessie S, Gauw SA, Spanevello A, et al. An electronic nose in the discrimination of patients with asthma and controls. J Allergy Clin Immunol. 2007;120(4):856-62.
36. Phillips M, Cataneo RN, Greenberg J, Gunawardena R, Naidu A, Rahbari-Oskoui F. Effect of age on the breath methylated alkane contour, a display of apparent new markers of oxidative stress. J Lab Clin Med. 2000;136(3):243-9.
37. Tirzite M, Bukovskis M, Strazda G, Jurka N, Taivans I. Detection of lung cancer in exhaled breath with an electronic nose using support vector machine analysis. J Breath Res. 2017;11(3):036009.
38. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. J Thorac Oncol. 2015;10(9):1243-60.
39. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Rev Cancer. 2014;14(8):535-46.
40. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455(7216):1069-75.

41. Herder GJ, van Tinteren H, Golding RP, Kostense PJ, Comans EF, Smit EF, et al. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. Chest. 2005;128(4):2490-6.
42. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung cancer risk prediction: Prostate, Lung, Colorectal And Ovarian Cancer Screening Trial models and validation. J Natl Cancer Inst. 2011;103(13):1058-68.
43. Duan X, Yang Y, Tan S, Wang S, Feng X, Cui L, et al. Application of artificial neural network model combined with four biomarkers in auxiliary diagnosis of lung cancer. Med Biol Eng Comput. 2017;55(8):1239-48.
44. Hart GR, Roffman DA, Decker R, Deng J. A multi-parameterized artificial neural network for lung cancer risk prediction. PLoS One. 2018;13(10):e0205264.
45. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. N Engl J Med. 2020;382(6):503-13.
46. National Lung Screening Trial Research T, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395-409.

## Summary

Lung cancer is the leading cause of cancer-related mortality worldwide. Lung cancer is subdivided in two major types: non-small cell lung cancer (NSCLC) and small-cell lung cancer (SCLC), respectively accounting for approximately 85% and 15% of the cases. Furthermore, NSCLC can be categorized in many subtypes, being the two most common adenocarcinoma and squamous cell carcinoma, each with different tumour characteristics, treatment options and prognosis. The 5-year survival rate for localized stage NSCLC approximates 60%, whilst the 5-year survival rate for metastatic disease equals 5%. In case of SCLC, 5-year survival rate for localized disease approximates only 30%, whilst the 5-year survival rate for metastatic disease conforms 3%. Despite substantial progress in the treatment options, such as targeted therapies with tyrosine kinase inhibitors (TKI's), immune therapy, improvements in surgical options, and personalized treatment, this high lung cancer related mortality reflects the fact that the majority of the patients present with advanced-stage disease, which is not curable.

In the past decades, various non-invasive technologies have been investigated as a potential tool to diagnose lung cancer. One of these technologies concerns exhaled breath analysis based on pattern recognition by electronic nose technology. Exhaled breath contains, besides inorganic compounds, such as water vapour, nitrogen, and carbon monoxide, also thousands of volatile organic compounds (VOCs) reflecting physiological and pathophysiological metabolic processes in the body. In case of a disease, metabolism alters leading to exhalation of a different composition of VOCs which can be captured by highly sensitive sensors and measured with artificial intelligence techniques. This type of technology mimics human olfaction in which one needs to be trained to recognize familiar smells and allows the electronic nose to recognize a 'smell' that matches lung cancer, or any other condition for which the electronic nose has been trained. In order to implement a new diagnostic tool to diagnose lung cancer, the technique should first be trained and validated to state whether the new technique is of sufficient additional value in clinical practice.

In this thesis, we investigated the potential of exhaled breath analysis to diagnose lung cancer by performing studies in which we trained and validated an electronic nose (Aeonose™) to distinguish patients with lung cancer from subjects without lung cancer. **Chapter 2** and **Chapter 3** mainly focus on methodological issues concerning exhaled breath analysis based on pattern recognition with machine learning techniques. **Chapters 4-6** show results of clinical studies in which the Aeonose™ is trained and validated.

In **Chapter 2** we outlined the proposed multicentre study design how to train the Aeonose™ as a diagnostic tool to diagnose lung cancer. This manuscript mainly focused on the technical working

mechanism of the device and the statistical analyses incorporating artificial intelligence and internal validation techniques to classify subjects as having lung cancer or not. We showed how a large amount of training data could be handled in such a way to prevent the risk of overfitting the prediction model.

As stated above, after a prediction model has been developed on training data, it is fundamental to validate this prediction model on new data in order to assess reproducibility and generizability of this prediction model in independent subjects. Since current diagnostic techniques rapidly evolve due to highly innovative technologies, inclusion of subjects for external validation often takes too long to properly assess the relevance and efficiency of the developed prediction model. In **Chapter 3** we proposed a methodological study design to simultaneously develop and validate prediction models based on machine learning techniques in general. We used our training study as published in 2018 **(Chapter 4)** as a demonstration for applying this proposed study design. This type of study design is especially suitable in case of an innovative, but highly relevant, diagnostic technique which can rapidly change due to technological developments, or in case of a rare disease where inclusion of subjects takes a lot of time.

**Chapters 4-6** show results of clinical multicentre studies where the Aeonose™ is trained and validated. The prediction model developed on the training data has been extended with clinical data to improve the diagnosis of lung cancer. In **Chapter 4** we performed an exploratory multicentre study to train the Aeonose™ to distinguish non-small cell lung cancer patients from subjects without lung cancer based on exhaled breath. Based on 290 subjects (144 NSCLC patients, 146 controls), the prediction model was able to discriminate both groups with a sensitivity of 94%, a specificity of 33%, a negative predictive value (NPV) of 86%, and an AUC of 0.76 (95% confidence interval (CI): 0.71-0.82). Since lung cancer is characterized by a high mortality when not timely remarked, we focused on a high negative predictive value, which was obtained. This high negative predictive value implies that a large number of subjects suspected of lung cancer could be prevented from undergoing unnecessary, probably invasive, interventions. Besides evaluation of the discriminative performance between NSCLC patients and non-NSCLC subjects, additional sub-analyses were performed on the two most common NSCLC histology types, i.e. adenocarcinoma, and squamous cell carcinoma. Squamous cell carcinoma showed an impressive high negative predictive value of 93% with an AUC of 0.78, indicating that in case of an Aeonose™ value lower than -0.015, there is a high certainty, with high clinical relevance, that squamous cell carcinoma is absent. Adenocarcinoma showed a slightly lower diagnostic accuracy with an AUC of 0.73, which might be explained by the heterogeneity of adenocarcinoma tumours. Also, in a small sub-analysis to evaluate differences in breath patterns between SCLC patients and non-lung cancer patients, we found promising results to exclude SCLC

with a high NPV of 97%, and an AUC of 0.86 (95% CI: 0.78-0.95). However, it must be noted that all sub-analyses were performed in a small group of subjects and further research is needed. Besides, all analyses in the training study have been performed with a CE-uncertified Aeonose™ device and have not been validated on independent data.

In **Chapter 5** we investigated the potential additional value of adding clinical parameters, which are also known to be predictive for lung cancer, to the obtained prediction model based on exhaled breath data from the training cohort in **Chapter 4**. We found that variables such as age, sex, smoking status, number of pack-years, presence of COPD, and the absolute classification value of the Aeonose™ were associated with the presence of lung cancer. Two types of multivariable statistical analysis were performed to assess the additional value of the extended prediction models. First, a multivariable logistic regression analysis, in which the absolute classification value of the Aeonose™ as obtained by the neural network analysis, was entered as an independent variable together with the univariately associated clinical variables. This model showed a substantial increase in diagnostic performance with an AUC of 0.86, a sensitivity of 96%, a specificity of 60%, and an NPV of 93%, compared to the original model based on exhaled breath data only with an AUC of 0.76. Second, we added the univariately independent variables *a priori* as an extension of the vector containing breath data defining the input of the artificial neural network, that was used in the training study to build the prediction model. This neural network model also showed great improvement of diagnostic performance to diagnose lung cancer with an AUC of 0.84, a sensitivity of 94%, a specificity of 49%, and an NPV of 90%.
Not only did sensitivity and NPV increase, also specificity increased in the multivariable models meaning a lower number of subjects that are incorrectly classified as having lung cancer.

Since the training studies, with and without clinical variables, indicated promising results of the Aeonose™ to diagnose lung cancer, we performed a large, multicentre, multinational validation study with multiple Aeonose™ devices to assess reproducibility and robustness of the obtained results. The results of this validation study are presented in **Chapter 6.** Due to the issue of continuous improvements in technology as mentioned in **Chapter 3**, and therefore the use of a second generation, CE-certified, Aeonose™, it was decided not to use the original data from the training cohort as described in **Chapter 4**. Instead, we recruited new subjects and performed a split-sample design which enabled development and subsequent validation of a new prediction model. The training set consisted of 376 subjects (160 lung cancer patients, 216 clinically relevant controls) and the validation set consisted of 199 subjects (79 lung cancer patients, 120 controls). We observed a moderate model performance to discriminate patients with NSCLC and controls based on exhaled-breath data only, at a cut-off probability of 20% for the diagnosis of lung cancer, with similar results

in the validation set including 'blind' subjects. This prediction model showed a sensitivity of 88%, a specificity of 48%, a PPV of 52%, an NPV of 87%, with an AUC of 0.79 (95% CI: 0.72-0.85) in the validation set. As seen in **Chapter 5**, adding relevant clinical variables that are also predictive for lung cancer substantially improved the diagnosis of lung cancer. Exhaled-breath data and clinical parameters from the training set were combined in a multivariable logistic regression analysis, maintaining a cut-off of 16% probability of lung cancer, resulting in a sensitivity of 95%, a specificity of 51%, and an NPV of 94%. This corresponded to an AUC of 0.87 (95% CI 0.83-0.90). When applying the identical multivariable logistic regression model on the validation set, maintaining the selected cut-off probability of 16%, we observed a sensitivity of 95%, a specificity of 49%, a PPV of 54%, and an NPV of 94%, with a corresponding AUC of 0.86 (0.81-0.91). This would mean, in case of this cut-off probability of 16%, that 63 of the 196 subjects (32%) were classified as "no lung cancer" and could, with high certainty, be prevented from undergoing unnecessary interventions.

In **Chapter 7** we place the main results of the performed studies in a broader context to discuss the relevance of the findings and future implications. Future research is needed to evaluate the value of exhaled breath analysis in lung cancer screening programmes, but also as an application to monitor treatment responses and detect early recurrence of the disease.

## Nederlandse samenvatting

Longkanker is wereldwijd de belangrijkste oorzaak van kanker-gerelateerde mortaliteit. Longkanker kan worden onderverdeeld in 2 typen: niet-kleincellig longcarcinoom (NSCLC) en kleincellig longcarcinoom (SCLC), welke respectievelijk circa 85% en 15% van de gevallen vormen. Daarnaast kan NSCLC verder worden onderverdeeld in meerdere subtypen, waarvan de twee meest voorkomende adenocarcinoom en plaveiselcelcarcinoom zijn, met elk hun eigen tumorkenmerken, behandelopties en prognose. De 5-jaars overleving van gelokaliseerd NSCLC ligt rond de 60%. Daarentegen ligt de 5-jaars overleving van gemetastaseerde ziekte slechts rond de 5%. In het geval van SCLC ligt de 5-jaars overleving van gelokaliseerde ziekte rond de 30% en bij gemetastaseerde ziekte is dit slechts 3%. Ondanks aanzienlijke verbeteringen in behandelopties de afgelopen jaren, zoals gerichte behandeling met tyrosinekinaseremmers (TKI's), immunotherapie, verbeteringen in chirurgische opties en gepersonaliseerde behandeling, reflecteert de hoge mortaliteit het feit dat de meeste mensen zich met reeds uitgebreide ziekte presenteren, waarbij curatie niet meer mogelijk is.

In de afgelopen decennia zijn er diverse niet-invasieve technologieën onderzocht als een potentieel diagnosticum voor longkanker. Dit betreft onder andere uitademingsanalyse gebaseerd op patroonherkenning door middel van een elektronische neus. Uitademingslucht bevat, naast anorganische componenten zoals waterdamp, koolstofdioxide, stikstof en koolstofmonoxide, ook duizenden vluchtige organische stoffen (VOC's) welke fysiologische en pathofysiologische processen in het lichaam reflecteren. In het geval van een ziekte verandert het metabolisme waardoor een andere samenstelling van VOC's wordt uitgeademd. Deze uitademingslucht kan worden geregistreerd door zeer gevoelige sensoren en vervolgens worden gemeten en geanalyseerd met verschillende kunstmatige intelligentie technieken. Dit type technologie komt overeen met de reuk bij mensen waarbij iemand eerst moet worden geleerd om een bepaalde geur te herkennen. In het geval van een elektronische neus wordt deze neus eerst geleerd en getraind welke 'geur' bij longkanker, of een andere ziekte past, waarna deze de volgende keer dezelfde geur kan matchen aan die specifieke ziekte. Om een dergelijk nieuw diagnosticum te implementeren, moet deze eerst getest en gevalideerd worden om te kunnen beoordelen of deze van voldoende additionele waarde is in de klinische praktijk.

In dit proefschrift hebben we de mogelijkheid van uitademingsanalyse om longkanker te diagnosticeren onderzocht, waarbij we een elektronische neus (Aeonose™) hebben getraind en gevalideerd om patiënten met longkanker te onderscheiden van personen zonder longkanker. **Hoofdstukken 2 en 3** richten zich met name op methodologische kwesties rondom uitademingsanalyses gebaseerd op patroonherkenning met kunstmatige intelligentie technieken.

**Hoofdstukkers 4-6** laten resultaten zien van klinische studies waarbij de Aeonose™ is getraind en gevalideerd.

In **hoofdstuk 2** hebben we in ons voorgestelde multicenter onderzoek uiteengezet hoe de Aeonose™ te trainen om longcarcinoom te diagnosticeren. Dit manuscript richt zich met name op de technologie achter de Aeonose™ en de statistische analyses gericht op kunstmatige intelligentie technieken en interne validatie technieken om personen te classificeren als het hebben van longkanker of niet. We laten zien hoe om te gaan met een grote hoeveelheid data om overfitting van een predictiemodel te voorkomen.

Zoals hierboven beschreven, nadat een predictiemodel ontwikkeld is op basis van training data, is het noodzakelijk om dit predictiemodel te valideren op nieuwe data om reproduceerbaarheid en generaliseerbaarheid van het model te beoordelen in een onafhankelijk groep mensen. Aangezien huidige diagnostische technieken zich snel ontwikkelen door uitermate innovatieve technologieën, duurt het soms te lang om personen voor een externe validatiestudie te includeren om goed de relevantie en efficiëntie van het ontwikkelde predictiemodel te kunnen beoordelen. In hoofdstuk 3 beschrijven we mogelijke studieopzet om tegelijkertijd een predictiemodel gebaseerd op kunstmatige intelligentie technieken te ontwikkelen en te valideren. Als voorbeeld gebruiken we onze training studie zoals gepubliceerd in 2018 als toepassing van deze voorgestelde studieopzet. Dit type studie design is voornamelijk geschikt in geval van een innovatieve, maar zeer relevante, diagnostische techniek die zich snel verder kan ontwikkelen, danwel in geval van zeldzame ziektes waarbij inclusie van studiedeelnemers erg lang duurt.

**Hoofdstukken 4-6** laten resultaten zien van klinische multicenter studies waarbij de Aeonose™ is getraind en gevalideerd. Het predictiemodel ontwikkeld op de training data is uitgebreid met klinische data om longkanker beter te kunnen diagnosticeren. In **hoofdstuk 4** hebben we een exploratieve multicenter studie uitgevoerd om de Aeonose™ te trainen om mensen met en zonder niet-kleincellig longcarcinoom van elkaar te onderscheiden op basis van uitademingsanalyses. Er werden 290 mensen geïncludeerd (144 NSCLC patiënten en 146 controles), waarbij het ontwikkelde predictiemodel beide groepen kon onderscheiden met een sensitiviteit van 94%, een specificiteit van 33%, een negatief voorspellende waarde (NPV) van 86%, en een oppervlakte onder de receiver operating characteristic curve (AUC) van 0.76 (95% betrouwbaarheidsinterval (CI): 0.71-0.82). Aangezien longkanker wordt gekenmerkt door een hoge mortaliteit wanneer dit niet vroegtijdig wordt ontdekt, zijn de predictiemodellen gericht op een hoge negatief voorspellende waarde. Deze aangetoonde hoge negatief voorspellende waarde impliceert dat bij een groot deel van de personen verdacht voor longkanker voorkomen kan worden dat ze onnodig, mogelijk invasief

vervolgonderzoek ondergaan. Naast evaluatie van het discriminerend vermogen van het predictiemodel tussen patiënten met en zonder niet-kleincellig longcarcinoom werden aanvullende sub analyses uitgevoerd gericht op de twee meest voorkomende subtypen van NSCLC: adenocarcinoom en plaveiselcelcarcinoom. Analyses gericht op plaveiselcelcarcinoom toonde een indrukwekkende hoge negatief voorspellende waarde van 93% met een AUC van 0.78, wat impliceert dat in geval van een absolute Aeonose™ waarde van lager dan -0.015, er met grote zekerheid en hoge klinische relevantie kan worden gesteld dat er geen sprake is van plaveiselcelcarcinoom. In geval van adenocarcinoom werd een iets lagere diagnostische nauwkeurigheid gevonden met een AUC van 0.73. Dit kan mogelijk verklaard worden door de heterogeniteit binnen adenocarcinomen. Ook werd in een kleine sub analyse gekeken naar verschillen in uitademingspatronen tussen SCLC-patiënten en personen zonder SCLC. Hierbij werden veelbelovende resultaten gevonden met een negatief voorspellende waarde van 97% en een AUC van 0.86 (95% CI: 0.78-0.95). Er moet echter worden benoemd dat deze sub analyses uitgevoerd zijn op datasets met een klein aantal inclusies. Daarnaast zijn alle analyses in de training studie uitgevoerd met een nog niet CE-gecertificeerd Aeonose™ apparaat en nog niet gevalideerd op onafhankelijke data.

In **hoofdstuk 5** is de potentiële waarde onderzocht van het toevoegen van klinische variabelen aan het reeds ontwikkelde predictiemodel gebaseerd op de uitademingsdata van het training cohort zoals beschreven in **hoofdstuk 4**. Resultaten lieten zien dat de variabelen leeftijd, geslacht, rookstatus, aantal gerookte pakjaren, aanwezigheid van COPD en de absolute classificatiewaarde van de Aeonose™ geassocieerd waren met het hebben van longkanker. Vervolgens zijn 2 typen multivariabele analyses uitgevoerd om de toegevoegde waarde van de uitgebreide predictiemodellen te onderzoeken. Enerzijds is een multivariabele logistische regressieanalyse verricht met als input de absolute classificatiewaarde van de Aeonose™, zoals verkregen door de neurale netwerkanalyse, samen met de onafhankelijke klinische variabelen. Dit model toonde een aanzienlijke verbetering in diagnostische nauwkeurigheid om personen met en zonder NSCLC van elkaar te onderscheiden met een AUC van 0.86, een sensitiviteit van 96%, een specificiteit van 60%, en een NPV van 93%, vergeleken met het training model gebaseerd op enkel uitademingsdata (AUC 0.76). Vervolgens is een analyse verricht waarbij de vector van de uitademingsdata, die als input voor het neurale netwerk diende in de training studie, uitgebreid werd met de onafhankelijk geassocieerde klinische variabelen. Dit neurale netwerkmodel toonde eveneens een evidente verbetering in diagnostische nauwkeurigheid met een AUC van 0.84, een sensitiviteit van 94%, een specificiteit van 49%, en een NPV van 90%. Beide uitgebreide modellen tonen niet alleen een verbetering van sensitiviteit en negatief voorspellende waarde, maar ook een toename van

specificiteit wat betekent dat er minder personen onterecht geclassificeerd worden als het hebben van longkanker.

Aangezien de training studies, met en zonder het toevoegen van klinische variabelen, veelbelovende resultaten hebben getoond ten aanzien van het diagnosticeren van longkanker met de Aeonose™, is vervolgens een grote, multicenter, multinationale validatie studie verricht met meerdere Aeonose™ apparaten om reproduceerbaarheid en robuustheid van de verkregen resultaten te beoordelen. De resultaten van deze validatiestudie worden gepresenteerd in **Hoofdstuk 6**. In verband met continue verbeteringen in de elektronische neus technologie, zoals beschreven in **Hoofdstuk 3**, is er gebruik gemaakt van een tweede generatie, CE-gecertificeerd, Aeonose™ apparaat. Dit heeft als gevolg gehad dat de originele data van de training studie uit **Hoofdstuk 4** niet konden worden gebruikt. In plaats daarvan zijn nieuwe proefpersonen geïncludeerd en is een split-sample studie design uitgevoerd om tegelijkertijd het nieuwe predictiemodel te ontwikkelen en te valideren. De training set bestond op 376 proefpersonen (160 NSCLC patiënten, 216 relevante controles) en de validatie set bestond uit 199 proefpersonen (79 NSCLC patiënten, 120 controles). In geval van een predictiemodel enkel op basis van uitademingsdata, werd een redelijke performance gezien om beide te groepen te onderscheiden met daarbij gelijke resultaten in de validatie set. Dit voorspellende model toonde, bij een afkapwaarde van 20% kans op longkanker, een sensitiviteit van 88%, een specificiteit van 48%, een positief voorspellende waarde van 52%, een negatief voorspellende waarde van 90% en een AUC van 0.79 (95% CI: 0.72-0.85) in de validatie set. Zoals ook gezien werd in Hoofdstuk 5, leidde het toevoegen van relevante klinische variabelen, die voorspellend zijn voor het hebben van longkanker, tot een aanzienlijke verbetering van diagnostische nauwkeurigheid van het predictiemodel. Uitademingsdata en klinische variabelen werden geanalyseerd middels een multivariabel logistisch regressiemodel op de training data, waarbij – bij een afkapwaarde van 16% kans op longkanker – een sensitiviteit van 95%, een specificiteit van 51% en een negatief voorspellende waarde van 94% werden gezien. Dit kwam overeen met een AUC van 0.87 (95% CI: 0.83-0.90). In geval van het toepassen van precies hetzelfde predictiemodel (gelijke formule met dezelfde B-coëfficiënten en handhaving van de afkapwaarde van 16% kans op longkanker) op de validatie set, zagen we sensitiviteit van 95%, een specificiteit van 49%, een positief voorspellende waarde van 54% en een negatief voorspellende waarde van 94% met een AUC van 0.86 (0.81-0.91). Dit zou betekenen, dat bij deze afkapwaarde van 16%, 63 van de 196 proefpersonen (32%) geclassificeerd zouden worden als het niet hebben van longkanker waarbij bij deze groep met grote zekerheid longkanker kan worden uitgesloten en onnodige invasieve onderzoeken kunnen worden voorkomen.

In **hoofdstuk 7** plaatsen we de belangrijkste resultaten van de verrichte studies in een bredere context waarbij de relevantie en toekomstperspectieven worden besproken. Vervolgonderzoek is

nodig om de waarde van uitademingsanalyse te beoordelen in longkanker screening programma's, maar ook als toepassing om behandelresponses te beoordelen en vroeg terugkeer van de ziekte op te sporen.

## Dankwoord

Het avontuur van dit promotieonderzoek begon in 2015 met mijn stage wetenschap voor de Master Geneeskunde in het Medisch Spectrum Twente waar ik een half jaar de tijd voor had. Wetenschap en het doen van onderzoek vond ik boven verwachting leuk en het verslag was, mede dankzij de goede begeleiding, na 10 weken al afgerond. Op dat moment kwam bij de afdeling Longgeneeskunde in MST het verzoek om een onderzoeksprotocol te schrijven voor een nieuw promotieonderzoek om longkanker op te sporen met een elektronische neus, toen kwam ik voor het eerst in aanraking met de Aeonose™. Het opstellen van dit studieprotocol vond ik wel een mooie uitdaging voor de resterende 16 weken, niet wetende dat ik 6.5 jaar later zelf dit promotieonderzoek zou afronden.

Met de finishlijn in zicht, realiseer ik me dat ik de route naar deze bekroning voor jarenlang hard werken nooit af had kunnen leggen zonder de hulp, support en uitgebreide inzet van anderen. Het dankwoord heb ik dan ook al meermaals geschreven in mijn gedachten en hierbij neem ik op papier graag de mogelijkheid om een aantal mensen in het bijzonder te bedanken.

Allereerst natuurlijk alle patiënten en vrijwilligers die deel hebben genomen aan het onderzoek. In 6 jaar tijd en in 9 ziekenhuizen hebben ruim 1300 mensen in de Aeonose™ geblazen. Wie had dat van tevoren kunnen bedenken?! Wat de setting ook was, na een goed- of slechtnieuwsgesprek, de patiënten waren altijd bereid om te blazen en tegelijkertijd ook nog interesse te tonen in het onderzoek. ‘Als het geen voordeel meer voor mijzelf oplevert, dan hopelijk wel voor iemand anders’, aldus een groot deel van de patiënten. Ook de vrijwilligers die via via bij het onderzoek terecht zijn gekomen, wil ik ontzettend bedanken voor hun deelname.

Dan mijn promotor prof. Dr. Job van der Palen. Job, vanaf het moment dat we elkaar ontmoetten heb jij volledig het vertrouwen gehad in mijn kwaliteiten als onderzoekster. Vanaf het begin heb jij jouw sprankelende enthousiasme over weten te brengen. Als het op bepaalde momenten even tegenzat, wist jij er altijd voor te zorgen dat enthousiasme het won van mijn onzekerheid en stress. Ik bewonder jouw positieve drive die je op iedereen uitstraalt enorm. Naast de goede hulp bij het analyseren en schrijven heb jij er ook mede voor gezorgd dat ik ben gegroeid als mens. Je hebt me met vertrouwen leren presenteren voor volle zalen en we hebben op veel mooie plekken aan diverse groepen ons onderzoek mogen toelichten. Samen met Marjolein staan jullie altijd voor iedereen klaar, zijn jullie ontzettend laagdrempelig benaderbaar en dit maakt jullie geweldige begeleiders voor vele studenten en promovendi. MST mag hier heel trots op zijn! Onder het genot van vele kopjes thee en regelmatig een lekkere taart, hebben we een prachtig proefschrift neergezet. Daarnaast hebben we de afgelopen jaren ook vele mooie en gezellige uitstapjes gemaakt. Hopelijk kunnen we

in de toekomst onze samenwerking voortzetten, want zoals je weet wil ik me graag verder ontwikkelen in het vak Epidemiologie.

Dan mijn eerste co-promotor Dr. Marjolein Brusse. Marjolein, jij hebt me tijdens mijn stage wetenschap in 2015 begeleid en mede dankzij jou ben ik destijds zo enthousiast geworden over wetenschap. Je hebt altijd een ontzettend goede kijk op de diverse onderzoekslijnen en raakt altijd net de kritische snaar waardoor je me goed aan het denken zet en de artikelen kwalitatief beter worden. Daarnaast was het natuurlijk ook altijd heel erg gezellig om de congressen samen te bezoeken.

Mijn andere co-promotor, Dr. Hugo Schouwink. Beste Hugo, ik vind het heel bijzonder dat jij mij als longarts, naast mijn klinische werk als AIOS, de afgelopen jaren ook hebt begeleid als co-promotor van mijn promotieonderzoek. Jouw jarenlange klinische ervaring heeft enorm veel bijgedragen aan de artikelen en je hebt me altijd goed laten nadenken over de relevantie van de inhoud. Jouw bevlogenheid voor oncologie als vak heeft er mede voor gezorgd dat ik ook erg enthousiast ben geworden en ik ben benieuwd of ik mijn pad hierin zal vervolgen.

Beste leden van de promotiecommissie, Prof. Dr. Baas, Prof. Dr. Ir. van Putten, Prof. Dr. Kusters, Prof. Dr. Sonke, Prof. Dr. Van den Heuvel en Dr. Brinkman. Hartelijk dank dat jullie bereid zijn geweest dit proefschrift kritisch te beoordelen. Ik voel mij vereerd dit proefschrift tegenover jullie te mogen verdedigen.

Emanuel, naast dat we tegelijkertijd aan onze promotie zijn begonnen, ben jij ook medeauteur op meerderheid van de artikelen en als longarts natuurlijk mijn dagelijkse begeleider in de klinische praktijk. Ik waardeer je kritische blik enorm, ook al leidde dat wel eens tot weer een tijdrovende aanpassing van een artikel, maar daardoor werd deze absoluut beter. Jouw kijk op de resultaten over hoe deze passen binnen de longoncologie, heeft ervoor gezorgd dat we continu goed zijn blijven nadenken over waar we naar streven en met beide benen op de grond bleven staan. Bedankt hiervoor! Ik hoop de komende jaren nog veel met je samen te werken en elkaar te mogen inspireren.

Beste Frans, nog voordat ik überhaupt van een elektronische neus had gehoord, hield jij je al jaren bezig met deze technische ontwikkeling. Je hebt ontzettend veel kennis van lopende onderzoeken en neemt altijd uitgebreid de tijd om, met een enorme drive en passie mij hierover te vertellen. Ik heb erg veel van je geleerd over ademanalyses, maar ook over het reilen en zeilen in de onderzoekswereld. Ontzettend bedankt hiervoor! Daarnaast heb ik natuurlijk het genoegen om als AIOS Longziekten in MST veel kennis op te doen van jouw expertise als longfysioloog, een echte pré van onze opleiding als je het mij vraagt.

De longartsen van MST wil ik bedanken voor hun uitgebreide interesse en steun de afgelopen jaren: Ilonka van Veen, Wendy van Beurden, Nicolle Hekelaar, Leonie Imming, Emanuel Citgez, Hugo Schouwink, Michiel Wagenaar, Albert Polman en Michiel Eijsvogel. Jullie hebben mij vanaf het begin dat ik bij jullie kwam werken een 'onderzoeksdag' gegeven en hebben altijd interesse getoond in de ontwikkelingen van het onderzoek. Soms waren er in een week wat meer onderzoeksdagen nodig, maar als de planning het toeliet, was dit altijd mogelijk. Via deze weg nogmaals bedankt hiervoor. Ook is het een fijne gedachte om te weten dat ik de komende jaren tijdens het afronden van de opleiding nog met jullie samen kan en mag werken. Ik kan niet anders zeggen dan dat het een (h)echte 'familie' bij de Longgeneeskunde MST is!

En natuurlijk niet te vergeten dr. Paul van der Valk. Beste Paul, in 2013 heb jij mij (destijds nog als Bachelor Geneeskunde student) al enthousiast gemaakt voor het vak Longgeneeskunde. Door jou en Job kwam ik bij het OBL terecht voor vakantiewerk. Hierbij heb ik voor de COPE III studie data verzameld en mocht ik als 'extraatje' ook aansluiten bij de overdrachten en de onderwijsmomenten. Vanaf dit punt is het zaadje voor mijn liefde voor de Longgeneeskunde geplant en ben ik, zoals we destijds afspraken, in 2016 als ANIOS bij jullie komen werken. Ik vind je een enorm inspirerend mens, vol energie, altijd vrolijk en tot op de dag van vandaag ontzettend ambitieus. Je hebt me zoveel geleerd over de Longgeneeskunde en wetenschap, maar ook over het leven en hebt mij altijd van goede tips voorzien hoe ik iets het beste aan kon pakken. Ik wil je hiervoor bedanken en hoop dat we elkaar de komende jaren nog regelmatig blijven zien.

Dan dr. Wouter van Geffen. Beste Wouter, als longarts in het MCL ben jij vanaf het validatie gedeelte bij dit onderzoek betrokken geweest. Daarvoor heb jij natuurlijk zelf promotieonderzoek gedaan met dezelfde elektronische neus bij patiënten met COPD. Ik heb jou leren kennen als een ontzettend gedreven, ambitieuze, maar ook zeer toegankelijke longarts. Vanaf het begin heb jij vertrouwen gehad in het onderzoek en je oppert regelmatig interessante ideeën die we nog kunnen uitwerken. Het is een eer dat jij zo betrokken hebt willen zijn bij mijn promotie en ik hoop dat we de komende jaren nog verder samen kunnen werken aan nog meer mooie artikelen.

Prof. Dr. Michel van den Heuvel. Beste Michel, jij bent als longarts vanuit het Radboud UMC bij het onderzoek betrokken geweest. Als hoogleraar in de longoncologie weet jij als geen ander wat nuttige en zinnige onderzoeken zijn en het is dan ook een enorme eer is dat jij het vertrouwen hebt gehad om bij te willen dragen aan dit onderzoek samen met het Radboud UMC. Ik waardeer je kritische blik, je realisme, maar vooral ook je doorzettingsvermogen om verschillende onderzoeken goed op te zetten en te laten slagen.

Dr. Van den Borne, Dr. Van Putten, Dr. Van den Bogart, Dr. Samii, Dr. Kastelijn en Dr. Van der Maten. Als longartsen van de deelnemende ziekenhuizen: Catharina Ziekenhuis Eindhoven, Martini Ziekenhuis Groningen, Bernhoven Ziekenhuis Uden, Deventer Ziekenhuis, Antonius Ziekenhuis Utrecht en Medisch Centrum Leeuwarden wil ik jullie bedanken voor het vertrouwen deel te nemen aan dit onderzoek. Jullie hebben ons de kans gegeven het onderzoek in samenwerking met jullie longfunctieanalisten en medewerkers van jullie researchafdelingen op te zetten en door het grote aantal deelnemende ziekenhuizen hebben we kunnen laten zien dat deze elektronische neus van grote waarde is. Het is geweldig om te zien dat we op deze manier een groot deel van Nederland hebben bereikt met het onderzoek. Ik hoop dat we in de toekomst in contact blijven en ik jullie over een paar jaar collega-longartsen mag noemen.

Dear Professor Dr. Daiana Stolz. It was such an honour to work with you and your team in Basel. Due to this multinational collaboration, our study is of increased quality and I want to thank you for your trust and support. Also, I got to meet the beautiful city of Basel and I will definitely come back. I also really want to thank Desiree who constantly helped me to overcome the distance when I needed some extra information on the inclusions. This surely saved a lot of travel time!

Beste Milou, ik wil jou ook ontzettend bedanken voor je uitgebreide hulp vanuit het Radboud UMC. Ik vind het heel bijzonder dat we allebei een eigen promotieonderzoek hebben gehad de afgelopen jaren, maar door het gemeenschappelijke onderwerp (biomarkers bij longkanker) ook samen een mooi artikel over de elektronische neus hebben kunnen schrijven. Heel erg bedankt dat je altijd weer snel de nodige informatie opzocht en we op die manier weer snel verder konden. Ik hoop dat we snel verder kunnen met ons volgende artikel en wens jou natuurlijk heel veel succes met het afronden van je eigen promotie en daarna als AIOS Longziekten.

Dan natuurlijk alle medewerkers van de Enose Company. Wat bijzonder om dit avontuur met jullie te mogen delen. Jullie hebben vanaf het begin vertrouwen in mij gehad en altijd geduld opgebracht gezien ook de combinatie met mijn klinische werk als AIOS. Naast dat jullie erg laagdrempelig benaderbaar zijn, staan jullie altijd direct klaar als er iets nodig is en is het altijd ook erg gezellig om jullie kantoor in Zutphen te bezoeken. Ik heb enorm veel respect voor jullie doorzettingsvermogen, kennis en manier van samenwerken in jullie relatief ‘kleine’ groep. Kort nog een paar extra woorden: Lieve Jetty, allereerst wil ik jou bedanken voor je onvoorwaardelijke inzet en hulp. Als er iets nodig was voor het onderzoek, waarbij het niet uitmaakte of het nou Leeuwarden of Eindhoven was, jij ging er direct heen en regelde het. Daardoor konden we altijd met volle vaart doorwerken. Daarnaast zal ik nooit meer onze geweldige trip naar Basel vergeten. Naast dat we in één dag honderden inclusies hebben gemonitord, was het vooral heel gezellig. We hebben ruim 15 uur samen in de auto gezeten

en volgens mij is het haast geen minuut stil geweest. Ik vind het heel bijzonder dat we dit samen hebben mogen doen.
Jan-Willem en Henny, ik wil jullie ook enorm bedanken voor jullie uitgebreide hulp, urenlange geduld als jullie voor de zoveelste keer weer moesten uitleggen hoe de analyses tot stand kwamen en jullie kritische blik op de inhoud. Ik vind het ontzettend leuk te zien hoe enthousiast jullie zijn en hoe jullie dit enthousiasme kunnen overbrengen op onder andere Job en mij.
André en Geeske, vanaf mijn eerste kennismaking met jullie in 2015 voelde onze samenwerking erg vertrouwd. Zonder jullie hulp was dit onderzoek nooit zo goed van de grond gekomen. Hulp bij het opstellen van het onderzoeksprotocol, de juiste contacten benaderen, maar ook regelmatig laagdrempelig face-to-face overleg heeft ervoor gezorgd dat we altijd goed door konden werken.
Ik kan niet anders zeggen dat ik een hele prettige samenwerking heb ervaren met jullie bedrijf en ik hoop dat we dit nog kunnen voortzetten in de toekomst.

Dan de longfunctieanalisten en het Onderzoeksbureau Longgeneeskunde (OBL) MST: Sylvia, Willemien, Monique, Hans, Marcel, Steffanie, Mirjam, Geke, Petra, Mardy, Merith, Ellen, José, Tanja, Anne en Joyce.
Lieve allen, het grootste deel van de patiënten in het MST zijn door jullie gemeten. Jullie maakten altijd tijd vrij als er weer een ‘onverwachte’ meting was en hebben er mede voor gezorgd dat zoveel mensen hebben kunnen blazen. Ik wil jullie hier enorm voor bedanken. Zonder jullie hadden we nooit zo’n uitgebreid onderzoek neer kunnen zetten in de afgelopen jaren!

Het grootste deel van de metingen in de overige ziekenhuizen zijn ook verricht door longfunctieanalisten en onderzoeksbureaus van de deelnemende ziekenhuizen. Ik wil jullie ook bedanken voor jullie hulp en inzet. Zonder jullie had ik alle metingen zelf moeten doen en dat was natuurlijk in zoveel ziekenhuizen, toch wel een eindje van Enschede, niet mogelijk. Jullie hebben de metingen snel meegenomen in jullie dagelijkse werkzaamheden en hebben in korte tijd veel patiënten gemeten. Bedankt hiervoor!

Nog een extra dankwoord voor Petra Hirmann (research manager MCL). Beste Petra, door jouw inzet hebben we in het MCL in korte tijd veel metingen kunnen verrichten. Jij gaf altijd direct aan wat er nodig was om het onderzoek zo voorspoedig mogelijk te kunnen laten verlopen en ik moet zeggen dat in het MCL altijd alles picobello op orde was als ik kwam monitoren. Ik kon zo achter een computer gaan zitten en alles was geregeld. Ik wil je hiervoor heel erg bedanken, het was altijd een pretje om Leeuwarden te bezoeken!

Daarnaast een speciale dank aan de diverse studenten die ik heb mogen begeleiden bij hun afstudeeronderzoeken of minor van de studies Geneeskunde en Gezondheidswetenschappen:

Mayke, Jesse, Simone, Harm en Melike. Door jullie te begeleiden heb ik zoveel extra geleerd de afgelopen jaren. Niet alleen over het onderzoek, maar ook qua organisatie en hoe studenten te begeleiden. Jullie eindverslagen waren stuk voor stuk van super goede kwaliteit en ik ben enorm trots op jullie dat jullie door mij begeleid wilden worden. Ik heb gemerkt dat het heel leuk is om met enthousiaste studenten te werken, dat geeft enorm veel energie!

Mijn collega AIOS en ANIOS wil ik bedanken voor hun steun en support de afgelopen jaren. Ik kreeg van jullie altijd de ruimte om een extra meting te doen, om naar een afspraak te gaan of een extra onderzoeksdag in te plannen. Jullie hebben regelmatig de telefoon voor mij vastgehouden, en we weten allemaal hoe vaak die gaat… Dus, heel erg bedankt! Ik vind het heel bijzonder deze promotie met jullie te mogen vieren. Zoals ik hierboven al schreef, we zijn een echte familie bij de Longgeneeskunde. Naast dat jullie collega's zijn, zijn jullie ook vrienden. Jullie bieden altijd een luisterend oor wanneer dit nodig is, maar zorgen vooral voor veel gezelligheid tijdens en buiten werktijd. Ik ben heel benieuwd hoe iedereens pad verder vervolgd zal worden, maar ik ben ervan overtuigd dat iedereen een hele mooie toekomst tegemoet gaat!

Beste Timon, samen met Emanuel zijn we tegelijkertijd onze promotie gestart, maar jij ging als een speer. Heel leuk dat we samen met de Aeonose™ hebben kunnen werken, maar dan allebei op een ander onderwerp. Ik bewonder jouw kennis, geduld en enthousiasme enorm. Je neemt uitgebreid de tijd om iets uit te leggen en je bent altijd als eerste op de hoogte van de meest recente artikelen, ook over Enose en longkanker als ik dit weer eens niet weet. Daarnaast ken ik niemand die zo enthousiast artikelen kan refereren als jij. Het is altijd een feest om naar een referaat van jou te mogen luisteren. Ik wil je bedanken voor je support en hulp de afgelopen jaren, dit deed je er 'maar' bij. Heel veel succes in de toekomst!

Lieve Jody, Masha danki!! voor dit prachtige artwork van het proefschrift. Jij hebt precies mijn passie en enthousiasme voor het onderzoek weten uit te drukken in deze mooi combinatie van afbeeldingen en kleuren. Ik hoop dat je hier nog veel meer mensen blij mee kan maken.

Dan mijn lieve paranimfen Joanne en Eveline. In 2016 zijn wij met zijn drieën begonnen als ANIOS Longgeneeskunde, de 3 musketiers. Wat een gezelligheid en wat hebben we veel aan elkaar gehad. Van hele drukke diensten tot gezellige avondjes en op de grond van het lachen om ons fictieve kindje 'Symbi Kort'. Heel bijzonder dat we na 6 jaar nog steeds dit contact hebben!
Lieve Joanne, naast dat jij in de dagelijkse praktijk echt mijn werk-maatje bent, ben je vooral de afgelopen jaren een super goede vriendin geworden. Als we op werk al veel gekletst hebben, kunnen we 's avonds zo weer aan de telefoon hangen. Jij biedt altijd een luisterend oor en weet altijd de rust terug te brengen. Daarnaast is het vooral erg gezellig en heb ik veel zin in toekomstige borrels,

fietstochtjes, koffietjes en onze toekomst als longarts.
Lieve Eveline, jou heb ik leren kennen tijdens ons M2 jaar in het MST. Ik weet nog goed hoe we voor het eerst gingen eten en borrelen bij Central Park en het was geen seconde stil. Jij bent pure gezelligheid en daarnaast sta jij altijd klaar, no matter wat de afstand is. Geweldig dat jij nu ook als AIOS Longziekten bent begonnen en ik hoop dat we ook snel van jouw promotie kunnen genieten!

Lieve schoonfamilie: Jan, Hettie en Ruud. Ik heb jullie in 2015 leren kennen en ben kort daarna met dit avontuur begonnen. Ondanks dat jullie niet precies kunnen weten wat het inhoudt, hebben jullie mij altijd gesteund. Als ik op de maandagen weer eens zei dat ik om 17.30 zou aanschuiven voor het eten en standaard om 18.00 (of later) binnen kwam lopen, was dit nooit een probleem. Jullie wachtten altijd.
Lieve Ruud, wat kan ik met jou goed reflecteren. Ook al hebben we beide een heel ander vak, jij kan HEEL goed luisteren en daardoor goed meedenken over een onderwerp waar jij niet eens helemaal in zit. Er zijn op deze manier veel goede ideeën naar boven gekomen en ik heb gemerkt dat we daardoor het beste in elkaar naar boven halen. Bedankt hiervoor! (Is de COVID tijd toch nog ergens goed geweest…).

Allerliefste Papa, Mama, Denice, Cherryl, Opa en Oma Kort, Robin, Martijn, James en Novi. Met zo'n familie kan ik me niet anders dan ontzettend gesteund, geliefd en dankbaar voelen. Jullie hebben altijd interesse getoond in het onderzoek, maar mij vooral continu laten voelen dat ik trots mag zijn op wie ik ben en wat ik doe. Soms moet ik even uitrazen en dat mag altijd bij jullie. Papa en mama, heerlijk hoe we altijd avonden lang kunnen kletsen over het leven. Zonder jullie hulp had ik überhaupt nooit aan de studie Geneeskunde kunnen beginnen. Ik ben zo trots op onze familie!
Een extra woordje voor mijn Oma van den Berg. Jij bent in 2010 overleden aan COPD. Wij hebben jou jarenlang meegemaakt met jouw ziekte en ik weet nog goed dat ik tegen je zei dat ik longarts zou worden zodat ik mensen zoals jij kon helpen. En kijk eens, nog even en dan ben ik echt een longarts!

En dan tot slot Rien, mijn alles! Rien, ik heb jou ontmoet in 2015 en kijk eens waar we nu staan: TEAM SHARIEN! Waar het personeelsrestaurant van MST wel niet goed voor is.. Jij bent mijn onvoorwaardelijke steun en relativeert voor mij wanneer dat nodig is. Naast dat ben je gewoon de allerliefste persoon die ik ken en maak je me altijd aan het lachen, ook al kost het de ene keer meer moeite dan de andere keer. Jij hebt er mede voor gezorgd dat ik hoogzwanger in mijn verlof dit proefschrift nog af heb gekregen en alle focus op ons prachtige gezin kan leggen. We zijn nu bijna 2 jaar getrouwd en samen met de allerliefste Isa gaat ons avontuur verder. Ik hou van jullie!

# Bibliography

## Publications


1. Koehorst- ter Huurne K, **Kort S**, Van der Palen J, et al. Quality of life and adherence to inhaled corticosteroids and Tiotropium are related. Int J Chron Obstr Pulmon Dis. 2016, Jul 26;11: 1679-88.

2. **Kort S**, Van der Palen J, Brusse-Keizer M. Data analysis of electronic nose technology in lung cancer: Generating prediction models by means of Aethena. J Breath Res. 2017;11(2):026006

3. **Kort S**, Tiggeloven MM, Brusse-Keizer M, et al. Multi-centre prospective study on diagnosing subtypes of lung cancer by exhaled-breath analysis. *Lung Cancer.* 2018;125:223-229.

4. **Kort S**, Brusse-Keizer M, Gerritsen JW, et al. Improving lung cancer diagnosis by combining exhaled-breath data and clinical parameters. *ERJ Open Res.* 2020;6(1).

5. Haalboom M, **Kort S**, Van der Palen J. Using a stepwise approach to simultaneously develop and validate machine learning based prediction models. J Clin Epidemiol. 2022; Feb;142:305-310

6. Haalboom M, **Kort S**, Van der Palen J. Letter to the editor regarding "Stepwise development and simultaneous validation to keep up with clinically relevant and rapidly changing diagnostic techniques. J Clin Epidemiol. 2022 Feb;142:332


## Presentations

1. *European Respiratory Society Congress, Milan, 2017*
   Poster discussion: Detection of non-small cell lung cancer by electronic nose.

2. *WEON Congress, Wageningen, 2018*
   Oral presentation: Detection of small cell lung cancer by electronic nose.
   Thematic poster: Detecting subtypes of non-small cell lung cancer by electronic nose.

3. *European Respiratory Society Congress, Paris, 2018*
   Poster discussion: Detecting subtypes of non-small cell lung cancer by electronic nose.
   Poster discussion: Detection of small cell lung cancer by electronic nose.

4. *European Respiratory Society Congress, Madrid, 2019*
   Oral presentation: Combining clinical parameters and exhaled breath analysis to improve the diagnosis of lung cancer.

5. *European Respiratory Society Congress, Digital, 2020*
   Oral presentation: External validation of exhaled-breath analysis to detect non-small cell lung cancer: a step-wise design to simultaneously develop and validate a prediction model.

## Curriculum vitae

Sharina Kort is geboren op 27 december 1991 in Enschede, als eerste dochter van Dick en Claudia. Na de basisschool ging zij in 2003 door naar het VWO, Tweetalig Onderwijs op het Stedelijk Lyceum Zuid in Enschede. Dit rondde zij in 2009 af waarna de keuze werd gemaakt om Geneeskunde te gaan studeren aan de Rijksuniversiteit Groningen. In de tussentijd werkte zij als bijbaantjes in de bediening bij een pizzeria en als maaltijdbezorgster. Na het behalen van de Bachelor (cum laude) duurde het een aantal maanden om te kunnen starten aan de Master en heeft zij enkele maanden gewerkt als medewerker van het Onderzoeksbureau Longgeneeskunde in Medisch Spectrum Twente (MST). Jaar 2 en 3 van haar Master werden ook in MST doorlopen waarna zij na het behalen van haar Masterdiploma in 2016 begon als ANIOS Longziekten in MST. Tegelijkertijd startte ook haar PhD project aan de Universiteit Twente onder leiding van Prof. Dr. Job van der Palen, Dr. Marjolein Brusse-Keizer en Dr. Hugo Schouwink, waar zij vanaf 2016 gedurende 1 dag in de week aan werkte. In juli 2017 begon Sharina aan de opleiding tot Longarts in MST waarvan de verwachting is dat zij deze eind 2024 afrondt. Sinds 2020 is zij getrouwd met Rien en in juni 2022 zijn zij trotse ouders geworden van dochter Isa. Sharina verdedigt haar proefschrift met als titel ‘Detection of lung cancer in exhaled breath by electronic nose technology’ op 7 oktober 2022 aan de Universiteit Twente.

Sharina Kort was born in Enschede, The Netherlands, on December 27th 1991 as the first daughter of Dick and Claudia. After finishing primary school, she started secondary school (VWO, bilingual education) at the Stedelijk Lyceum Zuid in Enschede. After her graduation in 2009, she decided to move to Groningen to study Medicine at the Rijksuniversiteit Groningen. In the meanwhile, she had two side jobs: as a waitress at a pizzeria and as a meal deliverer. In 2012 she obtained her Bachelor Degree cum laude. Due to some waiting time to start with her Master Degree, she worked a few months as an employee of the Onderzoeksbureau Longgeneeskunde in Medisch Spectrum Twente (MST). After graduating from her Master of Medicine in 2016, Sharina began to work as a resident at the Pulmonology department in MST. Simultaneously, she started as a PhD student for one day a week under the supervision of Prof. Dr. Job van der Palen, Dr. Marjolein Brusse-Keizer en Dr. Hugo Schouwink. In July 2017 she started to work as a resident to become a pulmonologist, which is expected to be completed at the end of 2024. She married Rien in 2020 and they became parents of their lovely daughter Isa in June 2022. Sharina defends her dissertation entitled ‘Detection of lung cancer in exhaled breath by electronic nose technology’ on October 7th 2022 at the University of Twente.