EDIAQI
Evidence Driven Indoor Air Quality Improvement

Deliverable D5.1

Report of existing cohort data and analysis regarding health effects of pollutants

Work Package 5
TOX & HEALTH
Version: Final

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Deliverable Overview

EDIAQI encompasses multiple cohorts with the objective of analysing the connections and associations between air quality and health. The substantial number of participants in these established cohorts offers the opportunity to uncover both linear and non-linear relationships in complex settings, extending beyond simple association cohorts' studies. Some of these cohorts are retrospective in nature. Before current cohorts and planned experimental research starts off, the project partners want to summarize research results in the project cohorts based on already published research preceding this project and a re-analysis of existing data. The aim of this deliverable is hence following; a) to investigate relationships within the retrospective cohorts and report the findings that will be continued and utilized in both newly planned and ongoing cohorts, and b) to document the known relationships as well as identify and analyse potential relationships based on pre-EDIAQI data. To do this, the pre-EDIAQI started cohorts, namely COPSAC2000 and COPSAC2010 provided by the partner RegionH in Denmark and the ATOPICA cohort provided by SCH in Croatia were taken into the re-analysis and reporting. Herein we report existing and new findings and ideas based on these cohorts.

Additional Information

Type: R – Document, report

Dissemination Level: PU - public

Official Submission Date: 31st of May 2023

Actual Submission Date: 31st of May 2023

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

Disclaimer

Document Revision History

Version	Date	Description	Partners
V0.1	8th of May 2023	1st draft of deliverable	REGIONH, SCH, ANT, KNOW
V0.2	14th of May 2023	2nd draft, added findings	REGIONH, SCH, ANT, KNOW
V0.3	30th of May 2023	Feedback internal revision	WINGS, ASC
V0.4	31st of May 2023	Revised version	REGIONH, SCH, ANT, KNOW
FINAL	31st of May 2023	Final quality check	REGIONH, SCH, ANT, KNOW, LC

Authors and Reviewers

Authors

- Astrid Sevelsted (RegionH)
- Michael Forsmann (RegionH)
- Mario Lovrić (KNOW)
- Ivana Banić (SCH)
- Mirjana Turkalj (SCH)
- Anja Bošnjaković (ANT)
- Iva Šunić (ANT)
- Kristina Pavlović (KNOW)

Reviewers

- Dejan Strbad (ASC)
- Vera Stavroulaki (WINGS)

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation, or both.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

Table of Contents

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

List of Figures

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

List of Tables

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

List of Terms and Abbreviations

Abbreviation	Description
EDIAQI	Evidence Driven Indoor Air Quality Improvement
ML	Machine Learning
IAP	Indoor air pollution
BC	Blackening (air pollutant)
$PM_{2.5}$	Particulate matter below 2.5um diameter
NO_2	Nitrogen dioxide
IAQ	Indoor air quality
RF	Random Forests algorithm
WPL	Work Package Leader
ATOPICA	An FP7 project and cohort https://cordis.europa.eu/project/id/282687
COPSAC	Copenhagen Prospective Studies on Asthma in Childhood Research unit withing RegionH
PM	Particulate matter (e.g. PM_{10})
NO	Nitrogen oxide(e.g. NO, NO_2)
FeNO	Fractional exhaled Nitric Oxide in the breath of patients
IgE	Immunoglobulin E, a type of antibody
WHO	World Health Organizsation
int	Integer (natural number)
TNF-alpha	Tumor necrosis factor alpha, a cytokine
Il	Interleukins are a group of cytokines
UHPLC-MS/MS	Ultra-high performance liquid chromatography with mass spectrometry
ECZ	Eczema, is a condition that causes dry, itchy and inflamed skin

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

Executive Summary

EDIAQI Mission

Indoor air pollution (IAP) and outdoor air pollution are substantial contributors to health issues, along with other environmental factors. The presence of indoor air pollutants originating from various sources such as building materials, furniture, dust, and cleaning as chemicals as well as outdoor sources negatively impacts both physical and mental well-being. It is worth noting that while the European Union has primarily emphasized outdoor air quality, there is limited knowledge and research on indoor air quality (IAQ). Key aspects that require further investigation include understanding the intricate relationships between indoor and outdoor pollution, identifying pollution sources and exposure pathways, examining the health effects of emerging pollutants, and evaluating the impact of indoor space ventilation.

Studying health effects

To increase the resilience of EU citizens regarding IAQ threats and to comprehend and promote living and working in healthy environments, project EDIAQI has started to conduct IAQ pilots and cohort studies. The research findings indicate relationships between indoor and outdoor air quality and health conditions in children. The study reveals that exposure to pollutants and allergens present in both indoor and outdoor environments can have a profound impact on the development and exacerbation of these conditions. In conclusion, the findings from the ATOPICA and COPSAC cohorts provide valuable insights into the relationships between environmental factors, indoor microbiomes, and the development of atopic allergies, respiratory symptoms, and asthma in children. The studies highlight the importance of exploring the impact of indoor and outdoor air pollution on asthma and allergy outcomes using various data sources. They also emphasize the need to analyse data using machine learning techniques to gain a comprehensive understanding of the complex relationships between air pollution and health outcomes. Furthermore, the studies shed light on the vulnerability of specific populations, such as young children, pregnant women, and individuals with low income and social status, to the adverse effects of poor air quality on their respiratory health. Additionally, the investigations into the relationships between

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

indoor microbiomes, airway microbiota, and asthma/allergy provide insights into the early-life programming of these conditions and the potential role of microbial exposures in their development. Finally, the studies underscore the importance of considering non-linear relationships between indoor air pollution and asthma/allergy outcomes to capture the full spectrum of effects. Overall, these findings contribute to the growing body of knowledge and provide a foundation for targeted prevention and management strategies to improve respiratory health and mitigate the risks associated with environmental determinants, particularly in vulnerable populations.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

1. Introduction

Asthma, allergies, and immunological issues are significant health concerns affecting children and adults worldwide[1] [2]. Environmental factors, particularly indoor and outdoor air pollution, play a crucial role in the development and exacerbation of these conditions [3]. Understanding the complex relationships between air pollution and respiratory health requires comprehensive analysis of large-scale cohort data. In recent years, advancements in machine learning and statistical techniques have provided new opportunities to uncover hidden patterns and associations within extensive datasets[4] [5]. This research aims to leverage existing cohort data from three diverse populations to investigate the impacts of indoor and outdoor air pollution on asthma, allergies, and immunological issues. The cohorts encompass individuals of different ages, geographical locations, and exposure profiles, offering a valuable opportunity to explore the multifaceted aspects of this complex relationship. By employing advanced ML algorithms and statistical models, this study aims to unravel the intricate interplay between air pollution, immunological responses, and the development of respiratory conditions. The primary objectives of this research are to identify and quantify the associations between indoor and outdoor air pollution and the prevalence, severity, and onset of asthma, allergies, and immunological issues. By harnessing the power of machine learning algorithms, we can uncover non-linear relationships, identify novel risk factors, and develop predictive models for these health outcomes. Additionally, statistical analyses will help elucidate the role of confounding

[1] K. F. Rabe *et al.*, 'Worldwide severity and control of asthma in children and adults: the global asthma insights and reality surveys', *J. Allergy Clin. Immunol.*, vol. 114, no. 1, pp. 40–47, Jul. 2004, doi: 10.1016/j.jaci.2004.04.042.

[2] A. J. Woolcock and J. K. Peat, 'Evidence for the Increase in Asthma Worldwide', in *Ciba Foundation Symposium 206 - The Rising Trends in Asthma*, John Wiley & Sons, Ltd, pp. 122–139. doi: 10.1002/9780470515334.ch8.

[3] C. Lu et al., 'Early life exposure to outdoor air pollution and indoor environmental factors on the development of childhood allergy from early symptoms to diseases', Environ. Res., vol. 216, p. 114538, Jan. 2023, doi: 10.1016/j.envres.2022.114538.

[4] M. Lovrić, I. Banić, E. Lacić, K. Pavlović, R. Kern, and M. Turkalj, 'Predicting Treatment Outcomes Using Explainable Machine Learning in Children with Asthma', Children, vol. 8, no. 5, p. 376, May 2021, doi: 10.3390/children8050376.

[5] M. K. Ross, J. Yoon, A. Van Der Schaar, and M. Van Der Schaar, 'Discovering pediatric asthma phenotypes on the basis of response to controller medication using machine learning', Ann. Am. Thorac. Soc., vol. 15, no. 1, pp. 49–58, Jan. 2018, doi: 10.1513/AnnalsATS.201702-101OC.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

variables, control for potential biases, and provide robust evidence supporting the observed associations. Through the integration of ML and statistical approaches with existing cohort data, this research has the potential to enhance our understanding of the complex interplay between air pollution, immunological factors, and respiratory health outcomes. The findings from this study can contribute to evidence-based interventions, policies, and strategies aimed at mitigating the adverse effects of indoor and outdoor air pollution, ultimately improving the respiratory health and quality of life for individuals affected by asthma, allergies, and immunological issues.

1.1 The challenges

Researching the associations between IAP and health effects in children poses several challenges. Some of these challenges include:

- Measurement and Exposure Assessment: Accurately measuring IAP and quantifying children's exposure to these pollutants can be challenging. IAP can vary in concentration and composition based on various factors. Collecting precise and representative data on IAP and exposure levels requires sophisticated measurement techniques and long-term monitoring.
- Multifactorial Nature: IAP is influenced by many factors, including occupants' behaviour, their number, heating and ventilation systems, building materials, and outdoor pollution sources. Determining the specific contribution of IAP to health effects in children requires disentangling these complex interactions and accounting for other confounding factors.
- Variability and Heterogeneity: Children's susceptibility to IAP can vary depending on age, genetic factors, gender, pre-existing health conditions, and socioeconomic factors. Designing studies that account for this variability and capturing diverse populations is essential for obtaining comprehensive and generalizable results.
- Ethical Considerations: Conducting research involving children necessitates careful consideration of ethical guidelines and obtaining informed consent from parents or guardians. Researchers must prioritize the safety and well-being of children and ensure that the study procedures and data collection methods are age-appropriate and ethically sound.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

1.2 Health related objectives

- Explore the relationship between indoor air and outdoor air pollution and asthma/allergy outcomes using various data sources
- Analyse data using machine learning techniques to understand the relationship between air pollution and health outcomes.
- Focus on vulnerable populations, including young children, pregnant women, and individuals with low income and social status, to understand the effects of air quality on their health.
- Understand the relationship between indoor microbiomes, airway microbiota, and asthma/allergy.

1.3 Ethics statement

The present data analysis study on children and indoor air pollution adheres to the highest ethical standards and prioritizes the well-being and rights of the participants involved. The ethical considerations were described in the grant agreement № 101057497.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

2. Materials and methods

In the following subsections, the cohorts and the methods for data and statistical analysis will be presented.

2.1 ATOPICA cohort

The ATOPICA project, "Atopic diseases in changing climate, land use, and air quality," investigates how factors such as climate change, the spread of invasive plant species, and land use affect the spatial and temporal distribution of inhalant allergens, particularly pollen. It aims to understand the impact of these changes on the prevalence of allergic diseases in Europe, focusing on sensitive populations. The project also examines the exposure to inhalant allergens in both outdoor and indoor environments, as well as the interaction between pollen and air pollutants and their influence on allergenicity. Finally, it explores the effects of climate change on pollen density and allergenicity, considering both the current situation and future projections. Children's Hospital Zagreb (SCH) conducted a 36-month investigation involving "case studies" in the paediatric population, encompassing three pollen seasons. The focus was on the plant species Ambrosia artemisiifolia L., commonly known as ragweed. 4015 children aged 4 to 9 years were recruited through kindergartens and primary schools. Exclusion criteria included the use of chronic immunosuppressants or antihistamine therapy. The cohort represents three areas with different concentrations of ragweed pollen: Slavonia region with high pollen concentration, Zagreb and surrounding areas with moderate pollen concentration and Dalmatia region with low pollen concentration area.

The data collection process encompassed several components. Firstly, skin prick tests (SPT) were conducted with parental consent using a standard panel of inhalant allergens: pollens (birch, hazel, pine, olive, and a five-species grasses mix), dog dander, cat dander, house dust mite (Dermatophagoides pteronyssinus; HDM), Parietaria, and Cladosporium. Additionally, basic information regarding the child, such as age, gender, and contact details, was recorded. If specific consent for blood sampling was obtained, peripheral blood IgE levels were measured. Information pertaining to the child's allergic symptoms from birth to the present was collected using the ISAAC phase II questionnaire. Furthermore, socio-economic

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

questionnaire data were gathered to provide additional contextual information for the study. Children sensitised to ragweed pollen also complete pollen diaries and measure indoor pollen concentrations and air pollutants. The initial goal of the project was to identify de novo allergies to ragweed during 2012, 2013 and 2014 pollen seasons.

1. Atopica [Internet]. [cited 2023 May 4]. Available from: https://www.bolnica-srebrnjak.hr/index.php/hr/znanost/192-znanstveno-istrazivacki-odjel/atopica/250-atopica
2. ATOPICA – Atopic diseases in changing climate, land use & air quality — English [Internet]. [cited 2023 May 4]. Available from: https://climate-adapt.eea.europa.eu/en/metadata/projects/atopica-2013-atopic-diseases-in-changing-climate-land-use-air-quality

In this study the authors aim at analysing relationships and associations which were not included in the original findings but concern IAQ. The cohort data set is enriched with determinants directly and indirectly affecting IAQ. The presence of such is tested against the observed symptoms and diseases.

2.2 COPSAC 2000 cohort

COPSAC stands for the Copenhagen prospective study on asthma in childhood. The COPSAC2000 cohort was established to study the relationships between genetic, environmental, and lifestyle factors in the development of asthma and related atopic diseases in children. Families were recruited in the third trimester of pregnancy from the greater Copenhagen area, inclusion criteria were maternal doctor diagnoses asthma and a threshold for treatment for asthma. During 1998-2001 information about the study was sent to pregnant women and eligibility was determined in phone interviews, and subsequent visits to the clinical research unit for informed consent. Offspring was included in the study at age 1 month. Of 1,476 pregnant women reporting a history of asthma and interest in the study, 798 (54%) were asthmatic after a telephone interview. Of these 798 women, 548 visited the CRU to receive detailed information, and mothers of 452 infants consented to participation. Before enrolment, 2 infants were excluded owing to congenital diseases (heart disease and microcephalia), and 39 infants failed to attend the enrolment visit 1 month after birth. Thus, 411 infants of 394 mothers (9 pairs of twins and 8 siblings)

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

were enrolled in the cohort [6]. Compared to the background population, mothers of the cohort were less likely to have given natural childbirth. The fathers were younger, and the households were slightly less affluent, with fewer children and fewer pets. The families live in greater Copenhagen area mainly in urban areas with different exposers from their home, school and outdoors. Children were followed in the research unit and all asthma and related diagnoses were set by research doctors at the research unit. The cohort is followed for planned visits throughout childhood in the clinical research to the final visit at age 18, as well as at all acute episodes of respiratory symptoms. Demographic information is obtained from parents at interviews and stored in the research database. The cohort primarily focuses on exposure from their home, inflammation markers and diagnosis of the infants in the first years of life and how it influences their health later in life. Below we detail the outcome data, such as the clinical measurements and diagnosis, as well as the exposure data.

2.2.1 Clinical measurements

Children were followed in the research unit for planned visits during childhood and at all acute episodes of respiratory symptoms. In the first 3 years of life parents were given an extensive educational session to keep a daily diary of respiratory symptoms, with emphasis on lower respiratory troubling symptoms. Asthma, eczema, and allergic outcomes were diagnosed prospectively in the research clinic by trained physicians and according to predefined algorithms. In short, asthma was diagnosed after 1) 5 episodes (defined as 3 consecutive days of lower respiratory symptoms) within 6 months, 2) rescue use of beta-2-agonist, 3) response to a 3-month trial of inhaled corticosteroids followed by relapse at the end of treatment. Eczema was defined according to the Hanifin Rajka criteria. Allergy and sensitization was determined repeatedly with skin prick test and serum IgE at multiple ages. Lung functions are measured repeatedly with whole body plethysmography until the children reach the ability to perform spirometry around age 6. The clinical measurements

[6] H. Bisgaard, 'The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study', Ann. Allergy Asthma Immunol. Off. Publ. Am. Coll. Allergy Asthma Immunol., vol. 93, no. 4, pp. 381–389, Oct. 2004, doi: 10.1016/S1081-1206(10)61398-1

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

and timestamps can be found in the reference[7], with the other data layers stemming from samples of blood, and faeces, physical health (eyesight, nose, ear and throat examination in great detail), history of medicine use and symptoms, lung function: Spirometer, FeNO; also being stored in RegionH biobank [8]. Omics data are given there as well [9].

2.2.2 Environmental exposures

Information regarding the children's environment, such as information about home exposures and details about school and day-care institutions are collected prospectively and stored in the research database. These are defined by their environment their home, school, and day-care. The thing which is monitored is the diet, exposure to passive smoking, allergies, indoor air pollution, dust, and exposures to the mom during pregnancy etc [10].

2.3 COPSAC 2010 cohort

Based on the experience from the COPSAC2000 cohort, the COPSAC2010 cohort was established in 2008, with the main aim to recruit families from the general population and perform pregnancy intervention, factorial design with either high-dose vitamin D or fish oil supplementation in third trimester with the aim of preventing offspring asthma.
Pregnant women from eastern Denmark were identified from the monthly reimbursement to general practitioners for the mandatory pregnancy visit were invited during 2008-2010. 54358 invitations were sent, 1876 (3.5%) contacted the clinic and 738 women were included in the study. 43 children were withdrawn before birth, and 700 children (5 pairs of twins) were included in the COPSAC2010 cohort [11]. Compared to the background population, women included in the COPSAC2010 cohort had more self-reported asthma, allergy and hay fever, and families were more affluent than the background population. Compared to the COPSAC2000 cohort, the families live in more diverse areas, as the catchment area was entire Zealand. Similarly, to the COPSAC cohort, the children are followed meticulously from

[7] https://copsac.com/home/copsac-cohorts/copsac2000-clinic/
[8] https://copsac.com/home/copsac-cohorts/copsac2000-biobank/
[9] https://copsac.com/home/copsac-cohorts/copsac2000-omics/
[10] https://copsac.com/home/copsac-cohorts/copsac2000-environmental-exposures/
[11] H. Bisgaard et al., 'Deep phenotyping of the unselected COPSAC2010 birth cohort study', Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol., vol. 43, no. 12, pp. 1384–1394, Dec. 2013, doi: 10.1111/cea.12213.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

birth with multiple planned visits especially in infancy and early childhood and visits at acute illness and symptoms related to the main disease's studied; asthma, allergies and eczema. With the increased emphasis on microbiome in the 2010 cohort, multiple microbiome samples have been collected throughout childhood including house dust, which may serve as a marker for indoor air quality.

2.3.1 Clinical measurements

Like the COPSAC2000 cohort children are exclusively diagnosed with asthma, allergy and eczema in the clinical research unit according to the same strict predefined algorithms detailed above. Parents are instructed to fill a daily symptom diary, and in the 2010 cohort, the daily diary includes more symptoms, such as fever and gastrointestinal symptoms, which were not recorded in the 2000 cohort. Immune functioning is investigated prospectively with cytokine profiling in the mucosal lining in infancy, at 18 months with whole blood phenotyping of immune profiles from various ligand stimulations. Neurodevelopment and cognition are additional outcomes in the 2010 cohort with prospective data collection on milestones, psychopathology and a large scale neurodevelopment evaluation at age 10 including MRI scans of the brains [12].

2.3.2 Environmental exposures

Child home exposures to land cover is based on address and extrapolation of land cover classes from the European Corine satellite images. Child environment in house, institution and school are collected prospectively from parents and stored in database, including information on smoking exposure, fireplace and house pets. House dust collected from the child's bed at age 6 months with a vacuum filter attachment were microbiome profiled with 16S RNA sequencing.

[12] P. Mohammadzadeh et al., 'Effects of prenatal nutrient supplementation and early life exposures on neurodevelopment at age 10: a randomised controlled trial - the COPSYCH study protocol', BMJ Open, vol. 12, no. 2, p. e047706, Feb. 2022, doi: 10.1136/bmjopen-2020-047706.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

2.4 Statistical analysis of cohort data

2.4.1 Machine Learning (ML)

ML involves training algorithms which learn from the data, identify patterns, and finally make predictions on desired target variable. Before training ML models, one defines usually train and validation sets or does cross-validation. Models are commonly analysed using certain metrics (error, accuracy etc.) which tell in which direction algorithm should further train until achieving best metrics. If evaluation metrics of prediction are satisfying, a trained model could be used for predicting outcomes on unseen data.

2.4.2 Random Forest (RF)

The RF algorithm, conceptualized by Breiman [13], creates a large collection of decorrelated decision trees using bootstrapping aggregation. The final prediction results are thereby averaged from an ensemble of weak learners, i.e. decision tree regressors or classifiers, with the ensemble reducing the bias in the models. The variance can be controlled by carefully optimizing weak learner hyperparameters, such as tree depth. Besides their good performance, RF and other decision tree-based learners accept many feature representations and are associated with reduced pre-processing efforts, making them convenient for use in many applications. When compared with boosting ensembles, RF has a significant advantage because trees get trained in parallel.

2.4.3 Logistic Regression (LR)

Logistic Regression[14] calculates the weighted sum from input variables which are transformed with sigmoid function to predict logarithmic odds of e.g. a binary target variable. Thanks to logarithmic transformation the LR algorithm captures a non-linear relationship between input and target variables. Calculations are fast and interpretation of the results is easy since the coefficients associated with each input variable provide insights into the direction and magnitude of their influence on the prediction. The LR algorithm can be used for one- or multi-class classification problems.

[13] L. Breiman, 'Random Forests', Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[14] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. in Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

2.4.4 Chi-Square Test

The chi-square test is used to determine the association between discrete/categorical variables, respectively if there is a significant difference between the observed and expected frequencies. The expected frequencies are calculated based on the null hypothesis which claims there is no connection between the variables. To determine if the difference is statistically significant, the calculated chi-square value is compared with the critical value. If the calculated chi-square value is greater, null hypothesis can be rejected and conclusion is that there is a connection between variables.

2.4.5 T-Test

The t-test is used to compare the means of two groups and determine if there is a significant difference between them, respectively if the difference in means is larger than expected resulting from random sampling variability. The t-test assumes normal distribution in the data and that group variances are equal. The calculated t-value is compared with the critical value to check if the difference is statistically significant.

2.4.6 Kolmogorov-Smirnov Test(ks_test)

Compared the distance between cumulative probability function between two samples which gives a statistical indication if the two distributions have a similar shape. The null hypothesis is that the two-distribution come from the same underlying distribution.

2.4.7 Cat boosting

CatBoost[15] is a machine learning algorithm that is specifically designed for gradient boosting on decision trees. It is known for its ability to handle categorical features in a highly efficient manner. CatBoost employs a combination of ordered boosting, customized learning rates, and advanced feature discretization techniques to achieve high accuracy with minimal hyperparameter tuning. It is suitable for a wide range of tasks, including classification, regression, and ranking. CatBoost is known for its robustness, scalability, and ability to handle large datasets, making it a popular choice in various applications.

[15] https://arxiv.org/abs/1810.11363

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

2.4.8 Weighted KNN (k-nearest neighbours)

Used the distance in the N-dimensional space to impute the values under the assumption that the point lying closer to you should have similar values. Weighted kNN is just the case where each Neighbour distance is the weight for the values of the point.

2.4.9 Mann-Whitney U test

The Mann-Whitney U test is usually used to test if there is a significant difference between the groups where the data is not normally distributed or assumptions for the t-test are not met. The data from both groups is combined and ranked in an order, without considering which group the data point comes from. Data points with the same value get the average of ranks assigned. The U value results from comparison of the sum of ranks for one group to the sum of ranks from the other group. To determine if the difference is statistically significant, the calculated U value is compared with the critical value.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

3. Results

3.1 Previous findings from the COPSAC 2000 cohort

3.1.1 Long-term exposure to indoor air pollution and wheezing symptoms in infants

Raaschou-Nielsen et al [16]nvestigated the relationship between long-term exposure to air pollution and symptoms of wheezing in the COPSAC2000 cohort. The main outcome of the study was episodes of wheeze defined from minimum 3 consecutive days of troublesome airway symptoms recorded in the daily diary. Air quality in the infant's bedroom measured $PM_{2.5}$, NO_x, NO_2, formaldehyde and black smoke. $PM_{2.5}$ and black smoke was averaged over a week; NO_x, NO_2 and formaldehyde were averaged over a month. All measurements were compiled at ages 6, 12 and 18 months. The concentration of $PM_{2.5}$, black smoke, and NO_x differed by season, and was transformed by the ratio between the means of the months of measurement with the mean of December. After the transformation, logistic regression between the particle concentrations to overall wheezing(yes/no) and a standard linear regression with the end-point number of wheezy episodes was applied to estimate the signal between concentrations and symptoms. There were no associations between air pollution concentrations and the number of wheezy episodes nor any wheezing. In conclusion, majority of children were exposed to low concentrations of the particles, and previous studies showing short term associations between higher concentrations and wheezing, could not be translated to associations of cumulative low-dose exposures to air pollution in a high-risk asthma cohort.

3.1.2 Predictors of indoor fine particulate matter in Infants' Bedrooms in Denmark

Raaschou-Nielsen et al investigated the main contributors to indoor $PM_{2.5}$ and black carbon. [17] Based on a questionnaire detailing the actions of the family while the air quality measurements were ongoing, the main contributors to air quality were investigated. Data on frequency of smoking; use of fireplace; household cleaning such as dusting, and

[16] O. Raaschou-Nielsen et al., 'Long-term exposure to indoor air pollution and wheezing symptoms in infants', Indoor Air, vol. 20, no. 2, pp. 159–167, 2010, doi: 10.1111/j.1600-0668.2009.00635.x.

[17] O. Raaschou-Nielsen et al., 'Predictors of indoor fine particulate matter in infants' bedrooms in Denmark', Environ. Res., vol. 111, no. 1, pp. 87–93, Jan. 2011, doi: 10.1016/j.envres.2010.10.007.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

vacuuming; airing out (opened windows); and types of cooking, e.g., oven, gas stove and cooker hood. In a mixed model, estimating log transformed $PM_{2.5}$ and black carbon each contributor was investigated. All sources showed linear association except gas stoves and smoking which were log linear. The main contributors for $PM_{2.5}$ were smoking (2.8 times higher $PM_{2.5}$ for smoking households); heavy traffic (1.77 times higher); winter (1.4 times higher than summer); but also vacuum cleaning and frying were significant contributors. For black smoke, the contributors were smoking (1.9 times higher); candles and fireplace use; season and urbanicity.

3.1.3 Childhood Asthma after Bacterial Colonization of the Airway in Neonates

In the paper by Bisgaard et al [18] an association between neonatal airway colonization with common pathogens is established to later development of asthma. Unique for the COPSAC2000 cohort all children underwent full anaesthesia at the age of 4 weeks for a lung function measurement (squeeze jacket). During the sedation children were sampled with a soft suction in the hypopharynx. Samples were cultivated and a surprising number of 21% of the healthy neonates were colonized with pathogenic bacteria: *S. pneumoniae*, *M. Catarrhalis*, or *H. influenzae.* Colonization was associated with an increased risk of persistent wheeze, acute severe exacerbation of wheeze and hospitalization due to wheeze. Interestingly colonization at age 12 months showed no significant association with any wheezing symptoms, emphasizing the early programming of asthma.

3.1.4 Associations of pre- and postnatal exposures with optic nerve status in young adults

In Zhu et al[19] the long term follow up of the COPSAC2000 cohort is utilized to investigate the long term effects of early life smoking and $PM_{2.5}$ exposure to optic nerve status at the age of 18. At the 18-year visit COPSAC2000 participants underwent eye examination, where peripapillary retinal nerve fibre layer thickness were measured. Primarily one eye was investigated per participant; in total 216 right eyes and 53 left eyes in a total of 269

[18] H. Bisgaard et al., 'Childhood asthma after bacterial colonization of the airway in neonates', N. Engl. J. Med., vol. 357, no. 15, pp. 1487–1495, Oct. 2007, doi: 10.1056/NEJMoa052632.

[19] L. Zhu et al., 'Associations of pre- and postnatal exposures with optic nerve status in young adults', Acta Ophthalmol. (Copenh.), Mar. 2023, doi: 10.1111/aos.15657.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

participants were investigated. Predictors for retinal nerve fibre layer thickness were birth weight and maternal smoking during pregnancy, and indoor $PM_{2.5}$. The association to indoor $PM_{2.5}$ did not survive multiple adjustment for indoor smoking.

3.2 Previous findings from the COPSAC 2010 cohort

The hallmark paper on neonatal colonization [20] with common bacterial pathogens in New England Journal of Medicine was a key driver for the establishment of the COPSAC2010 cohort, with an increased focus on early life microbiome and other prenatal and early life predictors for childhood asthma.

3.2.1 Environmental Shaping of the Bacterial and fungal community in infant bed dust

Gupta et al.[21] investigated the associations of infant's bed dust with airway microbiome in the COPSAC2010 cohort. In total 542 children with both dust samples from the infant's beds at age 6 month and airway sample at age 3 year were investigated. Both dust and airway samples were 16S sequenced to a depth of minimum 3000 reads. Correlations between fungal and bacterial communities in infants' bed dust versus their airway microbiome and the correlation in alpha and beta diversity between microbiota-microbiota, fungal-fungal but also fungal-microbiota were investigated. Multiple contributing factors, including the number of siblings, household furred animal, type of home, and season were considered. Potential transfer of the microbiome from bed dust to the airway for individual taxa was also investigated.

3.2.2 Correlation between fungal and bacterial

Spearman correlation between bacterial and fungal alpha diversity (diversity within one community) was 0.17 for the observed taxa and 0.2 for the Shannon index. At genus level for 173 samples for fungal-fungal correlation where most of the significant correlation was positive except for four genera with negative correlation. The number of significant

[20] H. Bisgaard et al., 'Childhood asthma after bacterial colonization of the airway in neonates', N. Engl. J. Med., vol. 357, no. 15, pp. 1487–1495, Oct. 2007, doi: 10.1056/NEJMoa052632.

[21] S. Gupta et al., 'Environmental shaping of the bacterial and fungal community in infant bed dust and correlations with the airway microbiota', Microbiome, vol. 8, no. 1, p. 115, Aug. 2020, doi: 10.1186/s40168-020-00895-w.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

bacteria-bacteria correlations was higher than fungal-fungal and with a negative range [-0.12 to -0.28] versus the positive range [0.15 to 0.72].

3.2.3 Influence of pets on bed dust microbiome

The bacterial alpha diversity was not significantly associated with the presence of cats or dogs, but co-presence of cats and dogs was significantly associated with bacterial alpha diversity Kruskal Wallis test (P_observed=0.0039 and P_shannon 0.066). Testing for fungal alpha diversity, there was no influence by dog or cat (P_observed=0.52, P_shannon=0.92). A small but significant effect was found looking at the beta diversity of the bacterial composition when they looked at families with and without cats and dogs. There was no difference in fungal beta diversity by pets.

3.2.4 Rural vs urban environment impact on bed dust microbiome

Urbanicity, defined from satellite images of surrounding land cover of the birth address, was associated with bed dust microbiome. 353 more genera were present in a rural environment compared to urban areas. There were no seasonal effects on bed dust fungal or microbial diversity.

3.2.5 Correlation between the bed dust and airway microbiota.

There was a higher alpha diversity in the bed dust compared to the infant's airways. There was no significant correlation between bed dust and airway microbiome, and similar values were found when bed dust and airways were assigned randomly. This indicates that airway and bed dust microbiomes do not interact. A significant positive correlation between some specific genera from the samples was found which could indicate that some of the bacteria in the airways and bed have similar environmental profiles.

3.2.6 Fungi and the bacteria in the beds of rural and urban infants correlate with later risk of atopic diseases. (Not published)

Based on Gupta et al [18] and a study by Lehtimäki et al [22] where urbanicity is associated with later risk of asthma, allergic rhinitis and aeroallergen sensitization through an altered gut

[22] J. Lehtimäki et al., 'Urbanized microbiota in infants, immune constitution, and later risk of atopic diseases', J. Allergy Clin. Immunol., vol. 148, no. 1, pp. 234–243, Jul. 2021, doi: 10.1016/j.jaci.2020.12.621.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

microbiome profile, a new study (unpublished) investigates the long term clinical effect of the bed dust microbiome, with a special focus on the urbanised effect in bed dust microbiome.

At age 6, 7.7% of the children had asthma, 7.0% had allergic rhinitis, 7.7% had eczema ongoing at age six and lastly, 23.5% were sensitized to aeroallergens in either skin prick test or specific IgE. Fungal diversity in bed dust from child's bed at 6 months, was associated with lower risk of asthma and allergic rhinitis at age 6, whereas risk of eczema was increased. Bacterial diversity was associated with higher risk of allergic rhinitis. There were no effects to aeroallergen sensitization at age six. When looking at a combination of both asthma and allergic rhinitis the children had both very low fungal and bacterial richness and diversity p=0.004. Children who had asthma or rhinitis with eczema had higher bacterial richness than children with asthma or rhinitis alone. Including farm living children did not alter the conclusions. Investigating specific genera abundances by disease status at age 6, several were found to be lower in children with atopic diseases. Asthma and allergic rhinitis overlapped in general whereas aeroallergen sensitization had independent. However, after FDR correction only a few specific genera remained statistically significant.

3.2.7 Bed dust microbiota protective effect of rural environment.

To study the urban vs rural environment, bed dust urbanicity microbiome scores (fungal and bacterial) were generated and associated to diseases at age 6. There were no associations between fungal environmental score with any disease, but bacterial environmental score was associated to allergic rhinitis (p=0.04) and borderline associated with asthma (p=0.06), in the direction where a rural environment imprinted bed dust microbiome had a protective effect against disease.

3.3 Previous findings from the ATOPICA cohort

There were three published scientific papers from the ATOPICA cohort reflecting the relationships of the development of atopic allergies and environmental determinants.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

Researchers in this consortium[23] explored the environmental and lifestyle factors that may contribute to common ragweed allergy and disease in children. The study is a case-control study, which means that it compares the characteristics of children with ragweed allergy or disease (cases) to those without (controls). The authors found that exposure to tobacco smoke, low levels of physical activity, and poor diet were associated with an increased risk of ragweed allergy and disease in children. The study highlights the importance of modifying these risk factors to prevent and manage ragweed allergy and disease in children. The researchers from SCH furthermore[24] investigated the prevalence of sensitization to the common ragweed (Ambrosia artemisiifolia) allergen in children from different regions of Croatia. The study is based on skin prick tests performed on children with suspected allergies, and the results show that the prevalence of sensitization to ragweed is higher in coastal regions compared to inland regions of Croatia. The authors suggest that this may be due to higher levels of ragweed pollen in coastal areas, as well as differences in environmental and lifestyle factors. The study highlights the need for targeted prevention and management strategies for ragweed allergy and disease based on regional differences in allergen sensitization. Furthermore, the researchers examined the factors that contribute to the higher prevalence of ragweed allergy and disease in coastal regions of Croatia[25], as found in a previous study. The authors hypothesize that differences in environmental and lifestyle factors, such as air pollution and diet, may also play a role in the higher prevalence of ragweed sensitization in coastal areas. The study is based on skin prick tests performed on children with suspected allergies, and the results support the authors' hypothesis that other factors besides pollen concentration contribute to regional differences in ragweed

[23] M. Agnew et al., 'Modifiable Risk Factors for Common Ragweed (Ambrosia artemisiifolia) Allergy and Disease in Children: A Case-Control Study', Int. J. Environ. Res. Public. Health, vol. 15, no. 7, p. 1339, Jul. 2018, doi: 10.3390/ijerph15071339.

[24] I. Banic et al., 'Regional differences in sensitization to Ambrosia in Croatian children', Eur. Respir. J., vol. 44, no. Suppl 58, Sep. 2014, Accessed: May 13, 2023. [Online]. Available: https://erj.ersjournals.com/content/44/Suppl_58/P1190

[25] I. Banic et al., 'Regional differences in sensitization to Ambrosia in Croatian children', Eur. Respir. J., vol. 44, no. Suppl 58, Sep. 2014, Accessed: May 13, 2023. [Online]. Available: https://erj.ersjournals.com/content/44/Suppl_58/P1190

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

sensitization. The study underscores the need for further research to identify and address modifiable risk factors for ragweed allergy and disease in children.

3.4 EDIAQI findings from joint COPSAC cohorts

In the following section the two COPSAC cohorts are combined. Although the cohorts are different, not least regarding selection, the purpose is to use the large databases to extrapolate and predict air quality data using machine learning models, to increase to statistical power to investigate IAQ to the well-established clinical data in both cohorts. The particles used in the report are $PM_{2.5}$, NO_2, formaldehyde and black smoke (black carbon). These were chosen due to extrapolation values for our boosting algorithm being above 0.5 R^2. The findings from this joint cohort are still under investigation, and presented results are preliminary and unpublished.

3.4.1 Introduction

Understanding air pollution's influence on health from birth to the second year of life is a complicated task, there are multiple things to consider social economic status, genetics and microbiome. In this section, a very simplistic approach to all the subjects is introduced using WHO air quality index to separate the children into groups based on the concentration levels of different pollutants. The pollutants included are $PM_{2.5}$, NO_2, formaldehyde and black smoke. The concentration of chosen pollutants was only measured in the 2000 cohort where only 390 bedrooms had been measured. In order to increase the number of samples the weighted kNN was chosen with five neighbours to impute missing values besides missing the air pollution concentration under the assumption that children who have similar feature profiles should be closely related in terms of pollution indoors. To do the extrapolation for the air pollution concentration cat boosting was chosen which is a gradient boosting algorithm optimized towards larger trees between 2-10 maximum depth of the regression trees since the data is quite complex. A successful extrapolation means similarity in patterns between the cohorts, so 2000 vs combined. A lower average concentration is expected from the 2010 cohort due to cohort effects and time, less smoking, gas stove and fireplaces etc. This can also be observed in the prediction of the 2010 cohorts $PM_{2.5}$, NO_2, formaldehyde and BC. This also means that the distribution of the cohorts should have

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

different distributions which can be observed in the two-samples t and Kolmogorov-Smirnov test in the following figure below.

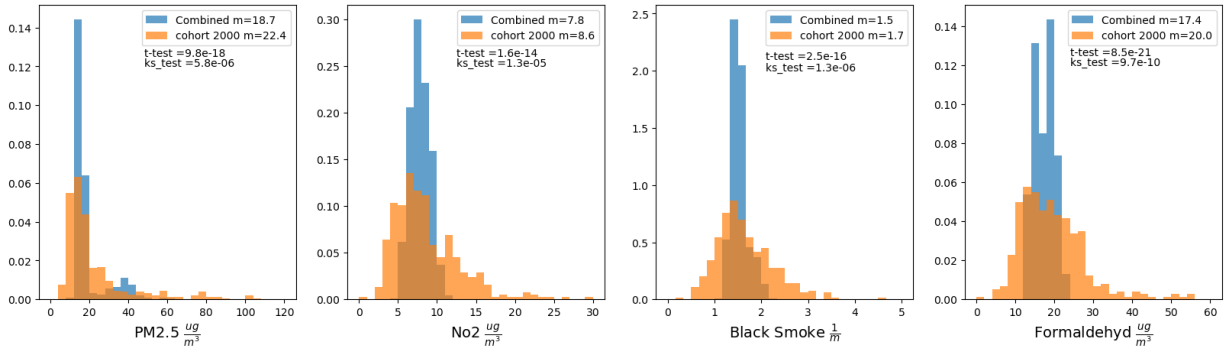

Figure 1 Kolmogorov-Smirnov test results and histograms of the cohort distributions regarding analysed pollutants PM2.5 , NO2, formaldehyde and BC.

To compare the concentrations to the international standard a plot showing the cohort's concentration plotted into WHO targets five target indexes with an addon of a 6-index target 5 and above for $PM_{2.5}$, and NO_2 into WHO interim targets[26] and put formaldehyde into the Canadian index[27] come from Canadian health at the workplace reports since it was the only one exiting for that kind of particle. The indices are presented in Table 1 and Figure 2.

Table 1 Categorization of Air Quality Levels based on Index Values for PM2.5, NO2, and Formaldehyde

source	Index_ $PM_{2.5}$	Index_ NO_2	Index_formaldehyd
1) Good	0-5	0-10	0-10
2) fair	5-10	10-20	10-20
3) moderate	10-15	20-30	20-40
4) poor	15-25	30-40	40-60
5) Very poor	25-35	40-80	60-80
6) Extremely poor	35-200	80-200	80-200

The combined index is calculated by means of following equations:

$$Index = int\left(\frac{\left(Index_{PM25} + Index_{no2} + Index_{formaldehyd}\right)}{3}\right)$$

Which should include the infant's worst air quality for all the particles.

[26] World Health Organization. (2021). WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization. https://apps.who.int/iris/handle/10665/345329. License: CC BY-NC-SA 3.0 IGO

[27] https://www.canada.ca/en/health-canada/services/environmental-workplace-health/reports-publications/air-quality/formaldehyde-indoor-air-environment-workplace-health.html

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

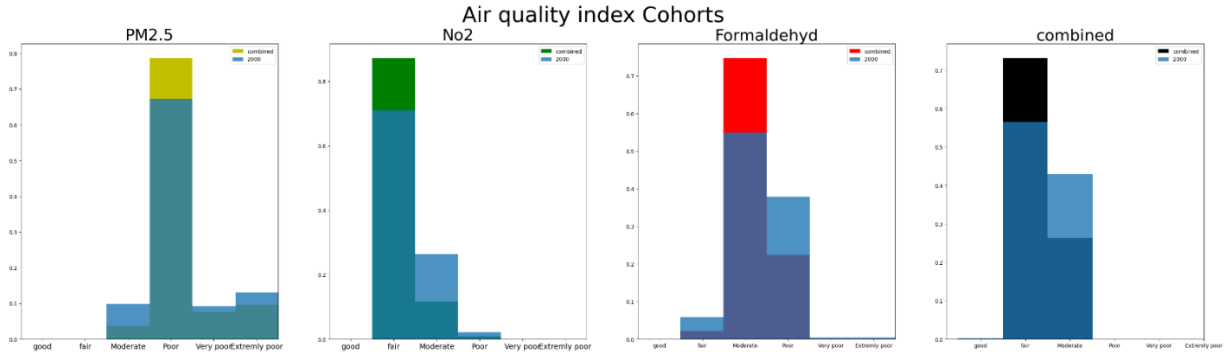

Figure 2 Distribution of the of the bedrooms air quality index

3.4.2 Data collection

The indoor sources, sinks and delayers are gas-stoves, fireplaces, indoor smoking, carpets cooker hoods which were estimated with a questionnaire about how many days a year they had it in their main house. We also asked the parents about household income, and total education a value from 2 to 10, two means both parents have finished school 10 means both have master's degrees. House categories were taken from the Danish building registry (BBR) with the use of the families addresses it returned: floor level, age of house, indoor house area, number of rooms and number of bathrooms. Based on the addresses geopandas[28] were used to estimate the coordinates of the houses to estimate the outdoor sources. The outdoor sources were estimated by using open-source map python package[29] which returned the area of the sources near the family's house in three circular areas with radius r1, r2 and r3 (Figure 3). The outdoor sources were measured indirectly as roads(cars), highways(cars), coastlines(ships), power plants, forests (sink effect), farmland (agriculture) and airports(planes). The three circle areas around the houses were used to estimate the area density of the sources. Then the total area density for each house the following way was calculated using following equation:

$$A_{source} = \frac{A_{bs}}{A_b} + \frac{A_{ys-}A_{bs}}{A_y} + \frac{A_{gs-}A_{gs}}{A_g}$$

A_{bs} =Source within the Blue circle

A_{ys} =Source within the yellow circle

A_{gs} =Source within the green circle

[28] https://geopandas.org/en/stable/

[29] https://wiki.openstreetmap.org/wiki/OSMPythonTools

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

A_b =Blue circular area

A_y =Yellow circular area

A_g =Green circular area

The plots show an example of the three circles with forest, roads and powerplants.

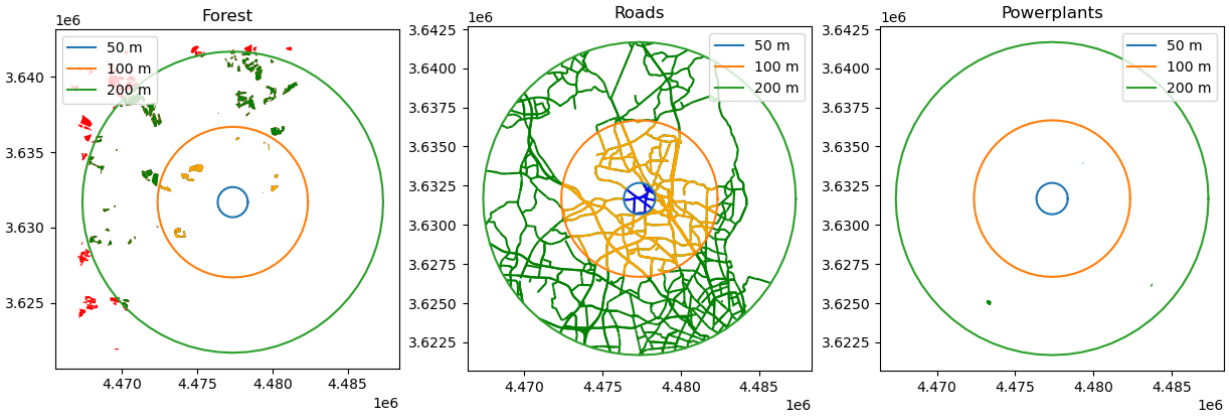

Figure 3 Example plots of the three circles with forest, roads and powerplants.

The three values for r based were estimated using prior papers which can be seen in Table 2.

Table 2 Proximity of Different Landmark in the models

source	Blue circle (m)	Yellow circle (m)	Green circle (m)
Forrest	50	100	200
Roads	50	100	200
Highway	50	100	200
Airport	1000	5000	10000
Coastline	50	100	200
Power plant	1000	5000	10000
Train track	50	100	200
Farmland	50	100	200

3.4.3 Computational methodology

A weighted KNN with N=5 was used to impute missing values resulting in a total of 411 infants in the 2000 cohort and 700 in the 2010 cohort. The 2000 cohort was used to train the cat boosting algorithm using cross-validation $PM_{2.5}$ had R^2=0.44, NO_2 R^2= 0.23 and both formaldehyde and black smoke had an R^2 of 0.15 estimate $PM_{2.5}$ well and the three others were better than random but not great prediction in general. The following hyper parameters was used.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 3 Performance Metrics for Particle Estimation Models with Different Features

particles	estimators	Max depth	Learning rate	subsample	Max features	R^2
$PM_{2.5}$	360	4	0.03	0.6	0.6	0.44±0.11
NO_2	360	7	0.03	0.5	0.5	0.21±0.04
Formaldehyde	360	7	0.015	0.5	0.5	0.15±0.04
blackening	360	4	0.015	0.7	0.7	0.15±0.04

The model was expected to work worse on the new cohort due to the year's gap between the cohorts, the education and social class differences and that 2000 was a high-risk cohort but the main sources are still comparable. The cohorts were then combined into a cohort of 1111 kids and look at them together but also separately to see if there were any significant signals in the data after extrapolation. The 2000 cohort was compared to the combined cohort since the 2000 cohort mainly includes direct measurements related to air pollution.

3.4.4 Rural vs urban area cohort 2000

One interesting aspect of this analysis is the urban vs rural component. The analysis was conducted by analysing distributions against these types of location (Figure 4).

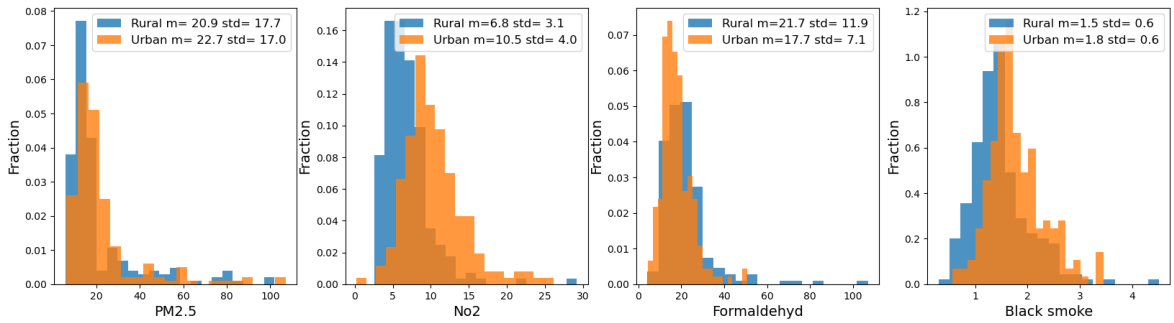

Figure 4 Urban vs rural areas in Denmark have different distribution for the particles

$PM_{2.5}$, NO_2 and black smoke are all related to traffic, smoking and gas-stove, this can also be observed in the distribution since Urban has a higher mean concentration. Formaldehyde is mostly associated with the age of the house the older the home the higher the concentration normally is, this can also be seen since houses in Rural areas tense to be older hence higher mean formaldehyde concentration.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

3.4.5 Results

To estimate if children with worse air quality in the 2000 cohort and the combined cohort have been influencing the infant's health, five clinical measures have been collected:

- the nasal cytokines and hence the immune responses
- the alpha diversity of bacteria in the airways
- the fev1 and FVC related to lung function
- several metabolites

3.4.6 Nasal cytokines

After reading multiple papers three cytokines related to inflammation was chosen il_1beta, Il_6 and TNF-alpha. Due to an extreme number of values, six tables are made including the means of the cytokines for different index values but also the standard deviation. There is a significant difference between the two indices if the distance between the means is more than two standard deviations from each other. All the cytokines are taken six months after the infant is born and are quite limited since the cytokines change quite fast it is only a snapshot of children's inflammation over time.

Table 4 TNF –alpha combined cohort analysis

	Index_ $PM_{2.5}$	Index_ NO_2	Index_formaldehyd	Index_Combined
1) Good	nan	2.8±0.9	2.1±0.6	nan
2) fair	2.5±0.7	2.4±0.7	2.7±0.9	2.8±0.9
3) moderate	2.7±0.9	2.5±0.9	2.7±1	2.6±0.9
4) poor	2.8±1	nan	2.5±0.5	2.7±0.7
5) Very poor	2.7±0.9	nan	2.13±0.04	nan
6) Extremely poor	2.5±0.9	nan	nan	nan

Table 5 Il_1-beta combined cohort analysis

	Index_ $PM_{2.5}$	Index_ NO_2	Index_formaldehyd	Index_Combined
1) Good	nan	0.04±0.12	0.2±0.5	nan
2) fair	0.1±0.3	0.1±0.2	0.04±0.1	0.04±0.1
3) moderate	0.04±0.1	0.2±0.5	0.04±0.1	0.05±0.1
4) poor	0.04±0.1	nan	0.08±0.09	0.1±0.2
5) Very poor	0.04±0.1	nan	0.06±0.08	nan
6) Extremely poor	0.06±0.1	nan	nan	nan

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 6 Il_6 combined cohort analysis

	Index_ $PM_{2.5}$	Index_ NO_2	Index_formaldehyd	Index_Combined
1) Good	nan	0.4±0.6	0.2±0.2	nan
2) fair	0.4±0.6	0.3±0.8	0.4±0.6	0.4±0.6
3) moderate	0.4±0.5	0.18±0.13	0.3±0.4	0.3±0.7
4) poor	0.4±0.7	Nan	0.4±0.4	0.3±0.3
5) Very poor	0.3±0.6	Nan	0.2±0.3	nan
6) Extremely poor	0.3±0.6	nan	nan	nan

Table 7 TNF –alpha COPSAC2000 cohort analysis

	Index_ $PM_{2.5}$	Index_ NO_2	Index_formaldehyd	Index Combined
1) Good	nan	2.5±0.9	2.1±0.6	nan
2) fair	2.5±0.7	2.4±0.8	2.5±0.9	2.5±0.9
3) moderate	2.4±0.8	2.5±0.9	2.5±0.9	2.5±0.9
4) poor	2.5±1	nan	2.5±0.5	2.7±0.7
5) Very poor	2.5±0.8	nan	2.13±0.04	nan
6) Extremely poor	2.5±0.9	nan	nan	nan

Table 8 Il_1-beta COPSAC2000 cohort analysis

	Index_ $PM_{2.5}$	Index_ NO_2	Index_formaldehyd	Index_Combined
1) Good	nan	0.1±0.2	0.2±0.5	nan
2) fair	0.1±0.3	0.1±0.3	0.1±0.2	0.1±0.3
3) moderate	0.1±0.3	0.2±0.5	0.1±0.1	0.1±0.2
4) poor	0.1±0.2	nan	0.008±0.009	0.1±0.2
5) Very poor	0.1±0.1	nan	0.006±0.009	nan
6) Extremely poor	0.1±0.2	nan	nan	nan

Table 9 Il_6 COPSAC2000 cohort analysis

	Index_ $PM_{2.5}$	Index_ NO_2	Index_formaldehyd	Index_Combined
1) Good	nan	0.4±0.5	0.2±0.2	nan
2) fair	0.4±0.6	0.3±0.3	0.4±0.5	0.4±0.6
3) moderate	0.4±0.6	0.18±0.13	0.3±0.5	0.3±0.3
4) poor	0.3±0.4	nan	0.3±0.4	0.3±0.3
5) Very poor	0.2±0.2	nan	0.2±0.3	nan
6) Extremely poor	0.3±0.3	Nan	nan	nan

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Looking at these tables one can see that no difference between means lies more than one sigma away hence only p values of p=0.36 which is not significant. This is for both 2000 cohort and the combined cohort which indicates that **air pollution is not one of the main contributors to these cytokines at this given timestamp.**

3.4.7 Alpha diversity

Alpha diversity is the number of specific species within a bacterial community and a healthy person has been **shown to have a more diverse bacterial community in general**. There are two measures of alpha diversity observed the total number of species and the Shannon index: $shannon = \sum p_i \ \ln \ \ p_I$ where p_i is the count of a specific bacteria divided by the total count of bacteria. This also means that the higher the observed taxa are the higher the alpha diversity and the lower the Shannon index is the higher the alpha diversity. Observed taxa are corrected with library size since a sample size gives an bacteria number of species by default. The alpha diversity is shown in Figure 5 against the air quality index for the 2000 and the combined cohort. Looking at Figure 5. two things are important; one 2000 cohort is mainly based on data which makes it more reliable and two looking for tendency does any of the plots show trends towards the same direction at all points. Observing **$PM_{2.5}$ and NO_2 for the combined cohort only NO_2 shows trend** but for the 2000 cohort, **NO_2 shows a clear trend the high concentration the lower the diversity**.

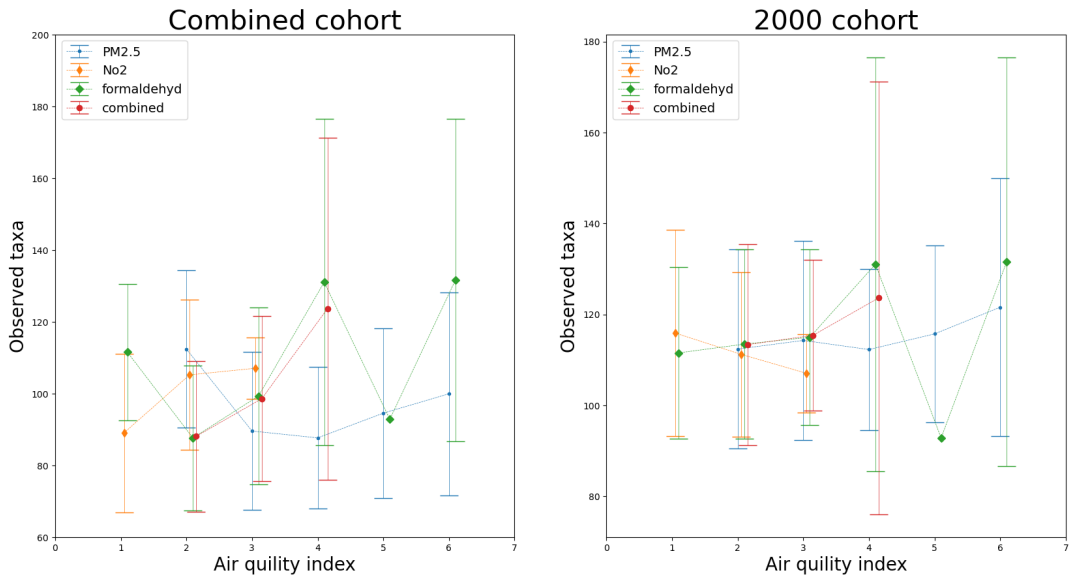

Figure 5 A comparison between the combined cohort and the 2000 cohort, examining the relationship between observed data and air quality index.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

The same data, however using the Shannon index are presented in Figure 6. Based on this one can observe that $PM_{2.5}$ show a strong trend for the 2000 cohort.

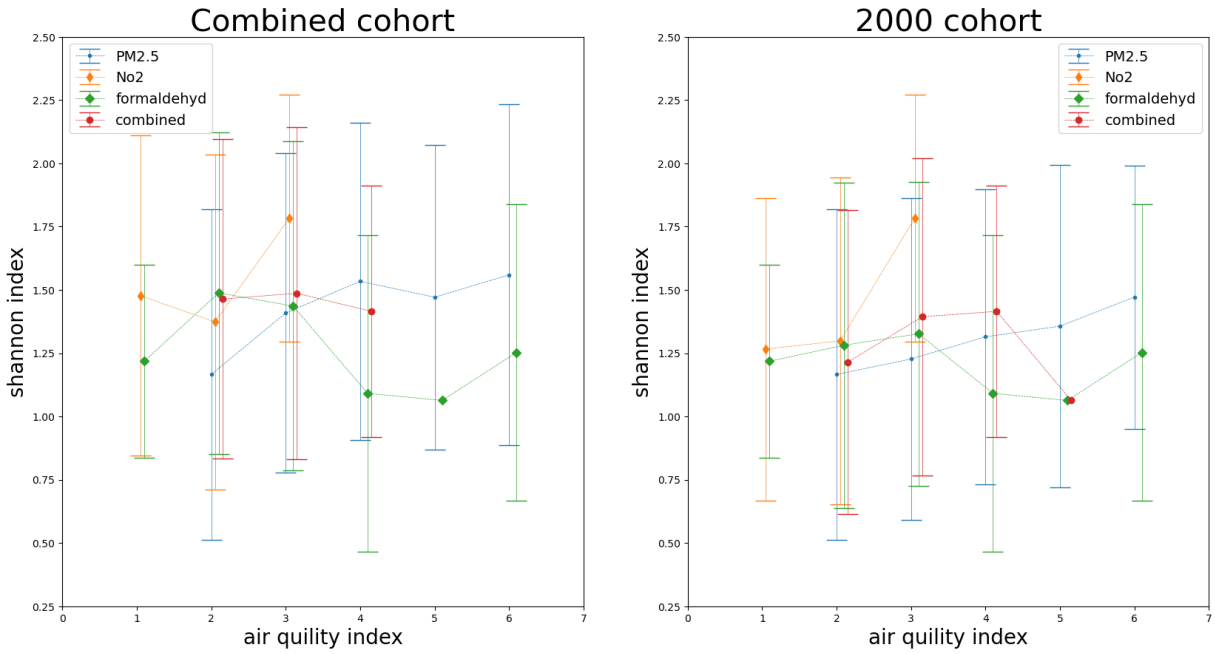

Figure 6 A comparison between the combined cohort and the 2000 cohort, examining the relationship between the Shannon index and air quality index.

3.4.8 Lung capacity

Multiple papers reported a decrease in FEV1 and FVC the maximum amount of air a person can inhale and exhale, hence such an analysis was chosen here. The values for exhaled air FEV1 are plotted against the air quality index in Figure 7 and 8. Observing the plots, it is clear that in both the combined cohort and in the 2000 cohort NO_2 and combined decrease the total inhaled amount of air there can be two explanations one NO_2 and combined do capture smokers the best or is a mixer of signal from air pollution together with smoking. For the inhaled air FVC a clear signal in NO_2 is still there but only for the combined cohort so the conclusion should be taken lightly.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

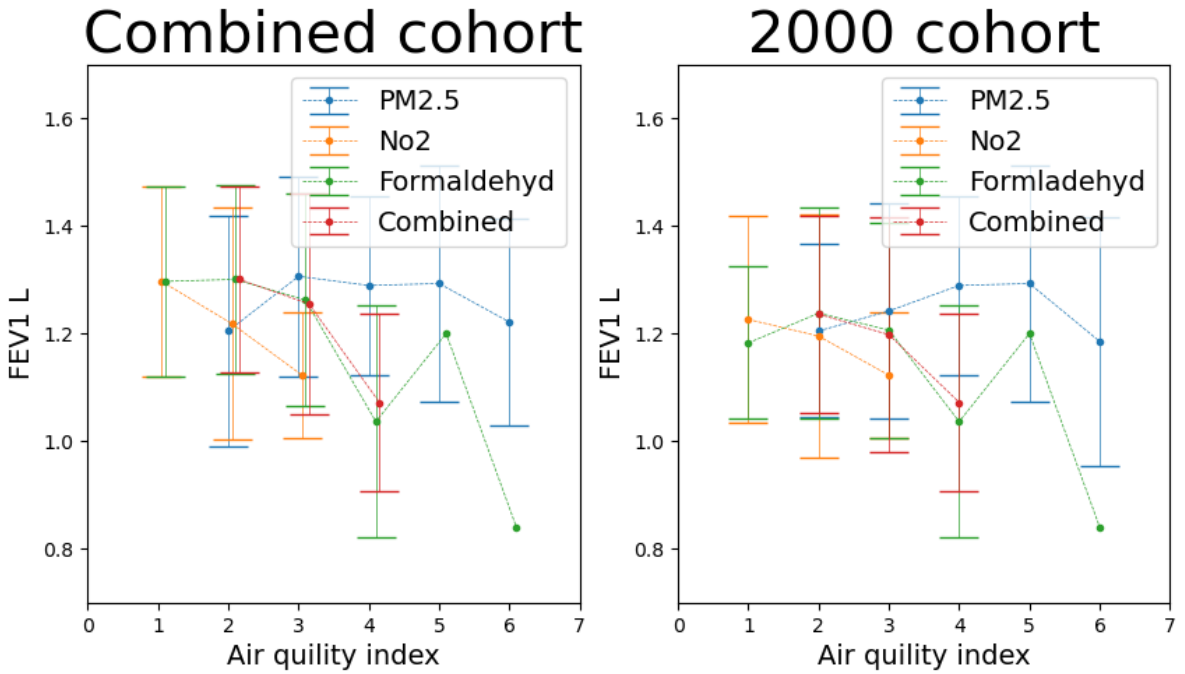

Figure 7 A comparison of the cohorts for NO2 and formaldehyde

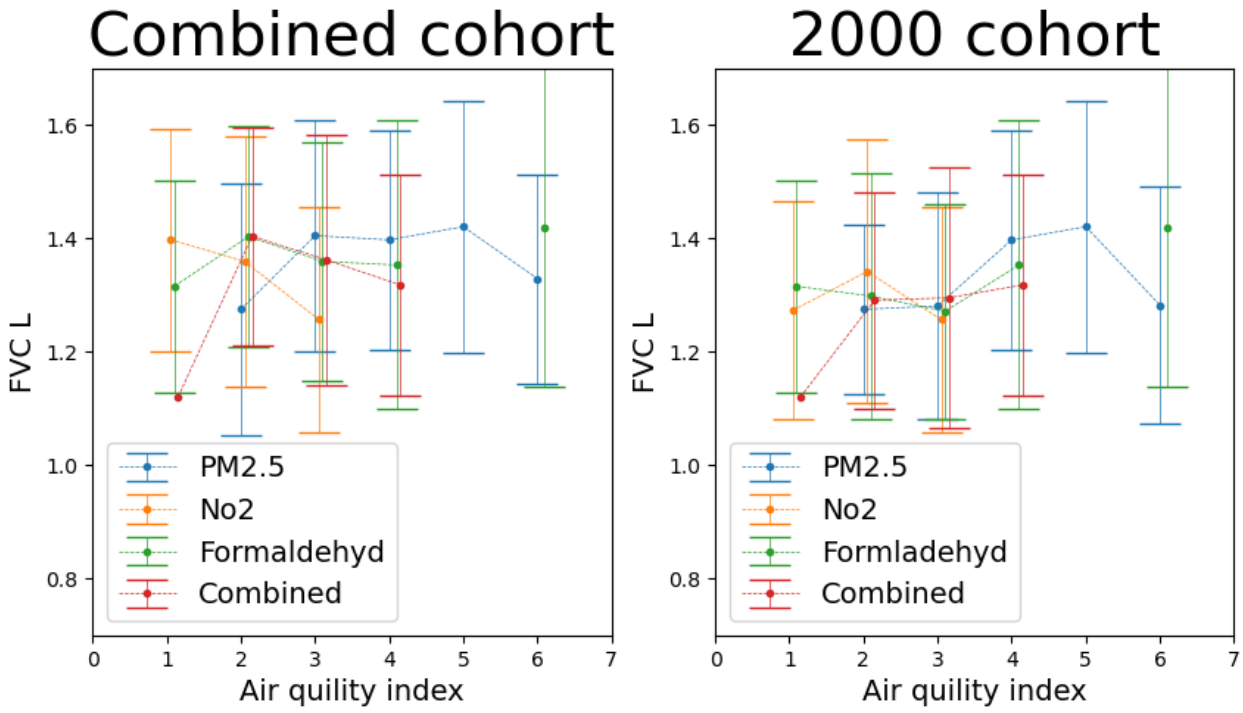

Figure 8 A comparison of the cohorts across all determinants

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

3.4.9 Metabolomics

In the COPSAC 2010 cohort, the metabolome of 660 mother-child pairs was studied. Details on sample preparation, UHPLC-MS/MS analysis, and quality control have already been outlined[30]. The list of metabolites taken for analysis was inspired by the research in [31] [2] where the researchers analysed metabolites against air pollutants the metabolites. They found significant associations between several metabolites and $PM_{2.5}$, NO_2, and temperature, and identified eight metabolic pathways, including those related to inflammation, oxidative stress, immunity, and nucleic acid damage and repair, which were perturbed due to long-term exposure to $PM_{2.5}$ and temperature. This comprehensive metabolomics study was the first to report a metabolomic signature of long-term temperature and air pollution exposure. The extracted metabolites alongside their pathways are given in Table 11. The metabolites were calculated as mean values in the mother child pairs (mother in 1st week after giving birth and child at 6 years of age). To avoid instable results, a filter was applied to take pairs which are in the lowest 25 quantiles of mother-child pairs standard deviation. Each of the metabolites was tested in a linear model against indoor air determinants from the data set:

1) area_average_room, 2) carpet, 3) construction_year, 4) cookerhood, 5) fireplace,
6) floor, 7) furred_days, 8) gasstove, 9) house_area, 10) number_of_kids,
11) passive_smoke_home, 12) pm, 13) no2, 14) blackening, 15) formaldehyde

with the hypothesis that there might an association with metabolite level following the model, where β are the coefficients for the intercept (0) and the 15 covariates:

$$E(Metabolite) = \beta_0 + \beta_{indoor\ factors1-15}$$

The results of each model for the metabolites from Table 10 where the R^2 is above 0.2 is shown in Tables 11- 18. In the tables variables with a p-value on the coefficients value less than 0.05 may be considered statistically significant in relation to the metabolite.

[30] Olarini, A., Ernst, M., Gürdeniz, G., Kim, M., Brustad, N., Bønnelykke, K., Cohen, A., Hougaard, D., Lasky-Su, J., Bisgaard, H., Chawes, B., Rasmussen, M.A., 2022. Vertical Transfer of Metabolites Detectable from Newborn's Dried Blood Spot Samples Using UPLC-MS: A Chemometric Study. Metabolites 12, 1–14. https://doi.org/10.3390/metabo12020094

[31] Nassan, F.L., Wang, C., Kelly, R.S., Lasky-Su, J.A., Vokonas, P.S., Koutrakis, P., Schwartz, J.D., 2021. Ambient PM2.5 species and ultrafine particle exposure and their differential metabolomic signatures. Environ Int 151, 106447. https://doi.org/10.1016/j.envint.2021.106447

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 10 Pathways and Metabolites

Pathway/metabolism	metabolite	pathway	metabolite
Alanine and Aspartate	aspartate	Phosphatidylcholine (PC)	1,2-dipalmitoyl-GPC (16:0/16:0)
Chemical	iminodiacetate (IDA)	Phosphatidylserine (PS)	1-stearoyl-2-oleoyl-GPS (18:0/18:1)
Food Component/Plant	ergothioneine, tartronate (hydroxymalonate)	Phospholipid Metabolism	glycerophosphoethanolamine
Gamma-glutamyl Amino Acid	gamma-* *-glutamylglycine, *-glutamylmethionine *-glutamylphenylalanine, *-glutamyl-alpha-lysine. *-glutamylvaline, *-glutamylthreonine, *-glutamylglutamine	Polyamine Metabolism	5-methylthioadenosine (MTA)
Glutathione	cys-gly, oxidized, cysteine-glutathione disulfide, cysteinylglycine	Purine Metabolism, (Hypo)Xanthine/Inosine containing	inosine, hypoxanthine, allantoin
Glycerolipid	glycerophosphoglycerol	Purine Metabolism, Adenine containing	AMP
Glycine, Serine and Threonine	sarcosine	Purine Metabolism, Guanine containing	N2,N2-dimethylguanosine
Glycolysis, Gluconeogenesis, and Pyruvate	lactate	Pyrimidine Metabolism, Orotate containing	dihydroorotate
Hemoglobin and Porphyrin	biliverdin, bilirubin (E,E)*, bilirubin, bilirubin (E,Z or Z,E)*	Pyrimidine Metabolism, Uracil containing	beta-alanine
Leucine, Isoleucine and Valine	alpha-hydroxyisocaproate	Sphingolipid Synthesis	sphinganine-1-phosphate
Lysophospholipid	1-stearoyl-GPI (18:0)	Sphingomyelins	sphingomyelin (d18:2/18:1)*
Medium Chain Fatty Acid	heptanoate (7:0)	Sphingosines	sphingosine

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Methionine, Cysteine, SAM and Taurine	taurine, cysteine s-sulfate	TCA Cycle	malate, succinate
Nicotinate and Nicotinamide	1-methylnicotinamide	Tocopherol Metabolism	alpha-tocopherol
Pentose	ribitol		

Table 11 Summary of the linear regression results for metabolite allantoin

Names	Coef	SE	P-Value	R^2
Intercept	0.16	0.03	0	0.2
area_average_room	-0.01	0.01	0.39	0.2
carpet	-0.02	0.01	0.04	0.2
construction_year	0.01	0.03	0.74	0.2
cookerhood	-0.01	0.02	0.8	0.2
fireplace	-0.03	0.01	0.02	0.2
floor	0	0.02	0.99	0.2
furred_days	0.01	0.01	0.46	0.2
gasstove	0	0.01	0.89	0.2
house_area	0.02	0.02	0.23	0.2
number_of_kids	0.01	0.01	0.57	0.2
passive_smoke_home	0.02	0.03	0.42	0.2
$PM_{2.5}$	0.07	0.05	0.21	0.2
NO_2	0.04	0.02	0.15	0.2
blackening	-0.1	0.04	0.02	0.2
formaldehyd	-0.03	0.02	0.17	0.2

For allantoin the variables "carpet," "fireplace," and "blackening" have p-values less than 0.05 and can be considered significant.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 12 Summary of the linear regression results for metabolites Taurine and Cysteine S-Sulfate.

Names	Coef	SE	P-Value	R^2
Intercept	0.09	0.04	0.02	0.27
area_average_room	-0.01	0.02	0.64	0.27
carpet	0.01	0.02	0.71	0.27
construction_year	-0.02	0.04	0.66	0.27
cookerhood	0.02	0.02	0.29	0.27
fireplace	-0.02	0.02	0.3	0.27
floor	-0.01	0.02	0.8	0.27
furred_days	0.02	0.01	0.13	0.27
gasstove	0	0.01	0.73	0.27
house_area	-0.01	0.03	0.79	0.27
number_of_kids	-0.01	0.02	0.74	0.27
passive_smoke_home	0.09	0.04	0.02	0.27
$PM_{2.5}$	-0.08	0.06	0.23	0.27
NO_2	-0.03	0.04	0.4	0.27
blackening	-0.03	0.06	0.61	0.27
formaldehyd	0	0.03	0.99	0.27

For Taurine and Cysteine S-Sulfate , one can observe that none of the variables have a p-value less than 0.05.

Table 13 Summary of the linear regression results for metabolite Cysteinylglycine.

Names	Coef	SE	P-Value	R^2
Intercept	0.04	0.03	0.16	0.21
area_average_room	0.02	0.02	0.21	0.21
carpet	-0.02	0.03	0.51	0.21
construction_year	-0.02	0.03	0.56	0.21
cookerhood	0.03	0.02	0.13	0.21
fireplace	-0.01	0.01	0.51	0.21
floor	-0.02	0.01	0.14	0.21
furred_days	0.01	0.01	0.18	0.21
gasstove	-0.01	0.01	0.41	0.21
house_area	-0.02	0.02	0.5	0.21
number_of_kids	-0.01	0.01	0.54	0.21
passive_smoke_home	-0.05	0.03	0.05	0.21
$PM_{2.5}$	0.04	0.05	0.44	0.21
NO_2	-0.02	0.03	0.52	0.21
blackening	0.04	0.05	0.43	0.21
formaldehyd	0.04	0.02	0.11	0.21

Based on the analysis for Cysteinylglycine, the variable "Passive_smoke_home" is the only variable that exhibits statistical significance at a significance level of 0.05.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 14 Summary of the linear regression results for metabolite Ergothione.

Names	Coef	SE	P-value	R^2
Intercept	0.06	0.02	0.02	0.27
area_average_room	-0.02	0.02	0.27	0.27
carpet	-0.04	0.02	0.07	0.27
construction_year	-0.03	0.03	0.20	0.27
cookerhood	0.01	0.02	0.77	0.27
fireplace	0.00	0.01	0.83	0.27
floor	0.01	0.01	0.30	0.27
furred_days	0.00	0.01	0.55	0.27
gasstove	0.01	0.01	0.12	0.27
house_area	0.04	0.02	0.08	0.27
number_of_kids	-0.03	0.01	0.01	0.27
passive_smoke_home	0.01	0.03	0.77	0.27
$PM_{2.5}$	-0.01	0.05	0.81	0.27
NO_2	0.03	0.02	0.20	0.27
blackening	-0.02	0.04	0.57	0.27
formaldehyd	0.01	0.02	0.67	0.27

Among the variables analyzed, only "Number_of_kids" (p-value = 0.01) demonstrate statistical significance.

Table 15 Summary of the linear regression results for metabolite gamma-glutamylglutamine.

Names	Coef	SE	pval	R^2
Intercept	0.35	0.05	0.00	0.21
area_average_room	-0.03	0.04	0.42	0.21
carpet	0.08	0.05	0.13	0.21
construction_year	0.10	0.06	0.08	0.21
cookerhood	-0.16	0.05	0.00	0.21
fireplace	0.02	0.02	0.37	0.21
floor	0.03	0.03	0.30	0.21
furred_days	-0.02	0.01	0.20	0.21
gasstove	0.03	0.02	0.11	0.21
house_area	0.02	0.05	0.67	0.21
number_of_kids	0.02	0.02	0.38	0.21
passive_smoke_home	-0.04	0.06	0.46	0.21
$PM_{2.5}$	0.13	0.12	0.30	0.21
NO_2	-0.06	0.05	0.29	0.21
blackening	-0.08	0.08	0.33	0.21
formaldehyd	-0.04	0.05	0.36	0.21

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Based on the p-values in the analysis, the variables that show statistical significance are ''Intercept''(p-value = 0.00) and ''Cookerhood'' (p-value = 0.00).

Table 16 Summary of the linear regression results for metabolite gamma-glutamylmethionine.

Names	Coef	SE	pval	R^2
Intercept	0.35	0.07	0.00	0.21
area_average_room	-0.03	0.05	0.45	0.21
carpet	-0.03	0.02	0.11	0.21
construction_year	0.09	0.07	0.15	0.21
cookerhood	0.03	0.06	0.59	0.21
fireplace	-0.03	0.03	0.28	0.21
floor	-0.04	0.05	0.43	0.21
furred_days	-0.03	0.02	0.22	0.21
gasstove	0.06	0.03	0.04	0.21
house_area	-0.07	0.05	0.13	0.21
number_of_kids	0.01	0.03	0.82	0.21
passive_smoke_home	0.03	0.07	0.61	0.21
$PM_{2.5}$	0.15	0.15	0.30	0.21
NO_2	-0.04	0.10	0.72	0.21
blackening	-0.23	0.15	0.11	0.21
formaldehyd	-0.09	0.06	0.12	0.21

The variable that show statistical significance is'' Gasstove'' (p-value = 0.04) .

Table 17 Summary of the linear regression results for metabolite glycerophosphoethanolamine.

Names	Coef	SE	pval	R^2
Intercept	0.10	0.02	0.00	0.20
area_average_room	0.01	0.01	0.45	0.20
carpet	0.02	0.01	0.00	0.20
construction_year	-0.01	0.02	0.65	0.20
cookerhood	0.01	0.01	0.58	0.20
fireplace	0.01	0.01	0.17	0.20
floor	0.01	0.01	0.29	0.20
furred_days	0.01	0.01	0.13	0.20
gasstove	0.00	0.01	0.54	0.20
house_area	-0.01	0.01	0.59	0.20
number_of_kids	0.00	0.01	0.76	0.20
passive_smoke_home	0.01	0.02	0.49	0.20
$PM_{2.5}$	-0.01	0.04	0.78	0.20
NO_2	0.00	0.02	0.95	0.20
blackening	-0.02	0.03	0.48	0.20
formaldehyd	0.00	0.01	0.65	0.20

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

The variable that shows statistical significance is Carpet (p-value = 0.00)

Table 18 Summary of the linear regression results for metabolite iminodiacetate (IDA).

Names	Coef	SE	pval	R^2
Intercept	0.16	0.03	0.00	0.27
area_average_room	0.00	0.02	0.90	0.27
carpet	-0.05	0.02	0.00	0.27
construction_year	0.05	0.03	0.11	0.27
cookerhood	-0.03	0.01	0.03	0.27
fireplace	0.02	0.01	0.29	0.27
floor	0.02	0.02	0.26	0.27
furred_days	-0.01	0.01	0.22	0.27
gasstove	0.00	0.01	0.88	0.27
house_area	0.00	0.03	0.96	0.27
number_of_kids	0.03	0.02	0.03	0.27
passive_smoke_home	0.01	0.03	0.75	0.27
$PM_{2.5}$	-0.04	0.07	0.51	0.27
NO_2	0.01	0.03	0.84	0.27
blackening	0.03	0.06	0.64	0.27
formaldehyd	-0.02	0.02	0.46	0.27

The variables that show statistical significance are “Carpet” (p-value = 0.00),Cookerhood (p-value = 0.03)and “Number_of_kids” (p-value = 0.03).

This project has received funding from the European Union’s
Horizon Europe Framework Programme under grant agreement № 101057497.

4. EDIAQI findings from the ATOPICA cohort

Table 19 provides an overview of the demographic characteristics of the study population, which consists of 4016 participants. The participants' average age is 7 years. Among the participants, there are 1819 males and 1767 females, while gender information is unavailable for 430 participants. The study includes participants from various regions of Croatia, namely Zagreb, Split, Slavonia, and the Zadar area. The majority of participants, over a thousand, come from Zagreb, Split, and Slavonia, while only 390 participants are from the Zadar area. Regarding the participants' residency, 3288 individuals are from urban areas, while 726 individuals are from rural areas. Therefore, it can be inferred that most of the participants reside in urban environments.

Table 19 Demographic characteristics and overview of the positive and negative signals in the outcomes.

Demographic Characteristics	N = 4016
Age (Mean)	7 years
Sex o **Male** o **Female**	 1819 1767
Area o **Zagreb** o **Split** o **Slavonia** o **Zadar**	 1324 1180 1122 390
Urban **Rural**	3288 726

For 13 different health conditions logistic regression, univariate analysis and random forest classifier were applied to inspect associations towards the outcome. Table 20 provides an overview of positive and negative outcomes present in each health condition, together with the condition meaning.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 20 Positive and negative outcomes characteristics for different health conditions.

Health Condition	Outcomes	Meaning
ASTHMA	o Positive 126 o Negative 3490	Child had asthma in the last 12 months
DUSTMITE_>2mm	o Positive 597 o Negative 3419	Sensitised dustmite SPT >= 2 mm
ECZ_EVER	o Positive 1003 o Negative 2575	Child ever had eczema
ECZ_ITCH	o Positive 660 o Negative 2941	Child had eczema itch in the last 12 months
ITCH_AWAKE	o Positive 176 o Negative 912	Child awake because of itch in the last 12 months
RHINITIS	o Positive 1388 o Negative 2180	Child had rhinitis in the last 12 months
RHINO_AMB	o Positive 350 o Negative 3209	Child had ambrosia rhinitis in the last 12 months
RHINO_CONJ	o Positive 721 o Negative 2843	Child had rhinoconjunctivitis in the last 12 months
SENS_>2mm_not AMBROSIA	o Positive 1093 o Negative 2923	Sensitised to one or more allergens other than ambrosia SPT >= 2
WHEEZE	o Positive 630 o Negative 2992	Child had wheezing or whistling in the chest in the last 12 months
WHEEZE_EVER	o Positive 1095 o Negative 2533	Child ever had wheezing or whistling in the chest
WHEEZE_ATTACKS	o Positive 500 o Negative 132	Child had wheezing attacks in the last 12 months
WHEEZE_SPEECH	o Positive 141 o Negative 495	Wheezing limited child's speech in the last 12 months

4.1 Logistic regression

The results of logistic regression are presented in Table 21 - 33, showcasing the outcomes of the analysis pertaining to asthma symptoms such as eczema, rhinitis, wheezing, sensitisation, and itch. The logistic regression model revealed that two predictor variables, namely "mattress older than 3 years" and "central heating," are statistically significant in predicting dust mite-specific skin prick test (SPT) results. Both predictors exhibit a negative correlation with the outcome, as evidenced by their coefficients of -0.5 and -0.48, respectively. Additionally, the variable "feather_pillow" emerged as statistically significant for itch symptoms, exhibiting a similar coefficient of 0.53. This suggests that if a child sleeps on a feather pillow, there is a 0.53 increase in the log-odds of experiencing itch symptoms.

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Similar results were observed for the variable "fossil_fuels," which exhibited a coefficient of 0.52, indicating its statistical significance in predicting ambrosia-related symptoms.
This finding suggests that the utilization or exposure to fossil fuels is associated with a 0.52 increase in the log-odds of experiencing symptoms related to ambrosia. Findings regarding the wheeze attack outcome reveal significant associations of variable "kindergarten months" with exhibited coefficient of 3.32, indicating its strong statistical significance. This suggests that for each additional month spent in kindergarten, there is a substantial increase of 3.32 in the log-odds of experiencing a wheeze attack. On the other hand, the variable "bird home" displayed a coefficient of -1.07, which also demonstrated statistical significance. This negative coefficient implies that having a bird at home is associated with a decrease of 1.07 in the log-odds of experiencing a wheeze attack. Notably, the variable "central heating" demonstrated a statistically significant coefficient of -0.50, indicating a negative relationship. This suggests that the presence of central heating is associated with a decrease of 0.50 in the log-odds of experiencing wheezing or whistling in the chest.
These results highlight the potential impact of both environmental and lifestyle factors on the occurrence of asthma symptoms, underscoring the importance of considering multiple variables when examining the complex nature of this condition.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

Table 21 Summary of the logistic regression analysis for asthma symptoms.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-3,01	0,00	-4,23	-1,78
mattres_older_than_3y	0,18	0,65	-0,60	0,95
electrical_heating	-0,16	0,68	-0,92	0,60
mattress_oldness_year1	-0,03	0,95	-0,90	0,84
cat_or_dog_home	1,01	0,13	-0,31	2,32
nurs_kind_%life	0,86	0,43	-1,26	2,98
parents	0,06	0,83	-0,51	0,64
kindergarten	-1,04	0,11	-2,33	0,25
cat_home	-0,52	0,33	-1,56	0,52
kindergarten_months	0,96	0,42	-1,36	3,27
carpet	0,13	0,67	-0,45	0,70
bird_home	-0,21	0,62	-1,04	0,62
dog_home	-0,33	0,56	-1,42	0,77
coal_or_wood	0,07	0,84	-0,63	0,77
feather_pillow	-0,44	0,19	-1,09	0,21
smoking_preg_or_home	-0,12	0,70	-0,71	0,48
central_heating	-0,41	0,30	-1,17	0,36
mattres_oldness_year2-7	0,06	0,87	-0,69	0,81
fossil_fuel	-0,78	0,10	-1,71	0,14

Table 22 Summary of the logistic regression analysis for dust mite sensitization.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-1,63	0,00	-2,35	-0,91
mattres_older_than_3y	-0,50	0,03	-0,96	-0,04
electrical_heating	0,02	0,92	-0,41	0,45
mattress_oldness_year1	0,45	0,08	-0,05	0,94
cat_or_dog_home	0,20	0,63	-0,60	0,99
nurs_kind_%life	0,62	0,29	-0,52	1,76
parents	-0,15	0,36	-0,48	0,17
kindergarten	0,15	0,67	-0,55	0,86
cat_home	0,04	0,91	-0,60	0,67
kindergarten_months	-0,63	0,32	-1,87	0,61
carpet	-0,01	0,96	-0,33	0,31
bird_home	-0,04	0,86	-0,49	0,41
dog_home	-0,27	0,40	-0,91	0,36
coal_or_wood	-0,01	0,95	-0,42	0,39
feather_pillow	-0,18	0,31	-0,52	0,16
smoking_preg_or_home	0,21	0,21	-0,12	0,53
central_heating	-0,48	0,03	-0,91	-0,05
mattres_oldness_year2-7	0,01	0,98	-0,46	0,48
fossil_fuel	-0,46	0,07	-0,95	0,03

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 23 Summary of the logistic regression analysis for eczema symptom.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-0,75	0,01	-1,28	-0,22
mattres_older_than_3y	-0,03	0,84	-0,36	0,30
electrical_heating	0,02	0,91	-0,30	0,34
mattress_oldness_year1	0,06	0,75	-0,32	0,45
cat_or_dog_home	0,11	0,72	-0,50	0,72
nurs_kind_%life	-0,38	0,42	-1,31	0,55
parents	-0,05	0,68	-0,30	0,20
kindergarten	-0,21	0,45	-0,74	0,33
cat_home	-0,07	0,78	-0,56	0,42
kindergarten_months	0,51	0,32	-0,49	1,51
carpet	-0,10	0,43	-0,34	0,15
bird_home	0,13	0,43	-0,20	0,46
dog_home	-0,27	0,28	-0,76	0,22
coal_or_wood	-0,07	0,65	-0,37	0,23
feather_pillow	0,18	0,17	-0,07	0,43
smoking_preg_or_home	-0,07	0,57	-0,32	0,18
central_heating	-0,01	0,97	-0,31	0,30
mattres_oldness_year2-7	-0,06	0,75	-0,39	0,28
fossil_fuel	-0,09	0,59	-0,44	0,25

Table 24 Summary of the logistic regression analysis for eczema itch symptom.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-1,10	0,00	-1,70	-0,49
mattres_older_than_3y	-0,08	0,67	-0,45	0,29
electrical_heating	0,10	0,59	-0,27	0,46
mattress_oldness_year1	0,03	0,89	-0,41	0,48
cat_or_dog_home	-0,08	0,82	-0,76	0,60
nurs_kind_%life	0,13	0,81	-0,91	1,17
parents	-0,05	0,71	-0,34	0,23
kindergarten	-0,33	0,29	-0,94	0,28
cat_home	0,09	0,75	-0,44	0,61
kindergarten_months	0,24	0,67	-0,88	1,37
carpet	-0,22	0,13	-0,50	0,07
bird_home	0,14	0,46	-0,23	0,52
dog_home	0,20	0,47	-0,35	0,75
coal_or_wood	-0,10	0,57	-0,45	0,25
feather_pillow	-0,07	0,65	-0,36	0,22
smoking_preg_or_home	0,03	0,85	-0,26	0,31
central_heating	-0,10	0,59	-0,45	0,26
mattres_oldness_year2-7	-0,22	0,27	-0,61	0,17
fossil_fuel	-0,28	0,17	-0,69	0,12

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 25 Summary of the logistic regression analysis for itch symptom.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-2,01	0,00	-3,24	-0,78
mattres_older_than_3y	0,57	0,11	-0,14	1,27
electrical_heating	-0,13	0,73	-0,89	0,62
mattress_oldness_year1	-0,76	0,10	-1,69	0,16
cat_or_dog_home	-0,36	0,60	-1,70	0,98
nurs_kind_%life	0,75	0,47	-1,28	2,79
parents	-0,31	0,27	-0,87	0,25
kindergarten	0,44	0,46	-0,74	1,61
cat_home	0,46	0,40	-0,61	1,52
kindergarten_months	-0,61	0,59	-2,82	1,61
carpet	-0,04	0,87	-0,58	0,49
bird_home	-0,68	0,11	-1,51	0,15
dog_home	0,02	0,97	-1,04	1,07
coal_or_wood	0,00	0,99	-0,67	0,67
feather_pillow	0,53	0,05	0,00	1,06
smoking_preg_or_home	0,41	0,14	-0,13	0,94
central_heating	-0,43	0,26	-1,18	0,32
mattres_oldness_year2-7	-0,37	0,32	-1,10	0,36
fossil_fuel	-0,42	0,33	-1,26	0,42

Table 26 Summary of the logistic regression analysis for rhinitis symptom.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-0,12	0,62	-0,61	0,36
mattres_older_than_3y	-0,04	0,80	-0,34	0,26
electrical_heating	0,19	0,21	-0,10	0,48
mattress_oldness_year	-0,16	0,37	-0,52	0,19
cat_or_dog_home	0,00	0,99	-0,54	0,55
nurs_kind_%life	-0,62	0,16	-1,48	0,24
parents	-0,05	0,68	-0,27	0,18
kindergarten	-0,34	0,17	-0,83	0,14
cat_home	0,19	0,39	-0,24	0,62
kindergarten_months	0,85	0,07	-0,07	1,77
carpet	-0,03	0,81	-0,25	0,20
bird_home	-0,02	0,91	-0,32	0,29
dog_home	-0,03	0,88	-0,47	0,40
coal_or_wood	-0,21	0,13	-0,49	0,06
feather_pillow	0,09	0,43	-0,14	0,32
smoking_preg_or_home	0,07	0,57	-0,16	0,29
central_heating	-0,13	0,37	-0,41	0,15
mattres_oldness_year2-7	-0,10	0,51	-0,41	0,20
fossil_fuel	0,17	0,28	-0,14	0,47

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 27 Summary of the logistic regression analysis for rhinitis for ambrosia symptom.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-2,34	0,00	-3,07	-1,61
mattres_older_than_3y	0,04	0,87	-0,44	0,52
electrical_heating	0,22	0,32	-0,22	0,67
mattress_oldness_year1	0,02	0,94	-0,51	0,55
cat_or_dog_home	-0,03	0,94	-0,87	0,80
nurs_kind_%life	-0,37	0,60	-1,77	1,02
parents	-0,10	0,60	-0,46	0,26
kindergarten	-0,47	0,21	-1,21	0,27
cat_home	0,34	0,32	-0,34	1,03
kindergarten_months	0,90	0,24	-0,59	2,39
carpet	0,08	0,65	-0,27	0,43
bird_home	-0,21	0,40	-0,71	0,28
dog_home	-0,15	0,64	-0,78	0,48
coal_or_wood	0,20	0,34	-0,21	0,61
feather_pillow	0,21	0,25	-0,14	0,56
smoking_preg_or_home	0,14	0,42	-0,21	0,50
central_heating	-0,07	0,73	-0,49	0,35
mattres_oldness_year2-7	0,17	0,49	-0,31	0,64
fossil_fuel	0,52	0,02	0,07	0,97

Table 28 Summary of the logistic regression analysis for conjunctivitis for ambrosia symptom.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-0,98	0,00	-1,54	-0,42
mattres_older_than_3y	-0,15	0,40	-0,50	0,20
electrical_heating	0,17	0,32	-0,17	0,51
mattress_oldness_year1	0,06	0,79	-0,36	0,47
cat_or_dog_home	-0,13	0,68	-0,77	0,50
nurs_kind_%life	-0,36	0,49	-1,36	0,65
parents	-0,05	0,73	-0,31	0,22
kindergarten	-0,30	0,29	-0,86	0,26
cat_home	0,34	0,20	-0,17	0,85
kindergarten_months	0,37	0,50	-0,71	1,45
carpet	0,02	0,86	-0,24	0,28
bird_home	0,14	0,43	-0,21	0,49
dog_home	-0,08	0,75	-0,58	0,42
coal_or_wood	-0,14	0,40	-0,46	0,19
feather_pillow	0,20	0,14	-0,06	0,47
smoking_preg_or_home	0,06	0,67	-0,21	0,33
central_heating	-0,16	0,34	-0,48	0,17
mattres_oldness_year2-7	-0,15	0,42	-0,51	0,21
fossil_fuel	0,33	0,07	-0,03	0,68

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 29 Summary of the logistic regression analysis for sensitivity to one or more allergens other than Ambrosia.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-0,90	0,00	-1,44	-0,37
mattres_older_than_3y	-0,17	0,33	-0,50	0,17
electrical_heating	-0,10	0,54	-0,42	0,22
mattress_oldness_year1	0,18	0,36	-0,20	0,55
cat_or_dog_home	0,17	0,58	-0,43	0,77
nurs_kind_%life	-0,02	0,96	-0,92	0,88
parents	-0,16	0,22	-0,40	0,09
kindergarten	0,30	0,26	-0,22	0,83
cat_home	0,00	0,99	-0,47	0,48
kindergarten_months	-0,33	0,50	-1,31	0,64
carpet	-0,04	0,72	-0,29	0,20
bird_home	-0,11	0,52	-0,45	0,23
dog_home	-0,23	0,35	-0,71	0,25
coal_or_wood	0,09	0,57	-0,21	0,39
feather_pillow	-0,17	0,18	-0,43	0,08
smoking_preg_or_home	0,21	0,10	-0,04	0,45
central_heating	-0,30	0,06	-0,61	0,01
mattres_oldness_year2-7	0,00	0,98	-0,33	0,34
fossil_fuel	-0,22	0,21	-0,56	0,13

Table 30 Summary of the logistic regression analysis for wheeze symptom.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-1,28	0,00	-1,90	-0,67
mattres_older_than_3y	-0,01	0,94	-0,39	0,36
electrical_heating	0,03	0,86	-0,33	0,40
mattress_oldness_year	-0,03	0,89	-0,48	0,42
cat_or_dog_home	0,24	0,48	-0,43	0,91
nurs_kind_%life	-0,80	0,15	-1,89	0,30
parents	0,21	0,14	-0,07	0,49
kindergarten	0,05	0,87	-0,55	0,66
cat_home	0,05	0,87	-0,48	0,57
kindergarten_months	0,44	0,46	-0,73	1,60
carpet	0,04	0,77	-0,24	0,32
bird_home	-0,01	0,96	-0,39	0,37
dog_home	-0,07	0,79	-0,61	0,46
coal_or_wood	-0,14	0,42	-0,49	0,21
feather_pillow	0,15	0,31	-0,14	0,43
smoking_preg_or_home	-0,24	0,11	-0,53	0,06
central_heating	-0,19	0,31	-0,55	0,17
mattres_oldness_year2-7	-0,24	0,23	-0,63	0,15
fossil_fuel	-0,18	0,39	-0,57	0,22

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 31 Summary of the logistic regression analysis for wheeze attacks symptom.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	0,94	0,23	-0,59	2,47
mattres_older_than_3y	-0,61	0,18	-1,51	0,29
electrical_heating	0,32	0,48	-0,58	1,23
mattress_oldness_year1	0,91	0,14	-0,30	2,12
cat_or_dog_home	-0,32	0,68	-1,88	1,23
nurs_kind_%life	-1,37	0,25	-3,70	0,96
parents	0,14	0,67	-0,49	0,77
kindergarten	-0,37	0,59	-1,71	0,98
cat_home	0,50	0,43	-0,74	1,74
kindergarten_months	3,23	0,02	0,52	5,94
carpet	-0,44	0,16	-1,06	0,17
bird_home	-1,07	0,01	-1,88	-0,25
dog_home	0,27	0,68	-1,01	1,55
coal_or_wood	0,22	0,59	-0,59	1,04
feather_pillow	0,36	0,29	-0,30	1,03
smoking_preg_or_home	-0,18	0,60	-0,85	0,49
central_heating	0,23	0,61	-0,66	1,11
mattres_oldness_year2-7	0,13	0,79	-0,81	1,06
fossil_fuel	0,04	0,93	-0,86	0,94

Table 32 Summary of the logistic regression analysis for wheezing or whistling in the chest.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-0,10	0,70	-0,62	0,42
mattres_older_than_3y	-0,08	0,62	-0,40	0,24
electrical_heating	-0,30	0,07	-0,61	0,02
mattress_oldness_year1	0,03	0,87	-0,34	0,41
cat_or_dog_home	0,41	0,16	-0,17	0,99
nurs_kind_%life	0,10	0,82	-0,77	0,97
parents	0,00	0,97	-0,23	0,24
kindergarten	0,01	0,97	-0,50	0,52
cat_home	-0,22	0,33	-0,68	0,23
kindergarten_months	-0,44	0,36	-1,38	0,50
carpet	0,07	0,56	-0,17	0,30
bird_home	0,07	0,67	-0,25	0,39
dog_home	-0,19	0,44	-0,66	0,29
coal_or_wood	-0,26	0,09	-0,56	0,04
feather_pillow	-0,01	0,92	-0,26	0,23
smoking_preg_or_home	-0,23	0,06	-0,48	0,01
central_heating	-0,50	0,00	-0,81	-0,19
mattres_oldness_year2-7	-0,25	0,14	-0,57	0,08
fossil_fuel	-0,36	0,04	-0,70	-0,02

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

Table 33 Summary of the logistic regression analysis for wheezing being limitation for child.

names	coef	pval	CI[2.5%]	CI[97.5%]
Intercept	-1,38	0,10	-3,02	0,26
mattres_older_than_3y	0,31	0,51	-0,60	1,21
electrical_heating	0,08	0,87	-0,84	0,99
mattress_oldness_year1	-0,64	0,26	-1,75	0,47
cat_or_dog_home	0,88	0,27	-0,67	2,42
nurs_kind_%life	-0,40	0,77	-3,05	2,25
parents	0,37	0,25	-0,26	0,99
kindergarten	0,09	0,91	-1,41	1,58
cat_home	-0,69	0,28	-1,93	0,55
kindergarten_months	-0,04	0,98	-2,89	2,82
carpet	-0,42	0,20	-1,07	0,23
bird_home	0,34	0,43	-0,51	1,19
dog_home	-0,40	0,54	-1,69	0,88
coal_or_wood	-0,39	0,35	-1,21	0,43
feather_pillow	0,10	0,76	-0,57	0,78
smoking_preg_or_home	0,12	0,73	-0,55	0,79
central_heating	-0,06	0,91	-0,98	0,87
mattres_oldness_year2-7	0,18	0,70	-0,73	1,10
fossil_fuel	-0,61	0,22	-1,60	0,37

4.2 Univariate statistical tests

We conducted a series of statistical tests to explore the associations between various independent variables and health outcomes. All p-values reported are two-sided, and p-values less than 0.001 are considered statistically significant. This criterion was set due to many tests conducted and can be deemed a correction regarding multiple testing. For several tests, boxplots are shown in Figures 10 and 11.

Regarding the outcome of asthma, gender was found to be significantly associated (Chi-square test, p=2.31e-05). Furthermore, gender also showed a significant relationship with dust mite allergy (Chi-square test, p=4.71e-04).

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

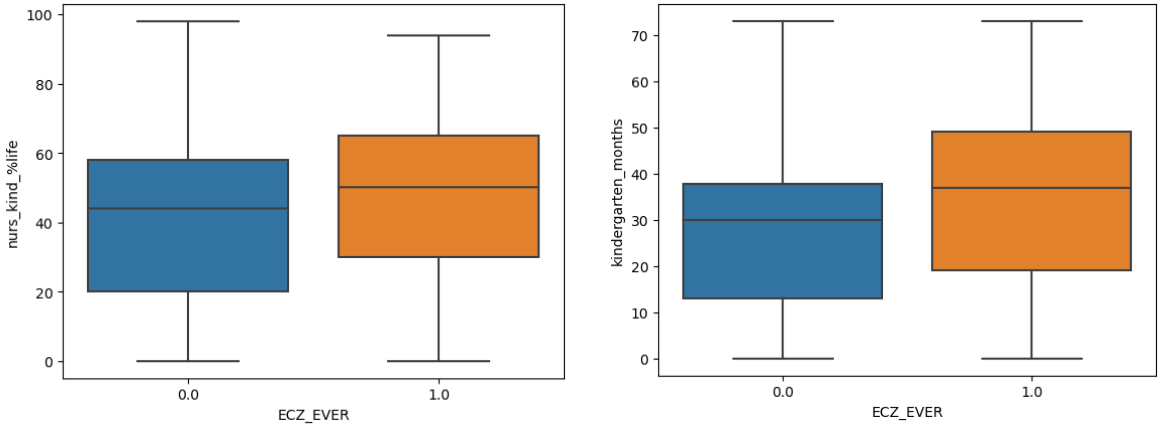

Figure 9 Boxplots of eczema symptoms against associated variables

For eczema, multiple variables showed significant associations. Kindergarten attendance was significantly related to ever having eczema (Chi-square test, p=6.63e-08), as was exposure to smoking during pregnancy or at home (Chi-square test, p=1.50e-04). The percentage of life spent in nursing or kindergarten (Mann-Whitney U test, p<0.0001), and having parents with history of the same condition (Chi-square test, p=1.95e-25) were also significantly associated with ever having eczema. Moreover, use of a feather pillow showed a significant association (Chi-square test, p=7.96e-04). The duration in months spent in kindergarten also demonstrated a significant association (Mann-Whitney U test, p<0.0001). For eczema with itching, the presence of parents with a history of the condition was significantly related (Chi-square test, p=1.96e-12). In terms of rhinitis, gender (Chi-square test, p=1.60e-04), parents with a history of the condition (Chi-square test, p=4.03e-19), and use of a feather pillow (Chi-square test, p=5.18e-05) were significantly associated. When considering rhinoconjunctivitis, gender (Chi-square test, p=2.80e-04), parents with a history of the condition (Chi-square test, p=9.26e-16), and use of a feather pillow (Chi-square test, p=7.25e-04) showed significant associations.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

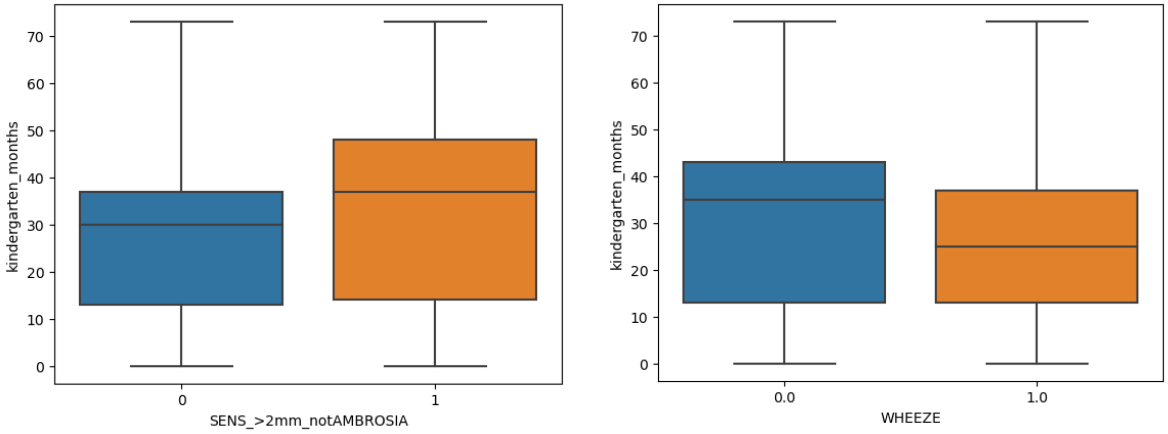

Figure 10 Boxplots for sensitization and wheeze symptoms against associated variables

For sensitivity to non-ambrosia allergens greater than 2mm in SPT, gender (Chi-square test, p=3.47e-09), parents with a history of the condition (Chi-square test, p=3.78e-06), and duration in months spent in kindergarten (Mann-Whitney U test, p<0.0001) were significantly associated.

Finally, concerning wheezing, gender was significantly associated with ever having wheezed (Chi-square test, p=2.65e-06), wheezing attacks (Chi-square test, p=1.55e-04), and wheezing that disrupted speech (Chi-square test, p=9.04e-04). Additionally, parents with a history of the condition were significantly associated with ever having wheezed (Chi-square test, p=1.28e-05) and duration in months spent in kindergarten showed a significant association with ever having wheezed (Mann-Whitney U test, p<0.0001). Exposure to smoking during pregnancy or at home was significantly related to ever having wheezed (Chi-square test, p=7.05e-04) and having a bird at home was significantly associated with wheezing attacks (Chi-square test, p=2.00e-04).

In summary, these findings provide evidence of significant associations between the variables examined and the health outcomes of interest. Further research is needed to explore these relationships in more depth and to investigate potential underlying mechanisms.

4.3 Random Forest Classifier

To predict the presence or absence of each outcome, cross-validation of Random Forest Classifier was performed with optimized classifier hyperparameters *n_estimators*, *max_depth*, and *min_samples_leaf*. The hyperparameters were optimized via Optuna in the

This project has received funding from the European Union's Horizon Europe Framework Programme under grant agreement № 101057497.

direction of maximization of resulting balanced accuracy score on the train set. Apart from the mentioned hyperparameters, classifier criterion was defined as logarithmic loss and class weight as balanced to account for the imbalanced data. Missing values were imputed in the independent variables with the most frequent value in each variable. Cross-validation was performed with 5 folds, and the resulting balanced accuracy scores for the test sets are reported in the Table 33 for each outcome. The balanced accuracy is a metric used for the evaluation of the classification model which calculates the average accuracy for each class considering the class imbalance. The main advantage of calculating balanced accuracy is fair evaluation of the classification model across all classes, regardless of class size which is very useful with imbalanced data. The balanced accuracy score ranges from 0 to 1, where 1 is the perfect model performance. When the score is around 0.5 it means that model performs no better than by random chance. For models given in Table 33 one can conclude that most are performing worse than random while only several are performing slightly better meaning that these data have no predictive power on the given outcomes.

Table 34 Overview of averaged balanced accuracy score for the outcomes.

Outcome	Balanced Accuracy Score
ASTHMA	0.50
DUSTMITE_>2mm	0.54
ECZ_EVER	0.47
ECZ_ITCH	0.49
ITCH_AWAKE	0.54
RHINITIS	0.53
RHINO_AMB	0.51
RHINO_CONJ	0.51
SENS_>2mm_notAMBROSIA	0.53
WHEEZE	0.54
WHEEZE_EVER	0.55
WHEEZE_ATTACKS	0.58
WHEEZE_SPEECH	0.56

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

5. Summary

In conclusion, the findings from the ATOPICA and COPSAC cohorts provide valuable insights into the relationships between indoor/outdoor environmental factors and the development of atopic allergies, respiratory symptoms, and asthma in children. These studies contribute to the objectives of exploring the relationship between indoor and outdoor air pollution and health outcomes, analysing data using machine learning techniques and statistical analysis, focusing on vulnerable populations, understanding the impact of air quality on their health, investigating the relationship between indoor and airway microbiota, and exploring non-linear relationships between indoor air pollution and asthma/allergy outcomes.

Three previous studies from the ATOPICA cohort examined the development of atopic allergies in children in relation to environmental determinants. The first study found that exposure to tobacco smoke, low physical activity, and poor diet increased the risk of ragweed allergy. The second study, conducted in Croatia, found a higher prevalence of ragweed allergy in coastal regions than inland, potentially due to higher pollen levels and varied environmental and lifestyle factors. The third study supported the idea that factors other than pollen concentration, such as air pollution and diet, contributed to the regional differences in ragweed sensitization. The studies emphasized the need for prevention strategies and further research to identify and address modifiable risk factors for ragweed allergy in children.

In current findings, a total of 13 different health conditions using logistic regression, univariate analysis, and random forest classifiers were analysed to predict positive or negative outcome for each condition and understand their relationships to indoor determinants. Logistic regression revealed several statistically significant predictor variables for different health conditions while however having low predictive power. Mattress age and central heating were negatively correlated with dust mite-specific skin prick test results, while feather pillow and fossil fuel exposure were positively associated with itch symptoms and ambrosia-related symptoms, respectively. The duration of kindergarten attendance showed a strong positive association with wheeze attacks. The presence of central heating was negatively associated with wheezing or whistling in the chest. Univariate analysis identified that gender has significant associations with asthma, dust mite allergy, rhinitis,

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

rhinoconjunctivitis, and wheezing outcomes and should be taken as confounder in future studies. Other significant associations included kindergarten attendance, exposure to smoking during pregnancy or at home, parents with a history of the condition, and the use of a feather pillow. The random forest classifier was optimized using hyperparameters and cross-validation. The balanced accuracy score, which considers class imbalance, was calculated, and ranged from 0.47 to 0.58 where higher scores indicate better model performance, but all results are close to 0.5 which indicates the model's performance is hardly better then random chance, showing that there is no true predictive power there. Previous research on COPSAC cohorts shed light on the impact of indoor air pollution and microbial exposures on respiratory health. The COPSAC 2000 cohort showed also that long-term exposure to indoor air pollution did not have significant associations with wheezing symptoms in infants which is in line with the ATOPIC findings for WHEEZE. This highlights the need to investigate other data sources to capture the complex dynamics between indoor air quality and health outcomes. Meanwhile in the metabolome analysis, some associations were presented which showcase the metabolites revealed in [31] and pointing toward future understanding of biological pathways

Moreover, the COPSAC cohorts provided insights into the relationship between indoor and airway microbiota and their associations with allergies. These findings emphasize the role of microbial exposures in respiratory health outcomes and highlight the need for further research in this area. In summary, the findings from these cohorts contribute to our understanding of the relationships between indoor and outdoor air pollution, asthma/allergy outcomes, and the influence of microbial exposures. They emphasize the importance of exploring these relationships using various data sources, applying machine learning techniques, considering vulnerable populations, and investigating non-linear relationships. This knowledge can inform the development of effective preventive measures and interventions to protect respiratory health, particularly in children and other vulnerable groups.

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.

Deliverable D5.1

Report of existing cohort data and analysis regarding health effects of pollutants

Work Package 5
TOX & HEALTH
Version: Final

This project has received funding from the European Union's
Horizon Europe Framework Programme under grant agreement № 101057497.