



Prenatal exposure to air pollutants and childhood atopic dermatitis and allergic rhinitis adopting machine learning approaches: 14-year follow-up birth cohort study

Yu Huang^{a,1}, Hui-Ju Wen^{b,1}, Yue-Liang Leon Guo^{b,c,d}, Tzu-Yin Wei^a, Wei-Cheng Wang^a, Shin-Fen Tsai^b, Vincent S. Tseng^{a,*}, Shu-Li Julie Wang^{b,e,f,**}

^a Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

^b National Institute of Environmental Health Sciences, National Health Research Institutes, Miaoli, Taiwan

^c Department of Environmental and Occupational Medicine, National Taiwan University Hospital, Taipei, Taiwan

^d Institute of Environmental and Occupational Health Sciences, National Taiwan University College of Public Health, Taipei, Taiwan

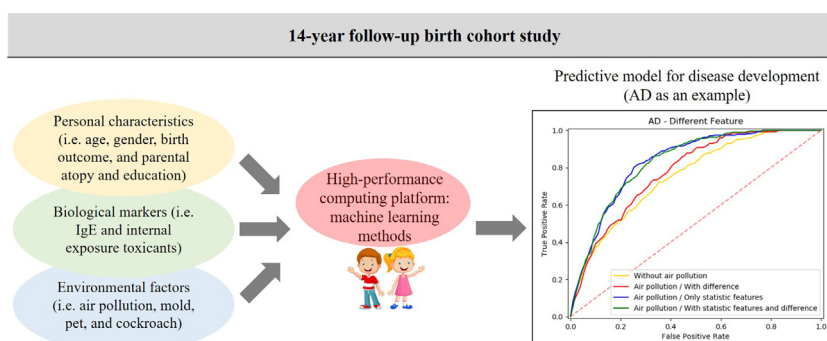
^e Department of Public Health, National Defence Medical Centre, Taipei, Taiwan

^f Department of Safety, Health, and Environmental Engineering, National United University, Miaoli, Taiwan

HIGHLIGHTS

- A 14-year birth cohort study of infants born at nine hospitals across Taiwan
- 1439 mother-infant pairs were investigated in machine learning analysis.
- Random forest model achieved higher predictive ability for childhood AD and AR.
- Prenatal air quality plays an important factor in predictive model for AD and AR.
- Prenatal NO₂ exposure and its concentration change were significant predictors.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 29 October 2020

Received in revised form 15 February 2021

Accepted 16 February 2021

Available online 20 February 2021

Editor: Damia Barcelo

ABSTRACT

The incidence of childhood atopic dermatitis (AD) and allergic rhinitis (AR) is increasing. This warrants development of measures to predict and prevent these conditions. We aimed to investigate the predictive ability of a spectrum of data mining methods to predict childhood AD and AR using longitudinal birth cohort data. We conducted a 14-year follow-up of infants born to pregnant women who had undergone maternal examinations at nine selected maternity hospitals across Taiwan during 2000–2005. The subjects were interviewed using structured questionnaires to record data on basic demographics, socioeconomic status, lifestyle, medical history, and 24-h dietary recall. Hourly concentrations of air pollutants within 1 year before childbirth were obtained from 76

Abbreviations: AD, atopic dermatitis; AR, allergic rhinitis; AUROC, area under the ROC curve; BMI, body mass index; CI, confident interval; CO, carbon monoxide; DL, detection limit; EPA, the Environmental Protection Administration; FEF, forced expiratory flow rate; FEV₁, forced expiratory volume in one second; FVC, forced vital capacity; GM, geometrical mean; IAQI, individual air quality index; LCA, latent class analysis; Ig, immunoglobulin; IQR, interquartile range; KNN, k-nearest neighbours; LOQ, the limit of quantitation values; LR, logistic regression; NB, naïve bayes; NN, neural network; NO₂, nitrogen dioxide; O₃, ozone; OR, odds ratio; PFASs, polyfluoroalkyl substances; PFDeA, perfluorodecanoic acid; PFDoA, perfluorododecanoic acid; PFHpA, perfluoroheptanoic acid; PFHxA, perfluorohexanoic acid; PFHxS, perfluorohexanesulfonic acid; PM₁₀, particulate matter with a 50% cut-off aerodynamic diameter of 10 mm; PFNA, perfluorononanoic acid; PFOA, perfluorooctanoic acid; PFOS, perfluorooctane sulfonate; PFUA, perfluoroundecanoic acid; RF, random forest; ROC curve, receiver operating characteristic curve; SD, standard deviation; SF-36, the short form 36 questionnaire; SO₂, sulphur dioxide; USD, US Dollar; XGB, XGboost.

* Correspondence to: V.S. Tseng, 1001 University Road, Hsinchu 30010, Taiwan.

** Correspondence to: S.-L.J. Wang, 35 Keyan Road, Zhunan, Miaoli County 35053, Taiwan.

E-mail addresses: vteng@cs.nctu.edu.tw (V.S. Tseng), slwang@nhri.edu.tw (S.-L.J. Wang).

¹ Authors contributed equally to this work.

Keywords:

Birth cohort
Allergic diseases
Machine learning
Prediction
Air pollution

national air quality monitoring stations in Taiwan. We utilized weighted K-nearest neighbour method ($k = 3$) to infer the personalized air pollution exposure. Machine learning methods were performed on the heterogeneous attributes set to predict allergic diseases in children. A total of 1439 mother-infant pairs were recruited in machine learning analysis. The prevalence of AD and AR in children up to 14 years of age were 6.8% and 15.9%, respectively. Overall, tree-based models achieved higher sensitivity and specificity than other methods, with areas under receiver operating characteristic curve of 83% for AD and 84% for AR, respectively. Our findings confirmed that prenatal air quality is an important factor affecting the predictive ability. Moreover, different air quality indices were better predicted, in combination than separately. Combining heterogeneous attributes including environmental exposures, demographic information, and allergens is the key to a better prediction of children allergies in the general population. Prenatal exposure to nitrogen dioxide (NO_2) and its concatenation changes with time were significant predictors for AD and AR till adolescent.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

In the past decades, the prevalence of childhood allergies has consistently increased in both developed and developing countries. Atopic dermatitis (AD), allergic rhinitis, (AR), and asthma are the most common chronic allergic diseases and now are recognized as global problems affecting children's health (Dogruel et al., 2017; Hansen et al., 2013; Henriksen et al., 2015; Patil et al., 2015). Appropriate identification of disease phenotypes is critical for the development of personalized prevention of allergic diseases. There is accumulating evidence that environmental factors play important roles in the rapid increase of the prevalence. It remains unclear to what extent and at which stages such environmental exposures can be combined to predict specific allergies in order to develop personalized health management plans.

Several approaches to predict such allergic diseases in children and adults have been proposed in recent years. Different population strata are one of the key reasons affecting the performance of the prediction models; for instance, Castro-Rodriguez et al. (2000) proposed a rule-based asthma predictive index for predicting subsequent asthma among infants who have a history of wheezing, attaining sensitivity of around 0.4 and 0.8 specificity at different time points. In 2013, Chatzimichail et al. (2013) devised a machine learning approach based on previous symptoms, medications, allergen sensitization, and lung function, which can be used to predict asthma outcome for symptomatic preschool children with sensitivity and specificity of 0.95, respectively. The recent work by Prosperi et al. (2014) combined demographic, clinical and laboratory features, genetic profiles and environmental exposures with a large attribute set, and used various machine learning techniques to build allergic phenotypes prediction models for adult patients. However, studies that are focused on predicting children allergies using pregnancy and children early life features, are lacking.

Emerging evidences (Glinianaia et al., 2004; Jackson-Browne et al., 2019; To et al., 2019) show that exposure to air pollution during pregnancy is associated with elevated risk of adverse birth outcomes with long-term health implications. Environmental exposure (Yeatts et al., 2006) is one of the main factors for the development of allergic diseases. In recent years, epidemiological studies (Mortimer et al., 2008; Sbihi et al., 2015; To et al., 2019) have demonstrated that air pollution has been linked with childhood allergies. Mounting evidence suggests that early-life exposure to environmental pollution, especially to air pollution, is an important determinant in the later development of allergic diseases in children. Despite the importance of early-life exposure in the development of allergy, information on environmental exposures, particularly during pregnancy, on allergic diseases early in life has rarely been considered to build childhood allergies prediction models. The scarce epidemiological evidence on the predictive ability of childhood allergic diseases in data mining methods among large heterogeneous attribute sets including air pollution, environmental exposures, and nutritional supplements, warrants further investigation.

In this work, using a rich dataset comprising a longitudinal birth cohort among pregnant women who had undergone prenatal examinations at nine selected private maternity hospitals across Taiwan, different prediction models for specific allergic diseases were developed, and we analyzed their prognostic factors. The aim of this research was to identify the extent to which such early-life heterogeneous attributes contribute and are combined toward the prediction of allergic diseases in children.

2. Material and methods

2.1. Data source

This study was based on a longitudinal follow-up birth cohort data collected between 2000 and 2005. A total of 2098 subjects, up to 14 years old, were included with 360 attributes from heterogeneous information sources, including questionnaires and doctors' diagnoses, and followed-up every 2–3 years.

2.2. Participants recruitment

In 2000, a longitudinal birth cohort study was conducted among pregnant women who had undergone prenatal examinations at nine selected maternity hospitals and medical centres located in eight areas, including one in Taipei, one in central Taiwan, one in eastern Taiwan, and six in the south part of Taiwan (Lin et al., 2011; Lin et al., 2004; Wen et al., 2011). Pregnant women who had undergone prenatal examinations at these hospitals and without pregnancy complications (such as eclampsia) were invited. Interviews were performed after pregnant women gave written informed consent to participate in the study. The infants born after July 2000 were consecutively recruited. Venous blood of the pregnant women and umbilical cord blood of their newborns were collected by a nurse and an obstetrician, respectively. In total, 2098 mother–newborn pairs were recruited between July 2000 to July 2005. Our protocols were approved by the Research Ethics Committee of the National Health Research Institutes in Taiwan, National Cheng Kung University Hospital, and Chung Shan Medical University Hospital, according to the respective responsibility for subjects' recruitment.

2.3. Questionnaire

Pregnant women were asked about their demographic and general characteristics, such as social economic status, life style, medical history, environmental factors at home (such as environmental tobacco smoke, presence of cockroaches, incense burning, carpets, pets, or fungi on walls), family history of allergic diseases (atopic dermatitis, asthma, and allergic rhinitis), diet during pregnancy, and reproduction history. Maternal self-reported mental status during pregnancy was also included in the prenatal questionnaire by using the modified Chinese version of Short Form 36 Health Survey (IQOLA SF-36 Taiwan standard

Version 1.0; Copyright© 1995 New England Medical Centre Hospital, Inc. All rights reserved.). Newborn birth outcome data (weight, height, head circumference, and gestational age) were collected from the hospital records by nurses after childbirth.

Children were followed up by well-trained interviewers after permission was obtained at the age of 1, 2, 7, 12, and 14 years. The mother or main caregiver was asked about the child's diet, life style, growth situation, medical history, and environmental exposure. Children were considered as having- or not having-allergic diseases according to the response to the question "Did your child ever have physician-diagnosed AD or AR?" A "Yes" response to the question in any one of the follow-up interviews was considered that the child had ever had AD or AR.

2.4. Blood sample measurement

The levels of cord blood immunoglobulin E (cblgE) and maternal total IgE (mIgE) were determined using the Pharmacia UniCap IgE assay system (CAP System, Pharmacia Diagnostics, Sweden). The limit of the detective value was 0.35 IU/mL for cblgE and 2.0 IU/mL for mIgE.

Cord plasma samples were also sent to the National Taiwan University for the measurement of nine polyfluoroalkyl substances (PFASs), including perfluorohexanoic acid (PFHxA), perfluoroheptanoic acid (PFHpA), perfluorohexanesulfonic acid (PFHxS), perfluorooctanoic acid (PFOA), perfluorooctane sulfonate (PFOS), perfluorononanoic acid (PFNA), perfluorodecanoic acid (PFDeA), perfluoroundecanoic acid (PFUA), and perfluorododecanoic acid (PFDoA). The experimental procedure is as described in a previous study (Lien et al., 2011). The measurement of the cord plasma PFAS concentration was performed by an Agilent-1200 high performance liquid chromatography system (Agilent, Palo Alto, CA, USA) coupled with a triple-quadrupole mass spectrometer (Sciex API 4000, Applied Biosystems, Foster City, CA, USA). The limit of quantitation values (LOQ) for the plasma PFASs were 0.25, 0.28, 0.08, 0.45, 0.10, 0.11, 0.19, 0.13, and 0.07 ng/mL for PFHxA, PFHpA, PFHxS, PFOA, PFNA, PFOS, PFDeA, PFUA, and PFDoA, respectively.

Participants' cord plasma concentrations of PFAS and IgE levels below the detection limit (DL) were replaced by half-of-detection-limit values. The percentage below DL ranged from 0.7% to 97.4%. The detection rates of cord blood measurements less than 60% were excluded from the analysis, i.e. PFHxA (2.6%), PFHpA (2.2%), and PFDeA (29.5%).

2.5. Construction of datasets for machine learning analysis

The dataset used to construct the models was exclusively composed of attribute data collected before 2 years of age. Before attribute data cleaning, we excluded the subjects whose percentage of missing variables were greater than 70%. The total retained data size came to 1439 samples. The following attribute data cleaning was conducted on this dataset.

Three types of attributes were included in this dataset, i.e. questionnaire attributes, clinical attributes, and air pollution attributes. Multiple choice responses from the questionnaire were cleaned by redefining the original features into one-hot features. Newly-added city attribute was extracted from the address in each object. The missing value filling methods were based on the calculation of statistic values in the training dataset and then applied on both the training and testing datasets. A total of 53 categorical attributes and 22 numerical attributes contained missing value. Missing values in the numerical attributes, such as cblgE, were filled using the average of each attribute. Missing values of categorical variables in the questionnaire were filled using the mode.

To obtain AR datasets, we excluded the subjects who were diagnosed before 2-years-old because they were hardly precisely diagnosed with AR at this early age. After data cleaning and selection of three target datasets, we obtained AD dataset ($n = 1399$) and AR dataset ($n = 975$) (Fig. 1).

2.6. Air pollution exposure assessment

2.6.1. Identification of air pollution for each subject

To identify historical air pollution exposure, the address of each subject was geocoded with Google Geocoding application program interface (Geocoding API) and three subjects whose address could not be identified by geocoder were excluded. Air pollution concentrations were retrieved from the Environmental Protection Administration, Executive Yuan, Taiwan (Taiwan EPA). By applying the latitude and longitude, we were able to infer the hourly air pollution concentration for each subject 1 year before childbirth by using weighted k -nearest neighbour (Macleod et al., 1987) ($k = 3$) to get the three nearest air pollution stations and assign a value by the weighted average. Total air pollution components included carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and particulate matter 10 (PM_{10}). For each subject in the original dataset, the collected air pollution information is linked by the patients' identification number.

2.6.2. Air pollution feature extraction methods

We collected air pollution values within 1 year before childbirth. Air pollution feature extraction was conducted by the following steps. First, we transformed the concentration of the five air pollutants into Individual Air Quality Index (IAQI) values according to the Air Quality Index (AQI) standard announced by Taiwan EPA (Table S1) via the equation.

$$IAQI = \frac{IAQI_{high} - IAQI_{low}}{C_{high} - C_{low}} (C - C_{low}) + IAQI_{low}$$

where:

IAQI = the individual air quality index,

C = the pollutant concentration,

C_{low} = the concentration breakpoint that is $\leq C$,

C_{high} = the concentration breakpoint that is $\geq C$,

$IAQI_{low}$ = the index breakpoint corresponding to C_{low} ,

$IAQI_{high}$ = the index breakpoint corresponding to C_{high} .

We then calculated the mean value within 1 month from the hourly IAQI, and computed monthly IAQI difference between each two nearest months ($n = 55$, 5 air pollutants multiply by the 11 monthly differences). Second, the monthly average values of IAQI values and IAQI differences across the 1 year for each air pollutant ($n = 10$, 5 air pollutants multiply by 2 average values). Third, we discretized the hourly IAQI value of each air pollution components, including CO, O_3 , NO_2 , SO_2 and PM_{10} , into values ranging between 1 and 5, denoting the severity of air pollution in the public IAQI and the AQI guide announced by Taiwan EPA (2021). Then, we calculated the monthly average IAQI class ($n = 5$, the mentioned 5 air pollutants). Finally, after air pollution feature extraction, we added 70 different air pollution features in AD and AR dataset, respectively.

2.7. Model construction

Among machine learning methods, logistic regression (LR), random forest (RF), k -nearest neighbours (KNN), XGboost (XGB), Naïve Bayes (NB), and neural network (NN) were applied to build the prediction models for the allergies. RF (Breiman, 2001) (400 trees, maximum depth of the tree: 5 and Gini function to split nodes), XGB (Chen and Guestrin, 2016) (400 trees, learning rate: 0.1 and maximum depth of tree: 5), Decision Tree, LR (Fan et al., 2008) and KNN ($k = 3$) were used to build the prediction models for AD and AR. Moreover, the Synthetic Minority Over-sampling Technique (Chawla et al., 2002) was conducted to solve the imbalance class problem in this research. For the evaluation metrics and comparison, four metrics including sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUROC), were used to evaluate these models. During

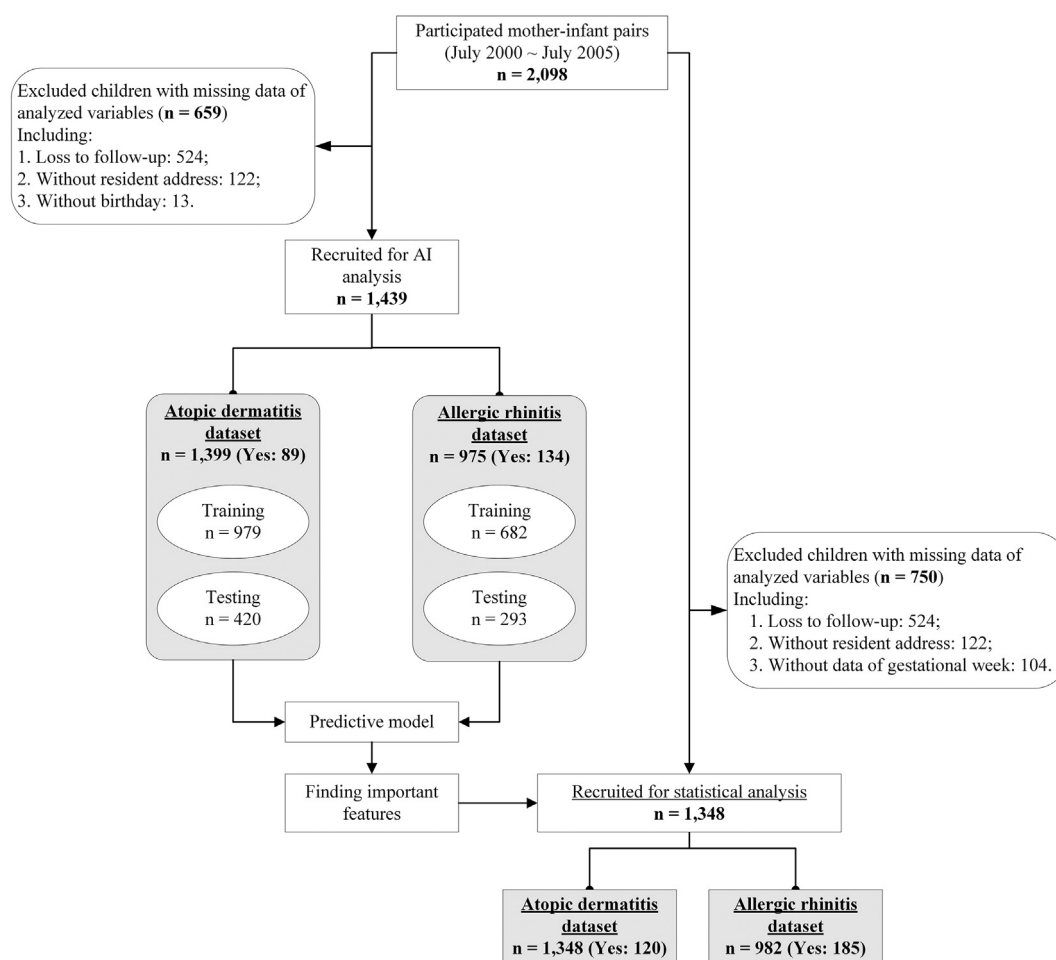


Fig. 1. Flow chart for participants' recruitment.

the evaluation, we randomly split the data using the “70/30” policy, and then ran the models 10 times to get the average results. In the training step, we split 20% of the data in the training set as the validation set to select the best prediction threshold to apply on the testing set, and to obtain the final comparison tables. All analyses were carried out using Python Software Foundation (Python Language, version 3.7.0). A *P*-value of <0.05 was considered statistically significant.

2.8. Statistical analysis

All the statistical analyses were performed using JMP, version 13 (SAS Institute Inc., Cary, NC, USA). Prenatal exposure pattern of the five air pollutants was identified by conducting latent class analysis (LCA). Odds ratios (ORs) and 95% confidence intervals (CIs) were applied to evaluate statistical associations between potential risk factors and childhood allergic diseases according to the machine learning results. Backward stepwise regression was applied to build the multiple regression model and to select the predictive factors into the model. Factors with the highest *p*-value were eliminated one by one till all the remaining factors had *p*-value <0.05 and were considered as significant predictors for childhood allergic diseases. A *p*-value of <0.05 in two-sided test was considered statistically significant.

3. Results

Overall, a total of 1439 mother-infant pairs with sufficient biological samples were recruited for the machine learning analysis after excluding the mother-infant pairs that were lost to follow-up (*n* = 659) (Fig. 1). The prevalence of allergic diseases in children was 6.4% (*n* =

89) for AD and 13.7% (*n* = 134) for AR. The characteristics of the children and parents for children with (*n* = 174) and without (*n* = 1265) any one of allergic diseases are shown in Table 1. Significant differences occurred with a higher prevalence of the first child in the birth order of the parents, parental education >12 years, family income ≥33.3 (10³USD), and parental atopy in children with diseases than in children without diseases (Table 1). Table S2 outlines the characteristics of children included (*n* = 1439) and excluded (*n* = 629) from the machine learning analysis. The percentages of the first child of parents, parental education ≥10 years and mother ever having atopic diseases were higher in included children as compared to excluded children. Lower percentages of mother smoking before pregnancy and during pregnancy were also observed in included children.

Table 2 demonstrated differences in air pollutant exposure in children with and without allergic diseases. Among children with AD, higher IAQI levels of NO₂ and CO, and lower IAQI levels of O₃ and PM₁₀ were observed, compared to children without AD within one year before childbirth. Similar results were also found in children with AR compared to children without AR (Table 2).

To demonstrate whether heterogeneous information can be combined to reveal three specific allergic diseases, the different learning methods (i.e. LR, RF, KNN, XGB, NB and NN) introduced above were applied to the full attribute set for predicting AD and AR, respectively. To evaluate the performance of the predictive models for each allergic disease, AUROC, sensitivity, specificity and accuracy were considered as the primary measurements for the high-class imbalance in the population data. Overall, RF yielded the most predictive model on AR prediction (average AUROC for AD and AR were 83% and 84%, respectively). Fig. 2 shows the averaged ROC curves for the different AD and AR

Table 1
Subjects' demographic and general characteristics in machine learning analysis.

Characteristics	N (%) or mean (SD)			p-Value [§]
	All (n = 1439)	AD/AR		
		With (n = 174)	Without (n = 1265)	
Children				
Sex				
Boys	706 (49.1)	75 (10.6)	631 (83.4)	0.545
Girls	644 (44.8)	62 (9.6)	582 (90.4)	
Birth weight	1332 (92.6)	3146.7 (429.3) [§]	3089.0 (393.8) [§]	0.123
Gestational age	1336 (92.8)	39.12 (1.27) [§]	38.89 (1.29) [§]	0.174
Birth order				
1st	641 (44.5)	84 (13.1)	557 (86.9)	0.042
2nd	456 (31.7)	47 (10.3)	409 (86.7)	
≥3rd	190 (13.2)	13 (6.8)	177 (83.2)	
Parents				
Maternal age at childbirth (year)	818 (56.8)	28.46 (3.90) [§]	27.45 (4.58) [§]	0.011
Maternal BMI before pregnancy	1174 (81.6)	20.55 (2.94) [§]	20.92 (3.15) [§]	0.271
Maternal weight gain during pregnancy	1230 (85.5)	12.56 (6.05) [§]	13.33 (5.36) [§]	0.185
Maternal education				
≤9 years	229 (15.9)	8 (3.5)	221 (96.5)	<0.0001
10–12 years	700 (48.6)	70 (10.0)	630 (90.0)	
>12 years	492 (34.2)	95 (19.3)	397 (80.7)	
Paternal education				
≤9 years	252 (17.5)	15 (5.6)	237 (94.1)	<0.0001
10–12 years	666 (46.3)	62 (9.3)	604 (90.7)	
>12 years	501 (34.8)	94 (18.8)	407 (81.2)	
Family income (10 ³ USD)				
<20	856 (59.5)	78 (9.1)	778 (90.9)	<0.0001
20–33.3	380 (26.4)	64 (16.8)	316 (83.2)	
≥33.3	135 (9.4)	27 (20.0)	108 (80.0)	
Maternal atopy				
Yes	252 (17.5)	50 (19.8)	202 (80.2)	0.0005
No	990 (68.8)	114 (11.5)	876 (88.5)	
Paternal atopy				
Yes	260 (18.1)	50 (19.2)	210 (80.8)	0.0003
No	963 (66.9)	105 (10.9)	858 (89.1)	
Smoking during pregnancy				
Yes	32 (2.2)	3 (9.4)	29 (90.6)	0.593
No	1364 (94.8)	171 (12.5)	1193 (87.5)	
Smoking before pregnancy				
Yes	87 (6.0)	10 (11.5)	77 (88.5)	0.82
No	1323 (91.9)	163 (12.3)	1160 (87.7)	

Abbreviation: AD, atopic dermatitis; AR, allergic rhinitis; BMI, body mass index; SD, standard deviation; USD, US Dollar.

Some numbers do not add up to total n because of missing values.

[§] P value was calculated by Kruskal-Wallis tests for continues variables and χ^2 test for categorical variables compared between children with and without allergic diseases.

[§] mean (SD).

Table 2

Monthly mean (SD) of air pollutant individual air quality index (IAQI) within one year before childbirth among children with and without allergic disease.

IAQI	Mean (SD)					
	AD			AR		
	Yes (n = 89)	No (n = 1310)	P-value	Yes (n = 134)	No (n = 841)	P-value
NO ₂	20.73 (4.08)	15.70 (4.73)	<0.0001	20.76 (4.21)	15.23 (4.57)	<0.0001
CO	8.26 (1.59)	6.47 (1.83)	<0.0001	8.12 (1.62)	6.33 (1.80)	<0.0001
SO ₂	4.04 (0.86)	4.09 (1.63)	0.965	4.07 (1.09)	4.07 (1.70)	0.926
O ₃	17.62 (3.91)	21.82 (3.95)	<0.0001	17.74 (4.21)	22.11 (3.90)	<0.0001
PM ₁₀	46.70 (5.34)	53.60 (8.63)	<0.0001	46.84 (5.40)	54.42 (8.54)	<0.0001

Abbreviation: AD, atopic dermatitis; AR, allergic rhinitis; PM10, particulate matter with a 50% cut-off aerodynamic diameter of 10 mm; SD, standard deviation.

disease outcomes, while Table S3 shows the average AUROC, specificity, sensitivity, and accuracy in detail; both were evaluated 10 times in the random “70/30 split” policy validation.

Based on these results, the RF method was retained and tested using different data pre-processing approaches with the environmental variables, because RF has good interpretability and predictability. Specifically, without environmental information and different environmental information processing methods to identify which method for processing environmental features was contributing to the increase in AUROC, sensitivity, or specificity with respect to each outcome prediction. Table S4 shows the sensitivity, specificity, AUROC, and accuracy of different data processing methods, both were estimated 10 times in the random “70/30 split” policy validation. For all the models, RF models using the whole environmental features, yielded the best performance. In detail, the model using environmental features outperformed the one without environmental features by up to 0.11 in AUROC. Fig. 3 shows the ROC curves of the RF model for the three diseases, considering each different data pre-processing methods.

When looking at the feature importance, the most relevant attribute reflected the overall performance of the RF trained, using the feature subsets. Table 3 and Table S5 reports the details of the feature

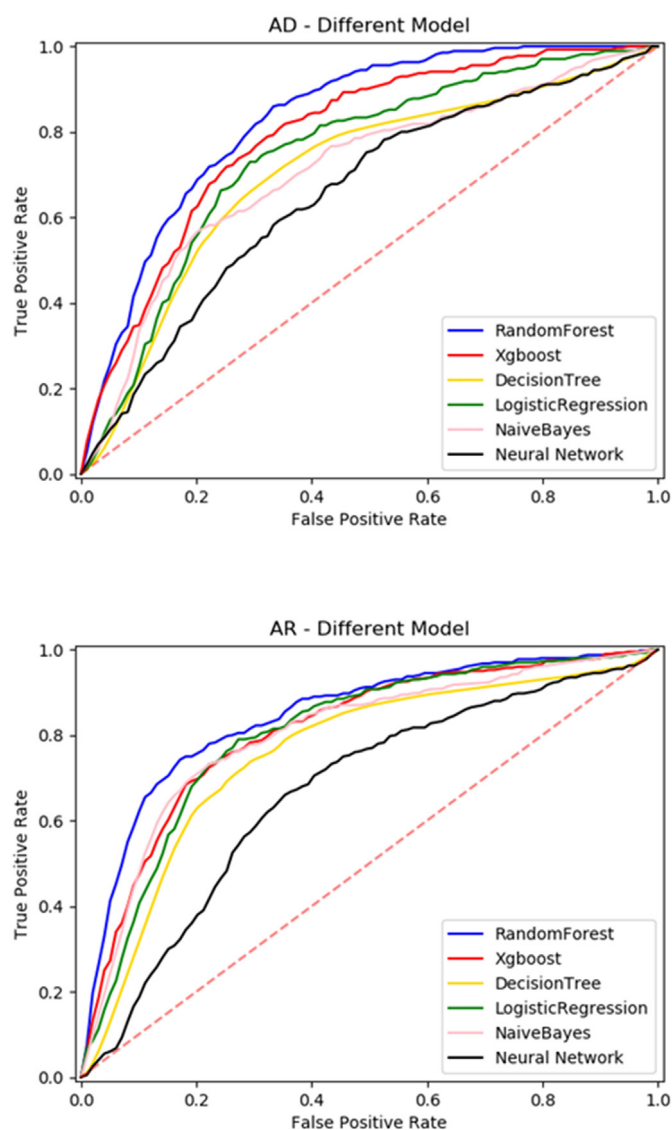


Fig. 2. Comparison of area under the receiver operating characteristic curve from different machine learning methods.

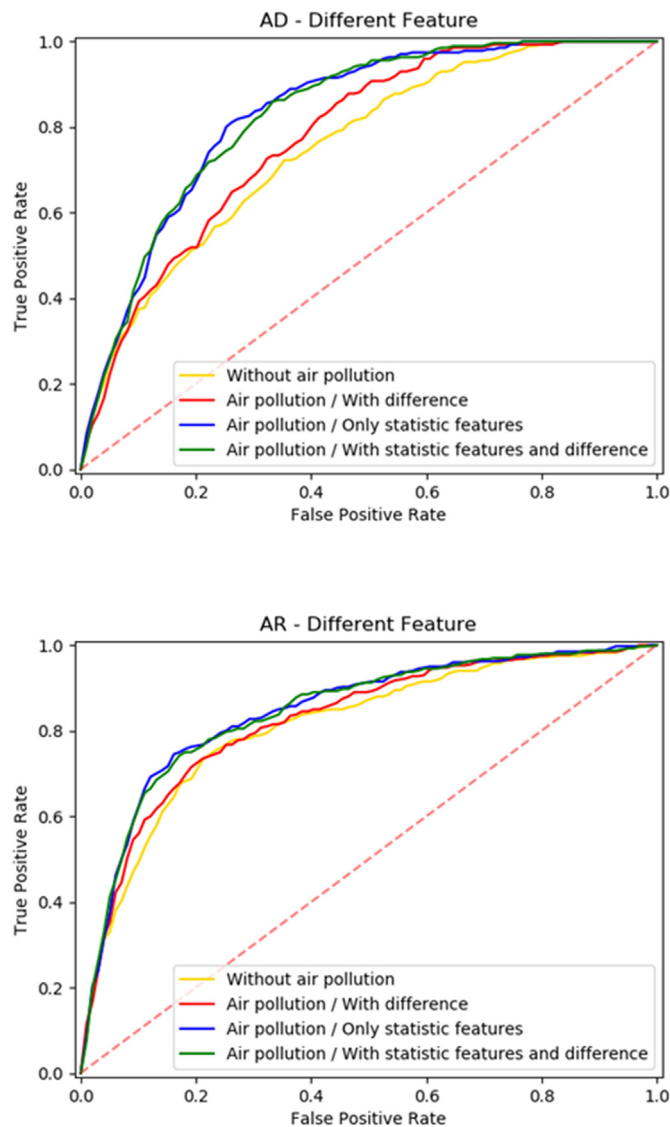


Fig. 3. Comparison of area under the receiver operating characteristic curve from random forest performance using various input features.

importance and correlation evaluation by RF with and without environmental factors, respectively. For Table 3 and Table S5, we highlighted that the features reached the significance level (P -value < 0.05), and we observed that the feature outputs by RF were useful. In Table S5, parents' educational level was the most important feature of AD and AR. Packing of hot food via plastic bags also showed positive correlation with AR, but negative correlation with AD. Of note, when considering environmental factors, the IAQI class of air pollutants ranked highest according to the RF. In Table 3, the most important features reported by RF were air pollutant features, such as the monthly average of IAQI classes of NO_2 , O_3 , and PM_{10} within 1 year before childbirth. It is worthy to notice that environmental factors (Left side) ranked the highest than those of the other features (Right side) (Fig. S1). The results showed that environmental factors contributed to the predictive model building for childhood allergic diseases to a certain extent.

On the basis of the RF results, we then used the typical statistical analysis in epidemiological studies to assess the association between prenatal ambient air pollutant exposure and allergic diseases in children. A total of 1348 mother-infant pairs were recruited for the final statistical analysis, including 1348 mother-infant pairs in the AD dataset and 982 pairs in the AR dataset (Fig. 1). The prevalence of allergic diseases was 8.9% ($n = 120$) for AD and 18.8% ($n = 185$) for AR in children,

before the age of 14 years. Table S2 also outlines the characteristics of children included ($n = 1348$) and excluded ($n = 750$) in the statistical analysis. In addition, the characteristics of the children and parents for children with ($n = 257$) and without ($n = 1091$) any one of allergic diseases are shown in Table S6. We then grouped air pollutant IAQI levels of AD and AR datasets by tertile, and applied the LCA to identify the potential exposure pattern to air pollutant, due to the high correlation between air pollutants (Table S7). The proportion of air pollutant IAQI groups among latent clusters was shown in Table S8. Three classes were identified via the LCA, including all the middle-level class (cluster 1), high O_3 and PM_{10} class (cluster 2), and high NO_2 , CO , and SO_2 class (cluster 3) according to the monthly mean and SD of air pollutant IAQI levels during maternal pregnancy as shown in Table 4.

The association between the three clusters of prenatal air pollutant exposure and childhood allergic diseases is shown in Table 5. Children in the cluster 3 had an increased odds of AD and AR development (OR [95% CI] = 2.98 [1.93–4.61] and 3.74 [2.53–5.52], respectively). After adjustment for potential covariates, children in the cluster 3 still had the increased odds of AD and AR development (OR [95% CI] = 3.11 [1.92–5.02] and 3.61 [2.38–5.48], respectively). No significant association was found in children in the cluster 2 compared to those in cluster 1.

4. Discussion

We investigated the ability of different types of machine learning models to predict AD and AR, using a heterogeneous set of attributes in children from a longitudinal birth cohort study. Models were compared in terms of performance and interpretability. Different feature inputs were evaluated to understand the determinants of different allergic diseases. The use of RF in model building yielded better AUROC, sensitivity, and specificity than other methods. This might be due to the ability of RF to model different variable interactions automatically. Moreover, the difference between RF and other methods was minimal in predicting AR phenotypes. However, the results showed that AR and AD can be predicted well by the RF model.

By extracting important features and correlation for different outcomes, we observed that environmental factors are very important in predicting childhood allergic diseases. Most air pollutant features were reported by the RF method, when we considered the environmental factors for building the prediction model, and these features were shown to reach significance level in most cases. However, regarding the importance and correlation of input features, our findings confirm the important contribution of air pollutants.

The relationship between air pollution and allergic diseases has been well reported. Air pollutant exposure was associated with the occurrence of respiratory allergic diseases (RAD) and could induce and aggravate the symptoms of RAD (Huang et al., 2015b; Wright and Brunst, 2013). Sbihi et al. (2015) found that air pollutant exposure, especially CO , NO , and NO_2 , was associated with the development of asthma in preschool children. CO , NO_x , and O_3 exposure were also associated with the occurrence of eczema, recurrence eczema, or flexural eczema in children (Lee et al., 2008; Lu et al., 2017). Children with higher prenatal CO exposure also had higher risk of AD development before 6 months of age (Huang et al., 2015a). Lu et al. (2017) reported that exposure to higher levels of traffic-related air pollution (i.e. NO_2 , SO_2 , and PM_{10}), both during pre-conceptional and perinatal period, was associated with increased risk of early childhood eczema. In the present study, our results were consistent with those of previous studies, and prenatal air pollutant exposure was associated with the development of allergic diseases in our long-term follow-up children.

In addition, we also found that children with higher O_3 and PM_{10} exposure levels was negatively associated with childhood AD and AR. In fact, the five observed air pollutants were highly correlated (Table S7). Among them, PM_{10} , O_3 were negatively correlated with NO_2 and CO . When we put all 5 air pollutants in the model, the O_3 became insignificant. Thereby we think the negative association between higher prenatal

Table 3

Feature importance (the top 20th) and correlation evaluation using random forest model with air pollutant data.

Features	AD			AR		
	Importance	Correlation	P-value	Importance	Correlation	P-value
The monthly average of IAQI class of PM ₁₀	0.137	−0.191	<0.0001	0.127	−0.311	<0.0001
Parental education	0.091	0.143	<0.0001	0.042	0.181	<0.0001
The monthly average of IAQI class of NO ₂	0.089	0.212	<0.0001	0.066	0.302	<0.0001
The monthly average of IAQI class of O ₃	0.077	−0.186	<0.0001	0.084	−0.283	<0.0001
Packaging hot food via plastic bag	0.048	−0.141	<0.0001	0.054	0.208	<0.0001
Maternal vigour in pregnancy	0.039	−0.093	0.001	—	—	—
Child sex	0.029	0.051	0.050	0.014	0.077	0.014
Pet at home	0.024	−0.097	<0.0001	0.036	−0.166	<0.0001
The monthly average of IAQI difference of CO	0.023	0.050	0.068	—	—	—
The monthly average of IAQI difference of NO ₂	0.020	0.120	<0.0001	0.011	0.153	<0.0001
Paternal atopy	0.016	0.054	0.081	—	—	—
Maternal vitality in pregnancy	0.016	−0.034	0.203	—	—	—
Cord blood IgE	0.013	0.044	0.239	0.014	0.007	0.794
Frequency of edible shellfish and oysters	0.011	−0.032	0.172	—	—	—
Intake vitamin supplement	0.011	0.058	0.051	0.009	0.086	0.015
Frequency of freshwater fish consumption	0.011	−0.105	0.0001	—	—	—
Birthweight	0.010	0.040	0.133	—	—	—
Number of pregnancy	0.008	−0.060	0.016	0.011	−0.105	0.0002
Frequency of edible squid fish	0.008	−0.027	0.243	—	—	—
Incense burning at home	0.008	0.051	0.018	—	—	—
Cord plasma PFOA	—	—	—	0.009	0.081	0.098
Cockroach at home	—	—	—	0.027	0.148	<0.0001
The monthly average of IAQI difference of O ₃	—	—	—	0.013	−0.179	<0.0001
Maternal IgE	—	—	—	0.016	−0.028	0.224
The monthly average of IAQI class of CO	—	—	—	0.012	−0.129	<0.0001
Maternal atopy	—	—	—	0.010	0.129	0.001
Birth height	—	—	—	0.010	0.090	0.003
Add vitamins to baby food	—	—	—	0.012	−0.143	<0.0001
Self-report air pollution surrounding residence at 2 years old	—	—	—	0.011	−0.112	<0.0001

“—” indicated that the feature was not the top 20th important features for the disease.

Abbreviations: AD, atopic dermatitis; AR, allergic rhinitis; IAQI, individual air quality index; IgE, immunoglobulin E; PFOA, perfluorooctane sulfonate; PM₁₀, particulate matter with a 50% cut-off aerodynamic diameter of 10 mm.

PM₁₀ and O₃ might result from the low prenatal exposure of NO₂ and CO in those with high O₃. For example O₃ might be high in the forest with sunshine but low traffic.

The biological mechanism of prenatal air pollution exposure on allergic diseases is still under investigation. Air pollutants can cross the placenta and affect the foetal development, and their health after birth (Luyten et al., 2018; Tsai et al., 2019). Wright and Brunst (2013) reported that prenatal and postnatal air pollutant exposure affects the development of immune system, nervous system, and neuroendocrine system in foetus and children. It also contributes to adverse health outcome in the respiratory tract of children after birth, including respiratory inflammation, lung function, sensitization of respiratory tract, and asthma occurrence (Wright and Brunst, 2013).

The effect of air pollution exposure on the immune system is one of the possible mechanisms of developing allergic disease (Luyten et al., 2018; Tsai et al., 2019). In a France birth cohort study (Baiz et al., 2011), women with air pollution exposure before and during pregnancy showed effect on the foetal immune function with altered distribution of the natural killer cells and T lymphocyte in the cord blood. In a

Canadian birth cohort study (Ashley-Martin et al., 2016), elevated concentrations of both interleukin (IL) 33 and thymic stromal lymphopoietin in the cord blood were observed in girls with higher maternal NO₂ exposure. Skewed T-helper 2 (Th₂) immune response with elevated levels of IgE and IL-4 characterized the immune response in allergic diseases (Akdis et al., 2020). In an animal study (Yue et al., 2017), maternal NO₂ exposure caused Th₂ polarization in offspring, including elevated IL-4 and IL-13 levels and suppressed interferon- γ expression. Airway inflammation with increasing eosinophil cells, an asthma related sign, was also observed in offspring with maternal NO₂ exposure. Moreover, IL-4 promoter demethylation that related to Th₂ polarization was also demonstrated in offspring with maternal NO₂ exposure and postnatal allergen sensitization (Yue et al., 2017).

Prenatal air pollution exposure effect on foetal development is another potential mechanism of allergic disease development. Respiratory airway development begins during the second trimester of pregnancy (Shi et al., 2007) and harmful effect occurring during this crucial period might have long-term effect on the respiratory health. Exposure to higher levels of NO₂ during pregnancy was associated with reduced

Table 4

Mean (SD) of average monthly air pollutant individual air quality index levels during pregnancy among latent clusters.

	AD (n = 1348)				AR (n = 982)			
	Cluster 1 (n = 566)	Cluster 2 (n = 454)	Cluster 3 (n = 328)	p-Value ^a	Cluster 1 (n = 417)	Cluster 2 (n = 333)	Cluster 3 (n = 232)	p-Value ^a
Conceived indicators	All mild levels	High O ₃ and PM ₁₀	High NO ₂ , CO, and SO ₂		All mild levels	High O ₃ and PM ₁₀	High NO ₂ , CO, and SO ₂	
NO ₂	15.34 (1.79)	10.58 (1.87)	22.38 (2.37)	<0.0001	15.27 (1.88)	10.52 (1.91)	22.22 (2.58)	<0.0001
CO	6.37 (1.53)	4.87 (0.64)	8.71 (1.48)	<0.0001	6.33 (1.51)	4.86 (0.65)	8.66 (1.50)	<0.0001
SO ₂	4.60 (1.24)	2.84 (0.93)	4.96 (1.87)	<0.0001	4.55 (1.26)	2.78 (0.91)	4.97 (1.92)	<0.0001
O ₃	22.83 (1.58)	24.75 (3.25)	16.85 (2.62)	<0.0001	22.79 (1.51)	24.72 (3.25)	16.85 (2.77)	<0.0001
PM ₁₀	54.59 (7.55)	56.33 (8.06)	46.58 (9.16)	<0.0001	55.08 (7.50)	56.18 (8.30)	47.47 (9.60)	<0.0001

Abbreviations: AD, atopic dermatitis; AR, allergic rhinitis; PM₁₀, particulate matter with a 50% cut-off aerodynamic diameter of 10 mm; SD, standard deviation.

^a p-Value was analyzed by Kruskal-Wallis Tests.

Table 5

Adjusted odds ratio (AOR) for childhood allergic diseases according to prenatal exposure cluster of air pollutant by simple and multiple logistical regression analysis

Latent clusters	AD		OR (95%CI)	AOR (95%CI) ^a	AR		OR (95%CI)	AOR (95%CI) ^b
	N (%)				N (%)			
	All (n = 1384)	Yes (n = 120)			All (n = 982)	Yes (n = 185)		
Cluster 1	566 (42.0)	38 (6.7)	Reference	Reference	417 (42.5)	55 (13.2)	Reference	Reference
Cluster 2	454 (33.7)	24 (5.3)	0.78 (0.45–1.31)	1.00 (0.57–1.74)	333 (33.9)	46 (13.8)	1.05 (0.69–1.61)	1.01 (0.64–1.61)
Cluster 3	328 (24.3)	58 (17.7)	2.98 (1.93–4.61)***	3.11 (1.92–5.02)***	232 (23.6)	84 (36.2)	3.74 (2.53–5.52)***	3.61(2.38–5.48)***
p for trend ^c			<0.0001				<0.0001	

Abbreviations: AD, atopic dermatitis; AOR, adjusted OR; AR, allergic rhinitis; OR, odds ratio.

^a AOR was adjusted by parental education and paternal atopy.^b AOR was adjusted by maternal atopy and parental education.^c p-Value was calculated by Cochran Armitage trend test.

lung function in children aged 4.5 years, especially among allergic children (Morales et al., 2015). In a 17-year follow-up birth cohort study in Hong Kong, lower forced expiratory volume in 1 s (FEV₁), forced vital capacity (FVC)/FEV₁ ratio, and forced expiratory flow at 25%–75% of the pulmonary volume (FEF_{25%–75%}) were found in adolescents with higher in utero, infancy, and toddlerhood NO₂ exposure (He et al., 2019). In addition, like the respiratory airway development, foetal skin barrier becomes functional and develop at about 20 weeks of pregnancy. The skin barrier dysfunction is considered the first step in AD development (Kim and Leung, 2018). Prenatal air pollution exposure induced oxidant stress might cause the damage to the skin barrier in neonates (Nagiah et al., 2015; Pastore and Korkina, 2010), and therefore increase the risk of AD occurrence.

The present study has some strengths. First, as a longitudinal follow-up birth cohort design, we could evaluate the temporal sequence between potential risk factors and health effect. Second, using machine learning approach we could consider heterogenetic factor together. Third, we established a predictive model for allergic diseases by prenatal and early postnatal factors. The public could evaluate the probability of developing allergic disease and therefore take action, such as avoiding air pollutant exposure, for early prevention of the development of allergic disease in children.

Our study had some limitations of concern. First, our study did not consider the in-house environmental factors and genetic attributes (i.e. genotypes and epigenome). This may decrease the discriminating ability of the predictive model of allergic disease due to genetic profile as one of the important risk factors for disease development (Ober and Yao, 2011). Parental history of allergic diseases was also considered in the present study and would likely represent the hereditary background of children. In addition, our analysis was not focused on genetic markers and in-door environmental attributes. In further study, collecting more data that include these attributes which will improve the predictive ability for allergic diseases in children is warranted. Second, childhood allergic diseases were determined based on the main caretaker or maternal reports. This might have caused misclassification of childhood allergic diseases. However, the interviewees were unaware of the main hypothesis, and the misclassification was likely non-differential among the different attributes groups. Thus, the bias related to the study results due to this misclassification was less likely. Third, participated children were from the general population and without allergic disease-related biomarker measurements (i.e. cytokines) in the model; this might have also resulted in decreased discriminatory ability of the predictive model. Fourth, personal air pollutant exposure level was evaluated based on the three nearest air pollution monitoring stations via weighted k- nearest neighbour method. This might not represent the accurate individual exposure levels as compared to a personal environmental monitoring, and was a surrogate of the true exposure levels. Fifth, due to Taiwan EPA PM_{2.5} monitoring data was only available after 2006, we could not investigate the effect of prenatal PM_{2.5} exposure on childhood AD and AR in

our study. Further study is necessary for birth cohort study established after 2006 when PM_{2.5} monitoring data is available. Sixth, a total of 75 attributes with missing value were filled by the mean for numerical attributes and the mode for categorical attributes. Among them, 15 attributes (>70%) out of 22 significant attributes had the missing percentage less than 20%. The conclusion was consistent no matter we put the feature in the model with the percentage of missing value greater than 30% or not. Future research is directed to apply this approach in large and complete long-term follow-up birth cohort study. Finally, participants were from nine medical centres, hospitals, or obstetrics and gynaecology clinics located in eight areas of Taiwan. Therefore, our findings, generalised to all children in Taiwan should be with caution.

5. Conclusions

We applied machine learning method to establish a predictive model for childhood allergic diseases. Among the recruited factors, our data demonstrated that incorporating environmental information into the model can boost the performance of the predicting ability of the phenotypes of AD and AR in children. Prenatal exposure to NO₂ and its concatenation changes over time were significant predictors for AD and AR till adolescence. Future studies might consider additional factors/attributes and harness this longitudinal data in the prediction of allergic diseases for children.

CRedit authorship contribution statement

Yu Huang: Methodology, Formal analysis, Resources, Data curation, Writing – original draft, Visualization. **Hui-Ju Wen:** Methodology, Formal analysis, Resources, Data curation, Writing – original draft, Visualization. **Yue Leon Guo:** Conceptualization, Resources, Writing – review & editing. **Tzu-Yin Wei:** Data curation. **Wei-Cheng Wang:** Data curation. **Shin-Fen Tsai:** Investigation, Project administration. **Vincent S. Tseng:** Conceptualization, Methodology, Resources, Writing – review & editing, Visualization, Supervision. **Shu-Li Julie Wang:** Conceptualization, Methodology, Resources, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank all the children and their main caretakers who participated in the present study, the interviewers who supported the data collection, and all the study groups who participated in our longitudinal birth cohort study. This study was supported by a research grant from the National

Health Research Institutes (PH-106-PP-09, EM-107-PP-05, EM-108-PP-05) and Ministry of Science and Technology (MOST 106-3114-B-400-002 -, MOST 107-2321-B-400-010 -, MOST 108-2321-B-400-007 -, MOST 109-2321-B-400-007 -), Taiwan. The authors have no conflict of interest to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.scitotenv.2021.145982>.

References

- Akdis, C.A., Arkwright, P.D., Bruggen, M.C., Busse, W., Gadina, M., Guttman-Yassky, E., Kabashima, K., Mitamura, Y., Vian, L., Wu, J., Palomares, O., 2020. Type 2 immunity in the skin and lungs. *Allergy*. <https://doi.org/10.1111/all.14318>.
- Ashley-Martin, J., Lavigne, E., Arbuckle, T.E., Johnson, M., Hystad, P., Crouse, D.L., Marshall, J.S., Dodds, L., 2016. Air pollution during pregnancy and cord blood immune system biomarkers. *J. Occup. Environ. Med.* 58, 979–986. <https://doi.org/10.1097/JOM.0000000000000841>.
- Baiz, N., Slama, R., Bene, M.C., Charles, M.A., Kolopp-Sarda, M.N., Magnan, A., Thiebaugeorges, O., Faure, G., Annesi-Maesano, I., 2011. Maternal exposure to air pollution before and during pregnancy related to changes in newborn's cord blood lymphocyte subpopulations. The EDEN study cohort. *BMC Pregnancy Childbirth* 11, 87. <https://doi.org/10.1186/1471-2393-11-87>.
- Breiman, L.J.M.L., 2001. Random forests. 45, 5–32.
- Castro-Rodriguez, J.A., Holberg, C.J., Wright, A.L., Martinez, F.D., 2000. A clinical index to define risk of asthma in young children with recurrent wheezing. *Am. J. Respir. Crit. Care Med.* 162, 1403–1406. <https://doi.org/10.1164/ajrccm.162.4.9912111>.
- Chatzimichail, E., Paraskakis, E., Sitzimi, M., Rigas, A., 2013. An intelligent system approach for asthma prediction in symptomatic preschool children. *Comput Math Methods Med.* 2013, 240182. <https://doi.org/10.1155/2013/240182>.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. <https://doi.org/10.1613/jair.953>.
- Chen T., Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, ACM, 785–794.
- Dogruel, D., Bingol, G., Altintas, D.U., Seydaoglu, G., Erkan, A., Yilmaz, M., 2017. The trend of change of allergic diseases over the years: three repeated surveys from 1994 to 2014. *Int. Arch. Allergy Immunol.* 173, 178–182. <https://doi.org/10.1159/000477726>.
- Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.J.J.o.m.l.r., 2008. LIBLINEAR: A library for large linear classification. 9, 1871–1874. doi:10.5555/1390681.1442794.
- Glinianaia, S.V., Rankin, J., Bell, R., Pless-Mulloli, T., Howel, D., 2004. Particulate air pollution and fetal health: a systematic review of the epidemiologic evidence. *Epidemiology* 15, 36–45. <https://doi.org/10.1097/01.ede.0000101023.41844.ac>.
- Hansen, T.E., Evjenth, B., Holt, J., 2013. Increasing prevalence of asthma, allergic rhinoconjunctivitis and eczema among schoolchildren: three surveys during the period 1985–2008. *Acta Paediatr.* 102, 47–52. <https://doi.org/10.1111/apa.12030>.
- He, B., Huang, J.V., Kwok, M.K., Au Yeung, S.L., Hui, L.L., Li, A.M., Leung, G.M., Schooling, C.M., 2019. The association of early-life exposure to air pollution with lung function at ~17.5 years in the “Children of 1997” Hong Kong Chinese Birth Cohort. *Environ. Int.* 123, 444–450. <https://doi.org/10.1016/j.envint.2018.11.073>.
- Henriksen, L., Simonsen, J., Haerskjold, A., Linder, M., Kieler, H., Thomsen, S.F., Stensballe, L.G., 2015. Incidence rates of atopic dermatitis, asthma, and allergic rhinoconjunctivitis in Danish and Swedish children. *J. Allergy Clin. Immunol.* 136 (360–6), e2. <https://doi.org/10.1016/j.jaci.2015.02.003>.
- Huang, C.C., Wen, H.J., Chen, P.C., Chiang, T.L., Lin, S.J., Guo, Y.L., 2015a. Prenatal air pollutant exposure and occurrence of atopic dermatitis. *Br. J. Dermatol.* 173, 981–988. <https://doi.org/10.1111/bjd.14039>.
- Huang, S.K., Zhang, Q., Qiu, Z., Chung, K.F., 2015b. Mechanistic impact of outdoor air pollution on asthma and allergic diseases. *J Thorac Dis.* 7, 23–33. <https://doi.org/10.3978/j.issn.2072-1439.2014.12.13>.
- Jackson-Browne, M.S., Henderson, N., Patti, M., Spanier, A., Braun, J.M., 2019. The impact of early-life exposure to antimicrobials on asthma and eczema risk in children. *Curr Environ Health Rep.* 6, 214–224. <https://doi.org/10.1007/s40572-019-00256-2>.
- Kim, B.E., Leung, D.Y.M., 2018. Significance of skin barrier dysfunction in atopic dermatitis. *Allergy Asthma Immunol Res.* 10, 207–215. <https://doi.org/10.4168/aiar.2018.10.3.207>.
- Lee, Y.L., Su, H.J., Sheu, H.M., Yu, H.S., Guo, Y.L., 2008. Traffic-related air pollution, climate, and prevalence of eczema in Taiwanese school children. *J Invest Dermatol.* 128, 2412–2420. <https://doi.org/10.1038/jid.2008.110>.
- Lien, G.W., Wen, T.W., Hsieh, W.S., Wu, K.Y., Chen, C.Y., Chen, P.C., 2011. Analysis of perfluorinated chemicals in umbilical cord blood by ultra-high performance liquid chromatography/tandem mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci.* 879, 641–646. <https://doi.org/10.1016/j.jchromb.2011.01.037>.
- Lin, Y.C., Wen, H.J., Lee, Y.L., Guo, Y.L., 2004. Are maternal psychosocial factors associated with cord immunoglobulin E in addition to family atopic history and mother immunoglobulin E? *Clin. Exp. Allergy* 34, 548–554. <https://doi.org/10.1111/j.1365-2222.2004.1928.x>.
- Lin, S., Ku, H.Y., Su, P.H., Chen, J.W., Huang, P.C., Angerer, J., Wang, S.L., 2011. Phthalate exposure in pregnant women and their children in central Taiwan. *Chemosphere* 82, 947–955. <https://doi.org/10.1016/j.chemosphere.2010.10.073>.
- Lu, C., Deng, L., Ou, C., Yuan, H., Chen, X., Deng, Q., 2017. Preconceptional and perinatal exposure to traffic-related air pollution and eczema in preschool children. *J. Dermatol. Sci.* 85, 85–95. <https://doi.org/10.1016/j.jdermsci.2016.11.004>.
- Luyten, L.J., Saenen, N.D., Janssen, B.G., Vrijens, K., Plusquin, M., Roels, H.A., Debacq-Chainiaux, F., Nawrot, T.S., 2018. Air pollution and the fetal origin of disease: a systematic review of the molecular signatures of air pollution exposure in human placenta. *Environ. Res.* 166, 310–323. <https://doi.org/10.1016/j.envres.2018.03.025>.
- Macleod, J.E.S., Luk, A., Titterton, D.M., 1987. A reexamination of the distance-weighted K-nearest neighbor classification rule. *Ieee Transactions on Systems Man and Cybernetics* 17, 689–696. <https://doi.org/10.1109/TSMC.1987.289362>.
- Morales, E., Garcia-Esteban, R., de la Cruz, O.A., Basterrechea, M., Lertxundi, A., de Dicastillo, M.D., Zabaleta, C., Sunyer, J., 2015. Intrauterine and early postnatal exposure to outdoor air pollution and lung function at preschool age. *Thorax* 70, 64–73. <https://doi.org/10.1136/thoraxjnl-2014-205413>.
- Mortimer, K., Neugebauer, R., Lurmann, F., Alcorn, S., Balmes, J., Tager, I., 2008. Air pollution and pulmonary function in asthmatic children: effects of prenatal and lifetime exposures. *Epidemiology* 19, 550–557 discussion 561–2. <https://doi.org/10.1097/EDE.0b013e31816a9dcb>.
- Nagiah, S., Phulukdaree, A., Naidoo, D., Ramcharan, K., Naidoo, R.N., Moodley, D., Chuturgoon, A., 2015. Oxidative stress and air pollution exposure during pregnancy: a molecular assessment. *Hum Exp Toxicol.* 34, 838–847. <https://doi.org/10.1177/0960327114559992>.
- Ober, C., Yao, T.C., 2011. The genetics of asthma and allergic disease: a 21st century perspective. *Immunol. Rev.* 242, 10–30. <https://doi.org/10.1111/j.1600-065X.2011.01029.x>.
- Pastore, S., Korkina, L., 2010. Redox imbalance in T cell-mediated skin diseases. *Mediat. Inflamm.* 2010, 861949. <https://doi.org/10.1155/2010/861949>.
- Patil, V.K., Kurukulaaratchy, R.J., Venter, C., Grundy, J., Roberts, G., Dean, T., Arshad, S.H., 2015. Changing prevalence of wheeze, rhinitis and allergic sensitisation in late childhood: findings from 2 Isle of Wight birth cohorts 12 years apart. *Clin. Exp. Allergy* 45, 1430–1438. <https://doi.org/10.1111/cea.12534>.
- Prosperi, M.C.F., Marinho, S., Simpson, A., Custovic, A., Buchan, I.E., 2014. Predicting phenotypes of asthma and eczema with machine learning. *BMC Med. Genet.* <https://doi.org/10.1186/1755-8794-7-S1-S7>.
- Sbihi, H., Allen, R.W., Becker, A., Brook, J.R., Mandhane, P., Scott, J.A., Sears, M.R., Subbarao, P., Takaro, T.K., Turvey, S.E., Brauer, M., 2015. Perinatal exposure to traffic-related air pollution and Atopy at 1 year of age in a multi-center Canadian birth cohort study. *Environ. Health Perspect.* 123, 902–908. <https://doi.org/10.1289/ehp.1408700>.
- Shi, W., Belluscio, S., Warburton, D., 2007. Lung development and adult lung diseases. *Chest* 132, 651–656. <https://doi.org/10.1378/chest.06-2663>.
- Taiwan EPA, 2021. Air Quality Index. Available: <https://airtw.epa.gov.tw/ENG/Information/Standard/AirQualityIndicator.aspx>.
- To, T., Zhu, J., Stieb, D., Gray, N., Fong, I., Pinault, L., Jerrett, M., Robichaud, A., Menard, R., van Donkelaar, A., Martin, R.V., Hystad, P., Brook, J.R., Dell, S., 2019. Early Life Exposure to Air Pollution and Incidence of Childhood Asthma, Allergic Rhinitis and Eczema. *Eur Respir J.* doi:10.1183/13993003.00913-2019 (Last Updated:2021-1-21).
- Tsai, M.S., Chen, M.H., Lin, C.C., Liu, C.Y., Chen, P.C., 2019. Children's environmental health based on birth cohort studies of Asia (2) – air pollution, pesticides, and heavy metals. *Environ. Res.* 179, 108754. <https://doi.org/10.1016/j.envres.2019.108754>.
- Wen, H.J., Wang, Y.J., Lin, Y.C., Chang, C.C., Shieh, C.C., Lung, F.W., Guo, Y.L., 2011. Prediction of atopic dermatitis in 2-yr-old children by cord blood IgE, genetic polymorphisms in cytokine genes, and maternal mentality during pregnancy. *Pediatr. Allergy Immunol.* 22, 695–703. <https://doi.org/10.1111/j.1399-3038.2011.01177.x>.
- Wright, R.J., Brunst, K.J., 2013. Programming of respiratory health in childhood: influence of outdoor air pollution. *Curr. Opin. Pediatr.* 25, 232–239. <https://doi.org/10.1097/MOP.0b013e32835e78cc>.
- Yeatts, K., Sly, P., Shore, S., Weiss, S., Martinez, F., Geller, A., Bromberg, P., Enright, P., Koren, H., Weissman, D., Selgrade, M., 2006. A brief targeted review of susceptibility factors, environmental exposures, asthma incidence, and recommendations for future asthma incidence research. *Environ. Health Perspect.* 114, 634–640. <https://doi.org/10.1289/ehp.8381>.
- Yue, H., Yan, W., Ji, X., Gao, R., Ma, J., Rao, Z., Li, G., Sang, N., 2017. Maternal exposure of BALB/c mice to indoor NO2 and allergic asthma syndrome in offspring at adulthood with evaluation of DNA methylation associated Th2 polarization. *Environ. Health Perspect.* 125, 097011. <https://doi.org/10.1289/EHP685>.

Prenatal exposure to air pollutants and childhood atopic dermatitis and allergic rhinitis adopting machine learning approaches: 14-year follow-up birth cohort study

Yu Huang^{a†}, Hui-Ju Wen^{b†}, Yue-Liang Leon Guo^{b,c,d}, Tzu-Yin Wei^a, Wei-Cheng Wang^a, Shin-Fen Tsai^b, Vincent S. Tseng^{a*}, Shu-Li Julie Wang^{b,e,f*}

^aDepartment of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

^bNational Institute of Environmental Health Sciences, National Health Research Institutes, Miaoli, Taiwan

^cDepartment of Environmental and Occupational Medicine, National Taiwan University Hospital, Taipei, Taiwan

^dInstitute of Environmental and Occupational Health Sciences, National Taiwan University College of Public Health, Taipei, Taiwan

^eDepartment of Public Health, National Defence Medical Centre, Taipei, Taiwan

^fDepartment of Safety, Health, and Environmental Engineering, National United University, Miaoli, Taiwan

[†]Authors contributed equally to this work.

*Corresponding authors.

Contents

Table S1. The Air Quality Index (AQI) standard and concentration breakpoints of five air pollutants from the Environmental Protection Administration, Executive Yuan, Taiwan (Taiwan EPA).

Table S2. Subjects' demographic and general characteristics in included and excluded children.

Table S3. Comparison of predictive allergic diseases among different machine learning methods.

Table S4. Comparisons of predictive model for allergic diseases by random forest performance using various input features.

Table S5. Feature importance (the top 20th) and correlation evaluation by means of random forest model without air pollutant data.

Table S6. Subjects' demographic and general characteristics in statistical analysis.

Table S7. The Spearman's correlation between monthly average of air pollutant IAQI during pregnancy

Table S8. The proportion (%) of air pollutant IAQI groups among three latent classes.

Figure S1. Feature importance evaluation using random forest method without (left side)/with (right side) air pollution.

Table S1**The Air Quality Index (AQI) standard and concentration breakpoints of five air pollutants from the Environmental Protection Administration, Executive Yuan, Taiwan (Taiwan EPA)¹**

AQI	Concentration breakpoints				
	C _{low} -C _{high}				
	O ₃ (ppm)	PM ₁₀ (µg/m ³)	CO (ppm)	SO ₂ (ppb)	NO ₂ (ppb)
0–50	-	0–54	0–4.4	0–35	0–53
51–100	-	55–125	4.5–9.4	36–75	54–100
101–150	0.125–0.164	126–254	9.5–12.4	76–185	101–360
151–200	0.165–0.204	255–354	12.5–15.4	186–304	361–649
201–300	0.205–0.404	355–424	15.5–30.4	305–604	650–1249
301–400	0.405–0.504	425–504	30.5–40.4	605–804	1250–1649
401–500	0.505–0.604	505–604	40.5–50.4	805–1004	1650–2049

Abbreviations: EPA, Environmental Protection Administration; AQI, air quality index; PM₁₀, particulate matter with a 50% cut-off aerodynamic diameter of 10 µm.

Table S2

Subjects' demographic and general characteristics in included and excluded children

Characteristics	N (%) or mean (SD)					
	Machine learning analysis			Statistical analysis		
	Included (n = 1439)	Excluded (n = 659)	p-value [§] -	Included (n = 1348)	Excluded (n = 750)	p-value [§]
Children						
Sex						
Boys	706 (49.1)	259 (39.3)	0.831	696 (51.6)	269 (35.9)	0.831
Girls	644 (44.8)	231 (35.1)		635 (47.1)	240 (32.0)	
Birth weight [¶]	3095.8 (397.8)	3107.5 (423.3)	0.408	3095.2 (397.0)	3109.5 (425.6)	0.366
Missing	107 (7.4)	167 (25.3)		10 (0.7)	264 (35.2)	
Gestational age [¶]	38.91 (1.29)	38.88 (1.35)	0.714	38.91 (1.29)	38.88 (1.36)	0.802
Missing	103 (7.2)	177 (26.9)		0	280 (37.3)	
Birth order						
1st	641 (44.5)	156 (23.7)	0.043	604 (44.8)	193 (25.7)	0.447
2nd	456 (31.7)	150 (22.8)		442 (32.8)	164 (21.9)	
≥3rd	190 (13.2)	62 (9.4)		185 (13.7)	67 (8.9)	
Parents						
Maternal age at childbirth (year) [¶]	27.59 (4.50)	27.65 (4.90)	0.903	27.49 (4.50)	27.81 (4.83)	0.231

Missing	620 (43.1)	287 (43.6)		613 (45.5)	295 (39.3)	
Maternal BMI before pregnancy [¶]	20.87 (3.13)	20.94 (3.47)	0.958	20.94 (3.14)	20.80 (3.38)	0.338
Missing	265 (18.4)	221 (33.5)		266 (19.7)	220 (29.3)	
Maternal weight gain during pregnancy [¶]	13.23 (5.46)	12.83 (5.91)	0.139	13.21 (5.51)	12.93 (5.77)	0.383
Missing	209 (14.5)	121 (18.4)		209 (15.5)	121 (16.1)	
Maternal education						
≤ 9 years	229 (15.9)	143 (21.7)	0.0004	224 (16.6)	148 (19.7)	0.087
10- 12 years	700 (48.6)	297 (45.1)		662 (49.1)	335 (44.7)	
> 12 years	492 (34.2)	181 (27.5)		444 (32.9)	229 (30.5)	
Paternal education						
≤ 9 years	252 (17.5)	146 (22.2)	0.002	245 (18.2)	153 (20.4)	0.049
10- 12 years	666 (46.3)	270 (41.0)		640 (47.5)	296 (39.5)	
> 12 years	501 (34.8)	183 (27.8)		446 (33.1)	238 (31.7)	
Family income (10 ³ USD)						
<20	856 (59.5)	348 (52.8)	0.228	814 (60.4)	390 (52.0)	0.028
20-33.3	380 (26.4)	165 (25.0)		344 (25.5)	201 (26.8)	
≥33.3	135 (9.4)	72 (10.9)		123 (9.1)	84 (11.2)	
Maternal atopy						
Yes	252 (17.5)	50 (7.6)	0.0005	234 (17.4)	68 (9.1)	0.009

No	990 (68.8)	349 (53.0)		937 (69.5)	402 (53.6)	
Paternal atopy						
Yes	260 (18.1)	67 (10.2)	0.566	240 (17.8)	87 (11.6)	0.670
No	963 (66.9)	271 (41.1)		920 (68.2)	314 (41.9)	
Smoking during pregnancy						
Yes	32 (2.2)	33 (5.0)	0.0002	29 (2.2)	36 (4.8)	0.0004
No	1364 (94.8)	571 (86.7)		1278 (94.8)	657 (87.6)	
Smoking before pregnancy						
Yes	87 (6.0)	77 (11.7)	<0.0001	79 (5.9)	85 (11.3)	<0.0001
No	1323 (91.9)	541 (82.1)		1242 (92.1)	622 (82.9)	

Some numbers do not add up to total n because of missing values.

[¶]Mean (SD).

[§]P value was calculated by Kruskal-Wallis tests for continues variables and χ^2 test for categorical variables as compared between children with and without allergic diseases.

Abbreviation: BMI, body mass index; SD, standard deviation; USD, US dollar.

Table S3**Comparison of predictive allergic diseases among different machine learning methods**

Outcome	Model	Sensitivity	Specificity	Accuracy	AUROC
AD	Random Forest	74.81	69.36	69.71	0.83
	Xgboost	76.3	67.4	67.98	0.79
	Decision Tree	50.74	76.95	75.26	0.7
	Naïve Bayes	65.56	65.8	65.79	0.71
	Neural Network	58.52	60.74	60.6	0.66
	Logistic Regression	70	66.92	67.12	0.74
AR	Random Forest	76.5	78.66	78.36	0.84
	Xgboost	72.75	74.15	73.96	0.81
	Decision Tree	65	69.45	68.84	0.76
	Naïve Bayes	71.25	73.68	73.34	0.8
	Neural Network	64.25	61.38	61.77	0.67
	Logistic Regression	74.25	74.9	74.81	0.8

Abbreviation: AD, atopic dermatitis; AR, allergic rhinitis; AUROC, area under the receiver operating characteristic curve.

Table S4**Comparisons of predictive model for allergic diseases by random forest performance using various input features**

Outcome	Model	Sensitivity	Specificity	Accuracy	AUROC
AD	Without air pollution	74.07	57.48	58.55	0.75
	Air pollution / With difference	77.04	58.32	59.52	0.78
	Air pollution / With statistic features	83.33	65.62	66.76	0.83
	Air pollution / With both statistic features and difference	74.81	69.36	69.71	0.83
AR	Without air pollution	73.75	72.25	72.46	0.81
	Air pollution / With difference	75.5	73	73.34	0.82
	Air pollution / With statistic features	75	81.3	80.44	0.85
	Air pollution / With both statistic features and difference	76.5	78.66	78.36	0.84

Abbreviation: AD, atopic dermatitis; AR, allergic rhinitis; AUROC, area under the receiver operating characteristic curve.

Table S5

Feature importance (the top 20th) and correlation evaluation by means of random forest model without air pollutant data

Features	AD			AR		
	Importance	Correlation	P-value	Importance	Correlation	P-value
Parental education	0.154	0.143	<0.0001	0.088	0.181	<0.0001
Maternal vigour in pregnancy	0.105	-0.093	0.001	0.021	-0.076	0.021
Packaging hot food via plastic bag	0.073	-0.141	<0.0001	0.085	0.208	<0.0001
Child sex	0.055	0.051	0.050	0.048	0.077	0.014
Maternal vitality in pregnancy	0.047	-0.034	0.203	—	—	—
Pet at home	0.044	-0.097	<0.0001	0.066	-0.166	<0.0001
Frequency of freshwater fish consumption	0.039	-0.105	0.0001	—	—	—
Number of pregnancies	0.026	-0.060	0.016	0.040	-0.105	0.0002
Birthweight	0.026	0.040	0.133	0.022	0.033	0.319
Paternal atopy	0.025	0.054	0.081	0.018	0.092	0.012
Maternal happiness in pregnancy	0.022	-0.046	0.110	—	—	—
Maternal weight after pregnancy	0.021	-0.062	0.020	—	—	—
Cord blood IgE	0.020	0.044	0.239	0.038	0.007	0.794
Intake vitamin supplement	0.019	0.058	0.051	0.025	0.086	0.015
Frequency of edible squid fish	0.018	-0.027	0.243	—	—	—
Frequency of edible shellfish and oysters	0.018	-0.032	0.172	—	—	—

Maternal atopy	0.017	0.042	0.160	0.032	0.129	0.001
Cord plasma PFDoA	0.015	-0.040	0.061	—	—	—
Cockroach at home	0.014	0.056	0.012	0.045	0.148	<0.0001
Maternal clam and peaceful in pregnancy	0.013	-0.036	0.193	—	—	—
Cord plasma PFOA	—	—	—	0.031	0.081	0.098
Add vitamins to baby food	—	—	—	0.030	-0.143	<0.0001
Maternal age at childbirth	—	—	—	0.028	0.097	0.003
Intake calcium supplement	—	—	—	0.025	0.056	0.052
Birth height	—	—	—	0.024	0.090	0.003
Maternal IgE	—	—	—	0.023	-0.028	0.224
Self-report air pollution surrounding residence at 2 years old	—	—	—	0.017	-0.112	<0.0001
Child height	—	—	—	0.014	0.079	0.003

“—” indicated that the feature was not the top 20th important features for the disease.

Abbreviations: IgE, immunoglobulin E; PFDoA, perfluorododecanoic acid; PFOA, perfluorooctane sulfonate.

Table S6**Subjects' demographic and general characteristics in statistical analysis**

Characteristics	N (%) or mean (SD)			p-value [§]
	All (n = 1348)	AD/AR		
		With (n = 257)	Without (n = 1091)	
Children				
Sex				
Boys	696 (51.6)	136 (19.5)	560 (80.5)	0.711
Girls	635 (47.1)	119 (18.7)	516 (81.3)	
Birth weight	1338 (99.3)	3111.9 (374.4) [¶]	3091.2 (402.3) [¶]	0.522
Gestational age	1348 (100.0)	39.02 (1.17) [¶]	38.89 (1.32) [¶]	0.306
Birth order				
1st	604 (44.8)	119 (19.7)	485 (80.3)	0.287
2nd	442 (32.8)	80 (18.1)	362 (81.9)	
≥3rd	185 (13.7)	27 (14.6)	158 (85.4)	
Parents				
Maternal age at childbirth (year)	735 (54.5)	28.22 (4.17) [¶]	27.29 (4.57) [¶]	0.012
Maternal BMI before pregnancy	1082 (80.3)	21.01 (3.16) [¶]	20.92 (3.13) [¶]	0.615
Maternal weight gain during pregnancy	1139 (84.5)	12.65 (5.15) [¶]	13.35 (5.59) [¶]	0.082
Maternal education				
≤ 9 years	224 (16.6)	21 (9.4)	203 (90.6)	<0.0001
10- 12 years	662 (49.1)	118 (17.8)	544 (82.2)	
> 12 years	444 (32.9)	117 (26.4)	327 (73.6)	
Paternal education				
≤ 9 years	245 (18.2)	24 (9.8)	221 (90.0)	<0.0001
10- 12 years	640 (47.5)	111 (17.3)	529 (82.7)	
> 12 years	446 (33.1)	121 (27.1)	325 (72.9)	
Family income (10 ³ USD)				

<20	814 (60.4)	129 (15.9)	685 (84.2)	<0.0001
20-33.3	344 (25.5)	81 (23.6)	263 (76.5)	
≥33.3	123 (9.1)	39 (31.7)	84 (68.3)	
Maternal atopy				
Yes	234 (17.4)	67 (28.6)	167 (71.4)	0.0002
No	937 (69.5)	165 (17.6)	772 (82.4)	
Paternal atopy				
Yes	240 (17.8)	59 (24.6)	181 (75.4)	0.028
No	920 (68.2)	168 (18.3)	752 (81.7)	
Smoking during pregnancy				
Yes	29 (2.2)	5 (17.2)	24 (82.8)	0.778
No	1278 (94.8)	247 (19.3)	1031 (80.7)	
Smoking before pregnancy				
Yes	79 (5.9)	10 (12.7)	69 (87.3)	0.13
No	1242 (92.1)	243 (19.6)	999 (80.4)	

Some numbers do not add up to total n because of missing values.

[§]P value was calculated by Kruskal-Wallis tests for continues variables and χ^2 test for categorical variables as compared between children with and without allergic diseases.

[¶] mean (SD).

Abbreviation: AD, atopic dermatitis; AR, allergic rhinitis; BMI, body mass index; SD, standard deviation; USD, US dollar.

Table S7

The Spearman's correlation between monthly average of air pollutant individual air quality index during pregnancy

	NO ₂	CO	SO ₂	O ₃	PM ₁₀
NO ₂	—				
CO	0.820***	—			
SO ₂	0.641***	0.395***	—		
O ₃	-0.692***	-0.495***	-0.358***	—	
PM ₁₀	-0.324***	-0.353***	0.018	0.343***	—

Abbreviation: PM₁₀, particulate matter with a 50% cut-off aerodynamic diameter of 10 mm.

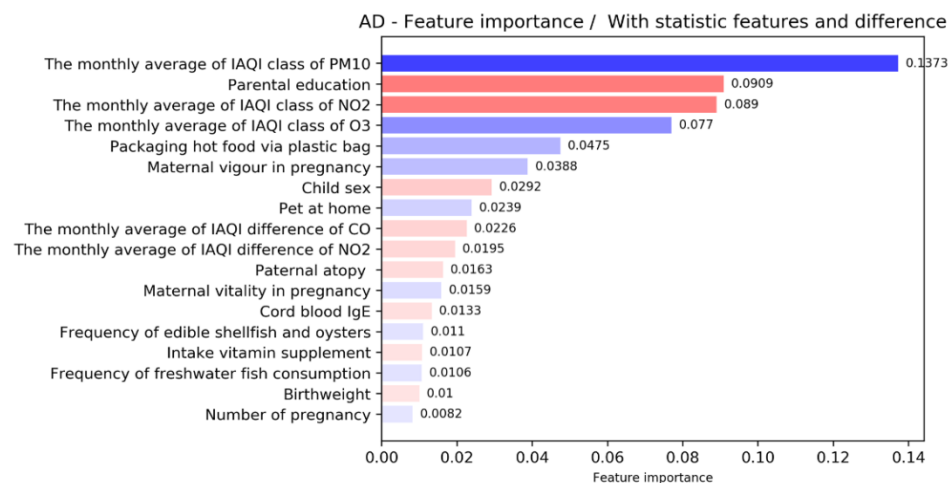
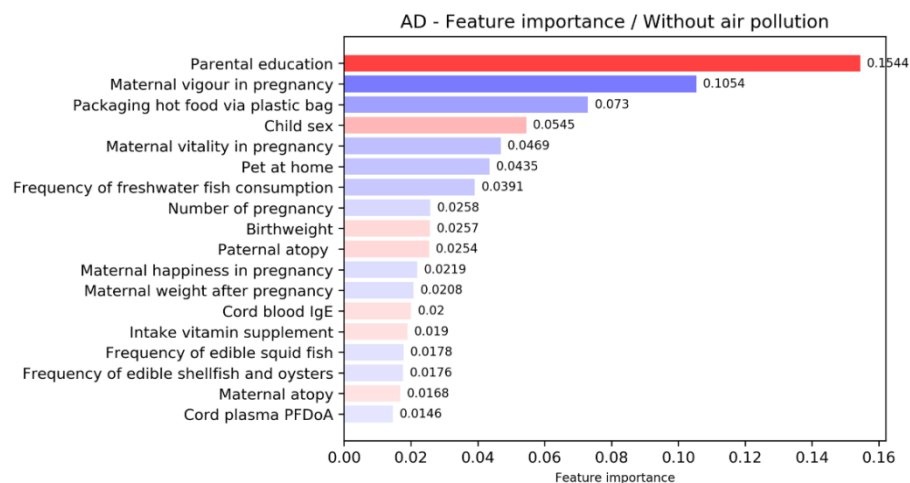
Table S8**The proportion (%) of air pollutant individual air quality index (IAQI) groups among three latent clusters**

AD				AR			
Groups of air pollutant IAQI (n)	Cluster 1	Cluster 2	Cluster 3	Groups of air pollutant IAQI (n)	Cluster 1	Cluster 2	Cluster 3
Overall (1348)	42	33.7	24.3	Overall (982)	42.5	33.9	23.6
NO ₂				NO ₂			
<12.7 (450)	0	99.8	0.2	<12.6 (328)	0	99.4	0.6
12.7-16.7 (449)	98.9	1.1	0	12.6-16.6 (327)	97.9	2.1	0
≥16.7 (449)	27.2	0	72.8	≥16.6 (327)	29.7	0	70.3
CO				CO			
<5.28 (450)	24	74.7	1.3	<5.28 (328)	22.6	76.2	1.2
5.28-6.72 (449)	72.4	26.1	1.6	5.28-6.56 (327)	73.4	25.1	1.5
≥6.72 (449)	29.6	0.2	70.2	≥6.56 (327)	31.5	0.3	68.2
SO ₂				SO ₂			
<3.6 (450)	14	77.3	8.7	<3.6 (328)	13.4	78.7	7.9
3.6-4.3 (449)	48.8	22	29.2	3.6-4.3 (327)	50.1	21.1	28.8
≥4.3 (449)	63.3	1.6	35.2	≥4.3 (327)	63.9	1.8	34.3
O ₃				O ₃			
<21.4 (450)	20.7	6.4	72.9	<21.5 (328)	22.6	6.7	70.7
21.4-23.6 (449)	69.3	30.7	0	21.5-23.6 (327)	69.1	30.9	0

≥23.6 (449)	36.1	63.9	0	≥23.6 (327)	35.8	64.2	0
PM ₁₀				PM ₁₀			
<47.9 (450)	31.8	12	56.2	<48.7 (328)	34.5	12.5	53
47.9-59.8 (449)	48.1	45.2	6.7	48.7-60.1 (327)	45	48.9	6.1
≥59.8 (449)	46.1	43.9	10	≥60.1 (327)	48	40.4	11.6

Abbreviation: AD, atopic dermatitis; AR, allergic rhinitis; PM₁₀, particulate matter with a 50% cut-off aerodynamic diameter of 10 mm.

(A) AD



(B) AR

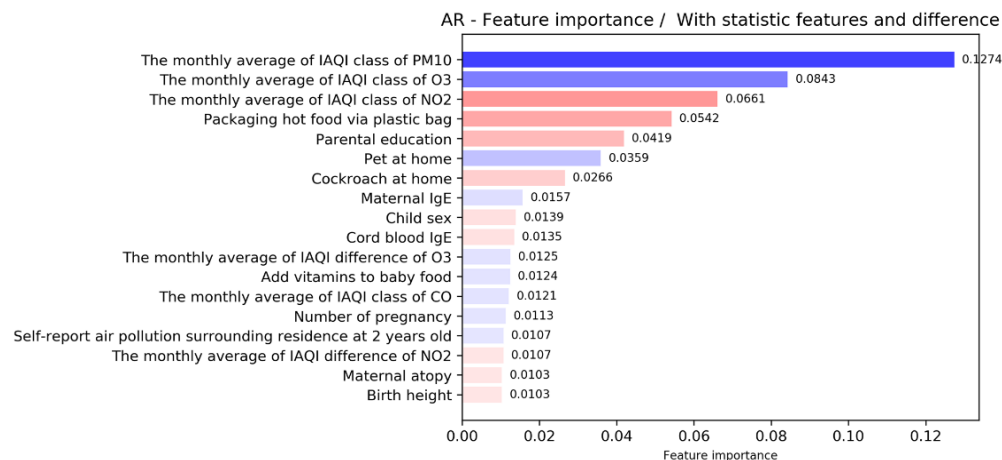
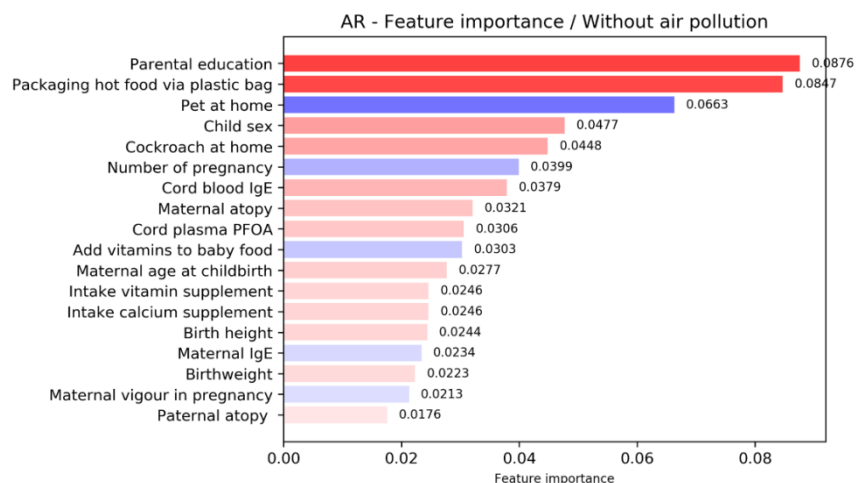


Fig. S1. Feature importance evaluation using random forest method without (left side)/with (right side) air pollution.

(A) AD; (B) AR.

Red bar, positive correlation; blue bar, negative correlation.

Reference:

1. The Environmental Protection Administration, Executive Yuan, Taiwan (Taiwan EPA). Air Quality Index. <https://airtw.epa.gov.tw/CHT/Information/Standard/AirQualityIndicator.aspx>