Evaluation of a prediction model based on ridge regression for asthma persistence in preschool children

Ioannis I. Spyroglou, Eleni A. Chatzimichail, E.N. Spanou, E. N. Paraskakis, and Alexandros G. Rigas

***Abstract*—**The accurate prediction of asthma persistence in childhood is one of the most significant issues about this chronic disease. Asthmatic children usually develop their first symptoms before the fifth year of age. The early detection of the preschoolers whose asthma persists after the age of five could lead into better treatment opportunities and disease long-term outcome. 148 patients with mean age (±SD) 9.9±2.7 years who received a diagnosis of asthma at the age under the five were used for the asthma prediction system. All children have performed spirometry pre and post bronchodilation and a detailed case history was obtained. In this work, since there is strong multicollinearity among the factors affecting the asthma persistence, a Logistic Ridge Regression model is proposed for its prediction. The estimated parameters of the proposed model are obtained by using a penalized likelihood function. Moreover, a test is developed for checking the validity of the fitted model based on the randomized quantile residuals. The 5% rejection regions of the randomized quantile residuals are constructed by using a proper bootstrap method and they are added in the QQ-plot. The prediction accuracy of asthma persistence was tested by a second group of 33 children aged 3.5-5 years who were reevaluated at the mean age (±SD) of 9.2±2.7 years. The proposed prediction system for asthma persistence evaluated in real life setting has shown an accuracy of 93.18%, while the positive predictive value of and negative predictive value 96.15% and 88.89% respectively. The experimental results of our study show that the proposed system can be a valuable tool in medical-decision making and clinical research.

***Keywords*—**Asthma outcome, Logistic Ridge regression, Penalized Likelihood Function, Randomized Quantile Residuals, Q-Q plot

Ioannis Spyroglou is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (corresponding author to provide phone: +306955954849 ; e-mail: ispyrogl@ee.duth.gr).

Eleni Chatzimichail is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (e-mail: echatzim@ee.duth.gr).

E.N. Spanou is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (e-mail: ispanou@ee.duth.gr).

Emmanouil Paraskakis, Prof., is with the Medical School , Democritus University of Thrace, Alexandroupolis, CO 68100 GRRECE (e-mail: eparaska@med.duth.gr).

Alexandros Rigas, Prof., is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (e-mail: rigas@ee.duth.gr).

I. Introduction

ASTHMA is a disease with polymorphic phenotype affected by several environmental and genetic factors which both play a key role in the development and persistence of the disease [1]. Among these factors seasonal symptoms, wheezing episodes during childhood and several prenatal and environmental factors are included [2].

Asthma usually presents in early childhood and is commonly associated with skin atopy and atopic disorders. Patients develop specific IgE antibodies after exposure to common environmental antigens and often have positive skin prick tests to several common allergens. Attacks of asthma may be provoked by allergen but are also provoked by various non-allergic factors such as infection, exercise, drugs and cold. The first and most obvious functional consequence of asthma is an increase in airway resistance and reduction in airflow. The airway changes result in changes in lung volumes, gas exchanges, exercise tolerance and work of breathing. Airflow obstruction is most conveniently measured as the FEV_1 (Forced Expiratory Volume in one second) or PEF (Peak Flow Rate). [3] The most common symptoms of asthma are wheezing, dyspnea, cough, particularly at night or after exercise, allergic rhinitis, allergic conjunctivitis and congestion.

Most children who suffer from asthma develop their first symptoms before the 5th year of age. Although the majority of the preschool children with asthma overcome their disease by the school age a substantial number of preschoolers exhibit a persistence of symptoms requiring early identification and treatment [4]. Predictive models, mostly based on simple clinical and laboratory parameters [5-11], aiming to identify children at risk of asthma persistence have been recently studied [12-16]. The Asthma Predictive Index [12] (API) has become the most popular among these predictive tools although the usefulness of this predictive tool in clinical practice has recently been questioned [17-18]. Different research groups have recently introduced similar predictive scores such as Isle of Whight score [14], ECA severity score [15] or PIAMA severity score [14], all based on simple laboratory and clinical parameters. A recent study evaluating the above tools exhibit practical limitations, are insufficiently validated, and they have limited predictive value of confirming or ruling out persistence of asthma symptoms

among preschool wheezers [19]. It should be mentioned that since asthma is a complex disorder with genetic heterogeneity and multiple clinical phenotypes [20-23] it is not surprising that the available predictive models using limited number of risk factors perform poorly in predicting the outcome of preschool wheezing disorders. The current lack of a gold standard tool for asthma prediction persistence have urged the need of the study of new applicable models taking into account multiple risk factors, exhibiting high predictive efficacy.

In preventive medicine, the value of a test lies in its ability to identify those individuals who are at high risk of an illness and who therefore require intervention while excluding those who do not require such intervention. The accuracy of the risk classification is of particular relevance in the case of asthma disease. Early identification of patients at high risk for asthma disease progression may lead to better treatment opportunities and hopefully better disease outcomes in adulthood [4].

In this paper a new dataset of 18 factors is used for modeling and prediction of asthma persistence. The strong correlations among the factors (Appendix), led us to use the Logistic Ridge Regression, which achieves much better results for the estimation of the model coefficients. The logistic ridge regression has never been used before, according to our knowledge, for asthma persistence prediction. In addition, the use of the penalized maximum likelihood improves the estimates of the coefficients, since the presence of multicollinearity increases the values of the estimated coefficients and makes their standard errors tend to very high levels. This method is also used for the first time in asthma persistence prediction and the prediction of the behavior of this disease is based on a new dataset which is not used for the estimation of the coefficients of the Logistic Ridge Model. Finally the construction of a validity test with the help of the bootstrap method based on the randomized quantile residuals, which takes into account the variability of the estimated parameters, is implemented for the first time as far as we know. As a result, the suitability of the logistic ridge model for the study of the behavior and evolution of this disease is confirmed in the best way.

As far as the comparison of the results of this paper with a previous work [56] is concerned, we must mention that the factors used in the construction of the logistic ridge model are not the same as in [56]. Moreover, in this work the factors that are found to be statistically significant are more than those in [56], resulting in very interesting medical conclusions also discussed in other research articles [57-61]. This happens because the ridge logistic regression deals with multicollinearity without reducing the model parameters. In contrast, the method of Principal Component Analysis used in [56] for dimension reduction makes the prediction model contain less information about factors affecting the asthma disease. However, the fact that some of the statistically significant variables of [56] are the same as in this paper confirms the validity of the model in [56] as well. Finally, the accuracy percentage of 95.48% in [56] cannot be compared with the accuracy of the logistic ridge model, because a new dataset of 33 patients was used for the examination of the performance of the proposed model which were not available in [56].

II. Materials And Methods

A. Clinical Data

Data from 148 patients were collected from the Pediatric Department of the University Hospital of Alexandroupolis, Greece during the period from 2008 to 2010. A group of 148 patients who were diagnosed for asthma were studied prospectively from the 7^{th} to the 14^{th} year of age. From this sample, 36 patients were removed because of missing data. The history of each case was obtained by questionnaire. A second group of 33 children was used for validation of the efficacy of the constructed model in real life. In this group of preschool children the proposed system was used to predict asthma persistence in school age. At mean age (±SD) of 9.2±2.7 years these children were revaluated, the diagnosis of asthma was based on case history, data on asthma control, the measurements of IgE and specific IgE (RAST) to ten common allergens and confirmed by pre and post bronchodilation spirometry. In this way an independent data set of 33 patients [21 positive (asthma persistence in school age) and 12 negative patients (asthma subsided in school age)] was constructed for comparison with the asthma persistence prediction and the proposed system was produced for these children when they were pre-schoolers. The new dataset has 18 available predictors which are going to be used in the logistic model. The 18 used prognostic factors have been derived by previous studies [1-4] and they are described in Table I. The encoding of the prognostic factor "seasonal symptoms" is presented in Table II.

TABLE I

Category	**Prognostic Factors**
Demographic	Age, height, weight, waist's perimeter
Bronchiolitis episodes	Until 3^{rd} year, between 3^{rd} – 5^{th} year
Symptoms	Wheezing, cough, allergic rhinitis, allergic conjunctivitis, dyspnea, congestion, runny nose, seasonal symptoms
Pharmaceutical therapy	Antileukotriene, antihistamine, corticosteroids inhaled
Asthma	Diagnosis of asthma (dependent variable), Treatment

The 18 used prognostic factors.

TABLE II

1 (none)	2 (Winter)	3 (Autumn)	4 (Spring)	5 (Summer)	6 (>2seasons)

The encoding of "seasonal symptoms".

B. Multicollinearity

Generalized Linear Models and Regression Analysis are two of the most important and popular statistical approaches

used in biomedical research [24]. In many cases it has been observed that medical data exhibit strong correlations between the predictor variables, a condition known as multicollinearity. Multicollinearity was introduced as a concept by Frisch [25], in order to illustrate a situation, where the variables are subject to two or more correlations.

One of the main consequences of multicollinearity is that the least squares estimates often do not make any sense, and the standard errors of the parameter estimates are very large or the t-ratios are very low. Therefore multicollinearity could lead into inaccurate results. For example, when the null hypothesis that the parameters of the model are zero is rejected and none of the estimated parameters have a p-value less than 0.5. One of several methods that have been used in order to overcome the multicollinearity problem is the Ridge Regression method that was introduced in [26]. When multicollinearity appears, the ridge estimator has a smaller total Mean Square Error (MSE) than the maximum likelihood estimator. Eigenvalues of the correlation matrix of the independent variables near zero indicate multicollinearity.

Ridge Regression (RR) is an alternative estimation method of the unknown parameters of the linear regression models and belongs to the category of biased regression methods [27-28]. This method introduces a bias in the regression equation in order to reduce the variance of the parameter estimates. This bias is entered with the ridge parameter, which determines the extent of the shrinkage of the least squares estimates. Also in [29] the ridge estimator was introduced for Logistic Regression, which is one of the most popular methods used for binary data modeling. Generally, this method is differentiated from the maximum likelihood as a penalty term is added, which includes the ridge parameter.

C. Ridge Regression

Let $y_i,\ i = 1, \dots, n$ be the binary responses of n random variables Y_i, where $Y_i \sim B(1, p_i)$, and $\boldsymbol{x}_i$ a vector of explanatory variables which consist of covariates (numerical or binary) and dummy variables corresponding to factor levels.

The logistic regression model is given by:

$$p_i = \frac{\exp(b\boldsymbol{x}_i)}{\{1 + \exp(b\boldsymbol{x}_i)\}}, \tag{1}$$

where b is the parameter vector [30-31]. This model is implemented without the use of a constant term.

Now, the maximum likelihood estimates of the parameters b_j, j=1,…,k and from them the probabilities p_i are obtained by maximizing the following likelihood function

$$L(b|y) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i},\ \ y_i = 0,1 \quad , \tag{2}$$

or by maximizing the log – likelihood function using a Newton – Raphson algorithm which is:

$$l(b|y) = \log L(b|y), \tag{3}$$

$$l(b|y) = \sum_{i=1}^{n} \left[y_i \log\left[\frac{1}{1 + \exp(-x_i^T b)}\right] + (1 - y_i)\log\left[1 - \frac{1}{1 + \exp(-x_i^T b)}\right]\right] \tag{4}$$

As it was mentioned before when multicollinearity exists, in order to obtain more stable estimates of the parameters, the logistic ridge regression is used. In ridge logistic regression the penalized maximum likelihood is used and is given by [32]:

$$l^{\lambda}(b|y) = l(b|y) - \lambda\|b\|^2 = l(b|y) - \lambda R, \tag{5}$$

and is known as restricted maximum likelihood function, whereas $l(b|y)$ is the unrestricted maximum likelihood and R is a penalty term of the following form [33]:

$$R = \sum_{j=0}^{k-1} (b_{j+1} - b_j)^2. \tag{6}$$

Generally the difference between this approach, and the approach of maximum likelihood function is the use of the penalty term which includes the ridge parameter. The ridge parameter is a positive number and its main role is the regulation of the significance of the penalty term R [33]. Therefore it is obvious that when $\lambda = 0$ the estimates produced are the same as the ones obtained by the unrestricted maximum likelihood function. The computational procedure of the penalized parameter estimates $\hat{b}^{\lambda}$ is based on the Newton – Raphson algorithm. However, a transformation of the linear estimates of the unrestricted logistic regression model is required, since the term R given by (6), should be in the form of (5). Therefore:

$$b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} = b_1 z_{i1} + (b_2 - b_1) z_{i2} + \cdots + (b_k - b_{k-1}) z_{ik}$$

$$= \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_k z_{ik}, \tag{7}$$

where

$$\gamma_1 = b_1, \dots,\ \gamma_j = b_j - b_{j-1},\quad j = 2, \dots, k \tag{7.1}$$

and

$$z_{ij} = \textstyle\sum_{u=j}^{k} x_{iu} \tag{7.2}.$$

Thus, the penalized maximum likelihood becomes as follows:

$$l^{\lambda}(\gamma|y) = l(\gamma|y) - \lambda\|\gamma\|^2. \tag{8}$$

The first derivative of equation (8) is now:

$$U^{\lambda}(\gamma) = \sum_{i=1}^{n} z_i\{y_i - p_i\} - 2\lambda\gamma = U(\gamma) - 2\lambda\gamma. \tag{9}$$

Then, calculating the negative second derivative we get:

$$\Omega^{\lambda}(\gamma) = \Omega(\gamma) + 2\lambda I, \tag{10}$$

where $\Omega(\gamma) = z^T W z$ and W is the $n \times n$ weight matrix which is diagonal with elements $W_{ii} = p_i(1 - p_i)$.

Applying the Taylor series expansion in the first derivative of the penalized maximum likelihood function, the properties that are valid for large sample can be obtained. Consequently:

$$U^{\lambda}(\gamma^{\lambda}) = U^{\lambda}(\gamma_0) - (\hat{\gamma}^{\lambda} - \gamma_0)\Omega^{\lambda}(\gamma_0) + o(\|\hat{\gamma}^{\lambda} - \gamma_0\|). \quad (11)$$

Using equations (9) and (10) in (11) and setting it equal to 0 it leads to:

$$\hat{\gamma}^{\lambda} = \{\Omega(\gamma) + 2\lambda I\}^{-1}\{U(\gamma_0) + \gamma_0\Omega(\gamma_0)\}. \quad (12)$$

The asymptotic variance of the estimated parameters $\hat{\gamma}^{\lambda}$, is given by[32]:

$$\{\Omega(\gamma) + 2\lambda \boldsymbol{I}\}^{-1}\Omega(\gamma)\{\Omega(\gamma) + 2\lambda \boldsymbol{I}\}^{-1}.$$

However according to [32] this approximation cannot be considered for the construction of the confidence limits of the estimated parameters because it does not take into account the bias of the estimates. Resampling methods such as bootstrapping could provide more information about the variability of the estimated parameters $\hat{\gamma}^{\lambda}$.

D. Choosing the ridge parameter

The most difficult task in RR is to determine the ridge parameter. In bibliography there are many methods proposed for choosing the ridge parameter [34-37]. One way of selecting an appropriate Ridge Parameter is the process of Cross Validation. In this direction it is possible to perform an estimate of the mean squared error of the cross validation set, which can be minimized to obtain the Ridge parameter. In this study, we chose the value of the ridge parameter that minimized the Mean Squared Error through 10-fold cross validation[32].

$$MSEcv = \frac{1}{n}\left(\sum_{i}\{Y_i - \hat{p}_i(X_i)\}^2\right) \quad (13)$$

The average value of the MSE was considered as the overall cross-validation error of the model. We selected the ridge parameter as the one with the minimum cross-validation error.

E. Residuals and bootstrapping

After fitting the model to the observed data, it is necessary to check if the fitted model is valid. A usual technique used for validity examination of the model is based on the residuals. In the case of logistic regression with binary response, the distributions of Pearson residuals which are defined by $r_{p,i} = (y_i - \hat{p}_i)/\sqrt{\hat{p}_i(1-\hat{p}_i)},\ i = 1, \dots, n$ and of deviance residuals which are defined by, $r_D = sign\{y_i - \hat{p}_i\}$ are far from normal. In addition, plots of the residuals against the explanatory variables, which are usually used in generalized linear models for model checking, are uninformative in a binary case and are not recommended. More details about the residuals are given in [38].

Let $F(y_i; p_i) = P(Y_i \le y_i) = \sum_{m=0}^{|y_i|} p_i^m (1-p_i)^{1-m}$ be the cumulative binomial distribution of the ith binary response, and $\lfloor y_i \rfloor$ is the greatest integer less than or equal to y_i, i.e. the 'floor' under y_i. Then the randomized quantile residuals for a logistic regression model are defined by

$$r_{rq,i} = \Phi^{-1}\{u_i\}, \quad (14)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, and u_i is a uniform random variable on the interval

$$(a_i, b_i] = \left(\lim_{y \uparrow y_i} F(y; \hat{p}_i), F(y; \hat{p}_i)\right]$$
$$\approx [F(y_i - 1; \hat{p}_i), F(y; \hat{p}_i)]$$

The randomized quantile residuals defined by (14) follow exactly the standard normal distribution, apart from sampling variability in $\hat{p}_i$. These residuals [39] can be used for any discrete distributed response. Thus, the validity of the model can now be tested by using goodness of fit tests for the normality of $r_{rq,i}$. A very strong method to test the null hypothesis that the randomized quantile residuals follow a standard normal distribution i.e. $\boldsymbol{r_{rq}} \sim \boldsymbol{N(0, I)}$ that is commonly used to check if a data sample comes from a normal distribution is the Anderson – Darling test [46].

A lot of tests have been developed in the past for measuring the linearity of the plotted data in Q-Q plots. Stephens (1986) gave a review of correlation and regression tests based on Q-Q plots [40] and Coin (2008) presented a goodness of fit test for the normality of the data based on polynomial regression methods [41]. Vassiliadis and Rigas (2009) proposed an alternative goodness of fit test which combines the known Q-Q plot with David's theorem for the asymptotic distribution of the quantile estimates [42].

Also the Q-Q plot of the randomized quantile residuals has been proposed by Dunn and Smyth [39] as a mean for checking the validity of the model. Here a method for constructing pointwise a $\times 100\%$ rejection regions around the Q-Q plot of any random sample is proposed by using bootstrapping [43-44]. Because of the large number of the estimated parameters, the additional uncertainty due to the estimation of the regression parameters must be taken into consideration. Therefore a proper bootstrap of the randomized quantile residuals must be used in order to take the above into account. Residual resampling is known to be an appropriate bootstrap process for studying the properties of the estimates [49-50]. Moreover this bootstrap is very important since the standard errors of the ridge estimated parameters can be obtained as it was mentioned before. The bootstrap is implemented with the next steps:

- Step 1: Obtain estimates of p_i, and randomized quantile residuals with the use of logistic ridge regression.
- Step 2: Bootsrapping 2000 times the randomized quantile residuals obtained by the logistic ridge model. So now we have $\boldsymbol{r_{rq,1}^T, \dots, r_{rq,2000}^T}$. We use the randomized quantile residuals because they have unit variance as they approximate standard normal distribution [39][47].
- Step 3: Apply logistic ridge regression 2000 times using as response the summations $\boldsymbol{\hat{p}^T + r_{rq,t}^T}$, $t =$

$1, \ldots, 2000$, where $\hat{\boldsymbol{p}}^{\boldsymbol{T}}$ are the estimated probabilities from Step 1. In this step if a sum $\hat{p}_i + r_{rq,i} > 1$ then it becomes 1. Also if sum<0 then it becomes 0 and finally we round to the nearest integer if 0<sum<1 [47-48]. Moreover 2000 samples of $\hat{b}^{\lambda}$ and $\hat{p}$ can be obtained.

- Step 4: The standard errors of the estimated parameters $\hat{b}^{\lambda}$ can be obtained by finding the standard deviation of the 2000 bootstrapped samples $\hat{b}_1^{\lambda}, \ldots, \hat{b}_{23}^{\lambda}$.
- Step 5: From the 2000 sets of estimated response variables $\hat{p}_t, t = 1, \ldots, 2000$, we calculate 2000 new sets of randomized quantile residuals which allows us to construct a $\times 100\%$ rejection regions around the Q-Q plot of the randomized quantile residuals.

III. Results

The correlations between some variables are very strong and statistically significant, indicating the presence of multicollinearity. As a first step it is necessary to transform the categorical variables with more than two categories into dummy variables. For the detection of multicollinearity we may use the Condition Indices, by calculating the eigenvalues of the correlation matrix and other similar procedures as in linear regression models [44-45]. The condition indices are shown in Table III. Very large values of the last two condition indices (>30) show that collinearity among the variables exists.

Another problem caused by the multicollinearity is the large values of the standard errors of the estimated parameters, which makes the model unstable. Moreover, while the model according to the F-test seems to be statistically significant against the null hypothesis ($b_1 = b_2 = \cdots = b_{23} = 0$), the p-values of the individual terms are all greater than 0.05 which suggests that none of the variables is statistically significant. The above are included in Table V which contains the estimates of the initial logistic model.

Thus the logistic ridge regression is applied to generate an improved model with more stable parameter estimates for a ridge parameter λ=0 to λ=0.5. Furthermore when collinearity exists there is always a model for λ>0 for which the MSE is less than the MSE of the unrestricted model [28][32].

For the calculation of p – values the following statistic is used:

$$T_{\lambda} = \frac{\hat{b}_j^{\lambda}}{se\left(\hat{b}_j^{\lambda}\right)} \qquad (15)$$

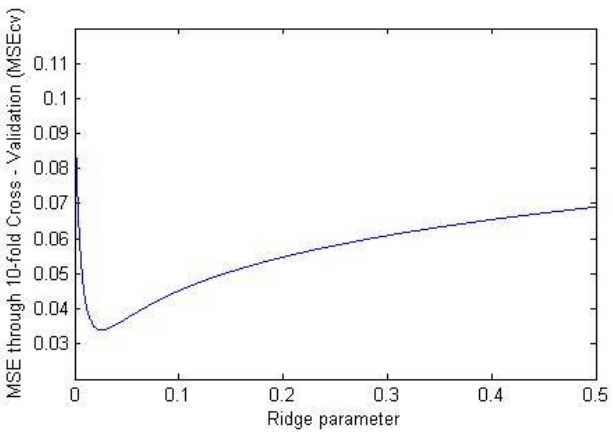

Fig. 1 Plot of MSEcv versus the Ridge Parameter λ

The standard errors were obtained by the bootstrap procedure that was described in section E. Thereafter we assume that under the null hypothesis $T_{\lambda} \sim N(0,1)$ to test the significance of the ridge coefficients [51].

TABLE III

1
2,2611
2,9374
3,0953
3,2883
3,3574
3,7078
4,0782
5,0419
5,0651
5,7507
5,9970
6,2390
7,4815
7,7832
9,4871
11,3555
13,9189
17,1612
17,5828
21,0211
43,0438
104,2085

Condition Indices for multicollinearity detection

For λ=0.0261 the minimum MSEcv is derived which is equal to 0.034 as shown in figure 1. The parameter estimates of the logistic ridge model are shown in Table VI. It now becomes clear that the prognostic factors Waist's perimeter, Congestion, Cough, Wheezing, Dyspnea, Bronchiolit1is

episodes until 3rd year are statistically significant. Moreover it is obvious that the estimates are now much more stable than the estimates of the initial unrestricted model and this is indicated from the two first columns of Table VI.

It is important to examine if the randomized quantile residuals have linear dependencies among themselves [52]. This can be done, if we plot the estimate of the autocorrelation coefficient. We obtain that the values are inside the confidence interval. Therefore, the randomized quantile residuals are uncorrelated. The autocorrelations of the randomized quantile residuals are described in Table IV and the plot of the estimate of the autocorrelation coefficients is shown in Figure 2.

TABLE IV

Lags	Autocorrelation	Bounds
1	0,0682	(-0.1990, 0.1990)
2	0,1883	(-0.1990, 0.1990)
3	0,0658	(-0.1990, 0.1990)
4	-0,1036	(-0.1990, 0.1990)
5	-0,0028	(-0.1990, 0.1990)
6	0,0668	(-0.1990, 0.1990)
7	-0,0187	(-0.1990, 0.1990)
8	0,0097	(-0.1990, 0.1990)
9	0,1268	(-0.1990, 0.1990)
10	0,073	(-0.1990, 0.1990)
11	0,0854	(-0.1990, 0.1990)
12	0,0108	(-0.1990, 0.1990)
13	0,0591	(-0.1990, 0.1990)
14	0,1368	(-0.1990, 0.1990)
15	-0,0586	(-0.1990, 0.1990)
16	-0,0667	(-0.1990, 0.1990)
17	-0,1024	(-0.1990, 0.1990)
18	-0,1055	(-0.1990, 0.1990)
19	0,0397	(-0.1990, 0.1990)
20	0,0497	(-0.1990, 0.1990)

Autocorrealtions of randomized quantile

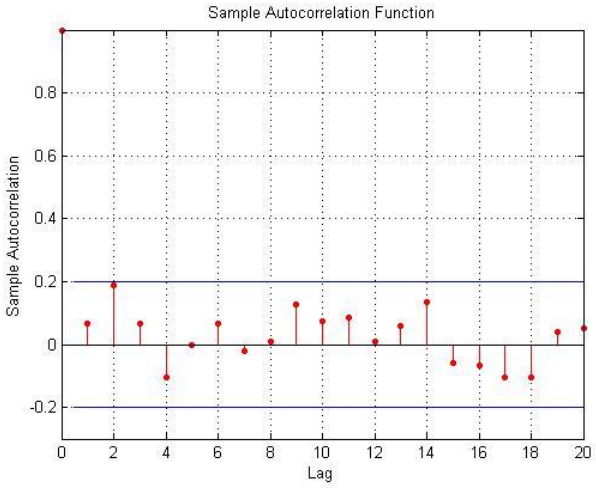

Fig. 2 Plot of autocorrelations of randomized quantile residuals

We now proceed to provide a test in order to check the validity of the model. Figure 3 shows the Q-Q plot of the randomized quantile residuals of the fitted logistic ridge model denoted with +. The 5% rejection regions were computed by the procedure described in Section E after 2000 bootstrap simulations. Only 1 (0.99%) of the 101 residuals lie outside the 5% rejection regions and generally the Q-Q plot does not present serious deviations from normality.

Here it is important to mention that if the percentage of the randomized quantile residuals that are outside the rejection regions is greater than 5%, then the model should be rejected.

In addition, the powerful Anderson-Darling test gives the value 0.3456 with a p-value 0.4810. Therefore, the null hypothesis that the randomized quantile residuals follow an approximate standard normal distribution cannot be rejected, which suggests that the fitted model is valid.

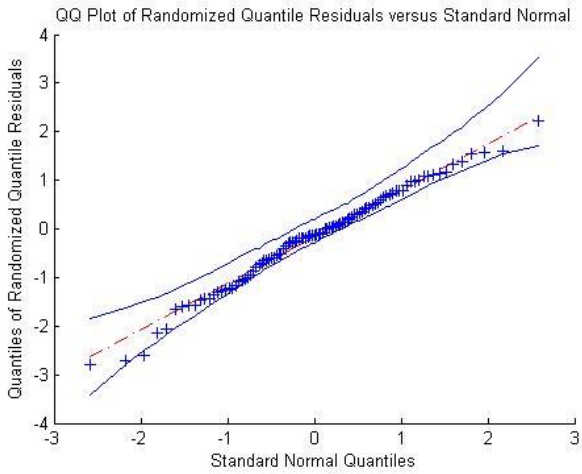

Fig. 3 QQ Plot of randomized quantile residuals versus Standard Normal

Finally we would like to examine the performance of the proposed model in new real data. These data were collected in a period after 2010 and refer to 33 new patients. Here, it is important that those new data have never been used before in any other study and this is the first study which makes a use of them. Regarding the group which was used to evaluate the prediction system in preschool patients who were re-evaluated in school age, the mean(±SD) age, FEV_1% predicted, FVC% predicted, FEF_{25-75}% predicted, IgE for the negative group [12 children with asthma subsided in school age] versus that of positive group [21 children with asthma persistence in school age] were 8.25±3.15 vs 9.5±2.48 years, 99.38±13.34 vs 102.4±11.9%, 94.13±13.34 vs 99±11.5%, 103.1±27.86 vs 114.3 ±24.66 %, 249.4±253.7 vs 225.6±64.45 IU/ml respectively. No difference was found regarding the age, FEV_1% predicted, FVC% predicted, FEF_{25-75}% predicted, IgE levels between the two groups (p=0.25, p=0.36, p=0.54 and p=0.86 respectively).

Based on the equation:

$$\hat{p}_{ridge} = \frac{1}{1 + \exp(-X_{new} * \hat{b}_{ridge})},$$

a prediction for the diagnosis of a new patient can be found. The positive predicted value, the negative predicted value and the accuracy of this model are estimated using false positive (FP), false negative (FN), true positive (TP), and true negative (TN) values. The test set consists of the new 33 patients and the 11 patients which were used for the cross – validation test.

$$Positive\ Pred.Value = \frac{N_{TP}}{N_{TP} + N_{FN}} \times 100,$$
$$Negative\ Pred.Value = \frac{N_{TN}}{N_{TN} + N_{FP}} \times 100, \quad (16)$$
$$Accuracy = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \times 100.$$

All the above are statistical measures of the performance of a binary classification test [53]. The obtained accuracy of asthma persistence in school age was 93.18%, while the positive predictive value and negative predictive value 96.15% and 88.89% respectively. The use of an independent data set is very important for the evaluation of generality of the constructed model.

IV. Discussion

Recent attention has turned toward alternative forms of analysis, including Support Vector Machines (SVMs), logistic regression analysis and neural networks (ANNs) which are commonly used statistical models in medical predictions. SVMs is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data [54]. The foundations of SVMs have been developed by Vapnik [55] and gained popularity due to many promising features. On the other hand, ANNs have an excellent capability of learning the relationship between the input-output mapping from a given dataset without any prior knowledge or assumptions about the statistical distribution of the data. This capability of learning from a certain dataset without any a priori knowledge makes the neural networks quite suitable for classification and prediction tasks in practical situations. Furthermore, neural networks are inherently nonlinear which makes them more practicable for accurate modelling of complex data patterns, as opposed to many traditional methods based on linear techniques.

One of the disadvantages of ANNs when compared to logistic regression models is that ANNs frequently have difficulty analysing systems which have a large number of inputs due to the large amount of time taken to train the system as well as possibly over-fitting the model during the training time. Linear and logistic regression models have less potential for over-fitting primarily because the range of functions they can model is limited. Moreover, ANNs are a relative "black box" in comparison to a logistic regression model. The network is trained itself and determines which input variables are the most important.

In the present study, we have evaluated a novel asthma persistence prediction system. To our knowledge this study is the first to evaluate a model based on ridge regression in asthma outcome prediction. Our results showed a high degree of accuracy (93.18 %). Accordingly previous studies of our group [53][56] have shown similar asthma persistence prediction accuracies experimentally.

In the study [53] the used method is based on Multi-Layer Perceptron Neural Networks and Probabilistic Neural Networks architectures. By employing Partial Least Square regression, 9 prognostic factors correlated to the persistent asthma have been chosen. Various Neural Networks topologies have been investigated in order to obtain the best prediction accuracy. Based on the results, it is shown that the proposed system is able to predict the asthma outcome with 96.77% success. Due to the fact that asthma is a serious condition, the various models that have been used to detect it, must have high sensitivity so that patients with asthma are not overlooked. An ANN that has been trained to predict 96.77% of patients with asthma may be very useful to physicians.

In the study [56] the proposed intelligent system consists of three stages. At the first stage, Principal Component Analysis is used for feature extraction and dimension reduction. At the second stage, the pattern classification is achieved by using Least Square Support Vector Machine Classifier. Finally, at the third stage the performance evaluation of system is estimated by using classification accuracy and 10-fold cross-validation. The proposed prediction system can be used in asthma outcome prediction with 95.54 % success.

In the present study, the prognostic factors waist's perimeter, nasal congestion, cough, wheezing, dyspnea, bronchiolitis episodes until 3rd year are statistically significant. The significance of these factors in asthma prediction have been indicated by previous studies [53][56] while bronchiolitis episodes and waist's perimeter were highlighted as important in study [53]. Particularly the number of wheezing episodes have been evaluated by the majority of predictive models indicating the importance of the duration of symptoms throughout childhood [4][57-59]. Interestingly the identification of the waist's perimeter as a prognostic factor enhance the recently studied link between asthma and obesity [60-61].

The comparison of the predictive accuracy of the proposed intelligent system with that of other group's studies [57-59][62] is quite difficult since there are numerous differences in study design and objectives. Most of the previous studies of asthma prediction resulted in a correlation of one or two predicting factors with asthma persistence. Although there is an increase scientific interest in asthma outcome after the age of five, and valuable studies have been published, the models using a small number of predicting factors are not able to achieve substantially high predictive accuracies. This outcome should probably be expected since asthma phenotypes are the result of the multi-factorial influence of the environment to a diverse genotype, so models using a limited number of factors for the prediction of a disease with multiple phenotypes such as asthma. usually end up having low predictive accuracy. It is therefore meaningful to utilize computational intelligence methods in order to include and validate many factors in asthma outcome prediction.

V. Conclusion

In this paper, a new intelligent method based on the Logistic Ridge Regression for asthma persistence prediction has been validated in preschool patients in real time data. The proposed model predicted the persistence of asthma at the approximate age of nine years with an accuracy of 93.18%, positive predictive value of 96.15% and negative predictive value of 88.89%. To conclude, despite the fact there are several limitation of this study since the number of patients used for validation in real life setting is small, the proposed method exhibits high accuracy in asthma persistence prediction and shows the importance priority of each factor in asthma persistence. A better prediction rate will be possible by increasing the patient data and further clinical evaluation may enhance the implications of the present study. Finally, for future research, we could collect data from different regions, with different environmental and climatic factors, to examine if asthma prediction is affected by them.

TABLE V

Covariates	Estimates			
	Parameter Estimates	*Standard Errors*	*t-stats*	*p-values*
Age	-0,42579	1,0983	-0,3877	0,6983
Treatment	-4,26963	63,1442	-0,0676	0,9461
Corticosteroids inhaled	6,2993	63,3275	0,0995	0,9208
Antileukotriene	-2,74189	4,6348	-0,5916	0,5541
antihistamine	-5,48657	8,9273	-0,6146	0,5388
height	-65,4392	73,1685	-0,8944	0,3711
weight	0,745009	0,7693	0,9684	0,3329
waist's perimeter	-0,2428	0,2026	-1,1982	0,2308
allergic rhinitis	-2,79072	6,3479	-0,4396	0,6602
allergic conjunctivitis	2,148512	10,3675	0,2072	0,8358
runny nose	1,70444	6,2093	0,2745	0,7837
congestion	4,430885	6,2917	0,7042	0,4813
Cough	18,79093	26,8595	0,6996	0,4842
Wheezing	-13,4146	23,2761	-0,5763	0,5644
dyspnea	9,81705	11,7808	0,8333	0,4047
seasonal symptoms (none)	48,86427	50,4889	0,9678	0,3331
seasonal symptoms (winter)	55,20393	55,0824	1,0022	0,3162
seasonal symptoms (autumn)	65,33002	93,9732	0,6952	0,4869
seasonal symptoms (spring)	74,89878	78,2987	0,9566	0,3388
seasonal symptoms (summer)	55,7921	60,1451	0,9276	0,3536
seasonal symptoms (>2 seasons)	67,24084	70,5941	0,9525	0,3408
Bronchiolitis episodes until 3rd year	1,620487	2,6564	0,6100	0,5418
Bronchiolitis episodes b/w 3rd – 5th year	-0,44338	1,0864	-0,4081	0,6832

The initial logistic regression model

TABLE VI

Covariates	Estimates			
	Parameter Estimates	*Standard Errors*	T_λ	*p-values*
Age	0,059821	0,1221	0,4900	0,6241
Treatment	0,531238	0,4817	1,1028	0,2701
Corticosteroids inhaled	0,889768	0,4942	1,8002	0,0718
Antileukotriene	-0,32763	0,5650	-0,5799	0,5620
Antihistamine	0,111097	0,6674	0,1665	0,8678
Height	0,600211	0,7353	0,8163	0,4143
Weight	-0,01082	0,0276	-0,3925	0,6947
Waist's perimeter	-0,06579	0,0216	-3,0455	0,0023
Allergic rhinitis	-0,06907	0,5733	-0,1205	0,9041
Allergic conjunctivitis	-0,72709	0,5819	-1,2494	0,2115
Runny nose	0,429069	0,6068	0,7071	0,4795
Congestion	1,096703	0,5424	2,0220	0,0432
Cough	1,640577	0,5871	2,7942	0,0052
Wheezing	1,719255	0,5874	2,9271	0,0034
Dyspnea	1,18429	0,5962	1,9865	0,0470
Seasonal symptoms (none)	1,26928	0,7360	1,7246	0,0846
Seasonal symptoms (winter)	1,148824	0,8505	1,3507	0,1768
Seasonal symptoms (autumn)	0,273617	0,8385	0,3263	0,7442
Seasonal symptoms (spring)	0,856916	0,9088	0,9429	0,3457
Seasonal symptoms (summer)	0,216933	0,8830	0,2457	0,8059
Seasonal symptoms (>2 seasons)	1,15655	0,8174	1,4149	0,1571
Bronchiolitis episodes until 3rd year	-0,21639	0,1089	-1,9866	0,0470
Bronchiolitis episodes b/w 3rd – 5th year	0,132402	0,0932	1,4212	0,1553

The ridge logistic regression model

CONFLICT OF INTERESTS

The authors report no conflict of interests.

References

[1] C. Porsbjerg, M.L. von Linstow, C. Ulrik, S. Nepper-Christensen, V. Backer, "Risk factors for onset of asthma: a 12-year prospective follow-up study," *Chest*, vol. 129, no. 2, pp. 309–16, 2006.

[2] N. N. Hansel, E. C. Matsui, R. Rusher, M. C. McCormack, J. Curtin-Brosnan, R. D. Peng, D. Mazique, P. N. Breysse, G. B. Diette, "Predicting future asthma morbidity in preschool inner-city children, " *Journal of Asthma*, vol. 48, no.8, pp. 797–803, 2011.

[3] *A. Tattersfield and M. McNicol, Respiratory Disease, Springer-Verlag London, 1987.*

[4] A. Bush, "Diagnosis of asthma in children under five, " *Prim Care Respir J*, vol. 16, pp. 7–15, 2007.

[5] Clough JB, Keeping KA, Edwards LC, et al. Can we predict which wheezy infants will continue to wheeze? Am J Respir Crit Care Med 1999;160:1473–80.

[6] Nickel R, Lau S, Niggemann B, et al. Messages from the German Multicenter Allergy Study. Pediatr Allergy Immunol 2002;13:7–10.
[7] Custovic A, Simpson BM, Murray CS, et al., NAC Manchester Asthma and Allergy Study Group. The National Asthma Campaign Manchester Asthma and Allergy Study. Pediatr Allergy Immunol 2002;13:32–7.
[8] Taussig LM, Wright AL, Holberg CJ, et al. Tuscon Children's Respiratory Study:1980 to present. J Allergy Clin Immunol 2003;111:661–75.
[9] Kurukulaaratchy RJ, Fenn M, Matthews S, et al. Characterisation of atopic and no-atopic wheeze in 10 year old children. Thorax 2004;59:563–8.
[10] Illi S, von Mutius E, Lau S, et al. Perennial allergen sensitisation early in life and chronic asthma in children: a birth cohort study. Lancet 2006;368:763–70.
[11] Savenije OE, Granell R, Caudri D, et al. Comparison of childhood wheezing phenotypes in 2 birth cohorts: ALSPAC and PIAMA. J Allergy Clin Immunol 2011;127:1505–12.
[12] Castro-Rodrıguez JA, Holberg CJ, Wright AL, Martinez FD. A clinical index to define risk of asthma in young children with recurrent wheezing. Am J Resp Crit Care Med 2000;162:1403–6.
[13] Guilbert TW, Morgan WJ, Krawiec M, et al. The Prevention of Early Asthma in Kids study: design, rationale and methods for the Childhood Asthma Research and Education network. Control Clin Trials 2004;25:286–310.
[14] Kurukulaaratchy RJ, Matthews S, Holgate ST, et al. Predicting persistent disease among children who wheeze during early life. Eur Respir J 2003;22:767–71.
[15] Devulapalli CS, Carlsen KC, Haland G, et al. Severity of obstructive airway disease by age 2 years predicts astma at 10 years of age. Thorax 2008;63:8–13.
[16] Caudri D, Wijga A, Schipper MA, et al. Predicting the long-term prognosis of children with symptoms suggestive of asthma at preschool age. J Allergy Clin Immunol 2009;124:903–10.
[17] Brand PL. The asthma predictive index: not a useful tool in clinical practice. J Allergy Clin Immunol 2011;127:293–4.
[18] Castro-Rodriguez JA, Cifuentes L, Rodriguez-Martinez CE. The asthma predictive index remains a useful tool to predict asthma in young children with recurrent wheeze in clinical practice. J Allergy Clin Immunol 2011;127:1082–3.
[19] Fouzas S, Brand PL. Predicting persistence of asthma in preschool wheezers:crystal balls or muddy waters? Paediatr Respir Rev. 2013 Mar;14(1):46-52.
[20] Sly PD, Boner AL, Bjorksten B, et al. Early identification of atopy in the prediction of persistent asthma. Lancet 2008;372:1100–6.
[21] Holt PG, Sly PD. Prevention of allergic respiratory disease in infants: current aspects and future perspectives. Curr Opin Allergy Clin Immunol 2007;7:547–55.
[22] Vercelli D. Discovering susceptibility genes for asthma and allergy. Nat Rev Immunol 2008;8:169–82.
[23] Sly PD, Holt PG. Role of innate immunity in the development of allergy and asthma. Curr Opin Allergy Clin Immunol 2011;11:127–31.
[24] R. K. Jain, "Ridge Regression and its Application to Medical Data, " *Computers and Biomedical Research*, vol. 18, pp. 363–368, 1984.
[25] R. Frisch, *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo: University Institute of Economics, Publication no. 5, 1934.
[26] A. E. Hoerl, "Application of ridge analysis to regression problems, " *Chemical Engineering Progress*, vol. 58, no.3, pp. 54–59, 1962.
[27] A.E. Hoerl and R. W. Kennard, "Ridge Regression: Applications to nonorthogonal problems," *Technometrics 1*\, vol. 12, no. 6, 1970.
[28] A.E. Hoerl and R.W. Kennard, "Ridge Regression: Biased estimates for nonorthogonal problems," *Technometrics*, vol. 12, no.55, pp.55–67 1970.
[29] R. L. Schaefer, L.D. Roi, and R.A. Wolfe, "A ridge logistic estimator, " *Communications in Statistics - Theory and Methods*, vol. 13, no. 1, pp. 99–113, 1984.
[30] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. London: Chapman & Hall, 1989.
[31] Annete J. Dobson, *An Introduction To Generalized Linear Models*, 2nd ed.: Chapman & Hall, 2002.
[32] S. Le Cessie and J. C. Van Houwlingen, "Ridge Estimators in Logistic Regression," *Journal of the Royal Statistical Society*, vol. 41, no. 1, pp. 191–201, 1992.
[33] D.R. Brillinger, K.A. Lindsay, and J.R. Rosenberg, "Combining frequency and time domain approaches to systems with multiple spike train input and output," *Biological Cybernetics*, vol. 100, pp. 459–474, 2009.
[34] A.V. Dorugade et al, "New ridge parameters for ridge regression," Journal of the Association of Arab Universities for Basic and Applied Sciences, vol. 15, pp. 94-99, 2014.
[35] Syaiba Balqish Arriffin and Habshah Midi, "Robust Logistic Ridge Regression Estimator in the Presence of High Leverage Multicollinear Observations," in Mathematical and Computational Methods in Science and Engineering, Kuala Lumpur, pp. 179-184, 2014.
[36] Hsiang - Chuan Liu, Chin - Chun Chen, Der - Bang Wu, and Tian - Wei Sheu, "Theory and Application of the Composed Fuzzy Measure of L-Measure and Delta-Measures," WSEAS TRANSACTIONS on SYSTEMS and CONTROL, vol. 4, no. 8, pp. 359-368, 2009
[37] Yongming Li, Qingming Gui, Yongwei Gu, Songhui Han, and Kai Du, "Ridge-Type Kalman Filter and Its Algorithm," WSEAS TRANSACTIONS on MATHEMATICS, vol. 13, pp. 852-862, 2014.
[38] D.A. Pierce and D.W. Schafer, "Residuals in Generalized Linear Models," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 977–986, 1986.
[39] P. Dunn and G. K. Smyth, "Randomized Quantile Residuals, " *J. Computat. Graph. Statist*, vol. 5, pp. 236–244, 1996.
[40] Stephens M.A.,. Test based on regression and correlation. In: D'Agostino, R.B., Stephens, M.A. (Eds.), Goodness-of-Fit Techniques, Marcel Dekker, New York, 195-234, 1986
[41] Coin D. A goodness-of-fit test for normality based on polynomial regression. Comp. Stat. and Data Anal. 52, 2185-2198, 2008.
[42] V.G. Vassiliadis and A.G. Rigas, "A new formulation of the Hinich's bispectral test for linearity based on a novel Q-Q plot for testing distributional hypotheses," in ICCRA 3, C. P. Kitsos, C. Caroni Proceedings, Porto Heli, Greece, 2009.
[43] B. Efron and R.J. Tibshirani, "An Introduction to the Bootstrap, " (Chapman & Hall, New York), 1993.
[44] E. Lesaffre, E. and B.D. Marx, "Collinearity in Generalized Linear Regression,"*Communications in Statistics – Theory and Methods*, vol. 22, no. 7, pp. 1933–1952, 1993.
[45] B. D. Marx and E. P. Smith, "Weighted Multicollinearity in Logistic Regression: Diagnostics and Biased Estimation Techniques with an Example From Lake Acidification," *Canadian Journal of Fisheries and Aquatic Sciences*, vol.47, no. 6, pp. 1128–1135, 1990.
[46] T.W. Anderson and D.A. Darling, "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes," *The Annals of Mathematical Statistics*, vol. 23, no. 2, pp. 193–212, 1952.
[47] H. Friedl and N. Tilg , "Variance estimates in logistic regression using the bootstrap, " *Communications in Statistics - Theory and Methods*, vol. 24, no. 2, pp. 473–486, 1995.
[48] D. Firth, J. Glosup, and D.V. Hinkley, "Model Checking with nonparametric curves," Biometrica, vol. 78(2), pp. 245-252, 1991.
[49] B. Efron, Bootstrap methods: "Another look at the jackknife., " *Ann.Statist.*, vol. 7, pp. 1–26, 1979.
[50] D.A. Freedman , " Bootstrapping regression models, " *Ann. Statist.*, vol. 9, pp. 1218–1228, 1981.
[51] Cule et al., "Significance testing in ridge regression for genetic data," *BMC Bioinformatics*, 12:372, 2011.
[52] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel, Time Series Analysis: Forecasting and Control, 4th ed.: Wiley Series in Probability and Statistics, 2008.
[53] E. Chatzimichail, E. Paraskakis, and A. Rigas, "Predicting Asthma Outcome Using Partial Least Square Regression and Artificial Neural Networks," *Advances in Artificial Intelligence*, vol. 2013, Article ID 435321, 7 pages, 2013. doi:10.1155/2013/435321
[54] A. Widodo, B. Yang, Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors, Expert Systems with Applications 33 (2007) 241–250.
[55] V. Vapnik, The Support Vector Method, In Proceeding of ICANN (1997) 263–271.
[56] E. Chatzimichail, E. Paraskakis, M. Sitzimi, A. Rigas, An intelligent system approach for asthma prediction in symptomatic preschool children, Comput Math Methods Med 2013, doi:10.1155/2013/240182
[57] Caudri D, Wijga A, Chipper CM, Hoekstra M, Postma DS, Koppelman GH, Brunekreef B, Smit HA, DE Jongste JC. Predicting the long-term prognosis of children with symptoms suggestive of asthma at preschool age. J Allergy Clin Immunol. 2009; 124(5):903-910
[58] Clough JB, Keeping KA, Edwards LC, Freeman WM, Warner JA , Warner JO. Can we predict which wheezy infants will continue to wheeze? Am J Respir Crit Care Med. 1999; 160:1473-1480

[59] Castro-Rodriguez JA, Holberg CJ, Wright AL, Martinez FD. A clinical index to define risk of asthma in young children with recurrent wheezing. Am J Respir Crit Care Med. 2000; 162:1403-1406

[60] McGarry ME, Castellanos E, Thakur N, Oh SS, Eng C, Davis A, Meade K, LeNoirMA, Avila PC, Farber HJ, Serebrisky D, Brigino-Buenaventura E, Rodriguez-Cintron W, Kumar R, Bibbins-Domingo K, Thyne SM, Sen S, Rodriguez-Santana JR, Borrell LN,Burchard EG. Obesity and Bronchodilator Response in African-American and HispanicChildren and Adolescents with Asthma. Chest. 2015 Mar 5. doi:10.1378/chest.14-2689.

[61] Spathopoulos D, Paraskakis E, Trypsianis G, Tsalkidis A, Arvanitidou V,Emporiadou M, Bouros D, Chatzimichael A. The effect of obesity on pulmonary lung function of school aged children in Greece. Pediatr Pulmonol. 2009;44(3):273-80

[62] Devulapalli CS, Carlsen KC, Håland G, Munthe-Kaas MC, Pettersen M, Mowinckel P, et al. Severity of obstructive airways disease by two years predicts asthma at 10 years of age. Thorax. 2008; 63: 8-13

APPENDIX
Correlation matrix of the explanatory variables

	Age	Treatment	Corticosteroids inhaled	Antileukotriene	Antihistamine	Height	Weight	Waist's perimeter	Allergic rhinitis	Allergic conjunctivitis	Runny nose	Congestion	Cough	Wheezing	Dyspnea	Seasonal symptoms (none)	Seasonal symptoms (winter)	Seasonal symptoms (autumn)	Seasonal symptoms (spring)	Seasonal symptoms (summer)	Seasonal symptoms (>2 seasons)	Bronchiolitis episodes until 3rd year	Bronchiolitis episodes b/w 3rd – 5th year
Age	1,00	-0,10	-0,14	-0,05	0,08	0,77	0,59	0,32	-0,01	0,18	0,05	-0,05	-0,17	-0,09	0,04	0,01	-0,22	0,02	0,16	0,08	0,10	-0,10	-0,07
Treatment	-0,10	1,00	0,88	0,46	0,40	0,07	0,16	0,12	0,35	0,35	0,41	0,43	0,55	0,43	0,48	-0,66	0,30	0,21	0,04	0,14	0,35	0,58	0,56
Corticosteroids inhaled	-0,14	0,88	1,00	0,51	0,44	0,01	0,11	0,07	0,27	0,34	0,34	0,50	0,49	0,46	0,50	-0,59	0,35	-0,13	0,06	0,16	0,34	0,61	0,57
Antileukotriene	-0,05	0,46	0,51	1,00	0,37	0,05	0,24	0,29	0,14	0,25	0,15	0,31	0,27	0,34	0,25	-0,26	0,16	-0,07	0,07	0,07	0,13	0,36	0,25
Antihistamine	0,08	0,40	0,44	0,37	1,00	0,15	0,14	0,05	0,29	0,33	0,23	0,20	0,15	0,27	0,20	-0,26	-0,09	-0,06	0,27	-0,07	0,37	0,25	0,17
Height	0,77	0,07	0,01	0,05	0,15	1,00	0,85	0,51	0,00	0,25	0,05	0,08	0,00	0,01	0,14	-0,10	-0,14	0,10	0,11	0,18	0,08	0,00	0,00
Weight	0,59	0,16	0,11	0,24	0,14	0,85	1,00	0,74	0,04	0,30	0,06	0,21	0,16	0,14	0,27	-0,18	-0,06	0,07	0,06	0,21	0,13	0,08	0,05
Waist's perimeter	0,32	0,12	0,07	0,29	0,05	0,51	0,74	1,00	0,07	0,16	0,03	0,17	0,14	0,10	0,27	-0,15	0,05	0,10	-0,07	-0,01	0,14	0,11	0,09
Allergic rhinitis	-0,01	0,35	0,27	0,14	0,29	0,00	0,04	0,07	1,00	0,66	0,69	0,49	0,52	0,52	0,35	-0,38	0,02	0,14	0,09	0,20	0,26	0,39	0,37
Allergic conjunctivitis	0,18	0,35	0,34	0,25	0,33	0,25	0,30	0,16	0,66	1,00	0,56	0,57	0,38	0,55	0,52	-0,39	-0,08	-0,09	0,15	0,28	0,42	0,46	0,42
Runny nose	0,05	0,41	0,34	0,15	0,23	0,05	0,06	0,03	0,69	0,56	1,00	0,60	0,41	0,54	0,41	-0,41	-0,07	0,15	0,10	0,21	0,39	0,41	0,38
Congestion	-0,05	0,43	0,50	0,31	0,20	0,08	0,21	0,17	0,49	0,57	0,60	1,00	0,50	0,61	0,59	-0,46	0,21	-0,12	-0,03	0,30	0,29	0,44	0,37
Cough	-0,17	0,55	0,49	0,27	0,15	0,00	0,16	0,14	0,52	0,38	0,41	0,50	1,00	0,71	0,54	-0,67	0,36	0,15	-0,03	0,17	0,35	0,69	0,71
Wheezing	-0,09	0,43	0,46	0,34	0,27	0,01	0,14	0,10	0,52	0,55	0,54	0,61	0,71	1,00	0,65	-0,54	0,05	-0,03	0,04	0,25	0,51	0,66	0,58
Dyspnea	0,04	0,48	0,50	0,25	0,20	0,14	0,27	0,27	0,35	0,52	0,41	0,59	0,54	0,65	1,00	-0,42	0,10	0,01	-0,14	0,08	0,46	0,56	0,58
Seasonal symptoms (none)	0,01	-0,66	-0,59	-0,26	-0,26	-0,10	-0,18	-0,15	-0,38	-0,39	-0,41	-0,46	-0,67	-0,54	-0,42	1,00	-0,52	-0,17	-0,20	-0,20	-0,46	-0,55	-0,57
Seasonal symptoms (winter)	-0,22	0,30	0,35	0,16	-0,09	-0,14	-0,06	0,05	0,02	-0,08	-0,07	0,21	0,36	0,05	0,10	-0,52	1,00	-0,09	-0,11	-0,11	-0,25	0,26	0,32
Seasonal symptoms (autumn)	0,02	0,21	-0,13	-0,07	-0,06	0,10	0,07	0,10	0,14	-0,09	0,15	-0,12	0,15	-0,03	0,01	-0,17	-0,09	1,00	-0,04	-0,04	-0,08	0,02	0,01
Seasonal symptoms (spring)	0,16	0,04	0,06	0,07	0,27	0,11	0,06	-0,07	0,09	0,15	0,10	-0,03	-0,03	0,04	-0,14	-0,20	-0,11	-0,04	1,00	-0,04	-0,09	-0,01	-0,04
Seasonal symptoms (summer)	0,08	0,14	0,16	0,07	-0,07	0,18	0,21	-0,01	0,20	0,28	0,21	0,30	0,17	0,25	0,08	-0,20	-0,11	-0,04	-0,04	1,00	-0,09	0,12	0,09
Seasonal symptoms (>2 seasons)	0,10	0,35	0,34	0,13	0,37	0,08	0,13	0,14	0,26	0,42	0,39	0,29	0,35	0,51	0,46	-0,46	-0,25	-0,08	-0,09	-0,09	1,00	0,37	0,36
Bronchiolitis episodes until 3rd year	-0,10	0,58	0,61	0,36	0,25	0,00	0,08	0,11	0,39	0,46	0,41	0,44	0,69	0,66	0,56	-0,55	0,26	0,02	-0,01	0,12	0,37	1,00	0,89
Bronchiolitis episodes b/w 3rd – 5th year	-0,07	0,56	0,57	0,25	0,17	0,00	0,05	0,09	0,37	0,42	0,38	0,37	0,71	0,58	0,58	-0,57	0,32	0,01	-0,04	0,09	0,36	0,89	1,00

Correlations between the predictor variables

	Age	Treatment	Corticosteroids inhaled	Antileukotriene	Antihistamine	Height	Weight	Waist's perimeter	Allergic rhinitis	Allergic conjunctivitis	Runny nose	Congestion	Cough	Wheezing	Dyspnea	Seasonal symptoms (none)	Seasonal symptoms (winter)	Seasonal symptoms (autumn)	Seasonal symptoms (spring)	Seasonal symptoms (summer)	Seasonal symptoms (>2 seasons)	Bronchiolitis episodes until 3rd year	Bronchiolitis episodes b/w 3rd – 5th year
Age	1,00	0,11	0,08	0,87	0,65	0,00	0,00	0,00	0,46	0,06	0,89	0,44	0,08	0,10	0,47	0,24	0,01	0,60	0,09	0,52	0,99	0,16	0,15
Treatment	0,11	1,00	0,00	0,00	0,00	0,90	0,82	0,82	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,11	0,46	0,42	0,00	0,00	0,00
Corticosteroids inhaled	0,08	0,00	1,00	0,00	0,00	0,80	0,97	0,89	0,01	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,24	0,38	0,37	0,00	0,00	0,00
Antileukotriene	0,87	0,00	0,00	1,00	0,00	0,41	0,17	0,13	0,09	0,02	0,11	0,01	0,00	0,00	0,00	0,00	0,03	0,51	0,19	0,48	0,03	0,00	0,00
Antihistamine	0,65	0,00	0,00	0,00	1,00	0,41	0,53	0,91	0,00	0,00	0,01	0,03	0,02	0,00	0,04	0,00	0,67	0,62	0,03	0,55	0,00	0,00	0,01
Height	0,00	0,90	0,80	0,41	0,41	1,00	0,00	0,00	0,41	0,02	0,90	0,90	0,54	0,35	0,91	0,63	0,03	0,30	0,17	0,26	0,90	0,54	0,47
Weight	0,00	0,82	0,97	0,17	0,53	0,00	1,00	0,00	0,67	0,01	0,97	0,30	0,44	0,84	0,25	0,71	0,28	0,39	0,60	0,13	0,69	1,00	0,82
Waist's perimeter	0,00	0,82	0,89	0,13	0,91	0,00	0,00	1,00	0,77	0,17	0,81	0,40	0,23	0,65	0,06	0,46	0,65	0,43	0,37	0,88	0,49	0,66	0,86
Allergic rhinitis	0,46	0,00	0,01	0,09	0,00	0,41	0,67	0,77	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,29	0,41	0,43	0,09	0,03	0,00	0,00
Allergic conjunctivitis	0,06	0,00	0,03	0,02	0,00	0,02	0,01	0,17	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,97	0,52	0,16	0,02	0,00	0,00	0,00
Runny nose	0,89	0,00	0,00	0,11	0,01	0,90	0,97	0,81	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,91	0,44	0,47	0,10	0,00	0,00	0,00
Congestion	0,44	0,00	0,00	0,01	0,03	0,90	0,30	0,40	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,06	0,35	0,61	0,01	0,00	0,00	0,00
Cough	0,08	0,00	0,00	0,00	0,02	0,54	0,44	0,23	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,18	0,74	0,10	0,00	0,00	0,00
Wheezing	0,10	0,00	0,00	0,00	0,00	0,35	0,84	0,65	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,17	0,82	0,41	0,04	0,00	0,00	0,00
Dyspnea	0,47	0,00	0,00	0,00	0,04	0,91	0,25	0,06	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,11	0,59	0,56	0,96	0,00	0,00	0,00
Seasonal symptoms (none)	0,24	0,00	0,00	0,00	0,00	0,63	0,71	0,46	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,17	0,03	0,09	0,00	0,00	0,00
Seasonal symptoms (winter)	0,01	0,01	0,00	0,03	0,67	0,03	0,28	0,65	0,29	0,97	0,91	0,06	0,00	0,17	0,11	0,00	1,00	0,46	0,24	0,36	0,00	0,00	0,00
Seasonal symptoms (autumn)	0,60	0,11	0,24	0,51	0,62	0,30	0,39	0,43	0,41	0,52	0,44	0,35	0,18	0,82	0,59	0,17	0,46	1,00	0,76	0,81	0,46	0,66	0,52
Seasonal symptoms (spring)	0,09	0,46	0,38	0,19	0,03	0,17	0,60	0,37	0,43	0,16	0,47	0,61	0,74	0,41	0,56	0,03	0,24	0,76	1,00	0,71	0,24	0,99	0,87
Seasonal symptoms (summer)	0,52	0,42	0,37	0,48	0,55	0,26	0,13	0,88	0,09	0,02	0,10	0,01	0,10	0,04	0,96	0,09	0,36	0,81	0,71	1,00	0,36	0,18	0,34
Seasonal symptoms (>2 seasons)	0,99	0,00	0,00	0,03	0,00	0,90	0,69	0,49	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,46	0,24	0,36	1,00	0,00	0,00
Bronchiolitis episodes until 3rd year	0,16	0,00	0,00	0,00	0,00	0,54	1,00	0,66	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,66	0,99	0,18	0,00	1,00	0,00
Bronchiolitis episodes b/w 3rd – 5th year	0,15	0,00	0,00	0,00	0,01	0,47	0,82	0,86	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,52	0,87	0,34	0,00	0,00	1,00

P-values of correlations