A Bayesian Logistic Regression approach in Asthma Persistence Prediction

Ioannis I. Spyroglou [1], Gunter Spöck [2], Eleni A. Chatzimichail [1], Alexandros G. Rigas [1], E.N. Paraskakis [3]

(1) Electrical Engineering Department, Democritus University of Thrace, Xanthi, Greece
(2) Statistics Department, Alpen – Adria Universität, Klagenfurt, Austria
3) Paediatric respiratory unit, Paediatric Department, Medical School, Democritus University of Thrace, Alexandroupolis, Greece

CORRESPONDING AUTHOR: *Ioannis I. Spyroglou - Telephone number: +306955954849, ORCID: 0000-0001-6680-3656, Address: Department of Electrical and Computer Engineering, Kimmeria Campus, Building B, Office 2.18, Xanthi 67100, Greece - e-mail: ispyrogl@ee.duth.gr*

DOI: ***10.2427/12777***
Accepted on March 8, 2018

ABSTRACT

Background: Previous models based on a limited number of clinical parameters that have been used so far failed to exhibit high accuracy of prediction of asthma persistence in children. The number and significance of factors that are used in a proposed model play a cardinal role in prediction accuracy. Different models may lead to different significant variables. In addition, the accuracy of a model in medicine is really important since an accurate prediction of illness persistence may improve prevention and treatment intervention for the children at risk. The aim of this study is to evaluate a model that could effectively and accurately predict asthma persistence in children.
Methods: Data from 147 asthmatic children were analyzed by a new method for predicting asthma outcome using Principal Component Analysis (PCA) in combination with a Bayesian logistic regression approach implemented by the Markov Chain Monte Carlo (MCMC). The use of PCA is required due to multicollinearity among the explanatory variables.
Results: This method using the most appropriate models seems to predict asthma with an accuracy of 84.076%, 84.924%, 86.3673% and 86.1951%, a Sensitivity of 84.96%, 85.49%, 87.25% and 86.38% and a Specificity of 83.22%, 84.37%, 85.52% and 86.02% respectively.
Conclusion: Our approach predicts asthma with high accuracy, gives steadier results in terms of positive and negative patients and provides better information about the influence of each factor (demographic, symptoms etc.) in asthma prediction.

Key words: Asthma outcome, Multicollinearity, Bayesian Logistic Regression, Markov Chain Monte Carlo, Principal Component Analysis

INTRODUCTION

Asthma is the most common chronic disease in childhood. The diagnosis of asthma at the preschool age and the prognosis of its persistence later in life is extremely challenging since wheezing is present in several heterogeneous disorders at this age [1,2,3]. It is now well known that approximately two thirds of the preschool wheezers no longer wheeze after the age of six years [2]. So far there are numerous efforts to identify factors related

with asthma persistence and create prediction models using simple clinical and laboratory parameters of the asthmatic children [4,5,6,7,8]. The asthma predictive index (API) was one of the first clinical indices used for the prediction of asthma persistence beyond the preschool age [4]. Since then similar predicting models such as Isle of Wright score [5], ECA score [6] and PIAMA [8] have been introduced as a result of recent studies with substantial number of asthmatic children. Despite the simplicity and popularity of these scores their clinical usefulness was questioned by a number of recent studies as most of them have not a sufficient negative predictive power [3, 7, 9, 10]. The poor clinical performance of the above simple predictive models is actually not surprising. Most of them use a limited number of simple parameters such as frequency, duration or severity of wheezing episodes, presence of nasal symptoms, eczema or parental atopic disease [3] for the prediction of persistence of a disease such as asthma which presents an outstanding phenotypic and genetic variability so the number of factors affecting its prognosis is vast. So it is expectable that the use of more parameters that take into account the interaction of the environment with various genetic and phenotypic features of the disease will lead us to more complicated but more effective predictive values. As previous studies on disease prediction have shown, the use of statistical methods or machine learning techniques may enable the validation of multiple factors contributing to more accurate prognosis [11,12,13].

Logistic regression is a widespread method for the analysis and prediction of binary or categorical responses in biomedical studies. In many cases, when several predictor variables are present, strong correlations, which imply multicollinearity among them, make the prediction of a binary response difficult [14].

One approach that is widely used for dealing with multicollinearity is Principal Component Analysis (PCA) [15]. This method is also used for asthma persistence prediction in [11] combined with Least Squares Support Vector machine (LSSVM) classifiers. PCA has been used in several medical studies as in the case of evaluating the multivariate association between functional microvascular variables and clinical-laboratorial-anthropometrical measurements [16].

In Bayesian inference there have been also developed other techniques for variable selection such as the spike and slab priors, the horseshoe prior and the Bayesian Lasso [17,18,19]. These techniques make some assumptions about the prior distribution of the predictor variables. It is known, however, that the wrong choice of a prior distribution could lead to inadmissible and false posterior distributions in many cases. The dimensionality reduction by the PCA approach facilitates the making of assumptions about the prior distribution with reduced variables (scores) coefficients and therefore it is preferred over the other methods.

A Bayesian approach such as Bayesian logistic regression with the use of MCMC, leads to interesting results based on statistically significant variables resulting from the posterior distributions of the model coefficients.

To our knowledge this study is the first to propose a model for predicting asthma persistence based on a Bayesian approach. Previous works [11,12,13] also predict asthma persistence accurately but the advantages of a Bayesian analysis may lead to even better results in the future. For example, an advantage of a Bayesian approach is the combination of prior information with data in such a way that we can include past information about a parameter and form a prior distribution for a future analysis with more observations [20]. Thus, when new observations become available, the previous posterior distributions can be used as priors. In this study a non - informative uniform prior and a weakly informative Cauchy prior have been selected due to lack of past information for the distribution of the coefficient values. Furthermore, Bayesian statistics have a straightforward way of dealing with nuisance parameters in such a way that if the posterior distribution of a parameter contains zero in the investigated credible interval then it can be extracted from the final model. Also, Bayesian methods have a single tool which is the Bayes theorem making them simpler, at least theoretically. On the opposite frequentist procedures may require a number of different tools [21]. The aim of the current study is to evaluate the accuracy of Bayesian analysis in asthma persistence prediction in children.

METHODS

Clinical Data

Data from 147 patients were gathered by the Paediatric Department of the University Hospital of Alexandroupolis, Greece during the period from 2008 to 2010. The history of each case was obtained by questionnaire. Those patients were diagnosed for asthma and were studied prospectively for seven years. 72 among them had persistent asthma. This dataset has 18 prognostic factors that have been derived by previous studies [22, 23, 24, 25, 26]. The prognostic factors and the patients' characteristics are described in Table 1. The 18 variables inevitably will become 23, as the factor "seasonal symptoms" is a dummy variable with 5 categories. The first category of seasonal symptoms can be removed because its information is included in the other 5 categories (if a patient has a value of zero in all five remaining seasonal symptoms then he has obviously none seasonal symptoms).

Principal Component Analysis

Principal Component Analysis (PCA) is a technique used when strong correlations exist among the explanatory variables. The result of using this method is a set of

uncorrelated variables called scores, which have the same information as the original data.

We select those variables which contribute for 95% of the total variability [27, 28].

Bayesian Logistic regression model

At first the creation of a probability model for the available data is required. In this case an appropriate probability model is a Bernoulli distribution model. Then it is necessary to select a prior distribution. Subsequently, the likelihood function is multiplied with the prior distribution to designate the posterior distribution. Finally as the posterior distribution is estimated (simulated) by MCMC, the calculation of the parameter estimates is available [20].

Prior distribution

We have no prior knowledge available for the parameters of the score – vectors. As a result the choice of the prior distribution becomes a challenge. In this case we can use a non – informative prior on the parameters of the score – vectors. Results of the Bayesian non – informative logistic regression approach tend to mimic a Maximum Likelihood approach, but we must observe that this non – informative approach on parameters of the scores is not non – informative on the parameters of the original variables. Also another choice can be a weakly informative Cauchy prior distribution with location parameter 0 and scale parameter 2.5 after the data matrix standardization and is proposed in [29]. This prior has many advantages that include the problem of dealing with complete separation

TABLE 1. Prognostic factors and Patient Demographic and Disease characteristics

VARIABLES - FACTORS	CHARACTERISTICS	
Demographic	**Median**	**IQR (Interquartile range)**
Age,	10	4
Height(m)	1.42	0.22
Weight(kg)	38	22
Waist's perimeter(cm)	69	14
Wheezing episodes		
Until 3rd year	6	9
Between 3rd – 5th year	6	10
Symptoms	**Absolute Frequency**	**Percentage**
Wheezing	63	42.86%
Cough	81	55.1%
Allerghic rhinitis	40	27.21%
Allerghic Conjuctivitis	28	19.05%
Dyspnea	49	33.3%
Nasal Congestion	54	36.73%
Runny nose	44	29.93%
Seasonal Symptoms		
None	70	47.62%
Winter	35	23.81%
Autumn	2	1.36%
Spring	8	5.44%
Summer	4	2.72%
>2 Seasons	28	19.05%
Pharmaceutical therapy		
Antileukotriene	31	21.09%
Antihistamine	19	12.93%
Corticosteroids Inhaled	64	43.54%
Asthma		
Treatment	77	52.38%
Diagnosis of Asthma	Patients with persistent asthma: 72(48.98%) Patients with non-persistent asthma: 75(51.02%)	

and the application of more shrinkage to higher – order interactions [29]. Applying two different prior distributions will also give clues about the robustness of the method when the prior distribution is changed.

Markov Chain Monte Carlo method and Metropolis – Hastings algorithm

Often, the integrals that have to be solved to calculate the posterior distribution pose major difficulties. As the complexity of a problem or the number of the parameters increase, it becomes more difficult to deal with them using direct techniques.

Therefore, Markov Chain Monte Carlo (MCMC) techniques are proposed [30, 31]. MCMC techniques simulate values of random variables from the posterior distribution. A Markov Chain is constructed. The property of Markov Chain processes allow the next value of each parameter vector to depend on the current value but not on the previous one. The advantage of these methods depends on the fact that the simulation algorithm is repeated multiple times and as a result the approximation of the posterior distribution is improved at every step. Thus the posterior distributions can be approximated with high accuracy by the histogram or kernel – density estimates of the simulated values.

One of the most popular MCMC algorithms is the random – walk Metropolis – Hastings Algorithm with Gaussian proposals [30, 31, 32, 33]. The whole procedure was conducted in RGui 3.3.3 with the use of the "MCMCpack" package [34].

RESULTS

As we explained previously, in order to deal with multicollinearity among the explanatory variables PCA is used before the implementation of the Bayesian logistic regression model. In the first case of centering the data matrix 3 principal components describe more than 95% of the variation and in the case of standardizing 16 principal components describe the same amount of variation. After each MCMC simulation the simulated coefficients $\boldsymbol{\theta}$ are transformed back to the original coefficients $\boldsymbol{\beta}$ with the following equations.

$$\boldsymbol{S}\,\theta = \boldsymbol{X_0}\beta$$

or

$$\boldsymbol{X_0 V}\theta = \boldsymbol{X_0}\beta$$

or

$$\beta = \boldsymbol{V}\theta \qquad (4)$$

where $\boldsymbol{X_0}$ is the data matrix, $\boldsymbol{V}$ is the loadings matrix and $\boldsymbol{S}$ is the scores matrix obtained by the PCA.

It is important to mention that the data table has to be preprocessed before the analysis. When the variables are measured in different units, (as in the case of the asthma dataset) it is usual to standardize each variable. This is obtained by subtracting the mean from each variable and then dividing each variable by its standard deviation [35].

Another important issue in Bayesian Logistic regression is which parameters are going to be kept in the model. In this study our approach leads to a model in which none of the estimated coefficients contains zero in the 95% credible interval (Figures A1-A4 are given in Supplementary files) of the coefficient posterior distribution, i.e. from the full model, we eliminate those variables whose 95% credible interval contains zero. Using the Metropolis – Hastings algorithm mentioned in the previous section with 100000 MCMC samples with a burn - in period of 25000 samples and centering or standardizing the data matrix before the use of the Principal Component analysis, the results are shown in the models of Tables 2 (Models A and Model B) and 3 (Models C and D). Also it must be mentioned that a thinning interval equal to 10 was used to remove dependencies between successive simulations. It can be observed (from Tables 2-3) that the four variables are included in all models, but the models of Table 3 have a reduction in the standard errors of the coefficients (posterior means) as well as in the odds ratio. This change is due to the covariance matrix being different in case of centering or standardizing before PCA.

The results of Table 2 and Table 3 are very similar leading us to the conclusion that the choice of the prior does not change them significantly and therefore the approach is robust.

In addition Figures A1-A4(see Supplementary files) show the posterior distribution of each coefficient (trace plot and density plot) that remains in the final form of the model in each case of the first approach described above. It seems that all the posterior distributions are approximate normal in both models.

After fitting the models it is necessary to examine the performance of each model in order to decide which one is the most appropriate for asthma persistence prediction and moreover to evaluate the importance of the factors that affect asthma. A very common way to examine the ability of a model to predict accurately is the 10–fold cross–validation [36]. In 10-fold cross-validation, the original data matrix $\boldsymbol{X}$ is randomly partitioned into 10 equal size subsamples. Afterwards 9 subsamples are used as the training dataset and the one subsample remaining is retained as the test dataset. The cross-validation process is then repeated 10 times, in such a way that all 10 subsamples are used both as training and test dataset. This procedure must be repeated several times. In this study it was repeated 100 times for each model as it is important for a prognostic model to work sufficiently for patients other than those used for the fitting of the model [37].

TABLE 2. Posterior means and their standard errors for model coefficients β with centered data matrix X

Coefficients of Model A (uniform prior)	Posterior mean	Posterior standard error
Treatment	1.496234	0.7122792
Corticosteroids Inhaled	1.434193	0.7192093
Cough	2.218498	0.601882
Dyspnea	1.761561	0.663517
Coefficients of Model B (Cauchy Prior)	**Posterior mean**	**Posterior standard error**
Treatment	1.476835	0.655431
Corticosteroids Inhaled	1.396371	0.6517608
Cough	2.129355	0.5730888
Dyspnea	1.671632	0.6323252

TABLE 3. Posterior means and their standard errors for model coefficients β with standardized data matrix X

Coefficients of Model C (uniform prior)	Posterior mean	Posterior standard error
Treatment	0.9550956	0.1895476
Corticosteroids Inhaled	0.9026634	0.2214718
Antihistamine	-0.7328311	0.3060576
Nasal Congestion	0.6755819	0.2944196
Cough	1.0100756	0.3191202
Dyspnea	0.9482515	0.3372774
Coefficients of Model D (Cauchy prior)	**Posterior mean**	**Posterior standard error**
Treatment	0.9269379	0.1824626
Corticosteroids Inhaled	0.8722235	0.215429
Antihistamine	-0.6900618	0.3004151
Nasal Congestion	0.6480334	0.2804873
Cough	1.0001017	0.3123239
Dyspnea	0.9143336	0.3279823

Then based on the following equation:

$$\hat{p} = \frac{\exp\left(\boldsymbol{X}_{test}\hat{\beta}\right)}{1 + \exp\left(\boldsymbol{X}_{test}\hat{\beta}\right)}, \qquad (5)$$

a prediction for the diagnosis of a new patient can be found. Since the dataset is balanced in terms of patients with persistent asthma (positive patients) and non-persistent asthma (negative patients) the threshold value for the predicted probabilities when the patients are classified is set 0.5. The sensitivity, the specificity, the positive predicted value (PPV), the negative predicted value (NPV) and the accuracy of the model are estimated using false positive (FP), false negative (FN), true positive (TP), and true negative (TN) values. The definitions of those accuracy measures are described in Table 4. [38, 39].

Those measures are very useful and provide us with important information about a patient. For example if a PPV of a disease prediction model is 90% then a patient with a positive test has a chance of 90% having the particular disease [38]. In addition, when the models have similar accuracy it is necessary to use extra criteria to decide which model is the most appropriate.

In Bayesian inference the most commonly used criteria are the Deviance Information Criterion (DIC) and the Widely Applicable Information Criterion (WAIC), which are defined in the following equations:

$$DIC = -2\log\left(p(y|\hat{\theta}_{Bayes}\right) + 2p_{DIC}, \qquad (6)$$

and

$$WAIC_1 = -2(LPPD\text{-}p_{WAIC1}), \qquad (7)$$

$$WAIC_2 = -2(LPPD\text{-}p_{WAIC2}), \qquad (8)$$

where p_{DIC} is the effective number of parameters in the model defined as:

TABLE 4. Definition of accuracy measures as they are applied in the asthma prediction models.

Sensitivity	Proportion of children with active asthma who were correctly classified by the model as being at risk.
Specificity	Proportion of children without active asthma who were correctly classified by the model as not being at risk.
Positive predictive value	Proportion of children who were classified by the model as being at risk and developed active asthma
Negative predictive value	Proportion of children who were classified by the model as not being at risk and did not develop active asthma.
Positive Likelihood Ratio (LR+)	The probability of a child with disease having a positive test divided by the probability of an individual without disease having a positive test.
Negative Likelihood ratio(LR-)	The probability of a child with disease having a negative test divided by the probability of an individual without disease having a negative test.

$$p_{DIC} = 2(\log(p(y|\hat{\theta}_{Bayes}) - E_{post}(\log(p(y|\theta)),$$

$\log(p(y|\hat{\theta}_{Bayes})$ is the log - likelihood of the data given the posterior means of the parameters $\hat{\theta}$. In equations 7 and 8, *LPPD* is the sum all over observations y of the logarithm of the expected value of the likelihoods for each sample from the posterior distribution of the parameters and $p_{WAIC1-2}$ are known as penalty terms that are estimates of the effective number of parameters and are defined by [40,41]:

$$p_{WAIC1} = 2\sum_{i=1}^{n}(\log(E_{post}p(y_i|\theta)) - E_{post}(\log p(y_i|\theta))).$$

$$p_{WAIC2} = \sum_{i=1}^{n} var_{post}(\log p(y_i|\theta)).$$

Two other measures that give some information about the model are the Pearson correlation coefficient between the actual values and the estimated probabilities of the model derived from equation (5) [42]and the root mean squared error of the validation sets. Those are defined as:

$$R = \frac{cov(y_{test}, \hat{p})}{\sigma_{ytest}\sigma_{\hat{p}}}, \qquad (9)$$

$$RMSE_{cv} = \sqrt{\frac{1}{N}\sum_{i}(y_{test_i} - \hat{p}_i)^2} \qquad (10)$$

Finally, it is useful to check how the model predicts the negative values against the positive ones by means of:

$$RMSE_0 = \sqrt{\frac{1}{N}(\sum_{i}\{y_{test_i}(y_{test_i} = 0) - \hat{p}_i(y_{test_i} = 0)\}^2}$$

$$RMSE_1 = \sqrt{\frac{1}{N}(\sum_{i}\{y_{test_i}(y_{test_i} = 1) - \hat{p}_i(y_{test_i} = 1)\}^2}$$

If the model predicts positive and negative values the same, $RMSE_0$ and $RMSE_1$ should be almost equal.

Another method of model selection can be based on Bayes factors. In statistics, Bayes factors are a Bayesian alternative to hypothesis testing. For comparing two models 1 and 2 it is calculated as [43]

$$B_{12} = \exp[-0.5(BIC_1 - BIC_2)], \quad for\ large\ N$$

and

$$BIC = -2\log(L) + k\log(N),$$

where k is the number of the estimated parameters in the model, L is the maximum value of the likelihood function and N is the number of the observations in the dataset.

In Table 5 are described all the statistical measures for the performance of the models including the root mean squared error. The results show that the most appropriate Bayesian model according to DIC and WAIC 1 and 2 is the model D of Table 3. Also in terms of $RMSE_0$ and $RMSE_1$, model D is again the most appropriate because the difference between those two values is the least. It is interesting to note that model B of Table 2 is the most appropriate according to the Bayes factors. This is of high significance because it shows the importance of the 4 variables that determine the model. The Bayes factor between models B and A is equal to 1.047, between models B and C 1.138 and between models B and D is equal to 1.077 which suggests that model B is slightly favored by the data.

Cough and dyspnea were also identified as significant in the ridge logistic regression model for asthma [13] revealing the importance of those symptoms combined in asthma persistence. In addition, residual diagnostics can be provided in order to check the validity of the models. For that reason the randomized quantile residuals (RQR) can be used because of their appropriateness in logistic regression problems [44]. The residuals for all models approximate the standard normal and this is confirmed by the powerful Anderson – Darling test (A-D test) [45].

TABLE 5. Accuracy measures for the models of Tables 2 - 3.

Repeats of 10-fold cross validation	100	100	100	100
Accuracy measures	**Model A**	**Model B**	**Model C**	**Model D**
Accuracy	84.07609	84.92339	86.36735	86.19505
PPV	82.94709	84.00492	85.26059	85.53547
NPV	85.21317	85.8365	87.47954	86.83853
Sensitivity	84.96158	85.49979	87.25	86.3764
Specificity	83.22572	84.36999	85.52	86.02151
LR+	5.065	5.470233	6.0255	6.179235
LR-	0.1807	0.1718645	0.1491	0.1583743
MSE over the test set	0.1242153	0.1155709	0.1071281	0.1048957
RMSE over the test set	0.3524419	0.3399573	0.3273043	0.3238761
DIC	99.18288	98.45763	91.15783	90.89677
WAIC 1	99.24836	98.4076	91.39442	90.98076
WAIC 2	99.89265	98.93421	92.47293	91.94984
BIC	111.0227	110.9308	111.1893	111.0788
R	0.7141177	0.7359201	0.7589249	0.7633664
RMSE0	0.3701722	0.3596769	0.3391169	0.332138
RMSE1	0.3329774	0.3181209	0.3145281	0.3149454

DISCUSSION

The main finding of our study is that Bayesian analysis exhibited high accuracy (approximately 86%) in asthma persistence prediction in children. Bayesian techniques are becoming very popular in the field of medicine [46]. Our study proposes a method based on Bayesian inference for asthma persistence prediction in real time data. The model, which according to DIC and WAIC is the most appropriate Bayesian model, predicts asthma with an accuracy of 86.19505% a PPV of 85.53547% and a NPV of 86.83853%.

Additionally, the novelty of this method is that the model coefficients contain information from all the available variables used, despite that we require only 4 to make a prediction. Also important results are the stability of the model between positive and negative prediction and the good values of LR+ and LR-.

It should be mentioned that it is quite difficult to compare the results of the prediction of the statistical models with that of the simple models used in clinical practice. However the inability of the existing simple clinical prediction models using four to eight disease parameters to achieve substantially high prediction accuracy [3] has led to the increasing scientific interest in alternative methods of analysis of asthma persistence.

Currently available asthma predictive models are based on non-invasive clinical and laboratory parameters [3]. The most popular of the predictive indices is the Asthma Predictive Index (API) introduced in 2000 by Castro-Rodriguez et. al. [4]. The use of the loose API seems that is not successful in either identifying or ruling out later asthma in children [3]. Additionally, recent validation of the stringiest API revealed that this index presented a LR- of 0.8-0.9 and a LR+ of 4.9-7.4, indicating that this index is also not appropriate to rule out later asthma in preschool wheezers [3]. Similarly the Isle of Wight score presented an acceptable LR+ but unacceptably high LR- and a false negative rate of 90.The performance of ECA severity score was similar while contrary to the previous models, PIAMA score performance seems to be grossly affected by the cut-offs used for the prediction [3]. The use of the lower cutoff of PIAMA score resulted in high false positive predictive rate but good ability to rule out later asthma while higher cut-offs of the same score resulted in good LR+ of 6.3 and a poor LR- of 0.8 indicating inability to exclude later disease [3]. As a result it is obvious that the presented models are much better in terms of the LRs both positive and negative.

Moreover studies that have validated the above mentioned predictive models in different cohorts (broad validation) concluded that the performance of clinical scores such as API [9] and PIAMA [10] in clinical settings were modest.

In recent studies for asthma persistence prediction alternative methods of analysis and prediction have been developed. Support Vector Machnes (SVM), RLR and Artificial Neural Networks (ANNs) are some methods based in statistical and artificial intelligence techniques used for the classification of asthma patients using multiple prognostic factors such as demographic, wheezing episodes, symptoms, pharmaceutical therapy, breathing

TABLE 6. Comparison of Bayesian Models with Machine Learning Techniques using accuracy measures.

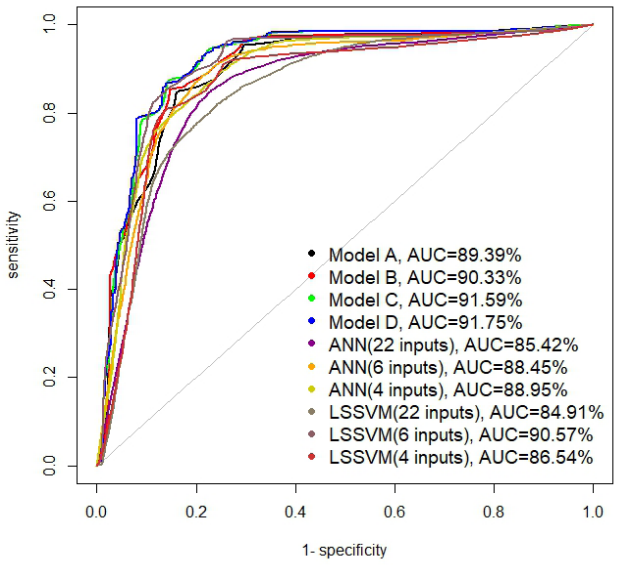
Model	Accuracy	Sensitivity	Specificity
Model A	84.07609	84.96158	83.22572
Model B	84.92339	85.49979	84.36999
Model C	86.36735	87.25	85.52
Model D	86.19505	86.3764	86.02151
ANN(22 inputs)	80.58	79.37	81.80
ANN(6 inputs)	83.46	82.37	84.58
ANN(4 inputs)	82.04	79.81	84.4
LSSVM (22 inputs)	79.70714	79.82011	79.59758
LSSVM (6` inputs)	84.8	87.54381	82.17282
LSSVM (4 inputs)	84.16405	86.59287	81.93451

tests and parental history but with a smaller number of observations. The similarity with our results follows from the fact that a subset of variables is included in the final models as well, which makes them to be sufficient [47, 48, 49].

One of the disadvantages of ANNs when compared to Bayesian models is related to the fact that ANNs frequently have difficulty analyzing systems with a large number of inputs due to the excess time required to train the system and possibly the over-fitting of the model during the training time [50].

SVM is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy [47]. An important advantage is that it is automatically implemented avoiding the problem of over - fitting [51]. Despite the fact that it may achieve high accuracy in asthma prediction [11], SVM has some major disadvantages. The most important are the selection of the parameters of the kernel function, the high algorithmic complexity and the extensive memory requirements [52, 53]. In addition a comparison with those machine learning techniques using the same dataset and with 100 repeats of 10 fold cross – validation is provided. The results of this comparison are presented in Table 6. The ANN that was used for this simulation has 1 hidden layer with 6 neurons with tan – sigmoid and saturated linear activation functions. The LSSVM method uses Radial Basis Function Kernels. A range of different kernel widths for the Radial Basis kernel function was used to obtain the best LSSVM classifier in each case. The results show that the Bayesian models are slightly superior in terms of accuracy measures. In addition the increase in the accuracy measures when only 6 or 4 variables are included in ANNs and SVMs shows that the selection of those variables from the Bayesian models is in good agreement. Finally in comparison with Ridge Logistic regression, the Bayesian Logistic model exhibits more stable results between negative (non – persistent asthma) and positive patients (persistent asthma) and fewer factors are included in the final form of the model. The discriminative power of the proposed models can also be seen in the Receiver Operating Characteristic (ROC)

FIGURE 1. ROC curves of the models of TABLE 6.



curves presented in Figure 1. This shows that Models C and D (green and blue line) have the highest area under curve (AUC equal to 91.59% and 91.75% respectively) and perform better at almost all regions.

In this study, the Bayesian PCA logistic regression model is applied. As mentioned above, the advantage of this model is its simplicity as it depends only on the Bayes formula. Furthermore, the fact that we do not have past information about the data in this study lead us to use a non – informative and a weakly informative prior distribution. Although this choice seems now to be a disadvantage, in the future this approach will become more interesting as it is possible, with the availability of new data, to use as prior the posterior distributions of this work leading to even better results.

Another advantage of great importance is that the Bayesian PCA logistic model exhibits high accuracy including only a small subset of explanatory factors. All the other models that have been used either are computationally

heavy with large memory requirements (SVMs, ANNs) or they include a large amount of information that may not be needed (RLR) [11, 12, 13].

Moreover, important advantages are that our model is more stable and balanced in terms of positive and negative patients which is indicated by the classification measures, is very simple to use (only 4 or 6 variables required) and has a high accuracy and sensitivity.

CONCLUSION

To conclude despite the fact that no prior knowledge about the distribution of the variables was available, this Bayesian model predicts asthma with high accuracy and gives many clues about the significance and importance of each factor in asthma persistence prediction. Given the known disappointing results of the existing clinical indices in asthma persistence prognosis, alternative models that utilize more effectively many disease parameters should be tested in clinical practice, since as we have shown, their accuracy seem to be extremely high despite the fact that these statistical models might present difficulties at the interpretation by the clinical doctor.

As far as for future research, an interesting prospect will be to check if there is an increase of the accuracy when we include more patients in the dataset, using the posteriors from this study as priors for the new dataset. Finally, it would also be important to check if asthma, from a statistical point of view, is affected by regional parameters, such as climate and environmental factors.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Stein RT, Martinez FD. Asthma phenotypes in childhood: lessons from an epidemiological approach. Paediatric Respiratory Reviews. 2004; 5(2):155–61.
2. Brand PL, Baraldi E, Bisgaard H, et al. Definition, assessment and treatment ofwheezing disorders in preschool children: an evidence-based approach. Eur Respir J. 2008; 32(4):1096–1110.
3. Fouzas S and Brand PLP. Predicting persistence of asthma in preschool wheezers: crystal balls or muddy waters? Paediatric Respiratory Reviews, 2013; 14(1):46-52.
4. 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 Resp Crit Care Med, 2000, 162(4):1403-6.
5. Kurukulaaratchy RJ, Matthews S, Holgate ST, et al. Predicting persistent disease among children who wheeze during early life. Eur Respir J. 2003; 22(5):767-71.
6. Devulapalli CS, Carlsen KC, Haland G, et al. Severity of obstructive airway disease by age 2 years predicts asthma at 10 years of age. Thorax. 2008; 63(1):8-13.
7. 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(5):903-10.
8. Brunekreef B, Smit J, De Jongste J, et al. The prevention and incidence of asthma and mite allergy (PIAMA) birth cohort study: design and first results. Pediatr Allergy Immunol. 2002; 13(s15):55-60.
9. Leonardi NA, Spycher BD, Strippoli MP, Frey U, Silverman M, Kuehni CE. Validation of the asthma predictive index and comparison with simpler clinical prediction rules. J Allergy Clin Immunol. 2011; 127(6):1466–72.
10. Rodriguez-Martinez CE, Sossa-Briceño MP, Castro-Rodriguez JA. Discriminative properties of two predictive indices for asthma diagnosis in a sample of preschoolers with recurrent wheezing. Pediatr Pulmonol. 2011; 46(12):1175–81.
11. Chatzimichail E, Paraskakis E, Sitzimi M, and Rigas A. An Intelligent System Approach for Asthma Prediction in Symptomatic Preschool Children. Computational and Mathematical Methods in Medicine. 2013; vol. 2013: 6 pages, doi:10.1155/2013/240182.
12. Chatzimichail E, Paraskakis E, and Rigas A. Predicting Asthma Outcome Using Partial Least Square Regression and Artificial Neural Networks. Advances in Artificial Intelligence. 2013; vol. 2013: 7 pages, doi:10.1155/2013/435321.
13. Spyroglou II, Chatzimichail E, Spanou EN, Paraskakis E, and Rigas A. Evaluation of a Prediction Model Based on Ridge Regression for Asthma Persistence in Preschool Children. International Journal of Mathematical Models and Methods in Applied Sciences. 2015; 9: p. 581-91.
14. Frisch R, Statistical Confluence Analysis by Means of Complete Regression Systems. Oslo: University Institute of Economics, 1934, Publication no. 5.
15. Bro R, Smilde AK. Principal component analysis. Analytical Methods. 2014; 6(9): 2812-31.
16. Panazzolo DG, Sicuro FL, Clapauch R, Maranha ˜o PA, Bouskela E and Kraemer-Aguiar LG. Obesity, metabolic syndrome, impaired fasting glucose, and microvascular dysfunction: a principal component analysis approach. BMC Cardiovascular Disorders. 2012; 12(1):102.
17. Mitchell TJ and Beauchamp JJ. Bayesian variable selection in linear regression. Journal of the American Statistical Association. 1988; 83(404):1023-32.
18. Carvalho CM and Polson NG. The horseshoe estimator for sparse signals. Biometrika. 2010; 97(2):465-80.
19. Hans C. Bayesian lasso regression. Biometrika. 2009; 96(4):835-45.
20. Glickman ME and van Dyk DA. Basic Bayesian Methods. Methods in Molecular Biology. 2007; 404:319-38.
21. Bolstad WM. Introduction to Bayesian Statistics. 2nd ed. Wiley, 2007.
22. Zahran HS, Person CJ, Bailey C and Moorman JE. Predictors of asthma self-management education among children and adults—2006-2007 behavioral risk factor surveillance system asthma call-back survey. The Journal of Asthma. 2012; 49(1):98–106.
23. Hansel NN, Matsui EC, Rusher R, et al. Predicting future asthma morbidity in preschool inner-city children. The Journal of Asthma. 2011; 48(8):797–803.
24. Nagel G, Weinmayr G, Kleiner A, et al. Effect of diet on asthma

and allergic sensitisation in the International Study on Allergies and Asthma in Childhood (ISAAC) Phase Two. Thorax. 2010; 65(6):516–22.
25. Lange NE, Rifas-Shiman SL, Camargo CA, Gold DR, Gillman MW and Litonjua AA. Maternal dietary pattern during pregnancy is not associated with recurrent wheeze in children. J Allergy Clin Immunol. 2010; 126(2):250–55.
26. Bracken MB, Belanger K, Cookson WO, Triche E, Christiani DC, and Leaderer BP. Genetic and perinatal risk factors for asthma onset and severity: a review and theoretical analysis. Epidemiologic Reviews. 2002; 24(2):176–89.
27. Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology. 1933; 24(6):417–41.
28. Jolliffe T. Principal Component Analysis. 2nd ed. New York: Springer Series in Statistics, 2002.
29. Gelman A, Jakulin A, Pittau MG and Su YS. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics. 2008; 2(4):1360-83.
30. Gamerman D and Lopes HF. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. 2nd ed. Chapman and Hall/CRC, 2006.
31. Gilks WR, Richardson S and Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. 1st ed. Chapman & Hall, 1996.
32. Metropolis N, Rosenbluth, AW, Rosenbluth MN, Teller AH and Teller E. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics. 1953; 21(6):1087-92.
33. Roberts GO, Gelman A and Gilks WR. Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann Appl Probab. 1997; 7(1):110–20.
34. Martin AD, Quinn KM and Park JH. MCMCpack: Markov chain Monte Carlo in R. Journal of Statistical Software. 2011; 42(9):1-21.
35. Abdi H and Williams LJ. Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010; 2(4):433-59.
36. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (San Mateo, CA: Morgan Kaufmann). 1995; 2(12):1137-43.
37. Altman DG and Royston P. What do we mean by validating a prognostic model?. Statistics in medicine. 2000; 19(4):453-73.
38. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatrica. 2007; 96(3):338-41.
39. Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre˜and posf˜test probabilities and their use in clinical practice. Acta Paediatrica. 2007; 96(4):487-91.
40. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 2010; 11(Dec):3571–94.
41. Gelman A, Hwang J and Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing. 2014; 24(6):997-1016.
42. McQuarrie ADR and Tsai CL. Regression and Time Series Model Selection. World Scientific, 1998
43. Berger J and Pericchi L. Bayes factors. Wiley StatsRef: Statistics reference Online. 2015; 1-14.
44. Dunn PK and Smyth GK. Randomized quantile residuals. Journal of Computational and Graphical Statistics. 1996; 5(3):236-44.
45. Razali NM, Wah YB, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics. 2011; 2(1):21-33.
46. Ashby D. Bayesian statistics in medicine: a 25 year review. Statistics in medicine. 2006; 25(21): 3589-631.
47. Vapnik V. The support vector method. Proceedings of the 7th International Conference on Artificial Neural Networks (ICANN '97); 1997 Oct 8-10; Lausanne, 1997:263–71.
48. Le Cessie S and Van Houwlingen JC. Ridge Estimators in Logistic Regression. Journal of the Royal Statistical Society. 1992; 41(1):191–201.
49. Rumelhart DE, Hinton GE and Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323(6088):533-36.
50. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol. 1996; 49(11):1225-31.
51. Widodo A, Yang BS. Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. Expert Systems with Applications. 2007; 33(1):241–50.
52. Burges CJ. A Tutorial on Support Vector Machines for Pattern. Data Mining and Knowledge Discovery. 1998; 2(2):121-67.
53. Suykens JAK, Horvath G, Basu S, Micchelli C and Vandewalle J. Advances in Learning Theory: Methods, Models and Applications, NATO-ASI Series Computer and Systems Sciences, IOS Press, 2003. IOS Press, 2003.

*

SUPPLEMENTARY FILES

1. TRACE AND DENSITY PLOTS

FIGURE A1. Trace and density plots of the regression coefficients β for model A.

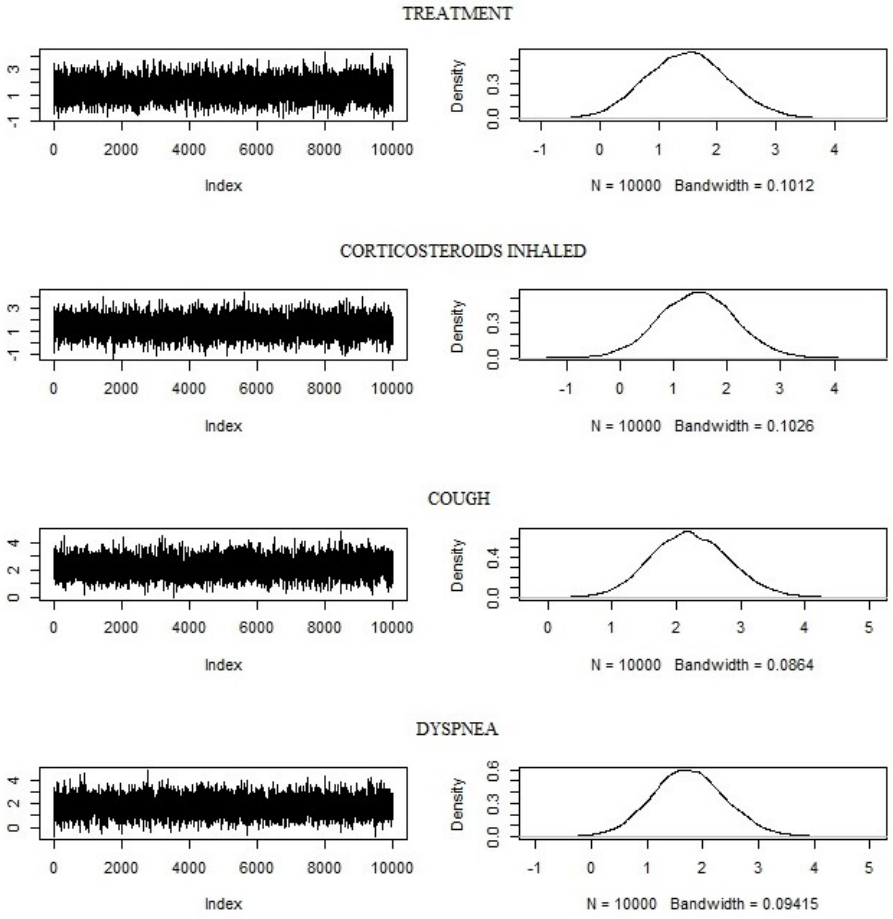


FIGURE A2. Trace and density plots of the regression coefficients β for model B.

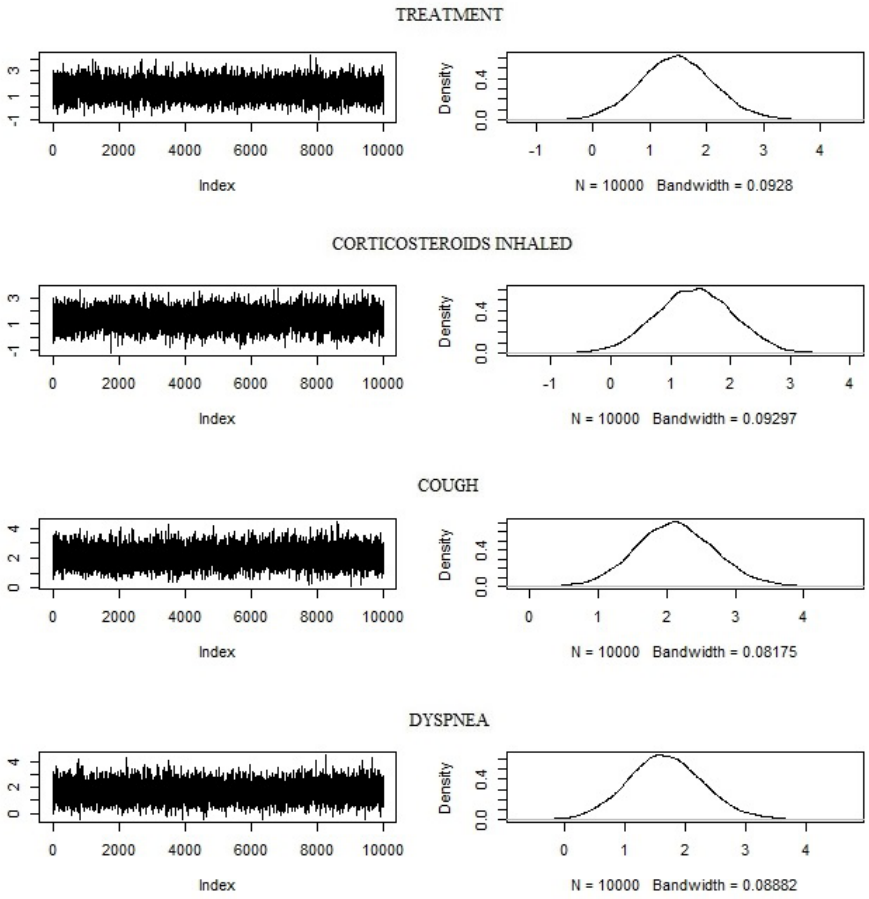


FIGURE A3. Trace and density plots of the regression coefficients β for model C

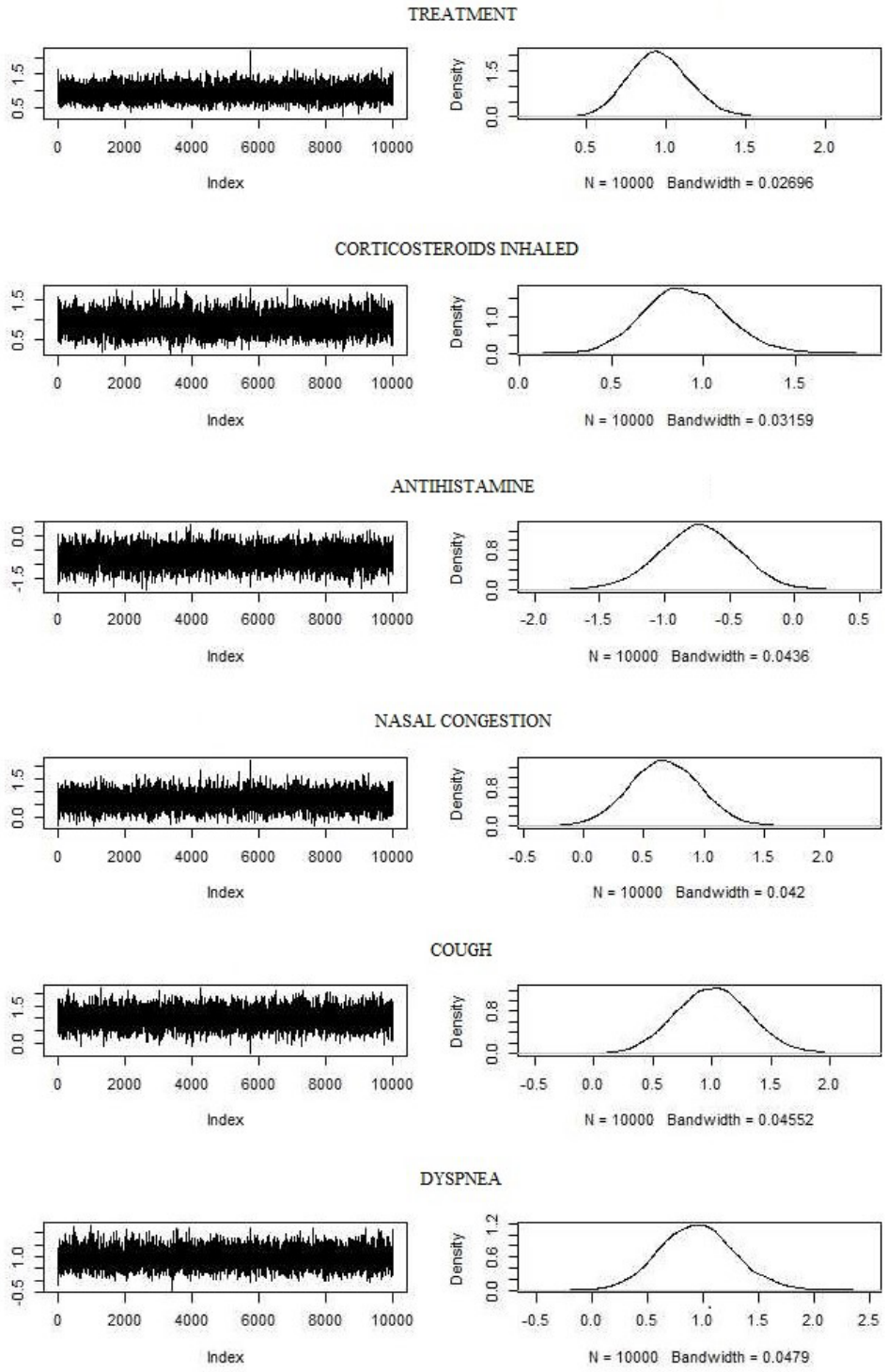


FIGURE A4. Trace and density plots of the regression coefficients β for model D.