



## ORIGINAL RESEARCH

## A Comparative Study of Three Classification Procedures: Asthma among Healthcare Professionals in Texas

Dejian Lai<sup>1,4\*</sup>, Ahmed A Arif<sup>2</sup>, Haiyun Xu<sup>3</sup> and George L Delclos<sup>4</sup>

<sup>1</sup>Department of Biostatistics and Data Science, The University of Texas School of Public Health, Houston, Texas, USA

<sup>2</sup>Department of Public Health Sciences, University of North Carolina at Charlotte, Charlotte, North Carolina, USA

<sup>3</sup>Faculty of Statistics, Jiangxi University of Finance and Economics, Nanchang, China

<sup>4</sup>Southwest Center for Occupational and Environmental Health, The University of Texas School of Public Health, Houston, Texas, USA

\*Corresponding author: Dejian Lai, Department of Biostatistics and Data Science, RAS 1008, The University of Texas School of Public Health, Houston, TX 77030, USA



### Abstract

Statistical classification analysis has been widely used in many fields. In this article, we applied and compared three different classification procedures: Logistic regression, Fisher's linear discriminant function and the second order Bahadur representation to two datasets from two surveys on asthma among healthcare professional in Texas. The first dataset contained 102 subjects and the second dataset 2963. The concordance of the classification from the three statistical procedures with possible asthma identified by physician and airway responsiveness to methacholine challenge was assessed through Cohen's  $\kappa$  statistic via a series of  $2 \times 2$  contingency tables.

### Keywords

Bahadur representation, Classification, Discrete discriminant analysis, Fisher's linear discriminant function, Logistic regression

### Introduction

The incidence and prevalence of asthma, a chronic inflammatory disease of airways, is on the rise in the US and has increased by 75% in the past two decades [1]. Estimates of the prevalence of asthma differ based on the definition used and range from 4.5% to as high as 16.4% [2]. It is estimated that more than 14 million persons in the United States suffer from asthma. Community-based studies have reported asthma incidence rates from 0.5 to 2.5 per 1000 [3,4]. Questionnaires have long

been a cornerstone of asthma epidemiology studies, and much work has gone into standardizing asthma questionnaires for use in the general population, by groups such as the British Medical Research Council (MRC) (1960), [5] American Thoracic Society (ATS) [6], and the International Union Against Tuberculosis and Lung Disease (IUATLD) [7]. However, in the absence of a gold standard, the definitions of asthma used in surveys vary and may not necessarily correspond to its clinical definition. Relatively few studies have been published with information on formal validation of asthma questionnaires [7-10]. Accurate detection of asthma in epidemiological studies is critical for the proper characterization of etiologic risk factors, triggers and the identification of prevention and intervention opportunities. There are many different ways that have been proposed and revised for the diagnosis of asthma. The current operational definition of asthma was given in the International Consensus Report on the Diagnosis and Treatment of Asthma, which is based on three components: Chronic airway inflammation, reversible airflow obstruction and enhanced bronchial reactivity that lead to symptoms of wheezing, breathlessness, chest tightness, cough, and sputum production [11].

The Southwest Center for Occupational and Environmental Health at The University of Texas School of Public Health conducted a two-phase survey of asthma among



**Citation:** Lai D, Arif AA, Xu H, Delclos GL (2019) A Comparative Study of Three Classification Procedures: Asthma among Healthcare Professionals in Texas. Int J Clin Biostat Biom 5:021. doi.org/10.23937/2469-5831/1510021

**Accepted:** April 27, 2019; **Published:** April 29, 2019

**Copyright:** © 2019 Lai D, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

healthcare professionals in Texas [12]. In the first phase, an initial questionnaire was given to a convenience sample of 102 subjects. A methacholine challenge was administered to the 102 subjects in addition to self-administered questions regarding asthmatic symptoms, environmental risk factors and basic demographic characteristics [13]. In Delclos, et al. logistic regression models were based on 118 subjects (16 subjects in the testing stage were included) [13]. However, in current article, the 16 subjects in the testing stage were excluded. For the second phase, the refined questionnaire was administered to a random sample of healthcare professionals in Texas. The second phase of the study consisted of a cross-sectional group-comparison study design, using a mail survey administered to a sample ( $n = 5600$ ) of four groups ( $n = 1400$  per group) of Texas healthcare workers: Physicians, nurses, respiratory therapists and occupational therapists. Questionnaires were received from 3528 participants, with an overall response rate of 63%. After removing subjects with missing values, we used 2963 subjects with complete responses for model-based classification analysis. In the second phase, no Methacholine challenge was given.

For an accurate estimation of prevalence, a proper diagnosis of asthma is necessary. Because of the multivariate nature of the risk factors and unknown etiology of asthma, there is always uncertainty for the diagnosis [14]. A reasonable diagnosis of asthma for a person by a medical doctor generally requires some period of follow up and sufficient clinical and physiologic information documented during this follow up. One of the purposes of developing the questionnaire from the surveys was to provide a useful instrument in assessing asthma burden to the healthcare professionals in Texas [13]. In the questionnaire, in addition to a sequence of questions on symptoms, environmental risk factors and demographic characteristics, subjects were also asked if they had ever been diagnosed as having asthma by a physician (MD asthma) [13]. Preliminary analysis based on logistic regression identified a subset of eight symptom items that exhibited the best combination of sensitivity and specificity when compared to MD asthma and  $PC_{20}4$  and  $PC_{20}8$ , where  $PC_{20}4 = 1$  denotes a  $\geq 20\%$  decline in the subject's  $FEV_1$  (forced exposure volume at one second) at  $\leq 4$  mg/ml methacholine challenge,  $PC_{20}8 = 1$  indicates an  $FEV_1$  fall of least 20% at  $\leq 8$  mg/ml for the challenge. The eight symptom items were: 1) *Have you ever had trouble with your breathing?* 2) *Have you had an attack of shortness of breath at any time in the last 12 months?* 3) *Have you had wheezing or whistling in your chest at any time in the last 12 months?* 4) *Have you been awakened during the night by an attack of cough in the last 12 months?* 5) *Have you been awakened during the night by an attack of chest tightness in the last 12 months?* 6) *When you are near animals, feathers or in a dusty part of the house, do you ever get itchy or watery eyes?* 7)

*When you are near animals, feathers or in a dusty part of the house, do you ever get a feeling of tightness in your chest?* 8) *When you are near tree, grass or flowers, or when there is a lot of pollen around, do you ever get itchy or watery eyes?*

There are many widely used classification procedures in the statistical literature [15]. Clinicians and statisticians have been engaged in discussions of the consistency and discrepancies of using various methods. In this article, we applied and compared three classification analysis methods for the diagnosis of asthma. Two methods (logistic regression model and Fisher's linear discriminant function) are commonly used and the third one (Bahadur model) is less commonly used. Our comparative study intends to highlight the utilities of these models through real datasets. These three methods were applied to two data sets. The first one was a small data set from our phase I survey with 102 subjects. The second was a large data set from our phase II survey with 2963 complete subjects. In Section 2, the three classification techniques studied in this article are briefly reviewed. The results from the three methods on the two data sets are tabulated in a series of  $2 \times 2$  contingency tables and the agreement among these three methods is quantified via  $\kappa$  statistic and presented in Section 3. Discussions and concluding remarks are given in Section 4.

## Classification Methods

The three classification analysis tools applied to the two data sets were logistic regression [16], Fisher's linear discriminant function [17] and the second order Bahadur model [18]. Logistic regression has been widely used to model binary dependent variable in response to risk factors in many fields [19]. In this study, we denoted the dependent variable being 1 as asthma positive and 0 as negative. As noted in the introduction, there is no gold standard for the detection of asthma. For phase I data, we modeled three dependent variables: Asthma diagnosed by a physician (MD asthma = 1), and two levels of response to methacholine challenge ( $PC_{20}4 = 1$  or  $PC_{20}8 = 1$ ). In the phase II survey, methacholine challenge testing was not performed.

### Logistic regression

Logistic regression models the probability of asthma in relation to symptoms (risk factors). In our setting, let  $y = 1$  be MD asthma = 1 or  $PC_{20}4 = 1$  or  $PC_{20}8 = 1$ . The eight symptoms variables were described in previous section. Mathematically, logistic regression establishes a generalized linear model:

$$p(y = 1 | x_1, x_2, \dots, x_k) = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}} \quad (1)$$

Where,  $x_j, j = 1, 2, \dots, p$ , denotes the  $p$  dichotomous symptom variables used in our study. In our case,  $p = 8$ . We used R 3.5.0 [20] to estimate the parameters in the model. In general, if we observed  $P(y = 1) > 0.5$ , we

would classify the subject with the given combinations of symptoms as being asthmatic. However, more careful assessment of the threshold value for classification may be needed in some cases as discussed in Section 4.

### Fisher's linear discriminant function

Fisher's linear discriminant function is another widely used technique in classification analysis. The simplest Fisher's linear discriminant function applied to the classification of two populations is based on two multivariate normal distributions with equal covariance [17]. In our two survey phases, the symptom variables were generally binary, with "yes" or "no" answers. It is then obvious that the application of Fisher's linear discriminant function to our data is questionable. Nevertheless, we included this method for comparison to the other two methods in our study. In applying Fisher's linear discriminant function, we may assume there was an underlying quantitative process of the symptoms. For example, subjects answering a question on shortness of breath would dichotomize the underlying obstruction of the airway into a "yes" or "no" response according to a subjective feeling.

Let  $N(\mu_1, \Sigma)$  and  $N(\mu_2, \Sigma)$  be the distribution of the asthmatic and nonasthmatic subjects, respectively, where  $\mu_1$  is the vector of proportions of positive responses ( $X_j = 1$ ),  $\mu_2$  is the vector of proportions of negative responses ( $X_j = 0$ ) and  $\Sigma$  is the common variance-covariance matrix for both populations. Let  $X = (X_1, X_2, \dots, X_p)$  be the vector of symptoms of an individual. The Fisher's linear discriminant function would classify a subject with  $X$  as an asthmatic if

$$x' \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2) - \frac{1}{2} (\hat{\mu}_1 + \hat{\mu}_2)' \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2) \geq \log(k) \quad (2)$$

Where  $x$  is the observed value of  $X$  and the hat on  $\mu_1$ ,  $\mu_2$  and  $\Sigma$  denotes the sample version of the parameters and

$$k = \frac{q_0 C(1|0)}{q_1 C(0|1)}$$

Where  $q_0$ ,  $q_1$  are prior probabilities of asthma or absence of asthma, respectively, and  $C(1|0)$  is the cost of misclassification of a nonasthmatic as asthmatic and  $C(0|1)$  is the cost of misclassification of an asthmatic as nonasthmatic. In our application, we assume  $k = 1$ , which is a commonly used criterion.

### Bahadur representation

The third method applied to our data sets was the second order Bahadur representation [18,21]. In our application, the symptom variables were all correlated and dichotomous. Let

$\theta_j = P(X_j = 1)$ ,  $j = 1, 2, \dots, p$ , where,  $X_j$  is one of the symptom variables of asthma such as cough, shortness of breath,  $X_j = 1$  denotes the presence of the symptom and  $X_j = 0$  for absence of the symptom. As mentioned previously, in our study, we identi-

fied eight symptoms for our comparative classification analysis.

Let  $X_j$  be a binary random variable. The standardized version of  $X_j$  is given by

$$Z_j = \frac{X_j - \theta_j}{\sqrt{\theta_j(1-\theta_j)}} \quad (3)$$

Define the expectations of their cross products as  $\rho_{jk} = E(Z_j Z_k)$ ,  $\dots$ ,  $\rho_{jk\dots p} = E(Z_j Z_k \dots Z_p)$ . Bahadur showed that the joint distribution of  $X = (X_1, X_2, \dots, X_p)$  can be written as (Goldstein and Dillon 1978)

$$f(x_1, x_2, \dots, x_p) = P(x_1, x_2, \dots, x_p) P_{[1]}(x_1, x_2, \dots, x_p), \quad (4)$$

Where

$$P(x_1, x_2, \dots, x_p) = 1 + \sum_{j < k} \rho_{jk} Z_j Z_k + \sum_{j < k < l} \rho_{jkl} Z_j Z_k Z_l + \dots + \rho_{12\dots p} Z_1 Z_2 \dots Z_p$$

And

$$P_{[1]}(x_1, x_2, \dots, x_p) = \prod_{j=1}^p \theta_j^{x_j} (1-\theta_j)^{1-x_j}$$

Assuming the correlation coefficients with order higher than 2 being zero and using the sample mean and sample Pearson correlation coefficients, we obtain the second (sample) Bahadur representation

$$\hat{f}(x_1, x_2, \dots, x_p) = \left( \prod_{j=1}^p \hat{\theta}_j^{x_j} (1-\hat{\theta}_j)^{1-x_j} \right) \left( 1 + \sum_{j < k} \hat{\rho}_{jk} \hat{z}_j \hat{z}_k \right), \quad (5)$$

Where the estimates of the mean, the standardized observation and the empirical pair wise correlation coefficient are calculated as following:  $\hat{\theta}_j = \frac{I(X_j=1)}{\sum_{j=1}^n 1}$

$$\hat{z}_j = \frac{x_j - \hat{\theta}_j}{\sqrt{\hat{\theta}_j(1-\hat{\theta}_j)}} \text{ and}$$

$$\hat{\rho}_{jk} = \frac{\sum_{j,k} I(X_j=1, X_k=1) / n - \hat{\theta}_j \hat{\theta}_k}{\sqrt{\hat{\theta}_j(1-\hat{\theta}_j) \hat{\theta}_k(1-\hat{\theta}_k)}}. \text{ Note that } I$$

(condition) is an indicator function that takes a value of 1 if the condition is true and 0 otherwise. The probability  $\hat{f}$  can be estimated based on the sample values of  $\theta$  and  $\rho$  from the asthmatic and the nonasthmatic group. Let  $\hat{f}_0(x_1, x_2, \dots, x_p)$  and  $\hat{f}_1(x_1, x_2, \dots, x_p)$

be the probability estimated from the asthmatic and the nonasthmatic group, respectively. We would classify a subject with symptom  $x = (x_1, x_2, \dots, x_p)$  into the asthmatic group if

**Table 1:** The agreement table for two classification methods.

Method B			
Method A	0	1	Total
0	$P_{00}$	$P_{01}$	$P_0$
1	$P_{10}$	$P_{11}$	$P_1$
Total	$P_0$	1	1

**Table 2:** Selected pairwise comparisons of three classifications procedures with MD asthma,  $pc_{20}4$  and  $pc_{20}8$  for phase 1 survey of 102 subjects.

Method A	Method B	$np_{00}$	$np_{01}$	$np_{10}$	$np_{11}$	$\kappa$	SD
MD Asthma	Logistic	82	2	2	16	0.8651	0.0659
MD Asthma	Logistic 4	63	21	4	14	0.3849	0.0938
MD Asthma	Logistic 8	47	37	2	16	0.2542	0.07
MD Asthma	Bahadur	76	8	5	13	0.5885	0.1023
MD Asthma	Fisher	79	5	3	15	0.7414	0.0864
MD Asthma	PC <sub>20</sub> 4	50	34	4	14	0.2254	0.0779
MD Asthma	PC <sub>20</sub> 8	45	39	3	15	0.2067	0.0695
PC <sub>20</sub> 4	Logistic 4	46	8	21	27	0.4207	0.0876
PC <sub>20</sub> 8	Logistic 8	34	14	15	39	0.4301	0.0895
Logistic	Bahadur	76	8	5	13	0.5885	0.1023
Logistic	Fisher	80	4	2	16	0.8061	0.0761
Bahadur	Fisher	77	4	5	16	0.7253	0.0861

Where  $np_{ij}$ s denote the observed counts in the cells of the contingency table,  $\kappa$  is the Kappa statistic and SD is the estimate of the standard deviation of  $\kappa$ .

**Table 3:** Pairwise Comparisons of Three Classifications Procedures with MD Asthma for Phase 2 Survey of 2963 Subjects.

Method A	Method B	$np_{00}$	$np_{01}$	$np_{10}$	$np_{11}$	$\kappa$	SD
MD Asthma	Logistic	2433	96	216	218	0.524	0.023
MD Asthma	Bahadur	2130	399	89	345	0.492	0.019
MD Asthma	Fisher	2331	198	131	303	0.583	0.021
Logistic	Bahadur	2219	430	0	314	0.522	0.019
Logistic	Fisher	2462	187	0	314	0.736	0.018
Bahadur	Fisher	2179	40	283	461	0.675	0.016

Where  $np_{ij}$ s denote the observed counts in the cells of the contingency table,  $\kappa$  is the Kappa statistic and SD is the estimate of the standard deviation of  $\kappa$ .

$$\delta \hat{f}_1(x_1, x_2, \dots, x_p) > (1 - \delta) \hat{f}_0(x_1, x_2, \dots, x_p), \quad (6)$$

where  $\delta$  is the prior probability of asthma. We assumed a  $\delta = 0.5$  in our comparative study of these three classification procedures.

## Results and Agreement Analysis

We applied the three classification methods to the two data sets from our surveys on Texas healthcare professionals. The pair wise comparisons of the classification results were based upon  $\kappa$  statistic via a series of  $2 \times 2$  contingency tables as that shown in Table 1 [22].

In Table 1,  $p_{ij}$ ,  $i = 0, 1$  and  $j = 0, 1$ , is the proportion of subjects in category  $i$  by method A and in category  $j$  by method B. The estimate of the  $\kappa$  statistic is defined in Equation (7). If two methods are in complete agreement,  $\kappa = 1$ . If  $\kappa \geq 0$ , the observed agreement is greater than chance, and if observed agreement is less than chance, then  $\kappa < 0$ . We used R 3.5.0 to compute the estimates of the  $\kappa$  statistic and its standard deviation (R Project 2019).

$$\hat{\kappa} = \frac{p_0 - p_e}{1 - p_e}, \quad (7)$$

Where  $p_0 = p_{00} + p_{11}$  and  $p_e = p_{01} + p_{10}$ .

The values of the  $\kappa$  statistic are shown in Table 2 and Table 3 for phase I and phase II data, respectively.

Results in Table 2 show that the classification based on logistic regression was highly concordant with a prior physician diagnosis of asthma (MD asthma) for the phase I data. We observed  $p_{01} = p_{10} = 2/102$  and the estimate of the  $\kappa$  statistic was 0.8651, with a standard deviation of 0.0659. Methacholine challenge was given during phase I and two indicator variables (PC<sub>20</sub>4 and PC<sub>20</sub>8) were generated from the outcomes as described in Section 2. The concordance between the logistic regression using the methacholine challenge and the MD asthma were low. Similar low concordance was observed between the MD asthma and the direct PC<sub>20</sub>4, PC<sub>20</sub>8 without using logistic regression. The methacholine challenge seemed much more sensitive than the physician's diagnosis. For phase I data, the concordance between MD asthma and the Fisher's linear discriminant function was high. The value of  $\kappa$  statistic was 0.7414 with a standard deviation a value of 0.0864. The results from the second order Bahadur representation and MD asthma produced 0.5885 for the  $\kappa$  statistic. For the three pair wise comparisons among logistic regression, second order Bahadur representation and Fisher's linear discriminant function, logistic regression

and Fisher's method had a high concordance with  $\kappa = 0.8061$  and standard deviation being 0.0761. Bahadur representation and Fisher's linear discriminant function had  $\kappa = 0.7253$  with a standard deviation 0.0861. The  $\kappa$  statistic for concordance between the logistic model and the Bahadur representation was 0.5885 with standard deviation 0.1023, which was the same as MD asthma compared directly to second order Bahadur representation.

In phase II of the survey, no methacholine challenge was given. The data set used in this comparative analysis consisted of 2963 subjects without missing values. Table 3 summarizes the  $\kappa$  and corresponding standard deviations for the six pair wise comparisons. Compared to MD asthma, the three statistical classification procedures showed relatively large  $\kappa$  values, although they were lower than those in phase I, ranging from 0.4917 to 0.5826. For the pair wise comparisons among the three statistical procedures, logistic regression produced overly sensitive classification since  $p_{10} = 0$ . The  $\kappa$  statistics were 0.5224 and 0.7362 when comparing results from logistic regression to the second order Bahadur representation and Fisher's linear discriminant functions, respectively. Comparison of the Bahadur representation to Fisher's method resulted in a  $\kappa$  statistic of 0.6749 with a standard deviation of 0.0164.

### Concluding Remarks

In this study, we applied three widely used statistical classification techniques to data obtained from two surveys on asthma in Texas healthcare professionals. All three procedures showed a high concordance with a prior physician diagnosis of asthma although concordance decreased as sample size increased. The Fisher's linear discriminant function used an assumption of normality for the explanatory variables that was clearly not true in our study. However, our study demonstrated its robustness when applied to dichotomous variables. For the classification based on logistic regression, we used a default cut-off value of 0.5 when classifying a subject as asthmatic or nonasthmatic, which may not be appropriate in other applications. A Bayesian approach, together with a cost function of misclassification, may add more insights in classification. However, this was beyond the scope of this study since it is hard to justify a particular cost function in a general setting. Due to the correlation among the dichotomous variable, we would expect use of Bahadur representation to produce better classification results. However, we were unable to confirm or reject this in the absence of a true gold standard. Comparing with the doctor's classification, logistic regression and Fisher's linear discriminant function provided stronger (higher)  $\kappa$  value than that of Bahadur representation. This is consistent with the current practice since Bahadur is less commonly used in data analysis. The estimates of  $\kappa$  statistic between the logistic regression and Fisher's linear discriminant function showed relatively

high consistency for these two methods. For simplicity only the second order Bahadur representation was used in this study. Analysis with the higher order representation was also beyond the scope of our current study.

Statistical classification tools have been widely applied for clinical evaluations. For example, multiple logistic regression was used to design and enroll patients for clinical trial of Early Treatment for Retinopathy of Prematurity study [23]. Gregori, et al. [24] have studied and compared statistical classifiers in evaluating coronary artery diseases.

### Acknowledgments

This study was partially supported by Grants No. 5R01OH03945-01A1, T42CCT610417 and 5T42OH008421 from the National Institute for Occupational Safety and Health/Centers for Disease Control and Prevention and by Grant No. 71161011/G0107 from National Natural Science Foundation of China.

### References

- Mannino DM, Homa DM, Pertowski CA, Ashizawa A, Nixon LL, et al. (1998) Surveillance for asthma-United States, 1960-1995. *MMWR CDC Surveill Summ* 47: 1-27.
- Arif AA, Delclos GL, Lee ES, Tortolero SR, Whitehead LW (2003) Prevalence and risk factors of asthma and wheezing among USA adults: An analysis of the third national health and nutrition examination survey (1988-94). *European Respiratory Journal* 21: 827-833.
- Kivity S, Shochat Z, Bressler R, Wiener M, Lerman Y (1995) The characteristics of bronchial asthma among a young adult population. *Chest*: 108: 24-27.
- Milton DK, Solomon GM, Rosiello RA, Herrick RF (1998) Risk and incidence of asthma attributable to occupational exposure among HMO members. *Am J Ind Med* 33: 1-10.
- MRC Medical Research Council (1960) Standardized questionnaires of respiratory symptoms. *British Medical Journal* 1665.
- Ferris BG (1978) Epidemiology Standardization Project (American Thoracic Society). *American Review of Respiratory Disease* 118: 1-120.
- Burney PG, Chinn S, Britton JR, Tattersfield AE, Papacosta AO (1989a) What symptoms predict the bronchial response to histamine? Evaluation in a community survey of the bronchial symptoms questionnaire (1984) of the International Union Against Tuberculosis and Lung Disease. *International Journal of Epidemiology* 18: 165-173.
- Burney PG, Laitinen LA, Perdrizet S, Huckauf H, Tattersfield AE, et al. (1989b) Validity and repeatability of the IUATLD (1984) Bronchial Symptoms Questionnaire: An international comparison. *European Respiratory Journal* 2: 940-945.
- Abramson MJ, Hensley MJ, Saunders NA, Wlodarczyk JH (1991) Evaluation of a new asthma questionnaire. *Journal of Asthma* 28: 129-139.
- Kongerud J, Boe J, Søyseth V, Naalsund A, Magnus P (1994) Aluminium potroom asthma: The Norwegian experience. *European Respiratory Journal* 7: 165-172.
- Sheffer AL, Bousquet J, Buss WW (1992) International consensus report on diagnosis and treatment of asthma. *European Respiratory Journal* 5: 601-641.

12. Delclos GL, Gimeno D, Arif AA, Benavides FG, Zock JP (2009) Occupational exposures and asthma in health-care workers: comparison of self-reports with a workplace-specific job exposure matrix. *American Journals of Epidemiology* 169: 581-587.
13. Delclos GL, Arif AA, Aday L, Carson A, Lai D, et al. (2006) Validation of an asthma questionnaire for use in healthcare workers. *Occupational and Environmental Medicine* 63: 173-179.
14. Douwes J, Pearce N (2002) Asthma and the westernization "package". *International Journal of Epidemiology* 31: 1098-1102.
15. Asparoukhov OK, Krzanowski WJ (2001) A comparison of discriminant procedures for binary variables, *Computational Statistics & Data Analysis* 38: 139-160.
16. Hosmer DW, Lemeshow S (1989) *Applied Logistic Regression*. Wiley, New York.
17. Anderson TW (1984) *An introduction to multivariate statistical analysis*. (2<sup>nd</sup> edn), Wiley, New York.
18. Goldstein M, Dillon WR (1978) *Discrete Discriminant Analysis*. Wiley, New York.
19. Agresti A (2002) *Categorical Data Analysis*. (2<sup>nd</sup> edn). Wiley, New York.
20. R Project (2019) R package version 3.5.0.
21. Bahadur RR (1961) A representation of the joint distribution of response to n dichotomous items. In: H Solomon, *Studies in Item Analysis and Prediction*. Stanford University, Palo Alto, CA, 158-168.
22. Fleiss JL (1981) *Statistical Methods for Rates and Proportions*. Wiley, New York.
23. Hardy RJ, Palmer EA, Dobson V, Summers CG, Phelps DL, et al. (2003) Risk analysis of prethreshold retinopathy of prematurity. *Archives of Ophthalmology* 121: 1697-1701.
24. Gregori D, Bigi R, Cortigiani L, Bovenzi F, Fiorentini C, et al. (2009) Non-invasive risk stratification of coronary artery disease: an evaluation of some commonly used statistical classifiers in terms of predictive accuracy and clinical usefulness. *Journal of Evaluation in Clinical Practice* 15: 1777-1781.