Predictive Modelling Strategies to Understand Heterogeneous Manifestations of Asthma in Early Life

Danielle Belgrave*, Rachel Cassidy*, Adnan Custovic
Department of Paediatrics, Imperial College,
London, United Kingdom
*Joint first-author

Daniel Stamate*
Department of Computing,
Goldsmiths, University of London
London, United Kingdom
*Joint first-author

Louise Fleming, Andrew Bush, Sejal Saglani
NHLI, Imperial College London, and Dept of Respiratory Paediatrics,
Royal Brompton Hospital London, United Kingdom

Abstract— **Wheezing is common among children and ~50% of those under 6 years of age are thought to experience at least one episode of wheeze. However, due to the heterogeneity of symptoms there are difficulties in treating and diagnosing these children. 'Phenotype specific therapy' is one possible avenue of treatment, whereby we use significant pathology and physiology to identify and treat pre-schoolers with wheeze. By performing feature selection algorithms and predictive modelling techniques, this study will attempt to determine if it is possible to robustly distinguish patient diagnostic categories among pre-school children. Univariate feature analysis identified more objective variables and recursive feature elimination a larger number of subjective variables as important in distinguishing between patient categories. Predicative modelling saw a drop in performance when subjective variables were removed from analysis, indicating that these variables are important in distinguishing wheeze classes. We achieved 90%+ performance in AUC, sensitivity, specificity, and accuracy, and 80%+ in kappa statistic, in distinguishing ill from healthy patients. Developed in a synergistic statistical - machine learning approach, our methodologies propose also a novel ROC Cross Evaluation method for model post-processing and evaluation. Our predictive modelling's stability was assessed in computationally intensive Monte Carlo simulations.**

Keywords— wheeze, pre-school, feature selection, predictive modelling, ROC analysis, model post-processing, Monte Carlo

I. Introduction

Wheeze, described as a high-pitched whistling sound emitted during expiration [1], [2], is common among children of pre-school age. It is thought that around 50% of pre-schoolers (up to age 6) will have at some point experienced at least one episode of wheeze [3]. Compared with older age groups, pre-schoolers with wheeze have been shown to have a 50% greater need for assistance by ambulance, just under twice the number of emergency response (ER) visits and almost three-times the rate of hospitalization [4]. With previous estimates placing the cost of care for pre-schoolers with wheeze at around 0.15% of the healthcare budget in the UK, further burden is being placed on already stretched healthcare resources [5]. In this study, we will examine a 'phenotype specific therapy' as one possible avenue of diagnosis of wheeze in children of pre-school age using meaningful symptom clusters and significant pathology and physiology to identify different groups of pre-schoolers with wheeze leading to appropriate treatment.

Diagnosing wheeze in pre-schoolers is fraught with difficulty due to the heterogeneity in the timing and manifestation of co-occurring symptoms. In surveys conducted within European populations, only 83.5% of parents could correctly identify an episode of wheeze [6]. Specifically, within the UK, 33% of parents after watching recordings of children with wheeze concluded their child did not actually suffer from episodes [7]. While the presentation of symptoms and family history of atopy are useful in an initial investigation, they are only an indication of diagnosis; a trial of steroids may be given but there is no certainty they will relieve the child of symptoms [13], [14]. Furthermore, lung function tests are not routinely performed in young children. While it is possible to undertake other measurements of pre-school children in a primary care setting, such as peak flow, these may require trained technicians and specialised laboratory equipment.

The diagnosis of wheeze in pre-school children is further complicated by the heterogeneity of wheeze development over time. There have been several approaches to classify wheeze phenotypes in pre-school age children with wheeze. One approach has been to divide children into atopic and non-atopic groups based on 'early aeroallergen sensitisation' [11]. Henderson et al's [8] study investigating the association between pre-school age wheezing phenotype and environmental influences on the development of asthma identified six wheeze phenotypes which reflect the heterogeneity in evolution of wheeze symptoms over time [9], [10].

Due to the complexities involved in asthma diagnosis in early childhood, conditions presenting with wheeze may lead to over- or under-diagnoses [15]. This has been acknowledged in global guidelines and reports on childhood asthma identifying the difficulty of diagnosis in children of pre-school age [1], [16]–[18]. The symptom may be considered indicative of asthma and, therefore, treated with courses of inhaled corticosteroids (ICS), resulting in the child developing significant steroidal side effects [19]. Under-diagnosis results in the absence of suitable and appropriate therapy [20].

Due to this heterogeneity of wheeze and difficulty in distinguishing patterns over time, 'phenotype specific therapy' [21] has been suggested as a possible approach for a more stratified approach to wheeze diagnosis and management strategies. The aim is to eventually phenotype children and drive individualised therapy based on the combination of the identification of meaningful symptom clusters and significant pathology and physiology. In this study, we hypothesise that using feature selection and prediction modelling techniques on biomarkers will enable us to identify groups regarding the presence and the stage of wheeze condition in pre-schoolers. We use data from a cohort of children with pre-school wheeze (PSW cohort) from the Royal Brompton Hospital (RBH) to test this hypothesis.

II. Methods

A. Description of study population

The RBH has collected clinical data from children of pre-school age since 2010, resulting in the PSW cohort. The dataset obtained from this group of children contains information on 150 patients for 636 variables. 61% of patients are male with a cohort mean age of 33.66 months (SD 16.82). 72% of patient data pertains to pre-schoolers who have been seen historically and are currently being followed-up. The remainder pertains to recently seen and documented patients.

B. Description of variables

The variables in the PSW cohort can be broadly separated into:

- Baseline Demographics (such as gender, age, ethnicity)
- Subjective clinical data obtained from the parents at the patients first research contact (such as current household smoking status, location of home and family history of asthma).
- Objective clinical data collected at the patients first research contact, and the first, one year and final follow-up appointments. These include whether the patient has recurrent chest infections, persistent cough or upper airway problems.
- Biological data collected at the patients first research contact (such as skin prick allergy test results, pH probe to detect level of acid present, blood tests).

In the PSW cohort, clinicians have identified 4 categories (referred to as 'Group' variable in the dataset) of patient based on clinical presentation: Wheeze (patient with wheeze), Wheeze+ (patient with wheeze and some other respiratory issue i.e. rattly chest), Clean Controls (patient with upper airway problem but no lower airway issue) and Diseased Controls (patient without wheeze but has other respiratory issue i.e. presumed chest infection). In this study, the Control categories have been merged in to a single Control category.

C. Statistical and machine learning methods

A synergistic combination of statistical and machine learning techniques was used in this study to identify and employ significant biomarkers in effectively distinguishing between different categories of the 'Group' variable. Techniques used to achieve this aim included feature selection methods such as univariate feature analysis, recursive feature elimination, principal component analysis (PCA), the RelieF method, as well as SMOTE, permutation tests, Chi-squared test, ROC analysis, etc. Supervised learning algorithms including feed-forward neural networks (NN), support vector machines (SVM), and random forests (RF), accompanied the above methods towards this aim.

D. Data cleaning

Extensive data cleaning was undertaken to ensure the harmonisation of data collected historically and more recently. For certain variables, the coding method between both data sets differed. For example, string notation such as 'moisturiser', 'steroid cream' and 'both' had been used to record eczema treatment for patients who had been seen historically, however, numeric coding (such as 0, 1, 2) had been implemented for recent patients.

Variables with less than 25% complete cases, describing open-ended questions, or categorical data where one response category contained less than 5% of the data were removed. A further 21 variables were discarded as they were considered less likely to provide useful or additional information during analysis. This resulted in a total of 89 variables being retained. As the sample size for this study was modest, variables with 3 or more categories were collapsed into dichotomous categories.

E. Imputation

Patterns of missing data were initially assessed using the 'misschk' package [22] in Stata [23]. Missing values were then imputed using the Stata package 'ice'[24], which performs multivariate imputation using chained equations. To implement multivariate imputation for missing data, 26 variables that were found to be collinear with other variables were identified and removed. After the data had been cleaned with missing values imputed, 63 variables including the 'Group' variable were retained for further analysis. Of the 63 retained, some variables relevant for this paper are described in Table I.

F. Feature selection

In order to identify variables with strong associations to the 'Group' variable, feature selection techniques including univariate feature analysis, recursive feature elimination and PCA were used on the PSW dataset. A chi-squared classification test for non-negative features was used on each variable in univariate analysis to determine which features of the dataset have been identified as having prominent relationships to the outcome 'Group' variable.

Four separate analyses were conducted with different outcome pairs; each Group class against the remaining classes referred to as ‘Other’ with a final test for Wheeze+ against Wheeze (see Table II). These four pairs of outcomes were also used in the recursive feature elimination analysis. Logistic regression models were created on a training set of data (that was then used on test data) before then discarding the 10% weakest features. This is then repeated until model accuracy significantly diminishes.

G. *Principal component analysis*

To reduce the number of redundant variables, PCA was implemented on the PSW dataset. As the dataset contained a large number of dichotomous and continuous variables, a variation of PCA was performed using the R CRAN package ‘PCAmixdata’[25]. This employs both PCA and multiple correspondence analysis (MCA) to accommodate both types of data. The principal component scores obtained were then regressed on Group classes Wheeze+, Wheeze and Control to determine if there was a significant association between patient categories and principal component (PC) scores.

TABLE I. DESCRIPTION OF SOME RELEVANT RETAINED VARIABLES A. NOMINAL VARIABLES. B. CONTINUOUS VARIABLES.

A.

Variable	Description	(%) in Categories
Group	'Control' (0), 'Wheeze' (1) or 'Wheeze+' (2)	0 (24.00%), 1 (36.67%), 2(39.33%)
Exposure to pets at home (currently)	No (0), Yes (1)	0 (64.67%), 1 (35.33%)
Family history of eczema	No (0), Yes (1)	0 (42.67%), 1 (57.33%)
Gender	Girl (0), Boy (1)	0 (39.33%), 1 (60.67%)
Ever admitted to hospital with wheeze	No (0), Yes (1)	0 (32.67%), 1 (67.33%)
Has your child ever wheezed	No (0), Yes (1)	0 (21.33%), 1 (78.67%)
How many episodes in the past 6 months	≤3 (0), >3 (1)	0 (38.00%), 1 (62.00%)
Infant feeding	Other (0), Breast (1)	0 (52.67%), 1 (47.33%)
Persistent cough	No (0), Yes (1)	0 (62.67%), 1 (37.33%)
PSW (Pre-school wheeze)	No (0), Yes (1)	0 (31.33%), 1 (68.67%)
Recurrent chest infections	No (0), Yes (1)	0 (72.00%), 1 (28.00%)

B.

Variable	Description	Mean	SD
BAL neutrophils (%) RBH	Neutrophils cell percentage (out of total cell count) from bronchoalveolar lavage undertaken by RBH	26.61	22.20
HDM iU	Skin prick allergy test	12.54	23.22
Macrophages (%) RBH	Macrophages cell percentage (out of total cell count) undertaken by RBH	54.12	23.08
MBL	Mannose-binding lectin	2569.69	1622.19
Pneumococcus result	Test for pneumococcal antibodies	108.91	102.19
Total IgE	Immunoglobulin E	118.39	227.77

H. *Predictive modelling*

The methodology developed for predicting a diagnosis for patients is discussed in this section, and was coded in R using ‘caret’ for parallel predictive model training and tuning, ‘pROC’ for ROC analysis, and other R libraries required by ‘caret’.

We built our methodology upon various algorithms which require and do not require external feature selection, such as back propagation for feed-forward neural networks and support vector machines, and random forests, respectively. As validation techniques, we employed simple and nested cross validation (CV), and in some variants of our methodology for the 2-class problems we performed post-processing of the predictive models based on alternative optimal cutoff points found on multiple ROC curves. These post-processing methods are also a solution for the class imbalance problem in the 2-class classification in our framework. A pre-processing method, SMOTE [26], which generates synthetic observations based on interpolations between an observation and its neighbours, is alternatively used to balance the training data, and is considered separately or in conjunction with the model post-processing methods. In addition to the data imbalance, a second issue in our framework is the relatively small size of the data (150 observations). This small size may become a challenge in a predictive modelling approach in which we employ two cross validations (as it is the case in nested CV) for model tuning and validation, and a model post-processing. As such, we propose a novel model post-processing and evaluation method, called ROC Cross Evaluation (ROC-CE), to address this issue.

Finally, we performed comparisons and studied the stability of the performances of various predictive models built in different variants of our methodology, via computationally intensive Monte Carlo (MC) simulations with 1000 experiments each.

Each of the classification problems were considered on two versions of the dataset. The first dataset, which we call the reduced dataset, comprises 47 variables containing objectively measured predictors, and the second dataset, which we call the mixed dataset, comprises 62 variables containing objective and subjective predictors. In particular, the reduced dataset was investigated in the context of the clinical lead to see how good a performance in diagnoses can be reached on objective variables only, and how it compares to diagnoses made on the objective and subjective variables from the mixed dataset.

III. RESULTS

A. *Missing data*

A preliminary investigation of the dataset found that only 2 patients (1.33%) had complete data, with over a third of patients (35%) missing data for 20 or more variables.

B. *Feature analysis*

Table III shows results from univariate feature analysis. MBL, Total IgE, Pneumococcus and BAL neutrophils gave the highest chi-squared statistic, indicating a strong relationship with each set of diagnostic outcome pairs.

The five strongest features identified with recursive feature elimination using the 62 retained variables from the PSW dataset against the four pairs of diagnostic outcomes are displayed in Table III. 'Recurrent chest infections' was identified as the strongest feature in the models with dependent outcome (Wheeze+, Other), (Wheeze, Other) and (Wheeze+, Wheeze) with 'PSW' the strongest feature with dependent outcome (Control, Other).

TABLE II. PAIRS OF OUTCOMES USED IN FEATURE ANALYSIS

Test	Outcome
Test 1	Wheeze+, Other
Test 2	Wheeze, Other
Test 3	Control, Other
Test 4	Wheeze+, Wheeze

TABLE III. UNIVARIATE FEATURE ANALYSIS AND RECURSIVE FEATURE ELIMINATION RESULTS USING 62 RETAINED VARIABLES FROM PSW DATASET

A. WHEEZE+, OTHER

Univariate Feature Analysis		**Recursive Feature Elimination**	
Variables	***Score***	***Variables***	***Score***
MBL	1154.00	Recurrent chest infections	1
Total IgE	685.10	PSW	2
HDM iU	134.40	IgM	3
Pneumococcus result	131.10	Persistent cough	4
BAL neutrophils (%) RBH	35.68	Exposure to pets at home (currently)	5

B. WHEEZE, OTHER

Univariate Feature Analysis		**Recursive Feature Elimination**	
Variables	***Score***	***Variables***	***Score***
MBL	13120.00	Recurrent chest infections	1
Total IgE	3480.00	PSW	2
Pneumococcus result	308.10	Persistent cough	3
BAL neutrophils (%) RBH	127.50	Infant feeding	4
Macrophages (%) RBH	46.60	Ever admitted to hospital with wheeze	5

C. CONTROL, OTHER

Univariate Feature Analysis		**Recursive Feature Elimination**	
Variables	***Score***	***Variables***	***Score***
MBL	8167.00	PSW	1
Total IgE	1341.00	Has your child ever wheezed	2
HDM iU	86.62	Recurrent chest infections	3
Pneumococcus result	45.02	How many episodes in the past 6 months	4
BAL neutrophils (%) RBH	34.90	Family history of eczema	5

D. WHEEZE+, WHEEZE

Univariate Feature Analysis		**Recursive Feature Elimination**	
Variables	***Score***	***Variables***	***Score***
MBL	7674.00	Recurrent chest infections	1
Total IgE	1990.00	Persistent cough	2
Pneumococcus result	271.90	IgM	3
BAL neutrophils (%) RBH	99.22	Ever admitted to hospital with wheeze	4
HDM iU	61.18	Gender	5

C. Principal component analysis

After conducting PCA on the PSW dataset, it was determined that 23 principal components had eigenvalues greater than (or equal to) one. Reducing the dataset to these principal components ensured that 79.5% of the original variance exhibited by the data was accounted for. Individuals were then assigned principal component scores.

The principal components (with their scores) were then used in three logistic regression analyses against the patient diagnostic outcomes (Control, Wheeze), (Control, Wheeze+), (Wheeze, Wheeze+). The key findings were:

- In regression analysis conducted with (Control, Wheeze) as the dependent variable, PC4 which loads heavily onto 'Eosinophils' and 'Eosinophils percentage' and PC21 which loads heavily on to 'Location of home' were found to be statistically significant risk ORs.
- When (Control, Wheeze+) was regressed on by each principal component, PC3 (risk factor) which loads heavily on to 'BAL neutrophils RBH' and 'Macrophages RBH' and PC14 (protective factor) which loads heavily on to 'Total IgG' was found to have produced statistically significant ORs, respectively.
- PC2 which loads heavily on to 'Height', 'Age (months)', 'Weight' and PC14 which loads heavily on to 'Total IgG' were found to produce statistically significant protective ORs with (Wheeze, Wheeze+) as the dependant variable.

D. Predictive modelling

In the 3-class problem based on Control, Wheeze and Wheeze+ classes, we tuned RF, NN and SVM models in 10-fold CV. Models such as NN and SVM are sensitive with a tendency of decreasing their performance in presence, in the dataset, of less-predictive or non-predictive variables in addition to the predictive ones, while RF models are robust from this point of view [29]. As such, when building the NN and SVM models, we applied a feature selection procedure based on the RelieF method [28] combined with a permutation test [27] conveniently implemented by using 2000 random permutations.

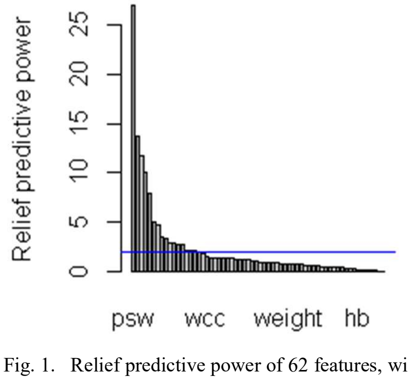

Fig. 1. Relief predictive power of 62 features, with 16 predictors crossing the horizontal blue line corresponding to 1.96 standard deviations, selected for model training.

The features having an observed Relief score at least 1.96 standard deviations away from the centre of the normal distribution of the Relief scores obtained, for each feature, by randomly permuting the classes of the observations, have been selected as predictors for model training. The number of standard deviations away of each feature's observed Relief score, from the centre of the above distribution, defines, in our approach, the 'Relief predictive power' of that feature. The 'Relief predictive power' is illustrated for all the 62 features in the bar chart in Fig. 1, with 16 predictors crossing the horizontal blue line corresponding to 1.96 standard deviations. The idea here, inspired by the permutation tests [27], is that the features having observed Relief scores larger than 95% of Relief scores (all in absolute value) obtained by randomly shuffling the classes, are considered predictive. Variables such as 'PSW' (Relief predictive power = 27) and 'Recurrent chest infections' (Relief predictive power = 11.8) are among the top 3 predictors.

By applying the feature selection method above, the best NN model was based on one single hidden layer with 20 nodes and a weight decay of 0.4 for L2 regularization, and led to an accuracy of 0.828 (SD 0.012) with a 95% confidence interval 95%CI [0.805, 0.849], and the kappa statistic of 0.736 (SD 0.019). The best SVM model was based on the radial kernel, had the hyper-parameters cost C=4, and gamma=0.002, and led to similar performances of accuracy of 0.830 (SD 0.011) with 95%CI [0.807 0.852], and of kappa statistic of 0.742 (SD 0.017). Finally, the best RF model, built on all the 62 features, was based on 1000 trees with mtry (the number of predictors competing in a tree node at a time) equal to 20, and achieved an accuracy of 0.817 (SD 0.015) with a 95%CI [0.787, 0.846] and a kappa statistic of 0.722 (SD 0.023). As mentioned above, the models' performances stability was studied with 1000 experiment based MC simulations. Given the 3-class classification and the high kappa values, these models with equivalent performance are judged as good, but perhaps not good enough to use them in a diagnosing process. As such we transformed the 3-class problem in multiple 2-class problems as mentioned above.

The methodology that we developed for the 2-class problems involved a higher complexity due to two joint aspects of the data, namely of being imbalanced (for instance 36 Control versus 114 Wheeze and Wheeze+), and of being relatively small (150 observations). We fully illustrate the methodology, with some of its developed variants, on the Control versus Other (i.e. Wheeze and Wheeze+) problem, which is one of the most important 2-class classification problems here. We then provide the results of the prediction modelling performed on the other 2-class classification problems.

Our approach consisted of tuning and evaluating NN, SVM and RF models in a nested CV, formed of a 10-fold inner CV, and 4-fold outer CV (results in outer 3-fold CV were obtained, but they were statistically comparable and not reported here), with extra operations encapsulated. The solution we adopted for the class imbalanced data was based on optimising the area under curve, AUC, in the inner CV, corroborated with applying two optimisation methods on ROC curves in the outer CV. The first such a method is based on determining the point on the ROC which is the closest to the top-left point of coordinates (0,1) which represents an ideal model with both sensitivity and specificity equal to 1 (see Fig. 2 for illustrated ROC curves). The second optimisation method is based on the Youden's J statistic [31] whose largest value corresponds to the cutoff point on the ROC, maximizing the sum of the sensitivity and specificity, or maximizing the distance to the main diagonal. Both cutoff points represent an alternative to the default 0.5 cutoff point for the classification probability, which usually leads to a disproportion between sensitivity and specificity in case of class imbalance. Using the closest top-left cutoff or youden cutoff balances sensitivity and specificity [29]. For instance, Table IV summarises a MC simulation consisting of 1000 experiments, each of which producing a tuned RF model with AUC optimisation in the inner 10-fold CV, which was post-processed using the closest top-left cutoff method and evaluated in the outer 4-fold CV.

As such, according to our methodology, a validation fold in the outer CV is used in both – the post-processing of the model by determining and applying a new cutoff point for classification probabilities, and in evaluating the model by applying a tailored method that we introduce here, called ROC Cross Evaluation, or ROC-CE. It is a recommended requirement the model evaluation to be performed on a dataset which is different from the dataset used for generating the ROC curve employed in determining an optimal cutoff.

TABLE IV. BEST MODEL IDENTIFIED FOR CONTROL VS OTHER WITH MIXED DATASET OF 62 FEATURES: OPTIMISED RF'S PERFORMANCES WITH 95% CONFIDENCE INTERVAL'S LEFT AND RIGHT ENDS

	Mean	SD	ci95left	ci95right
AUC	0.981	0.007	0.965	0.991
Sens	0.927	0.018	0.895	0.965
Spec	0.915	0.030	0.861	0.972
Accuracy	0.924	0.017	0.893	0.960
Kappa	0.802	0.043	0.723	0.892

ROC-CE is designed to make the ROC based model post-processing and evaluation also possible for small datasets, although in such cases usually data is insufficient to detach from it a new dataset for model evaluation. In ROC-CE we take a validation fold D in the outer CV, and split it into a number of k sub-folds. In order to score the observations in a hold-out sub-fold L of D, we use all the other k-1 sub-folds of D to build a ROC curve, on which we determine the closest top-left or youden cutoff point, that we then apply to get the post-processed model that in turn is used to produce predictions for the observations in L. We repeat this for each hold-out sub-fold in D, and aggregate predictions to finally produce the performance evaluation on fold D. The principle of ROC-CE is somehow similar to that of cross validation, but instead of model training and validation, the model post-processing and evaluation take place based on determining and using optimised cutoffs on multiple ROCs. In the case of our dataset, which is small, each hold-out sub-fold L of D was formed of one observation only.

Table IV shows a good balance between the sensitivity of 0.927 (the detection rate of Wheeze and Wheeze+ patients) and a specificity of 0.915 (the detection rate of Controls) of the best RF model tuned and evaluated in a nested CV with the ROC-CE method based on the closest top-left cutoff point (similar performances have been obtained with the youden cutoff). The stability of the performances of this best model overall, was as usually studied within 1000 MC experiments. SVM and NN best models were obtained in the same way. RF was better with 2% in average accuracy and 5% in average kappa than the SVM model, and with 1% in average accuracy and 3% in average kappa than the much more computationally expensive NN best model. The use of the SMOTE method [26] as an alternative to ROC-CE for the class imbalance, led to a 4% imbalance between sensitivity (0.936) and specificity (0.9) on tuned RF models, with comparable performances to those in Table IV. We conclude that the performance levels of RF model in Table IV makes it suitable in establishing if a patient suffers of a wheeze condition.

The same methodology was applied to the other 2-class classifications with analyses on Wheeze VS Other, Wheeze+ VS Other, and Wheeze+ VS Wheeze. In the first 2-class classification, the best model was an optimised SVM with radial kernel tuned in a nested CV with 10-fold inner CV, 4-fold outer CV, and post-processed and estimated in ROC-CE with the closest top-left method. Variation/ stability of the model performances were investigated in a MC simulation of 1000 experiments, with performances shown in Table V (Part A). The best model with the second 2-class classification was an optimised SVM with radial kernel, tuned in a nested CV with 10-fold inner CV, 4-fold outer CV, and post-processed and estimated in ROC-CE with the closest top-left method. Variation/ stability of the model were investigated with MC (1000 experiments), with performance shown in Table V (Part B). With the third 2-class classification, the best model was RF with mtry=20 obtained and evaluated in 10-fold CV with SMOTE technique applied on the training folds. MC performances are shown in Table V (Part C).

TABLE V. BEST MODELS IDENTIFIED FOR 2-CLASS CLASSIFICATIONS WITH MIXED DATASET OF 62 FEATURES. A. OPTIMISED SVM'S PERFORMACES FOR WHEEZE VS OTHER. B. OPTIMISED SVM'S PERFORMANCE FOR WHEEZE+ VS OTHER. C. RF'S PERFORMANCES FOR WHEEZE+ VS WHEEZE.

A.

	Mean	SD	ci95left	ci95right
Sens	0.796	0.029	0.745	0.855
Spec	0.836	0.027	0.779	0.884
Accuracy	0.821	0.023	0.773	0.860
Kappa	0.622	0.046	0.526	0.704

B.

	Mean	SD	ci95left	ci95right
Sens	0.805	0.030	0.746	0.864
Spec	0.821	0.028	0.758	0.879
Accuracy	0.815	0.023	0.767	0.853
Kappa	0.618	0.046	0.523	0.7

C.

	Mean	SD	ci95left	ci95right
Sens	0.871	0.015	0.843	0.900
Spec	0.822	0.025	0.770	0.863
Accuracy	0.848	0.015	0.816	0.875
Kappa	0.694	0.029	0.631	0.747

The results show good performance levels in distinguishing each single group Wheeze or Wheeze+ from each other, or from the other 2 groups together. In particular, Wheeze+ patients are detected in proportion of 87% by excluding controls from the analysis (see sensitivity in Table V Part C). In this case, we achieve an 85% accuracy. The detection of Wheeze+ patients among the whole cohort of subjects is 81% with an accuracy of 82%, and the detection of wheeze patients among the whole cohort is 80% with an accuracy of 82% (see Table V Parts B and A, resp.). Naturally, the highest performances were obtained in Control VS Other, whose results were presented in Table IV.

A question of clinical interest was if at least the same level of prediction performance and performance stability can be achieved by utilising only the objective features. To investigate this problem, we applied the same methodology (with its variants) to the Control VS Other classification problem (which led to the best prediction performance as shown in Table IV) on the reduced dataset with objective predictors. The resulting best model was obtained again with RF, that was tuned and evaluated in a nested CV with the ROC-CE method based on the closest top-left cutoff point. Comparing the performances of the best models built on the mixed and reduced datasets, which are illustrated in Tables IV and VI, respectively, we conclude that all the prediction performances decreased drastically, namely with 22% for AUC, 23% for sensitivity, 27% for specificity, 24% for accuracy, and 52% for kappa when using the objective features only. The model stability decreased also as suggested by the increase of the standard deviations of performances, or by the length of the 95% confidence intervals estimated in the 1000 MC experiments aggregated in Tables IV and VI. In conclusion, subjective variables are extremely useful in accurately classifying patients when added to the objective variables.

TABLE VI. BEST MODEL IDENTIFIED FOR CONTROL VS OTHER WITH RDUCED DATASET OF 47 OBJECTIVE FEATURES: OPTIMISED RF PERFORMANCES

	Mean	SD	ci95left	ci95right
AUC	0.758	0.029	0.701	0.812
Sens	0.695	0.049	0.603	0.793
Spec	0.644	0.054	0.528	0.750
Accuracy	0.683	0.041	0.599	0.763
Kappa	0.281	0.070	0.142	0.421

Let us note also that, just removing the subjective 'PSW' predictor alone from the mixed dataset, (feature which was scored with the highest Relief predictive power as illustrated in Fig.1), led, for the RF model built for the 3-class problem, to a decrease of accuracy with 11%, and a decrease of kappa statistic with 18%. On the other hand, results show that in the 2-class Control VS Other problem, by removing the 'PSW' predictor alone, the prediction performances of the best RF models decreased with 10% for AUC, 9% for sensitivity, 17% for specificity, 10% for accuracy, and 26% for kappa. The triple ROC chart in Fig. 2 illustrates a loss of prediction pattern, in the Control VS Other problem, of the 62 feature mixed dataset (black ROC) by removing first the 'PSW' predictor (blue ROC), and then removing all the other subjective features (red ROC).

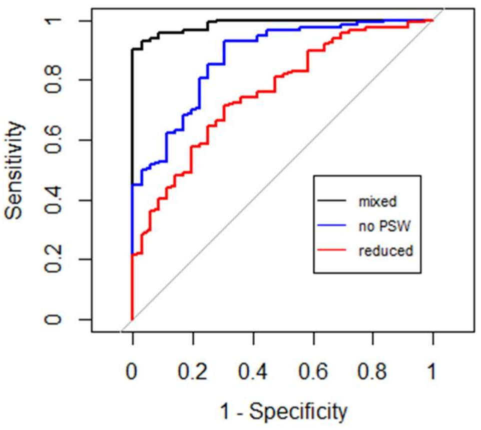

Fig. 2. Decrease in area under curve AUC for best RF models from using the mixed dataset (black ROC) to removing the 'PSW' predictor (blue ROC), then to removing all subjective variables and using the reduced dataset (red ROC).

IV. DISCUSSION AND CONCLUSION

In this study, we proposed that by utilising a variety of different statistical, feature analysis and predictive modelling techniques, it would be possible to identify groups within the PSW dataset by their pathology and clinical markers.

By first comparing the results from the univariate feature analysis and recursive feature elimination, it is clear that each provides a unique insight into which biomarkers are thought to be strongly associated with each pair of patient diagnostic outcomes. While MBL features highly in the univariate feature outcome tables, it is noticeably absent from the list of strongly associated variables that were produced through recursive feature elimination. Recurrent chest infection appears to be strongly associated with all pairs of 'Group' class outcomes in the recursive feature results but comparatively are not seen in the highest univariate feature chi-squared classification scores. The most stark difference is apparent when comparing each set of results between both feature selection analysis where there are no shared variables between the top five strongest feature lists.

When regressing the 23 PCs identified against the pairs of 'Group' classes, PC2 was found to be a protective factor and PC17 a risk factor for dependent variable (Wheeze+, Wheeze). These principal components load heavily on to variables pertaining to Height, Age (months), Weight and MBL inferring that these biomarkers could be useful in distinguishing between Wheeze and Wheeze+ patient classes.

A key finding from the prediction modelling performed on the mixed dataset and the reduced dataset containing only objective variables was the dramatic drop in prediction performance in the latter set. This suggests that the subjective variables are important in distinguishing ill patients from controls, particularly the 'PSW' variable which, when removed alone, resulted in a significant drop in prediction power.

To our knowledge, the study has used the largest number of clinical and pathology orientated variables using unsupervised and predictive modelling techniques in an attempt to distinguish if any markers are significant in determining pre-school wheeze class. By confirming physician-diagnosed groups or, alternatively, identifying new clusters of patients, care and therapy for these patients could in the future be tailored using significant biomarkers as a guide for treatment. A clear strength of this study is its breadth of analysis; in using multiple investigative methods to assess the dataset, various features associated with different wheeze classes can be determined.

An important limitation of this study was the level of missing data. After cleaning the data, patients were, on average, missing 25% of data. Although this is a common issue with clinical datasets and was addressed through multivariate imputation, the severity of missing data must be taken into consideration when analysing and interpreting results. As a result of discarding variables with less than 25% cases or in a format unsuitable for analysis, all data relating to follow-up appointments for patients was withheld from analysis (as only 18 of the 135 follow up variables fulfilled the data retention criteria, they were not included).

While PCA is a useful tool for reducing the number of variables while retaining a high level of overall data variability, this does result in a steep reduction of variability among certain variables which are not largely accounted for among the generated PCs. This could result in certain patterns and relationships between variables and, consequently, patient classes becoming lost or misinterpreted in results.

Our predictive modelling best results show we achieved 90%+ performance in AUC, sensitivity, specificity, and accuracy, and 80%+ in kappa statistic in distinguishing ill from healthy patients (Control VS Other classification). The predictive modelling was developed in a synergistic statistical - machine learning approach, and our methodologies incorporate a novel method that we proposed for predictive model post-processing and evaluation, called ROC Cross Evaluation. The latter works on all dataset sizes, but it is particularly useful when there is not enough data for model training, tuning, post-processing and/or evaluation/testing on independent data, as in the case of our relatively small dataset of 150 instances. The predictive models we developed were based on algorithms such as random forests, support vector machines, and back propagation for feed-forward neural networks. We performed comparisons and studied the stability of the performances of various predictive models built in different variants of our methodology, via a computationally intensive Monte Carlo simulation with 1000 experiments (each of which comprising all the phases of the methodology including model training, tuning and post-processing). This large volume of computation has been achieved by performing a parallel processing in R on a computer cluster formed of 11 servers based on Xeon processors and 832GB of fast RAM.

Ongoing work concerns, on one hand, the extension of our methodologies with a clustering approach. On the other hand, one of the aims of this work was to optimise prediction performance rather than develop supervised models with explanatory power, and as such, in this study we favoured algorithms with larger flexibility in adapting to the data, even though they were based on black box approaches. A natural extension of this work is to develop supervised models with explanatory power that would attempt to match, as much as possible, the prediction performance of the black box supervised models developed here. This is a natural extension as clinicians favour models with explanatory power in certain aspects of their research, especially when understanding the link between predictors and outcome is the primary goal.

References

[1] P. L. P. Brand, E. Baraldi, H. Bisgaard, A. L. Boner, J. A. Castro-Rodriguez, A. Custovic, J. de Blic, J. C. de Jongste, E. Eber, M. L. Everard, U. Frey, M. Gappa, L. Garcia-Marcos, J. Grigg, W. Lenney, P. Le Souef, S. McKenzie, P. J. F. M. Merkus, F. Midulla, J. Y. Paton, G. Piacentini, P. Pohunek, G. A. Rossi, P. Seddon, M. Silverman, P. D. Sly, S. Stick, A. Valiulis, W. M. C. van Aalderen, J. H. Wildhaber, G. Wennergren, N. Wilson, Z. Zivkovic, and A. Bush, “Definition, assessment and treatment of wheezing disorders in preschool children: an evidence-based approach,” Eur. Respir. J., vol. 32, no. 4, pp. 1096–1110, May 2008.

[2] F. D. Martinez, A. L. Wright, L. M. Taussig, C. J. Holberg, M. Halonen, and W. J. Morgan, “Asthma and Wheezing in the First Six Years of Life,” N. Engl. J. Med., vol. 332, no. 3, pp. 133–138, Jan. 1995.

[3] L. Tenero, G. Tezza, E. Cattazzo, and G. Piacentini, “Wheezing in preschool children,” Early Hum. Dev., vol. 89, pp. S13–S17, 2013.

[4] A. Beigelman and L. B. Bacharier, “Management of Preschool Children with Recurrent Wheezing: Lessons from the NHLBI’s Asthma Research Networks,” J. Allergy Clin. Immunol. Pract., vol. 4, no. 1, pp. 1–8, Jan. 2016.

[5] C. A. Stevens, D. Turner, C. E. Kuehni, J. M. Couriel, and M. Silverman, “The economic impact of preschool asthma and wheeze.,” Eur. Respir. J., vol. 21, no. 6, pp. 1000–6, Jun. 2003.

[6] G. Michel, M. Silverman, M.-P. F. Strippoli, M. Zwahlen, A. M. Brooke, J. Grigg, and C. E. Kuehni, “Parental understanding of wheeze and its impact on asthma prevalence estimates,” Eur. Respir. J., vol. 28, no. 6, pp. 1124–1130, Dec. 2006.

[7] H. E. Elphick, “Survey of respiratory sounds in infants,” Arch. Dis. Child., vol. 84, no. 1, pp. 35–39, Jan. 2001.

[8] J. Henderson, R. Granell, J. Heron, A. Sherriff, A. Simpson, A. Woodcock, D. P. Strachan, S. O. Shaheen, and J. A. C. Sterne, “Associations of wheezing phenotypes in the first 6 years of life with atopy, lung function and airway responsiveness in mid-childhood,” Thorax, vol. 63, no. 11, p. 974 LP – 980, Aug. 2008.

[9] O. E. Savenije, R. Granell, D. Caudri, G. H. Koppelman, H. A. Smit, A. Wijga, J. C. de Jongste, B. Brunekreef, J. A. Sterne, D. S. Postma, J. Henderson, and M. Kerkhof, “Comparison of childhood wheezing phenotypes in 2 birth cohorts: ALSPAC and PIAMA,” J. Allergy Clin. Immunol., vol. 127, no. 6, pp. 1505–1512.e14, Jun. 2011.

[10] D. C. M. Belgrave, A. Simpson, A. Semic-Jusufagic, C. S. Murray, I. Buchan, A. Pickles, and A. Custovic, “Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing,” J. Allergy Clin. Immunol., vol. 132, no. 3, pp. 575–583.e12, Sep. 2013.

[11] A. Bush, J. Grigg, and S. Saglani, “Managing wheeze in preschool children,” BMJ, vol. 348, no. feb04 16, pp. g15–g15, Feb. 2014.

[12] S. Illi, E. von Mutius, S. Lau, B. Niggemann, C. Grüber, and U. Wahn, “Perennial allergen sensitisation early in life and chronic asthma in children: a birth cohort study,” Lancet, vol. 368, no. 9537, pp. 763–770, Aug. 2006.

[13] M. L. Everard, “Pre-school wheeze: How do we treat and how do we monitor treatment?,” Paediatr. Respir. Rev., vol. 7, pp. S112–S114, Jan. 2006.

[14] J. A. Castro-Rodriguez and G. J. Rodrigo, “Efficacy of Inhaled Corticosteroids in Infants and Preschoolers With Recurrent Wheezing and Asthma: A Systematic Review With Meta-analysis,” Pediatrics, vol. 123, no. 3, pp. e519–e525, Mar. 2009.

[15] M. Weinberger and M. Abu-Hasan, “Pseudo-asthma: When Cough, Wheezing, and Dyspnea Are Not Asthma,” Pediatrics, vol. 120, no. 4, pp. 855–864, Oct. 2007.

[16] “National Asthma Education and Prevention Program. Expert Panel Report III: Guidelines for the Diagnosis and Management of Asthma.,” 2007.

[17] S. E. Pedersen, S. S. Hurd, R. F. Lemanske, A. Becker, H. J. Zar, P. D. Sly, M. Soto-Quiroz, G. Wong, and E. D. Bateman, “Global strategy for the diagnosis and management of asthma in children 5 years and younger,” Pediatr. Pulmonol., vol. 46, no. 1, pp. 1–17, Jan. 2011.

[18] L. B. Bacharier, A. Boner, K.-H. Carlsen, P. A. Eigenmann, T. Frischer, M. Götz, P. J. Helms, J. Hunt, A. Liu, N. Papadopoulos, T. Platts-Mills, P. Pohunek, F. E. R. Simons, E. Valovirta, U. Wahn, and J. Wildhaber, “Diagnosis and treatment of asthma in childhood: a PRACTALL consensus report,” Allergy, vol. 63, no. 1, pp. 5–34, Dec. 2007.

[19] F. Thomson, I. B. Masters, and A. B. Chang, “Persistent cough in children and the overuse of medications.,” J. Paediatr. Child Health, vol. 38, no. 6, pp. 578–81, Dec. 2002.

[20] W. E. Molis, S. Bagniewski, A. L. Weaver, R. M. Jacobson, and Y. J. Juhn, “Timeliness of diagnosis of asthma in children and its predictors,” Allergy, vol. 63, no. 11, pp. 1529–1535, Nov. 2008.

[21] S. Saglani and A. Bush, “Asthma in preschool children: the next challenge,” Curr. Opin. Allergy Clin. Immunol., vol. 9, no. 2, pp. 141–145, Apr. 2009.

[22] S. Long and J. Freese, Regression Models for Categorical Dependent Variables Using Stata, Second. Stata Press, 2006.

[23] StataCorp, “Stata Statistical Software: Release 14.” College Station, TX: StataCorp LP., 2015.

[24] P. Royston and I. White, “Multiple Imputation by Chained Equations (MICE): Implementation in Stata,” J. Stat. Softw., vol. 45, no. 4, 2011.

[25] M. Chavent and V. Kuentz, “Multivariate Analysis of Mixed Data.” CRAN Repository, 2014.

[26] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, “SMOTE: Synthetic Minority Over–Sampling Technique”, Journal of Artificial Intelligence Research, 16(1), pp. 321–357, 2002.

[27] P. Good, "Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses", Springer, 2000.

[28] I. Kononenko, “Estimating Attributes: Analysis and Extensions of Relief”, Machine Learning: ECML–94, vol 784, pp. 171–182, Springer, 1994.

[29] M. Kuhn, K. Johnson, "Applied Predictive Modeling", Springer, 2013.

[30] M. Kuhn, “The caret Package Homepage”, URL http://caret.r-forge.r-project.org/. 2016.

[31] W.J.Youden, "Index for rating diagnostic tests", Cancer, 3: 32–35, 1950.