2014•2015
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Diagnosing pneumonia, influenza and obstructive pulmonary diseases in adult patients presenting to primary care with acute cough: a multinomial logistic regression analysis

Supervisor :
Prof. dr. Marc AERTS

Supervisor :
Prof.dr. SAMUEL COENEN

Pavlina Mesiri

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2014•2015
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Diagnosing pneumonia, influenza and obstructive pulmonary diseases in adult patients presenting to primary care with acute cough: a multinomial logistic regression analysis

Supervisor :
Prof. dr. Marc AERTS

Supervisor :
Prof.dr. SAMUEL COENEN

Pavlina Mesiri
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

universiteit hasselt | Maastricht University

Abstract

Background:
Community-acquired lower respiratory tract infection (LRTI) is a broad term which describes airways/pulmonary tissue inflammations due to viral and/or bacterial infection, below the level of the larynx. Known infections that can affect the lower respiratory tract, among others, are: pneumonia, exacerbation of chronic obstructive airway disease (e.g. asthma or COPD) and influenza (can affect both the upper and lower respiratory tracts). The objective of this report is to assess the diagnostic value of signs and symptoms and added value of biomarkers, e.g. C-Reactive Protein, of adult patients from 16 primary care networks from 12 European countries who presented to primary care with acute cough for GRACE studies in the diagnosis of pneumonia, influenza and obstructive pulmonary diseases treating the diagnoses in parallel rather than in series.
Methods:
Missing information on patients diagnosis' symptoms and signs were imputed using Multiple Imputation Chained Equations. Candidate clinical predictors able to discriminate patients' condition (i.e. Pneumonia, Influenza, Asthma, COPD or "Other") were chosen through Random Forest approach. The contribution to the log odds of a category versus the baseline was then estimated for the selected clinical predictors with a Multinomial Logistic Regression(MNL) model. The diagnostic accuracy and prediction power of the model were assessed for each pair of categories by fitting a Logistic Regression(LR) and estimating the area under each curve (AUC) for each models' Receiver Operating Characteristic (ROC) Curve.
Results:
The pairs with the highest AUC were : Influenza when reference is Pneumonia, Asthma when reference is Pneumonia, COPD when reference is Pneumonia and Influenza when reference is Asthma. Influenza when reference is Pneumonia model revealed an AUC value equal to 0.92 i.e. a randomly selected patient from the Influenza group has a predicted probability for influenza larger than that for a randomly chosen patient from the Pneumonia group 92 percent of the time. Retrofitting the same logic, 0.86, 0.79 and 0.85 were the identified AUC values for Asthma when reference is Pneumonia, COPD when reference is Pneumonia and Influenza when reference is Asthma respectively. The added diagnostic value of CRP was quantified by fitting the logistic regression models without CRP and calculating the AUC value of their ROC curve. All models where CRP was significant (i.e. Influenza when reference is Pneumonia , Asthma when reference is Pneumonia and COPD when reference is Pneumonia) had a lower AUC value after we omitted CRP. The difference proved insignificant.
Conclusions:
Based only on symptoms and signs taken into consideration in this report, we can conclude that differentiating between: pneumonia, influenza when reference is other diagnoses as well as between each pair diagnosis (Influenza, Asthma or COPD when reference is Pneumonia and Influenza when reference is Asthma) at first day of patient consultation, can be adequately accurate.

Acknowledgement

I would like to express my great appreciation to my Supervisors Prof. Marc Aerts , Prof. Samuel Coenen whose guidance, ideas and suggestions have been determinative through the course of this Thesis. I would also like to thank Robin Bruyndonckx for her help and encouragement and GRACE project group for allowing us to work with the data they have collected and processed.

I am especially grateful to all staff members of I-BioStat for the enlightening courses, invaluable experience through lectures and team projects and for the overall atmosphere and spirit which have made my stay at Belgium most pleasant.

Finally, I would like to thank my father, mother and brother for their amazing support and encouragement during the years I have been studying in Belgium and my dear friends George and Dimitra.

Contents

1 Introduction

Acute cough is one of the most frequent complaints of patients who visit primary health care centers and one of the early symptoms of community-acquired lower respiratory tract infections (LRTI). Pneumonia, Influenza and Pulmonary Obstructive Diseases are three examples of LRT infections.

The proper antimicrobial treatment or the necessity for an antimicrobial treatment itself is a matter of contention when it comes to LRTI's. For the former it all comes to the fact that not all type of LRTIs are bacterial. For the latter, severity of the symptoms and the need for a prompt treatment invariably account for overly broad spectrum antibiotic prescription which accelerates antimicrobial resistance (AR) emphasizing the importance of the proper use of antimicrobials.

The matter of antimicrobial resistance (AR) has been studied extensively within GRACE (Genomics to combat Resistance against Antibiotics in Community-acquired LRTI in Europe; (www.grace-lrti.org), a Network of Excellence funded by the European Commission (2006-2011). Since 2006, GRACEs' role has been definitive in the battle against the threat of resistance against antibiotics. GRACE has integrated centers of excellence across 12 European Countries and its research program was composed of 4 platforms (GRACE-COMIT, GRACE-TECH, GRACE-PAT, GRACE-EDUT) and 12 workpackages (WP1 - WP12). GRACE-PAT focused on patients and through its observational Studies (e.g. GRACE-01) described the presentation, diagnosis, investigation, management and outcomes for people with cough / chest infection [5]. As regards acute cough, the coordinated action has been focused on incorporating all patient information which can add valuable knowledge to diagnosing its causal reason and therefore its proper treatment. A related paper by Prof. Samuel Coenen et. al., stresses the GPs diagnostic uncertainty in patients with acute cough and its significant effect on the antibiotic prescribing decision [6].

The reference standard for diagnosing pneumonia is the chest-x-ray. Additionally for patients with pneumonia symptoms present for more than 24 hours a test for the serum level of biomarker C-reactive protein (CRP) can be done. CRP level$<$ 20 mg/L, makes the presence of pneumonia highly unlikely whereas a level of CRP $>$ 100 mg/L is an indicator of pneumonia being present. Influenza can be diagnosed via a microbiological blood test analysis. Chronic obstructive pulmonary disease (COPD) is a type of Obstructive Pulmonary diseases which is characterized by airflow obstruction [8]. Since COPD has no direct cure yet, bronchodilator or bronchodilator and an inhaled steroid may be prescribed. People who have COPD are at higher risk for pneumonia or influenza than people who don't have COPD. Obstructive Pulmonary diseases can be diagnosed via lung function tests to identify asthma and chronic obstructive pulmonary disease. Because of the high interconnection between pneumonia, influenza, asthma and COPD symptoms, the discrimination prior to laboratory or chest x ray results, can be difficult.

The etiology, diagnosis prognosis and treatment of lower respiratory tract infections (LRTI) have been some of the many objectives of Grace Research Network. Traditionally, diagnoses of pneumonia, influenza and obstructive pulmonary diseases are analyzed separately, whereas in reality they are analyzed by clinicians in parallel rather than in series. Through this report, we present our work i.e. to assess the diagnostic value of signs and symptoms and added value of biomarkers, e.g. C-Reactive Protein (CRP), of adult patients from 16 primary care networks from 12 European countries who presented to primary care with acute cough for GRACE studies in the diagnosis of pneumonia, influenza and obstructive pulmonary diseases treating the diagnoses in parallel rather than in series.

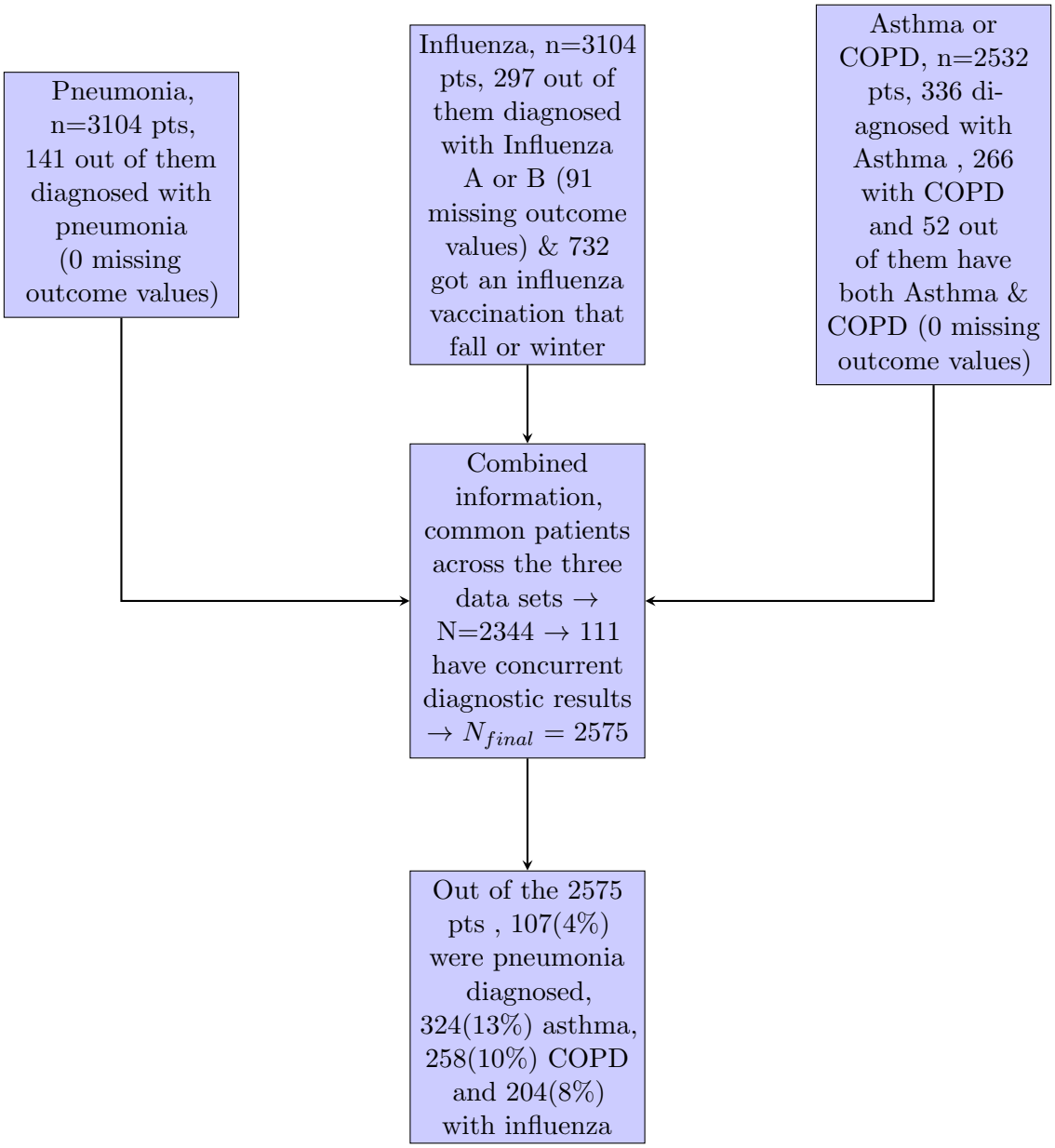
2 Data Description

The data belong to GRACE group project (Genomics to combat Resistance against Antibiotics Community-acquired LRTI in Europe, www.grace-lrti.org), a Network of Excellence funded by the European Commission. Observational data from primary care centers (Cardiff, Southampton, Utrecht, Barcelona, Mataro, Rotenberg, Antwerpen, Lodz, Milano, Jonkoping, Nice, Jesenice, Bialystok, Sczecin, Gent and Bratislava) in 12 countries (UK, Netherlands, Spain, Germany, Slovenia, Belgium, Poland, Italy, France, Sweden, Finland and Slovakia) were collected between 2007 and 2010.

Patients presenting to primary care with acute cough had their history, symptoms, clinical findings and CRP measurement taken the first day of consultation. Pneumonia was defined by radiologists' judgment of the Chest radiographs which were performed within 7 days after consultation [16]. Airway obstruction was diagnosed by spirometry results analysis 28-35 days after inclusion. Asthma was diagnosed if recurrent (more than 1 episode last year) complaints of wheezing, cough or chest tightness were present, in combination with an increase in FEV1 of $\geq$ 12% or more than 200 milliliters (ml) after bronchodilation [14]. According to the European Respiratory Society (ERS), COPD was confirmed when a "fixed" FEV1/FVC ratio was below 0.7 according to GOLD (obstruction GOLD), and a FEV1/FVC ratio was below the lower limit of normal (LLN) [14]. For the diagnosis of influenza, nasopharyngeal swabs were taken within 24 hours after consultation. The swabs were analyzed for Influenza A and B by reverse transcriptase polymerase chain reaction (RT-PCR). Influenza was considered present if the PCR was positive for influenza A or B [18].

Eligible patients were all at least 18 years old and all presented to primary care with acute cough or LRTI-like symptoms consulting for the first time for that illness episode. For Asthma & COPD, diagnosis was based on results of the lung function tests. For influenza only patients with symptom onset $\leq$ 7 days prior to consulting were considered in the analysis, because studies show a positive Influenza PCR up to 7 days after onset of symptoms [18].

Figure 1: Flow chart of the study and participants



- patients with missing data on any diagnostic outcome were excluded (figure 1). From the combined final data set 43 of the 91 initial influenza patients had missing outcome values and thus were omitted.

The number of patients' combinations with concurrent diagnoses was the following:

- 15 were COPD & influenza diagnosed
- 4 patients were pneumonia & influenza diagnosed
- 14 patients were pneumonia & COPD diagnosed
- 19 patients were influenza & asthma diagnosed
- 43 patients were asthma & COPD diagnosed
- 2 patients were influenza & asthma & COPD diagnosed
- 7 patients were asthma & pneumonia diagnosed
- 2 patients were influenza & asthma & pneumonia diagnosed
- 5 patients with were asthma & pneumonia & COPD diagnosed

This equals to 111 patients and 231 diagnoses. For the variables initially included, clinical indicators from already published diagnostic models for each diagnosis separately, were considered [14, 16, 18]. Van Vugt et al. studied the diagnostic accuracy of symptoms and signs as well the accuracy of selected inflammatory markers for predicting pneumonia on the scope of the high importance of an accurate diagnosis of pneumonia in primary care, focused on the prevalence of airway obstruction and bronchodilator responsiveness as well as at the high risk of undiagnosed asthma and COPD in adults with acute cough and investigated the validity of a clinical model to predict influenza in patients presenting with symptoms of lower respiratory tract infection in primary care with main goal to access the validity of an existing diagnostic model symptoms to accurately differentiate those with influenza on clinical grounds from the other LRTI's in primary care.

3 Methods

3.1 Multiple Imputation

Multiple imputation (Rubin 1987) is a useful mechanism to handle missing data that occur in more than one variable. Each missing value is replaced by a set of m plausible values(m is the number of imputations, m$\geq$ 1). The result of each replacement represents a complete data set, thereby generating m complete data sets. Multivariate imputation by chained equation is one of the two general approaches of multiple imputation. Multiple imputation by chained equations, selects imputed values by iterating over conditional distribution, $P(Y_1|Y_{-1}, \theta_1), \cdots P(Y_p|Y_{-p}, \theta_p)$, of each partially observed variable given the other variables in the data set [26]. In the presence of continuous , binary or categorical ($j > 2$) variables, this is a very useful approach since an appropriate regression model can be selected for each variable (e.g. linear regression for continuous variables, logistic regression for binary variables).

Important Notation:

- Let Y_j be the jth partially observed variable, $j = 1, 2, \cdots, p$.
- Let $Y_j^{obs} = (Y_1^{obs}, Y_2^{obs}, \cdots, Y_p^{obs}), Y_j^{mis} = (Y_1^{mis}, Y_2^{mis}, \cdots, Y_p^{miss})$ the obsereved and missing data.
- Let R be the vector of observation indicators being equal zero or one depending on whether the variable Y is missing or observed.
- $\hat{Q_1}, \hat{Q_2}, \cdots \hat{Q_m}$ stand for the coefficient(s) of interest as estimated from each imputed data set $(Y_1^1, Y_2^2, \cdots, Y_p^m)$.

Missing data mechanisms, i.e. Missing Completely at Random (MCAR), Missing not at Random (MNAR) and Missing at Random (MAR), reflect one's belief about whether the univariate or marginal distributions of the complete case and observed data are expected to be the same or not. Therefore, MCAR would e.g. assume that the distribution of a variable Y_j is the same regardless whether $Y_{-j} = (Y_1, \cdots, Y_{j-1}, Y_{j+1}. \cdots, Y_p)$, is observed or missing. Therefore, because MCAR is a very strong assumption to rely and because patients' symptoms and signs are hardly unaffected of each other, throughout this report, we assume that the mechanism behind missing data is MAR i.e. $P(R|Y, Y^{obs}, X) = P(R|Y, Y^{miss}, X)$ where Y the observed outcome and X the fully observed variables ($X = X_1, \cdots X_p$).

The complete m data sets can then be combined for inference(pooling). Rubin (1987, pp. 76-77), describes the method of pooling estimates from a data analysis performed m times. In our case, m=5 imputations. The overall estimate is the average of the individual estimates of each parameter but the overall standard error requires a bit more calculating. We first calculate the within-imputation variance $\bar{U} = \frac{1}{m}\sum_{j=1}^{m} U_j$ and between-imputation variance

$\bar{B} = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_j - \bar{Q})^2$. Then the total variance is equal to $T = \bar{U} + (1 + \frac{1}{m})B$ the square root of which is the overall standard error [25].

3.2 Variable Selection

Random forests (Breiman, 2001) is a powerful method when it comes to selection of important predictors. Initially, bootstrap samples of size N(b = 1 to B) are drawn from the original sample. A random-forest tree T_b is grown for b = 1 to B and at each terminal node of the tree the best split from a subset of predictor variables (m) is used to split the node [24]. The process is repeated for each T_b until minimum node size nmin is reached. Finally for classification purposes, the status of the response variable at a new point x is predicted as the majority vote of the predictions for all trees.

The subset of predictor variables (m) is randomly selected, constant during the whole process and its typical value for classification is $\sqrt{p}$, where p stands for the number of total variables. One can tune the subset m to find the optimal value which minimizes the out-of-Bag error estimate (OOB). Now suppose that we collect all subsets that do not include the particular (X_j, Y_j) in the construction of the k-th tree(oob samples). Let k be the class appearing most frequently every time (X_j, Y_j) is left out. The proportion(i.e. over n) of times k does not equal the true class is the OOB error estimate. The ideal number of trees to grow can be found in terms of oob error stabilization [24].

Mean decrease in accuracy can be used as selection index for measuring importance of a variable j . Approximately one third of each bootstrap sample cases are left out each time a tree is constructed (OOB sample). This is a way of "internal" cross validation. Mean decrease in accuracy is a measure of decrease in prediction accuracy between trees at the bottom of which OOB samples are placed and OOB samples with randomly permuted j_{th} variable values. So the more important the variable the larger the accuracy decrease it causes. Consequently predictors with large mean decrease in accuracy(averaged over all trees) are the most important in terms of correctly classifying the data.

3.3 Multinomial Logistic Regression analysis

Multinomial Logistic Regression generalizes logistic regression by simultaneously describing the log odds for all pairs $\binom{J}{2}$ of categories [10]. If each observation is independent (see Assumptions below), $y_{ij} = 1$ if population i has outcome in category j and $y_{ij} = 0$ otherwise and $\pi_{ij} = P(y_{ij} = 1)$ beeing the probability of outcome j, then the probability mass function follows a multinomial $(n, \pi i)$ distribution characterized by the sample size n and the probabilities π_i [28]:

$$p(n_1, n_2, \cdots, n_c) = \frac{n!}{n_1!, n_2!, \cdots, n_c!} {\pi_1}^{n_1} {\pi_2}^{n_2} \cdots {\pi_c}^{n_c} \qquad (1)$$

$$log(\frac{\pi_{ij}}{\pi_{iJ}}) = log(\frac{\pi_{ij}}{1 - \sum_{j=1}^{J-1} \pi_{iJ}}) = \sum_{k=0}^{K} x_{ik}\beta_{kj} \quad \binom{i=1,\cdots,N}{j=1,2,\cdots,J-1} \qquad (2)$$

$$\leftrightarrow$$

$$\pi_{ij} = \frac{e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}{1 + \sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}} \qquad (j < J) \quad (3)$$

$$\pi_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}} \qquad (4)$$

In our case, N=2575 and j=1,2,3,4 for pneumonia, asthma, copd or influenza, respectively and J=5 representing the reference category i.e the other diagnoses.

A model is then used to predict the probabilities of the different possible outcomes of the categorical response, given a set of independent variables (qualitative and/or quantitative) equations. The parameter estimation is done through Newton Raphson iterative method for nonlinear systems. The log likelihood function for the multinomial logistic regression model is :

$$l(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{J-1} (y_{ij} \sum_{k=0}^{K} x_{ik}\beta_{kj}) - n_i log(1 + \sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}) \qquad (5)$$

Where : i represents each patient (often called population), n is a column vector with n_i the number of observations for population i ($\sum_{i=1}^{N} n_i = N$), y is a column vector with y_i the observed numbers of successes for each population i, J is the number of discrete categories of the dependent variable (J $\geq$ 2), X is the design matrix with N rows and K+1 columns where K is the number of independent variables specified and the parameter vector β the K+1 length column parameter vector. Differentiating two times equation (5) with respect to each β_k we obtain the first and second order derivatives of the log likelihood and if the matrix of second partial derivatives is negative definite, and solution is the global maximum rather than a local maximum, we obtain the parameter estimates and their variance-covariance matrix iteratively through Newton-Rapson

method [28].

Assumptions of Multinomial Logistic Regression:

- Independence from irrelevant alternatives property (IIA) i.e. the odds of choosing a category over the reference do not depend on the other alternatives [27]. The assumption is tested when Multinomial models represent alternative specific variables i.e. variables which depend on the category chosen (Discete choice models). All variables in our case, are individual specific.
- Log of the odds ratio and the measurement variables have a linear relationship (Eq. 2).
- Multinomial logistic regression does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables.

Sample Size requirements :
The minimum number of outcomes per independent variable must be 10 [11].

3.4 Cluster-Robust Inference

When observations are grouped into clusters, model errors for individuals in the same cluster may be correlated. Ones' choices to account for clustering are : Introducing random effects (random effects models) or ignoring clustering but ensure that standard errors are based on so-called "sandwich" variance estimator(marginal models). For this thesis we proceed with the latter choice i.e correcting the standard errors to account for heterogeneity between countries and similarity within countries(clusters). This method is also called quasi likelihood approach (QL) for the univariate case (single response) or Generalized Estimating Equations (GEE) for a multivariate response (repeated measurements). Agresti (2002), mentions that the "sandwich" variance adjustment requires a rather large number of clusters n for the asymptotic covariance matrix of the QL estimator $\hat{\beta}$ to be unbiased. A way to address the issue as it is proposed by literature is to use small sample modifications of the cluster-robust variance matrix estimate.

In the presence of two way clustering or multi-way clustering one simply clusters at the highest level of aggregation [23].

Unlike maximum likelihood method(ML), quasi likelihood approach does not inflict a certain distribution for the response (Y_i) but instead it assumes a mean-variance relationship. The mean variance relationship is connected via a chosen variance function (υ), $Var(Y_i) = \upsilon\mu_i$. As showed by Agresti (2002), assuming that the true variance is the same as the one chosen (QL same as ML estimates) then the covariance matrix (McCullagh, 1983.) can be approximated by :

$$V = [\sum_i (\frac{\partial \mu_i}{\partial \beta})'[u(\mu_i)]^{-1}](\frac{\partial \mu_i}{\partial \beta})]^{-1} \qquad (6)$$

However, if the true variance is different from the initial "guess" that is if $Var(Y_i) \neq \upsilon\mu_i$ then the actual asymptotic covariance matrix of the QL estimator is the so called sandwich estimator:

$$V[\sum_i (\frac{\partial \mu_i}{\partial \beta})'[u(\mu_i)]^{-1}]var(Y_i)[u(\mu_i)]^{-1}(\frac{\partial \mu_i}{\partial \beta})]V \qquad (7)$$

3.5 Diagnostic Accuracy

Let x describe the explanatory variables. The effects of x on the J-1 logits, for one covariate x, are described simultaneously from the equation :

$$log\frac{\pi_j(x)}{\pi_5(x)} = \alpha_j + \beta'_j x \qquad j = 1,2,3,4 \qquad (8)$$

$$\leftrightarrow$$

$$\pi_j(x) = \frac{\exp(\alpha_j + \beta'_j x)}{1 + \sum_{h=1}^{4} \exp(\alpha_h + \beta'_h x)} \qquad (9)$$

where J is the baseline category (J=5) $\rightarrow \alpha_J = \beta_J = 0$.

Multinomial logistic regression fits one model simultaneously accounting for the different levels in the outcome variable such that the total probability of all five outcomes adds to one. From Equation (8), suppose we would like to examine category j=1 when the reference is j=2, i.e. suppose a different pair of categories inside the multinomial logistic regression model, is of interest. Then the log of odds for category 1 (j=1) when reference is category 2 (J=2) for one covariate would be:

$$log\frac{\pi_1}{\pi_5} = \alpha_1 + \beta_1 x \qquad (10)$$

and

$$log\frac{\pi_2}{\pi_5} = \alpha_2 + \beta_2 x \qquad (11)$$

so the difference :

$$log\frac{\pi_1}{\pi_2} = (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x \qquad (12)$$

For this thesis, we based the assessment of diagnostic accuracy of a multinomial model on the diagnostic accuracy of each binary logit model separately for the J-1 pairings of responses. From equation (8) we first began with category j=1 alone, using only observations in category 1 or 5(J) of the response variable to obtain the α_1, β_1. Then categories 2 and 5(J) to obtain estimates of α_2, β_2, categories 3 and 5 for α_3, β_3 and categories 4 and 5 for the α_4, β_4. We expect the estimates to differ from the ones obtained from a simultaneous multinomial logistic regression fit (Eq. 10-12). Acknowledging that differences may arise, we will compare the separate estimates (log of odds) with the simultaneous log off odds from the multinomial logistic regression model (j=1,2,3,4).

Receiver Operating Characteristic Curve (ROC) is a technique used to summarize the predictive power of a classifier in terms of sensitivity (the ability of the model to predict an event correctly) versus 1-specificity for the possible cut-off classification probability values π_0 [10]. Predicted value is classified as $\hat{y} = 1$ if the probability of belonging to i-th category ($\hat{\pi}$) is larger than the specified cut off probability point and $\hat{y} = 0$ otherwise. So, sensitivity is the probability that we predict the value is one if the true outcome value is one $P(\hat{y}_i = 1|y = 1)$ and specificity is the probability that we predict the value is zero if the true outcome value is zero $P(\hat{y}_i = 0|y = 0)$. From Equations (10-12), a ROC curve for j=1 when reference category is J=5, coming from a multinomial logistic regression when reference is category 5(J=5), would imply :

$$log\frac{\pi_1/(\pi_1+\pi_5)}{\pi_5/(\pi_1+\pi_5)} \Rightarrow \frac{\pi_1}{\pi_1+\pi_5} = \frac{e^{\alpha_1+\beta_1 x}}{1+e^{\alpha_1+\beta_1 x}} \qquad (13)$$

and for

$$\hat{y} = 1 \Leftrightarrow \hat{\alpha_1} + \hat{\beta_1} x > \kappa \qquad (14)$$

Similarly, a ROC curve for j=1 when reference category is J=2, would imply :

$$\hat{y} = 1 \Rightarrow (\hat{\alpha_1} - \hat{\alpha_2}) + (\hat{\beta_1} - \hat{\beta_2})x > \kappa \qquad (15)$$

where κ the cut off point.

The area under the ROC curve (AUC) can be expressed as $\int_0^1 f_1(t_1)dt_1$ where t_1, t_2 are the correct classification probabilities which can be described by a functional relationship, $t_2 = f_1(t_1)$ [29].

In our analysis, the AUC for the models was recalculated for the data set where the patients' combinations with concurrent diagnoses ($n_{concurrent} = 231$) were removed ($N_{final2} = 2344$).

As mentioned above, ROC curve is actually a plot of sensitivity(TPR) versus 1-Specificity (FPR) at different thresholds(cut off points). Of course, ones' interest is finding the best trade off between specificity and sensitivity. In our analysis, in order to take into account that we quantify the diagnostic accuracy of a multinomial response model from binary response models separately (J=5 categories) we specify the cut off point to be 0.5 with the following logic. In the presence of five categories the by chance selecting one out of five is 0.2. Taking into consideration the equivalence between a binary logit and a multinomial logit with information from only two categories (Eq. 8-15) we begin with a cut off point of 0.5 ($\pi_j(x)/\pi_j(x) + \pi_J(x) = 0.2/0.2 + 0.2$). Additionally, we present a 3-fold cross validation estimate of the ROC curve. Each cross validation fold is randomly removed from the test data and the remaining data set (training) is used to plot the TPR and FPR values, together with the average value across the 3-folds and box plots.

Hypervolume under Manifold (HUM) is an extension of the ROC curve for multi class categories. HUM estimator counts the proportion of subsets of M individuals in which each of the M persons is correctly classified [29].

$$\widehat{HUM} = \frac{1}{\prod_{h=1}^{M} n_h} \sum_{k1=1}^{n1} \sum_{k2=1}^{n2} \cdots \sum_{kM=1}^{nM} CR(\widehat{p1}k1, \cdots, \widehat{pM}kM) \qquad (17)$$

where $\widehat{p}_{ij}$ are the estimated probabilities for the j_{th} individual of class i ($i = 1, \cdots, M$), obtained from a multinomial logistic regression model. For more technical details and theory, we refer to paper by Li, J., Fine, JP. [29].

4 Results

4.1 Multiple Imputations by Chained Equations

Out of the 2575(N_{final}) individuals we have 1125 (44%) complete cases i.e. patients whose variables were all observed and recorded. A bit more than half of the patients (56%) have missing values. The total number of missing values is equal to 2896(4%). Because of the objective of the analysis only variables with missing values were imputed and participants with missing outcome values were excluded.
Variables with the highest number of missing data were: CRP measurement, Days feeling unwell, Recurrent cough(more than 3 cough episodes during last year) and Asthma present in family. We look at the marginal plot of each pair of those variables. The observed data are colored blue whereas the missing, red. There are 904 records in which CRP is missing (red dots left margin), 489 were Days feeling unwell are missing (red dots bottom margin) and 184 were both are missing (figure 2). For the other pairs of variables, the plots can be found at Appendix, Figures 12 - 16.

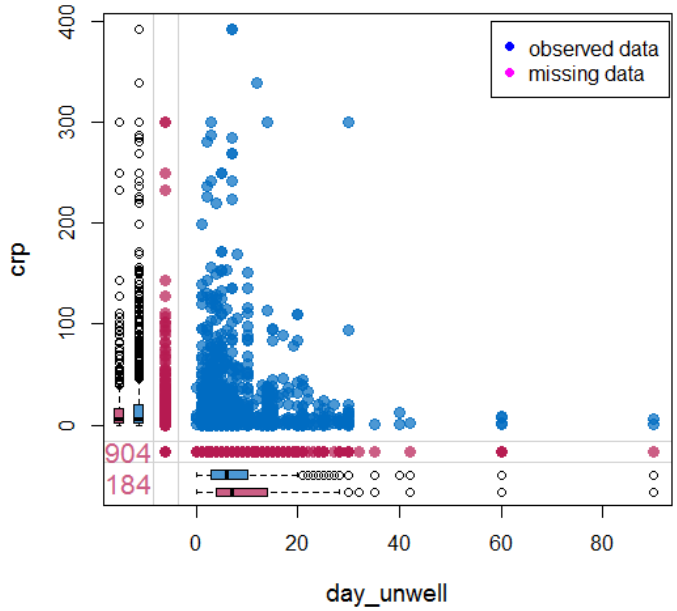


Figure 2: Marginal plot of CRP versus is Days feeling unwell

4.2 Exploratory Data Analysis

We look at the association between diagnostic variables(as identified by literature) and outcome : i.e. Pneumonia, Influenza, Asthma, COPD and other outcomes, separately. The values shown at Tables 1, 2 and 3 are the average results of five times imputed variables for each pair of diagnosis separately i.e. each response category versus the other diagnosis category (reference category). The total number of patients is n=2575. We have considered 32 clinical indicators for all diagnoses combined. As mentioned in Section 4.1., 4% of the data are missing. Because of the objective of the analysis only variables with missing values were imputed and participants with missing outcome values were excluded. Nevertheless, as mentioned in the data description section only influenza diagnosis had missing outcome (n=91) which were therefore excluded. 347 (13%) out of the total number of patients were under antibiotic treatment during the previous six months. Patients at high risk were patients with other lung disease (e.g. fibrosis or bronchiectatis, etc), patients with heart disease (e.g. valvularlesions, cardiomypoathy etc) and diabetic patients. Among the 106 diabetic patients, 8 (7 %), 23 (22 %), 21 (20%) and 5 (5%) were pneumonia, asthma, COPD and influenza diagnosed, respectively. Among patients who suffered from heart diseases (n=89), 8 (7 %), 4 (1.2 %), 9 (3.5 %) and 4 (2 %). Among patients with other lung diseases (n=48), 4 (8.3 %), 7 (15 %) , 9 (19 %) and 1 (2.1 %). 558 patients (22%) got an influenza vaccination during that fall or winter. Twenty nine out of them were influenza diagnosed.

Table 1: Association between diagnostic variables and Pneumonia (n = 1789)

Diagnostic Variable	Value(%) or mean(SD)	Pneumonia present (n = 107)	OR (95% CI)
Age, mean(SD)	50(17)	107	1.011(1.001; 1.017)
Men	1030(58%)	47	1.219(1.021;1.453)
Current Smoker	445 (25%)	4	1.586(1.275;1.97)
Days ill prior consult, mean (SD)	10 (8)	5	0.968(0.954;0.98)
Severe Cough	582(33%)	42	1.011(0.444;2.912)
Phlegm	1420(79%)	91	1.513 (1.193;1.942)
Severe Breathlessness	92 (5%)	13	3.812 (2.453;6.052)
Runny nose	1276 (71%)	55	0.399(0.335;0.476)
Fever	596(33%)	63	3.087(2.585;3.692)
Severe Chest Pain	80 (4.5%)	9	2.246(1.492;3.36)
Diarrhea	125(7%)	10	1.405(1.026;1.884)
Any heart comorbidity (valvularlesions, cardiomypoathy etc)	72(4%)	8	2.043(1.433;2.839)
Diabetes	111(6.2 %)	8	1.239(0.874;1.71)
General Toxicity	482 (27%)	34	1.281(1.06;1.544)
Diminished vesicular breathing	199 (11%)	24	2.407(1.931;2.98)
Tachycardia ($> 100 beats/min$)	853 (48%)	54	1.375(1.154;1.64)
Tachypnoea ($> 24 breaths/min$)	40 (2.2 %)	6	2.862(1.888; 4.198)
Systolic blood pressure	128(18)	107	0.986(0.981;0.991)
Diastolic blood pressure	79 (11)	107	0.982(0.974;0.99)
Oral temperature (> 37.8)	66 (3.7%)	14	4.682(3.52;6.16)
CRP(mg/L), mean(SD)	24(46)	-	1.012(1.011;1.013)
> 20	1363(76%)	48	2.84(2.173;3.675)
> 30	171 (10%)	16	3.301(2.459;4.374)
> 50	120 (6.7%)	13	7.607(6.076;9.5)
> 100	134(7.5 %)	29	Ref.cat

Table 2: Association between diagnostic variables and Influenza(n=1886)

Diagnostic Variable	Value(%) or mean(SD)	Influemza present (n=204)	OR (95% CI)
Age mean(SD)	49(16)	204	0.978(0.975;0.983)
Men	733 (39%)	75	0.905(0.79;1.035)
Current Smoker	492 (26%)	50	1.017(0.867;1.191)
Days ill prior consult., mean (SD)	10(8)	204	0.723(0.702;0.743)
Severe Cough	606 (32%)	66	NA
Phlegm	1462(78%)	134	0.505(0.439;0.581)
Severe Breathlessness	87(5 %)	8	0.802(0.538;1.175)
Runny nose	1365 (72%)	50	1.157(0.997;1.347)
Fever	677 (0.36%)	144	5.223(4.534;6.03)
Severe Chest Pain	93 (5 %)	22	4.066(2.984;5.564)
Myalgia	970 (51%)	155	0.3004(0.258;0.348)
Headache	1080 (57%)	153	2.425(2.094;2.816)
General feeling unwell	1434 (76 %)	190	4.636(3.658;5.969)
Interference with daily activities	1183(63 %)	172	3.541(2.985;4.227)
Abnormal auscultation *	616 (33%)	139	0.949(0.825; 1.09)
Tachycardia (pulse $> 100 beats/min$)	867 (46 %)	2	0.626(0.547; 0.716)
Tachypnoea (pulse $> 24 breaths/min$)	37 (2 %)	3	0.719(0.405;1.182)
CRP(mg/L), mean(SD)	21(39)	-	1.003(1.001;1.004)
> 20	1447 (77 %)	132	2.648(2.223;3.144)
> 30	196 (10 %)	41	1.862(1.475;2.329)
> 50	127 (7 %)	20	1.025(0.758;1.359)
> 100	116 (6 %)	11	Ref.cat.

* Abnormal auscultation breath sounds, wheeze, rhonchi, crackles.

Table 3: Association between clinical indicators and Asthma or COPD (n= 2006 & n = 1940)

Diagnostic Variable	Value(%) or mean(SD)	Asthma present (n = 324)	COPD present (n = 258)	$OR_{asthma}(95\%CI), OR_{COPD}(95\%CI)$
Age mean(SD)	50(17)	324	258	0.986, (0.983;0.989),0.975(0.971;0.979)
Men	783 (40 %)	818	748	1.518(1.364;1.689),1.199(1.062; 1.356721)
Current Smoker	521(26%)	538	504	1.874(1.659;2.117),0.557(0.485; 0.639)
Days ill prior consult., mean(SD)	10(8)	324	258	0.989 (0.982;0.996),1.013(1.01,1.021)
Severe Cough	644 (33 %)	654	633	0.687(0.421;1.172), 0.889(0.427; 1.664)
General feeling unwell	1464 (74%)	1492	1435	1.149(1.015;1.303),1.004(0.878;1.15)
Interference with daily activities	1227 (62%)	1298	1155	1.03(0.924;1.148),1.185(1.053;1.333)
Abnormal auscultation[1]	669 (34%)	681	658	1.351(1.211;1.507),0.702(0.623;0.792)
Diminished vesicular breathing	211(11%)	210	212	1.048(0.881; 1.241),0.69(0.583;0.820)
Severe Wheeze	47(3%)	52	42	6.212(4.493;8.615), 0.293(0.199; 0.434)
Severe Chest Pain	85 (4%)	87	82	1.555(1.153;2.085), 0.82(0.589;1.154)
Allergic disease	287(15%)	294	279	1.045(0.899;1.211),1.123(0.948; 1.338)
Recurrent (> 3) cough episodes last year	226(12%)	242	209	1.99(1.726;2.291),0.9463(0.788;1.144)
Asthma in the family	475(24%)	955	904	1.142(1.011;1.288),1.019(0.888;1.168)
Phlegm Colour (green, yellow or bloodstained)	930 (47%)	161	110	1.11(0.999;1.235),1.195(1.062; 1.345)
CRP(mg/L), Mean(SD)	21(40)	-	-	1.001(1;1.003),0.997(0.995; 0.998)
> 20	1527 (77%)	238	187	1.456(1.23;1.715),0.7699(0.638;0.935)
> 30	190(10%)	40	24	1.251(1.013;1.533),0.673(0.545;0.837)
> 50	130 (7%)	24	20	1.117(0.894;1.382),0.746(0.598;0.938)
> 100	126 (6%)	21	20	Ref.Cat

[1] Abnormal auscultation breath sounds, wheeze, rhonchi, crackles.

Asthma or COPD invariably have the same symptoms. Therefore, discriminating patients' condition to be Asthma, COPD or both, when they present to primary care with acute cough can be difficult prior to the lung function test results. All variables considered in our case, are variables / clinical predictors which were measured at day one the patient consulted the primary care center (Case Report Form). According to literature, the cough of COPD patients is frequently associated with chronic sputum production [19, 20]. Our analysis shows that 50% of asthmatic patients and 43% of COPD patients have phlegm production. The symptoms and signs are outwardly similar between COPD and asthmatic patients. Often the age of initial presentation is the distinguishing factor between COPD and asthma. Asthma typically appears in childhood whereas COPD over the age of 40.

4.3 Variable Selection

The variable importance list from the random forest process is given in Figures 3, 4, 5, 6 and 7, separately for each imputed data set with information on how important that variable is in classifying the data. We keep all variables which have a positive mean decrease in accuracy in at least one imputed data set. The mean OOB error estimate is 0.342(34.2%) and the optimal number of variables per tree level i.e. the ones that minimize the OOB error found to be 12 at three out of five imputed data sets and 6 for the rest. The number of trees to grow at each iteration (bootstrap replicates) are set to 500, as from 500 trees on wards the OOB error seems to stabilize (figures App. 17, 18, 19, 20 and 21).

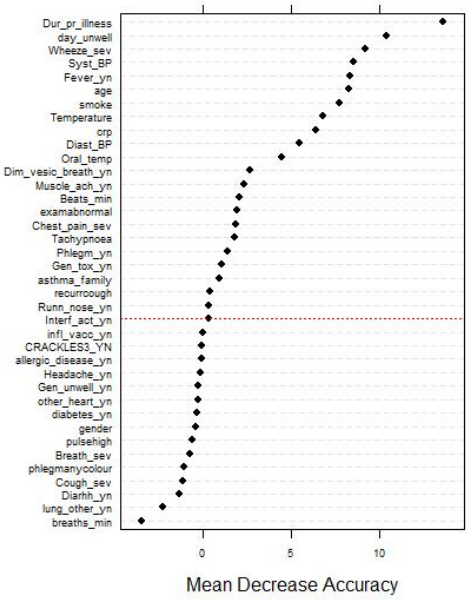


Figure 3: Imputed data 1

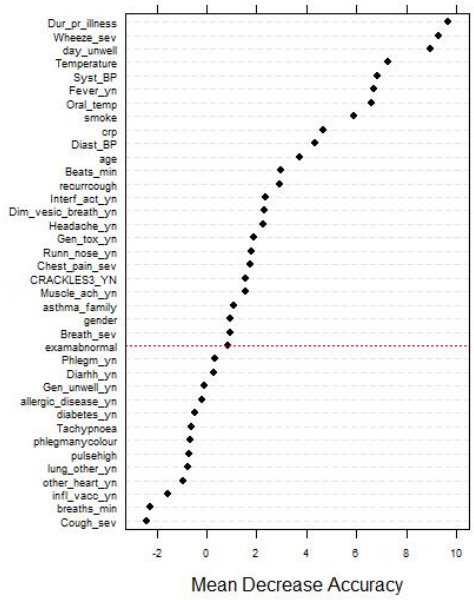


Figure 4: Imputed data 2

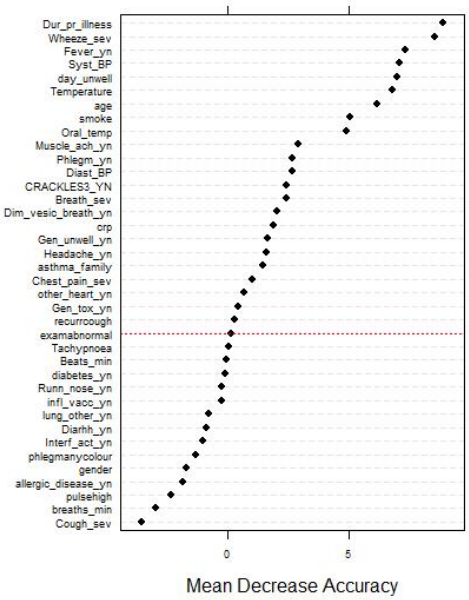


Figure 5: Imputed data 3

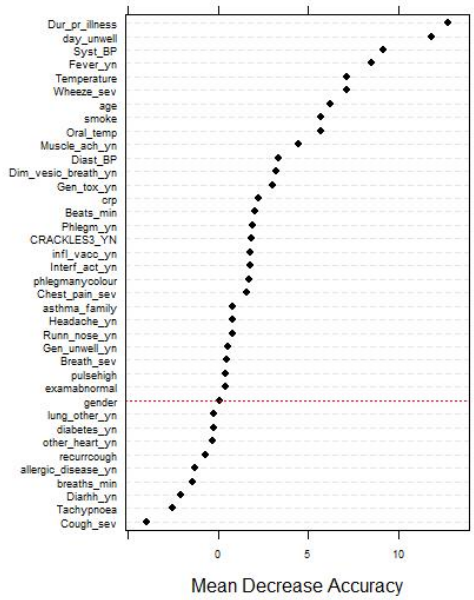


Figure 6: Imputed data 4

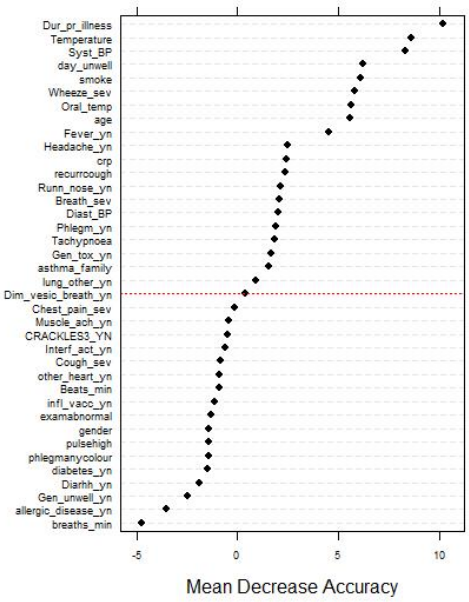


Figure 7: Imputed data 5

4.4 Multinomial Logistic Regression

From the importance lists above (figures 3- 7), we chose the clinical predictors with positive mean decrease in accuracy. Using the selected set of predictors we fit a multinomial logistic regression. The baseline category is the "Other" diagnosis. Below we look at the significant quasi likelihood estimators across the five imputed data sets. CRP measurements tend to vary depending on days of illness and therefore the interaction between them was tested. Additionally given that we re modeling patients whose age ranges from 18-92 years old, we considered the possibility that some clinical predictors may vary depending on age. Nevertheless, that was not the case in our analysis as no interactions proved to be significant, the smallest p_{value} for the 1st imputed data set is 0.06 (interaction age with phlegm production) and the highest 0.892 (interaction age with severe breathing difficulties). Below, we look at the significant parameter estimates (Tables 4,5,6 and 7).

Table 4: Parameter estimates and standard errors for the Multinomial model, effect on log of odds for Pneumonia when reference is other diagnosis.

	Analysis by imputation				
Parameter*	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5
Intercept	-3.399(0.529)	-1.918(0.7813)	-	-	-
Dur.pr.illness	-0.035(0.012)	-0.035(0.015)	-0.033(0.0122)	-	-0.041(0.013)
Wheeze.sev3†	-	-	0.353(0.45)	-	-
Syst.BP	-	-0.022(0.009)	-0.024(0.01)	-0.2445(0.01)	-0.023(0.009)
Day.unwell	0.017(0.008)	-	-	-	-
Fever.yn1	0.976(0.259)	1.112(0.301)	0.973 (0.259)	1.008(0.282)	1.066(0.299)
Age	-	0.02(0.008)	-	-	-
smoke2$^\bullet$	0.591(0.234)	-	-	0.5(0.247)	-
smoke3$^\bullet$	-	0.431(0.208)	0.441(0.22)	-	-
CRP	0.01(0.001)	-	0.009(0.001)	0.009(0.0007)	-
Muscle.ach.yn1	-	0.465(0.209)	-	0.009(0.0007)	-
Runn.nose.yn1	-0.837(0.148)	-	-0.859 (0.161)	-0.87(0.153)	-1.023(0.142)
Crackles.yn1	1.12(0.437)	1.497 (0.508)	1.125(0.412)	1.122(0.316)	1.407(0.512)
Breath.sev4‡	0.785 (0.376)	1.288(0.424)	-	0.897(0.393)	1.126(0.381)
Gen.unwell.yn1	-0.847(0.266)	-0.684(0.299)	-0.764(0.289)	-0.789(0.325)	-0.666(0.301)
Dim.vesic.breath.yn1	0.692(0.31)	-	-	-	-
Lung.other.yn1	-	0.778(0.336)	-	-	0.648(0.287)
other.heart.yn1	-	0.651(0.251)	-	-	-

* All variables significant at 5% level of significance ($p_{value} < 0.05$).
† Wheeze.sev3→ Moderate Problem as compared to No problem(Wheeze.sev1).
$^\bullet$ smoke3=Current smoker,smoke2=Past smoker as compared to Never smoked(smoke1).
‡ Breath.sev4 = Severe Breathing difficulty as compared to no problem.

Table 5: Parameter estimates and standard errors for the Multinomial model, effect on log of odds for Influenza when reference is other diagnosis.

	Analysis by imputation				
Parameter*	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5
Dur.pr.illness	-0.248(0.033)	-0.258(0.034)	-0.24(0.03)	-0.259(0.027)	-
Day.unwell	-0.046(0.021)	-	-	-	-0.039(0.146)
Wheeze.sev4†	1.461(0.426)	1.189(0.481)	1.381 (0.386)	1.425(0.41)	1.449(0.419)
Fever.yn1	1.167(0.168)	-	1.158(0.173)	1.175(0.173)	1.219(0.154)
Syst.BP	-	-	-0.008(0.004)	-0.009(0.004)	-
asthma.family1	-	-	-	-0.422(0.183)	-0.333(0.146)
Chest.pain.sev4*	-	0.492(0.22)	-	-	-
Interf.act.yn2$^{\bullet}$	-0.523(0.148)	-	-0.534(0.199)	-0.558(0.192)	-0.537(0.177)
Muscle.ache.yn1	0.445(0.198)	0.465(0.209)	0.482 (0.171)	0.519(0.159)	0.444(0.189)
Phlegm.yn1	-	-	-0.619(0.292)	-0.635(0.287)	-0.581(0.282)
other.heart.yn1	-0.629(0.227)	-0.651(0.251)	-0.631(0.269)	-	-0.643(0.267)
Breath.sev4‡	-0.957(0.362)	-0.858(0.366)	-0.941(0.363)	-0.896(0.362)	-0.907(0.359)
Breath.sev3‡	-	-	-0.643(0.309)	-	-
Gen.unwell.yn1	0.433(0.211)	0.448(0.181)	-	0.438(0.222)	0.4176(0.206)

* All variables significant at 5% level of significance ($p_{value} < 0.05$).

$^{\bullet}$ Interf.act.yn2→No problem of interference with daily activities versus problem.

† Wheeze.sev4→ Severe Wheeze as compared to No problem(Wheeze.sev1).

* Chest.pain.sev4→ Severe Chest paim as compared to No problem(Chest.pain.sev1).

‡ Breath.sev3 (Breath.sev4) = Moderete(Severe) Breathing difficulty as compared to no problem.

Table 6: Parameter estimates and standard errors for the Multinomial model, effect on log of odds for Asthma when reference is other diagnosis.

	Analysis by imputation				
Parameter*	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5
Intercept	-2.179(0.47)	-2.351(0.454)	-2.385(0.488)	-2.407(0.457)	-2.146(0.449)
Day.unwell	-	-0.024(0.009)	-0.021(0.009)	-	-0.021(0.009)
Wheeze.sev2 †	0.914(0.231)	0.905(0.203)	0.947(0.222)	0.520(0.430)	0.969(0.209)
Wheeze.sev3 †	1.061(0.371)	1.069(0.316)	1.096(0.352)	1.090(0.374)	1.106(0.338)
Wheeze.sev4†	1.721(0.386)	1.669(0.329)	1.727(0.345)	1.756(0.364)	1.720(0.348)
Fever.yn1	-0.219(0.093)	-0.256(0.090)	-0.227(0.0853)	-0.234(0.091)	-0.223(0.087)
Age	-0.011(0.004)	-0.011(0.004)	-0.013(0.004)	-0.013(0.004)	-0.013(0.004)
smoke3$^{\bullet}$	0.366(0.141)	0.371(0.152)	0.382(0.141)	0.387(0.140)	0.360(0.143)
CRP	0.002(0.001)	-	-	-	-
Gen.tox.yn1	-0.341(0.142)	-0.333(0.147)	-0.346(0.131)	-0.314(0.148)	-0.315(0.130)
recurrcough1	0.554(0.129)	-	0.566(0.122)	0.600(0.148)	0.731(0.165)
gender1*	0.489(0.141)	-	0.492(0.122)	0.461(0.131)	0.522(0.144)

* All variables significant at 5% level of significance ($p_{value} < 0.05$).

† Wheeze.sev2$\rightarrow$ Mild Wheeze as compared to No problem(Wheeze.sev1).

† Wheeze.sev3$\rightarrow$ Moderate Wheeze as compared to No problem(Wheeze.sev1).

† Wheeze.sev4$\rightarrow$ Severe Wheeze as compared to No problem(Wheeze.sev1).

$^{\bullet}$ smoke3=Current smoker, smoke2=Past smoker as compared to never smoked(smoke1).

* gender1 $\rightarrow$ male.

Table 7: Parameter estimates and standard errors for the Multinomial model, effect on log of odds for COPD when reference id other diagnosis.

	Analysis by imputation				
Parameter*	Imp. 1	Imp. 2	Imp. 3	Imp. 4	Imp. 5
Intercept	-4.472(0.829)	-4.663(0.871)	-4.895(0.871)	-4.824(0.841)	-4.676(0.824)
Dur.pr.illness	-	-	-	-0.019(0.009)	-
Age	0.035(0.010)	0.034(0.011)	0.032(0.010)	0.032(0.010)	0.034(0.011)
Smoke3$^{\bullet}$	0.839(0.233)	0.858(0.244)	0.856(0.234)	0.855(0.231)	0.848(0.229)
Muscle.ache.yn1	0.37(0.095)	0.353(0.099)	0.368(0.094)	0.363(0.099)	0.351(0.101)
lung.other.yn1	-	0.738(0.247)	0.627(0.272)	-	0.696(0.245)
CRP	0.003(0.001)	-	0.003(0.0007)	-	-
gender1 *	-0.318(0.110)	-0.318(0.105)	-0.323(0.117)	-0.343(0112)	-0.311(0.114)
Breath.sev4 †	0.711(0.312)	0.706(0.265)	0.749(0.281)	0.673(0.321)	0.764(0.312)
Gen.tox.yn1	-0.724(0.165)	-0.711(0.142)	-0.707(0.169)	-0.680(0.179)	-0.724(0.163)
Runn.nose.yn1	-	-	-	-0.198(0.099)	-

* All variables significant at 5% level of significance ($p_{value} < 0.05$).

† Breath.sev4 $\rightarrow$ Severe Breathing difficulty as compared to No problem(Breath.sev1).

$^{\bullet}$ smoke3=Current smoker, smoke2=Past smoker as compared to never smoked(smoke1).

* gender1 $\rightarrow$ male.

For the combined results across imputations, we followed the procedure outlined by Rubin, (see 4.1. Multiple Imputation by Chained Equations). None of the pooled estimates was significant. We discuss this issue later (see 5. Discussion).

Comparing with tables 14, 15, 16, 17 at Appendix, we could say that the estimates are similar to each other preserving the directionality. For example, the log of odds for Influenza for Dur.pr.illness variable, in the separate logistic regression, when reference is pneumonia, is equal to 4.722 -0.415 $\times$ xDur.pr.illnessi and for the multinomial model (see Eq. 10-12) 3.399 + (-0.248 + 0.035) = 3.399 - 0.213 $\times$ xDur.pr.illnessi.

We know examine three out of the five categories. We fit a Multinomial model for Pneumonia (j=1), Influenza (j=2) and Other diagnosis (reference). From this point on wards, there are some features changed in our analysis. Firstly the 111 patients with combined results are now removed. Secondly we ignore the heterogeneity between countries and similarity within countries (clusters) since we have reasons to believe that the small number of clusters mistakenly inflates the standard errors of the estimated coefficients. The significant estimated relative risks (ratio of the probability of outcome category j=1 or j=2 over the probability of J=3, reference category) from the first imputed data set, are presented at tables 8, 9 .

Table 8: Parameter estimates and standard errors for the 2nd Multinomial model, effect on log of odds for Pneumonia when reference is other diagnosis

Parameter*	Estimate	Standard error	p_{value}
Syst.BP	-0.027	0.008	0.020
Fever.yn2†	-0.800	0.258	$< .001$
CRP	0.01	0.002	$< .001$
Runn.nose.yn2†	0.984	0.245	$< .001$
Crackles.yn1	1.201	0.339	$< .001$
Gen.unwell.yn2†	0.959	0.324	0.003

* All variables significant at 5% level of significance ($p_{value} < 0.05$).
† yn2 $\rightarrow$ no problem when reference is problem.

Table 9: Parameter estimates and standard errors for the 2nd Multinomial model, effect on log of odds for Influenza when reference is other diagnosis.

Parameter*	Estimate	Standard error	p_{value}
Dur.pr.illness	-0.252	0.043	$< .001$
Fever.yn2†	-1.228	0.219	$< .001$
Diast.BP	-0.025	0.012	0.037
Chest.Pain.sev4*	0.853	0.437	0.051
Phlegm.yn1	-0.776	0.186	$< .001$
Interf.act.yn2†	-0.563	0.294	0.055
Breath.sev.3‡	-0.748	0.369	0.043
Breath.sev.4 ‡	-1.002	0.531	0.058

* All variables significant at 5% level of significance ($p_{value} <$ 0.05 except Breath.sev.4, Chest.Pain.sev4 $\rightarrow$ borderline significant).
† yn2 $\rightarrow$ no problem when reference is problem.
* Chest.pain.sev4$\rightarrow$ Severe Chest paim as compared to No problem(Chest.pain.sev1).
‡ Breath.sev3 $\rightarrow$ Moderate Breathing difficulty as compared to No problem(Breath.sev1).
‡ Breath.sev4 $\rightarrow$ Severe Breathing difficulty as compared to No problem(Breath.sev1).

4.5 Diagnostic Accuracy

The performance of the multinomial model was first assessed for each type of diagnosis versus reference(other diagnoses) by fitting separate Logistic Regressions. The ROC curves of the pairs had all extremely small AUC values even when the patients' combinations with concurrent diagnoses ($n_{concurrent} = 231$) were removed ($N_{final2} = 2344$). (Discussion part). Next, at figures 8, 9, 10 and 11 we look at the ROC curves and the discriminating power of contrasts which appeared to have the best discriminating ability in terms of Area Under the Curve (AUC) value. Variables that did not appear significant at any imputed data set from the multinomial (1st model) process are now removed (i.e. Asthma.family, Influenza.vaccination, Tachypnoea, Diastolic BP, phlegm.color, headache and interference.with.daily.activities). A forward selection procedure is applied to each Logistic Regression model to identify significant predictors for each pair of diagnoses. All significant predictors for each pair of diagnosis can be found at tables 14 -17 at Appendix. For example, the final model for Influenza when reference category is pneumonia, for patient i = 1,$\cdots$, 1789, will be :

$$
\begin{aligned}
log(\frac{\pi_4}{\pi_1}) = 4.722 + &-0.415 \times x_{Dur.pr.illnessi} - 0.019 \times x_{CRPi} - 2.557 \times x_{Gen.unwell.yn2i} \\
&-1.931 \times x_{Dim.vesic.breath.yn1i} - 1.821 \times x_{Crackles.yn1i} - 1.097 \times x_{Runn.nose.yn2i} \\
&-1.517 \times x_{Phlegm.yn1i} - 0.036 \times x_{agei} + 3.522 \times x_{lung.other.yn2i} \\
&-1.414 \times x_{smoke3i} + 2.193 \times x_{Wheeze.sev2i} \\
&+3.915 \times x_{Wheeze.sev3i} + 4.75 \times x_{Wheeze.sev4i} \\
&-1.937 \times x_{iBreath.sev2} - 2.104 \times x_{iBreath.sev3} \\
&-3.671 \times x_{iBreath.sev4}
\end{aligned}
$$

Table 10: Likelihood Ratio test (df) & p_{value}

Model	Chi-square (df)	p_{value}
1	200(19)	$2.775 * 10^{-32}$
2	149(18)	$1.279 * 10^{-22}$
3	94(12)	$7.923 * 10^{-15}$
4	233(10)	$1.614 * 10^{-44}$

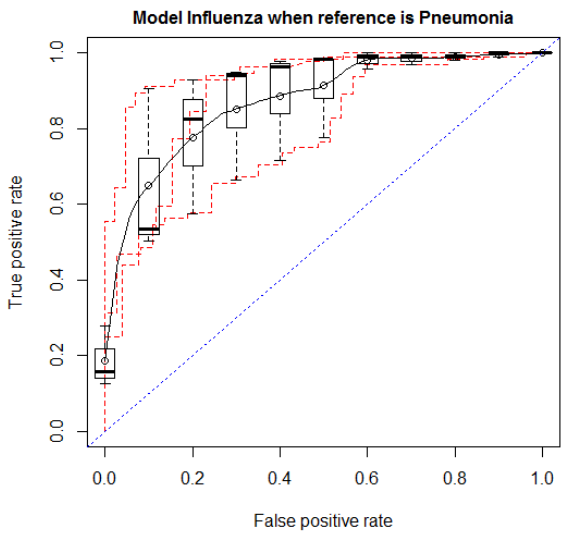


Figure 8: ROC curve.AUC=0.92

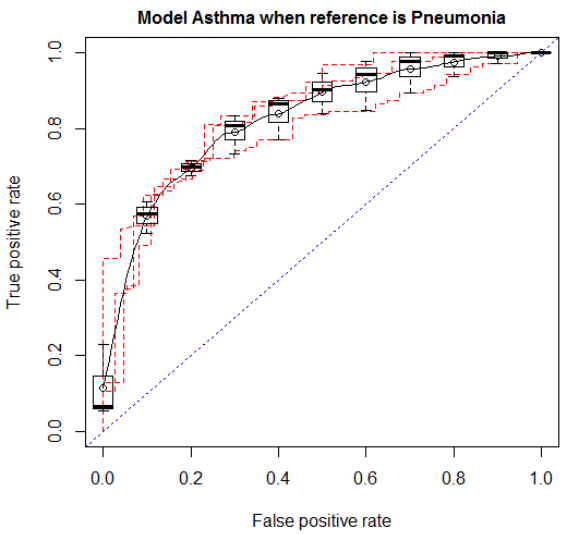


Figure 9: ROC curve.AUC=0.86

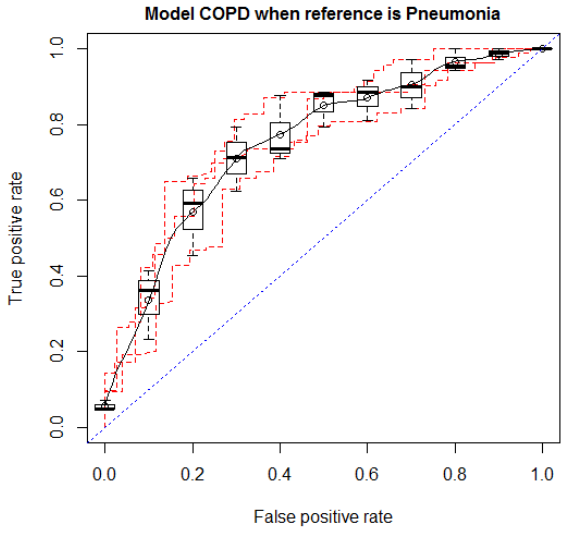


Figure 10: ROC curve.AUC=0.79

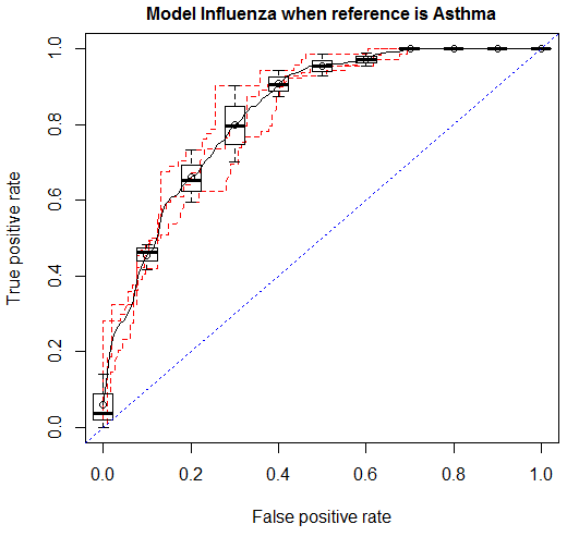


Figure 11: ROC curve.AUC=0.85

At figure 8 we observe the connected pairs of predicted probabilities at specific cut off points, as they were estimated from a Logistic Regression model for Influenza when Pneumonia is the reference category. The calibration of all models was good (figures Appendix 22, 23) , but not realistic enough since models are fitted to the actual data. We therefore proceed to a 3-k cross validation plot. Unlike models Asthma when Pneumonia is the reference category, COPD when Pneumonia is the reference category and Influenza when Pneumonia is the reference category, the 3-k ROC curves for Influenza when reference is Pneumonia (figure 8), are not very close to each other. The internal estimate of accuracy ranged from 0.751 to 0.814 and the cross-validation estimate from 0.737 to 0.789, for the 4 models. Likelihood ratio tests $LR = 2(loglikelihood(full_{model}) - loglikelihood(null_{model}))$, compare the log likelihoods of two nested models. The test statistic LR is chi-square distributed, with degrees of freedom equal to the difference between degrees of freedom of full and null model. All likelihood ratio tests (LR) showed that all models fit significantly better than a model without any predictors (null model). At table 10 we look at those differences in terms of residual deviance and degrees of freedom as well as the associated p_{value}.

The AUC value for each ROC curve above is again calculated with the difference that now we omit CRP for the 3 first models, in order to gain an insight on its diagnostic discriminating power. The AUC_{noCRP} for model Influenza when reference is Pneumonia, Asthma when reference is Pneumonia and COPD when reference is pneumonia is 0.839, 0.774, 0.774, in all cases slightly decreased, indicating its added value. The difference between comparisons AUC and AUC_{noCRP} proved insignificant for all models. In particular the alternative hypothesis: H_α = true difference in AUC and AUC_{noCRP} is not equal to 0, could not be rejected

$$\leftrightarrow$$

$$z = \frac{AUC - AUC_{noCRP}}{\sqrt{SE^2_{AUC} + SE^2_{AUCnoCRP} - 2rSE_{AUC}SE_{AUCnoCRP}}} \geq 1.96 \qquad (17)$$

The differences (95% CI) between AUC and AUC_{noCRP} for the first imputed data set, were 0.034 (-0.009;0.078), -0.036 (-0.081;0.009) and 0.023(-0.027;0.073).
The z_{scores} and p_{value} were: z = 1.5243, p_{values} = 0.1274, z = -1.5683, p_{values} = 0.1168 and z = 0.8957, p_{value} = 0.3704 for models Influenza when reference is Pneumonia, Asthma when reference is Pneumonia and COPD when reference is Pneumonia.

For the derivation of $SE_{AUC}, SE_{AUCnoCRP}, r$, i.e. the standard errors for the two AUC and the between area correlation, we refer to paper by Hanley, J. A. and McNeil, B. J. [30] .

With regard to comparisons with category other diagnoses, because of the high imbalance between the number of patients with the diagnosis of interest and other diagnosis (reference category) beside the classification accuracy ((tp + tn) / (tp+tn+fp+fn)) we take a look at the positive agreement as well. Positive agreement for a 2X2 table is calculated as the ratio PA=2tp/(2tp+fp+fn) with tp the true positive , fp the false positive and fn the false negative. On the other hand negative agreement can be found by NA=2tn/(2tn+fp+fn) where tn stands for the true negative. For pneumonia the positive agreement ranged from 25%-29% while the correct classification percentage for all imputed data sets was approximately 95%. Of course the huge correct classification arises from the fact that only 107 out of the 1789 are pneumonia diagnosed(204, 324 and 258 influenza,asthma and copd). For influenza the positive agreement was 31% with 90% diagnostic accuracy, for asthma 2% and 84% and for COPD 2% and 87%.

Below we see the 2×2 contingency tables (first imputed data set) for Pneumonia when reference is other diagnosis category and Influenza when reference is Pneumonia, both at a 0.5 cut off point. The tables for the other comparisons can be found at Appendix (tables 18,19,20 and 21,22,23).

		test value: p	test value: n	total
actual value	p′	tp= 19	fn= 88	P′=107
	n′	fp= 9	tn= 1673	N′=1682
total		P=28	N=1761	

Table 11: Contingency 2×2 table, pneumonia when reference is other diagnosis, tp: true positive, fp: false positive, fn : false negative, tn: true negative

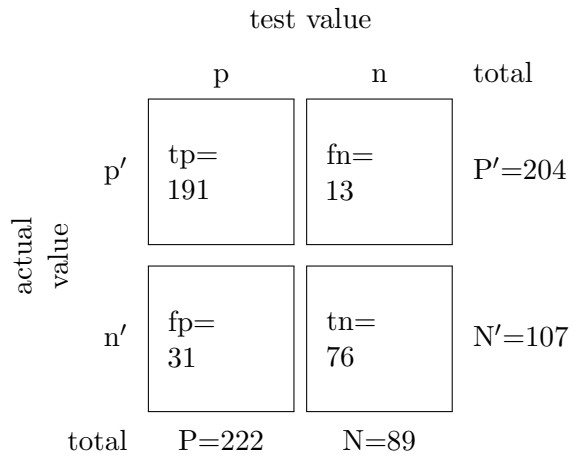


Table 12: Contingency 2×2 table, influenza when reference is pneumonia, tp: true positive, fp: false positive, fn : false negative, tn: true negative

Among the comparisons of interest, models Influenza and Asthma, both versus Pneumonia(baseline) had:

- The highest sensitivity (tp/(tp+fn)) and specificity (tn/(tn + fp)), i.e. the highest probability that Influenza or Asthma will be positive when the disease is present and the highest probability that the patient will not have Influenza or Asthma when the test is indeed negative, respectively. Sensitivity was 0.6, 0.69 and Specificity 0.86, 0.79.
- The highest positive likelihood ratios (sensitivity/(1-specificity)), the highest ratio between the probability of a positive test given the presence of the condition and the probability of a positive test given the absence of the disease. Positive LR was 4.29 and 3.29.

The negative likelihood ratio measures the difference between the probability of a negative test result when the patient has the condition and the probability of a negative result given that the patient does not have the condition. Asthma when reference is Influenza model revealed the smallest Negative LR, 0.42.

The above approach has certain limitations. The estimates fitted by separate logistic regressions will normally differ from those obtained simultaneously by a multinomial logit. The diagnostic accuracy was accessed for the second multinomial model i.e for the three categories simultaneously . The overall HUM is 0.57, and the variables with the highest HUM are crackles.yn ($\widehat{HUM} = 0.87$) and breathlessness severity ($\widehat{HUM} = 0.68$).

The overall HUM estimator suggests that each of the patient is correctly classified with 0.57 probability. The accuracy is good, since a non informative value for HUM would had by chance probability of occurring $0.17(M!)^{-1}$.

5 Discussion

In the observational study performed between October 2007-July 2010, out of the 2575 patients, 4% had Pneumonia on chest radiography, 13% Asthma based on an FEV1 $\geq$ 12% or more than 200 milliliters (ml) after bronchodilation, 10% COPD confirmed with an FEV1/FVC below 0.7 or below the lower limit of normal (LLN) and 8% Influenza confirmed as PCR testing of nasopharyngeal swab samples. The laboratory reference standard results were available after inclusion of the patient with spirometry results(FEV, FVC) e.g. 28-35 days after inclusion of the patient.

This report attempts to quantify the accuracy of a multinomial model in diagnosing pneumonia, influenza, asthma and COPD with "other" diagnoses as a baseline reference. Most commonly imported other diagnoses were: acute bronchitis, chronic bronchitis, nasopharygitis, chest infection, tracheitis and tracheobronchitis.

In order to understand why our combined results from all 5 multiple imputations were insignificant, we first calculated the within and between imputation variance(B) for each variable which was exceptionally small indicating that estimated coefficients as we can see from tables 4-7 are more or less similar across imputations without big discrepancies. Another explanation comes from the fact that we took into consideration heterogeneity between countries which implies that standard errors will be inflated in comparison to what we would have expected if we would have ignored it. Excluding the suspicion that insignificance comes from the variability between imputations, observing how small the odds of each category versus the reference (other) are and keeping in mind that the "other" category occupies 65% of the sample size, we have reasons to believe that pneumonia, asthma, copd and influenza are hard to differentiate versus other diagnoses if based only on sign and symptoms at first day of consultation. For that reason we also tried to change the reference category and see what happens when e.g. we test each category versus asthma. The AUC then, for all possible combinations was approximately 0.77 except for : Pneumonia when reference is Asthma (AUC=0.25), Pneumonia when reference is COPD(AUC=0.28), Asthma when reference is COPD (AUC=0.36), Pneumonia when reference is influenza(AUC=0.22), Asthma when reference is Influenza(AUC=0.14) and COPD when reference is Influenza (AUC=0.16). This doesn't strike as a surprise for two reasons: First patients with asthma and/or COPD are more susceptible to pneumonia and both can arouse as a flu-related complication. Secondly in our analysis patients with concurrent results were included as separate units. So to speak e.g. the 4 patients with pneumonia and influenza are included 2 times, the first as pneumonia and the second time as influenza. This action has certain limitations since the assumption that each independent variable has a single value for each case is violated. Nevertheless even when we removed the 111 patients with concurrent results, the AUC values for other diagnoses as reference, remained still low($AUC \approx 0.20$). For the diagnostic accuracy of the model we looked any two category subsets since any pair of a multinomial model is a valid parametrization i.e. any multinomial pair is conditionally binomial [10]. The pairs with the highest AUC, that is the categories with the highest discriminating ability are Influenza when reference is Pneumonia, Asthma when reference is Pneumonia, COPD when reference is Pneumonia and Influenza when reference is Asthma. The correspondent significant clinical predictors as were specified by forward selection procedure are shown at tables 14, 15 ,16 and 17(Appendix).

We then constructed the ROC curves of the specified contrasts along with a 3-fold cross validation estimate of the ROC curve, their average estimate and the corresponding box plots (figures 8,

9, 10, 11). The added diagnostic value of selected inflammatory markers, CRP was quantified by running each model again without CRP and calculating the AUC value of its ROC curve. The AUC values when CRP was ommited were lower but insignificant (Hanley McNeil test).

Finally, the HUM estimator showed that the multinomial logistic model (tables 8, 9) for predicting patients conditions i.e. Pneumonia, Influenza and other diagnoses had a good classification accuracy.

References

[1] WHO (2015) *Antimicrobial resistance.* Assessed :2015-04-10. Available at : http://www.who.int/mediacentre/factsheets/fs194/en/

[2] Wikipedia *Antimicrobial resistance.* Assessed :2015-04-11. Available at: *https* : *//en.wikipedia.org/wiki/Antimicrobial*$_r$*esistance*

[3] Lieberman, P. B., Wootan, M. G. (1998) *Protecting the Crown Jewels of Medicine,A strategic plan to preserve the effectiveness of antibiotics.* Assessed :2015-04-11. Available at : https://www.cspinet.org/reports/abiotic.htm

[4] CDC (1999) *Achievements in Public Health, 1900-1999: Control of Infectious Diseases.* Assessed :2015-04-11. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm4829a1.htm

[5] GRACE (2015) *Genomics to combat Resistance against Antibiotics in Community-acquired LRTI in Europe.* Assessed :2015-04-10. Available at : http://medicine.cf.ac.uk/primary-care-public-health/research/common-infections/applied-infection-completed-projects/grace/

[6] Coenen, S., Michiels, B., Renard, D., Denekens, J., & Van Royen, P. (2006). *Antibiotic prescribing for acute cough: the effect of perceived patient demand.* The British Journal of General Practice, 56(524), 183190.

[7] MedScape (2015) *Community-Acquired Pneumonia.* Assessed :2015-04-11. Available at : http://emedicine.medscape.com/article/234240-overview

[8] Wikipedia *Chronic obstructive pulmonary disease.* Assessed :2015-04-11. Available at : *https* : *//en.wikipedia.org/wiki/Chronic*$_o$*bstructive*$_p$*ulmonary*$_d$*isease*

[9] Goossens, H., Little, P. (2006). *Community acquired pneumonia in primary care: Doctors cannot target antibiotics and reduce resistance until new diagnostic tests prove feasible and affordable at the point of care.* BMJ: British Medical Journal, 332(7549), 10451046.

[10] Agresti A. (2002). *Categorical data analysis.* Wiley: New York.

[11] Agresti A. (1996). *Introduction to Categorical Analysis.* Wiley: New York.

[12] Liang, K. Y., and P. McCullagh. 1993. *Case studies in binary dispersion.* Biometrics 49: 623-630.

[13] Angeles Marcos M, Camps, M., Pumarola, T. , Antonio Martinez, J., Martinez, E., Mensa, J., Garcia, E., Penarroja, G., Dambrava, P., Casas, I., Jimnez de Anta, MT., and Torres, A. *The role of viruses in the aetiology of community-acquired pneumonia in adults.* AntivirTher2006, 11: 351-359.

[14] van Vugt, S., Broekhuizen, L., Zuithoff, N., Butler, C., Hood, K., Coenen, S., Goossens, H., Little, P., Almirall, J., Blasi, F., Chlabicz,S., Davies, M., Godycki-Cwirko, M., Hupkova,H., Kersnik, J., Moore, M., Schaberg, T., De Sutter, A., Torres, A., and Verheij, T.; GRACE Project Group. *Airway obstruction and bronchodilator responsiveness in adults with acute cough.* The Annals of Family Medicine. 2012; 10(6): 523-529

[15] Carrat F, Vergu E, Ferguson NM, Lemaitre M, Cauchemez S, Leach S, Valleron AJ. *Time lines of infection and disease in human influenza: a review of volunteer challenge studies.* Am J Epidemiol. 2008; Apr 1;167(7):775-85.

[16] van Vugt SF, Broekhuizen BD, Lammens C, Zuithoff NP, de Jong PA, Coenen S, Ieven M, Butler CC, Goossens H, Little P , Verheij TJ. *Use of serum C reactive protein and procalcitonin concentrations in addition to symptoms and signs to predict pneumonia in patients presenting to primary care with acute cough: diagnostic study.*BMJ 2013; 346.

[17] Athanazio R. *Airway disease: similarities and differences between asthma, COPD and bronchiectasis.* Clinics. 2012;67(11):1335-1343.

[18] van Vugt SF, Broekhuizen BD, Lammens C, Zuithoff NP, de Jong PA, Coenen S, Ieven M, Butler CC, Goossens H, Little P , Verheij TJ. *Validity of a clinical model to predict influenza in patients presenting with symptoms of lower respiratory tract infection in primary care.* Family Practice .2015 Aug;32(4):408-14.

[19] WHO(2013)*Asthma.* Assessed :2015-07-20. Available at: http://www.who.int/mediacentre/factsheets/fs307/en/

[20] WHO (2015)*Chronic obstructive pulmonary disease (COPD).* Assessed :2015-07-20. Available at : http://www.who.int/mediacentre/factsheets/fs315/en/

[21] Agresti A (2015). *Foundations of linear and generalized linear models.* Wiley: New York.

[22] Hosmer D.W.,Lemeshow S (1989). *Applied Logistic Regression.* Wiley: New York.

[23] Pepper J.V. (2002) *Robust Inferences from Random Clustered Samples: An Application using Data from the Panel Study of Income Dynamics*,Economics Letters, 75, 341-5.

[24] Hastie T., Tibshirani R., and Friedman J.(2009). *The Elements of Statistical Learning: Prediction, Inference and Data Mining Second Edition*. Springer Verlag.

[25] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

[26] Van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/.

[27] Long, J.S., Freeze, J. (2014). *Regression Models for Categorical Dependent Variables Using Stata*. Third Edition. College Station, TX: Stata Press.

[28] Czepiel, S. *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*.

[29] Li, J., Fine JP. (2008) *ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies*. Biostatistics, 9(3), 566-576

[30] Hanley, J. A., McNeil, B. J. (1982). *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. Radiology, 143, 29-36

A Appendix

Missing values per variable.

Table 13: Missing values per variable.

Variable	Missing Values	Variable	$MissingValues$
Dur. pr.illness	24	General.Toxicity	3
Cough	14	Dim.vesic.breath.	16
Phlegm	4	Tachycardia	37
Severe Breathlessness	6	Tachypnoea	61
Runny nose	3	Systolic BP	51
Fever	3	Diast BP	51
interference .every.day.activities	2	examabnormal	13
Allergic disease	1	Headache	2
Oral temperature	23	CRP	904
crackles	15	age	0
Days.unwell	480	gender	0
Myalgia	3	Diarrhea	2
reccurcough	541		
asthma .family	449		
phlegm.Colour	167		
smoke	1		
Severe Wheeze	4		
General feeling unwell	3		
Severe.Chest.Pain	11		
Any.heart.comorbidity	0		

Marginal distributions of pairs of variables with the highest missing values.

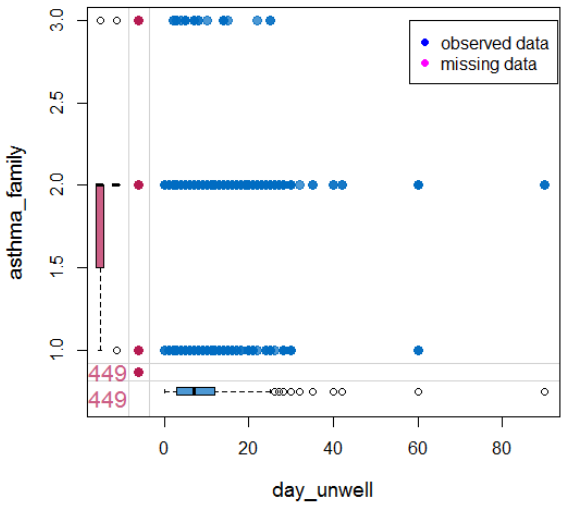


Figure 12: Marginal plot Days unwell versus Asthma.family

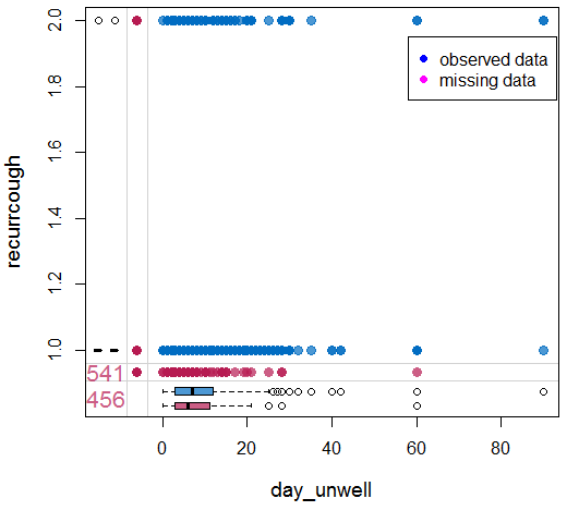


Figure 13: Marginal plot Asthma.family versus Days un.

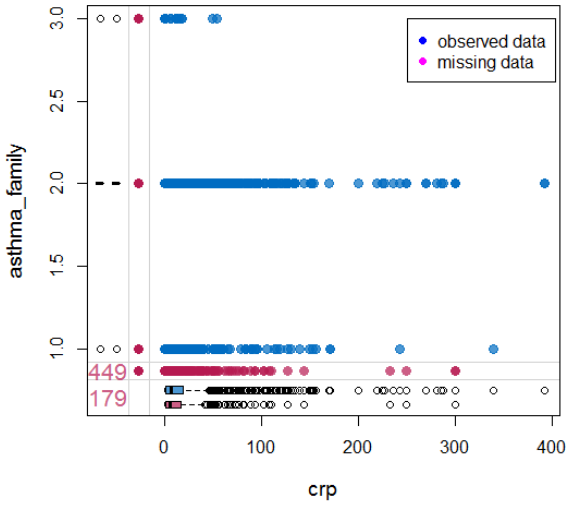


Figure 14: Marginal plot Asthma.family versus CRP

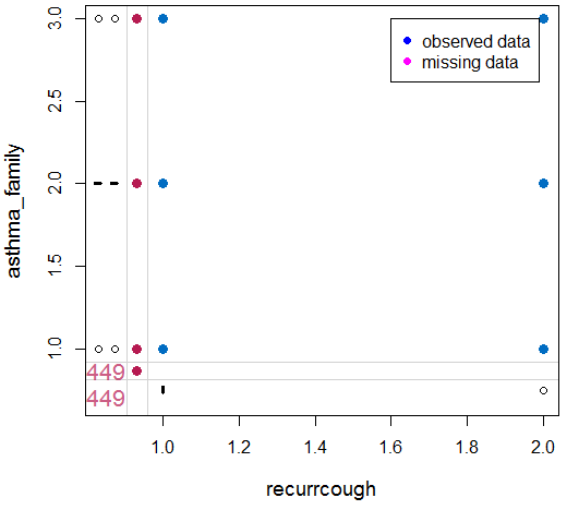


Figure 15: Marginal plot Asthma.family versus Recur.cough

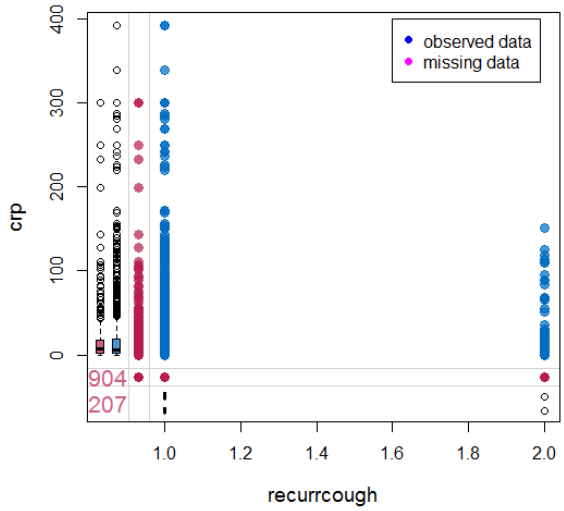


Figure 16: Marginal plot CRP versus Recurr.cough

Stabilization of OOB error at 500 trees, plot for each imputed data set.

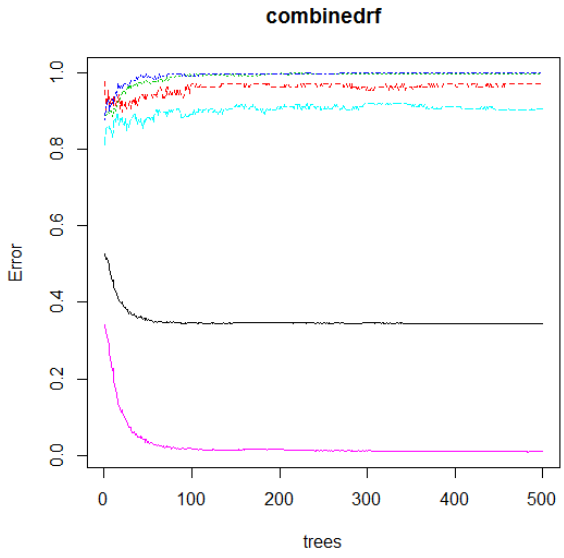


Figure 17: Imputed data 1

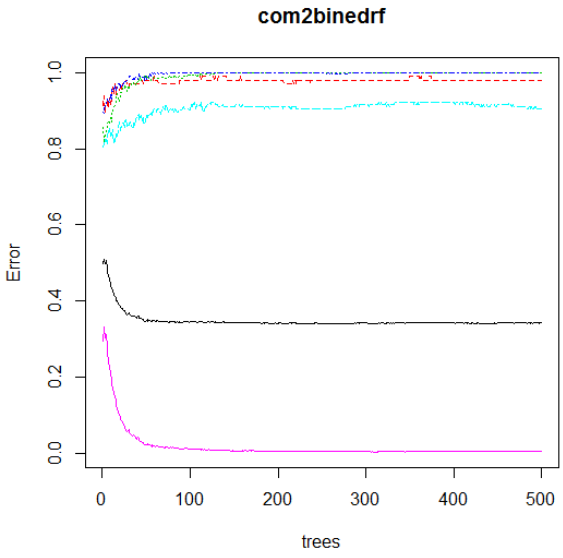


Figure 18: Imputed data 2

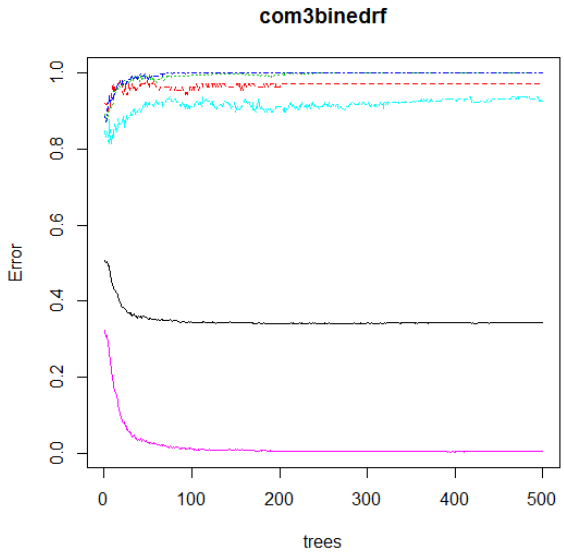


Figure 19: Imputed data 3

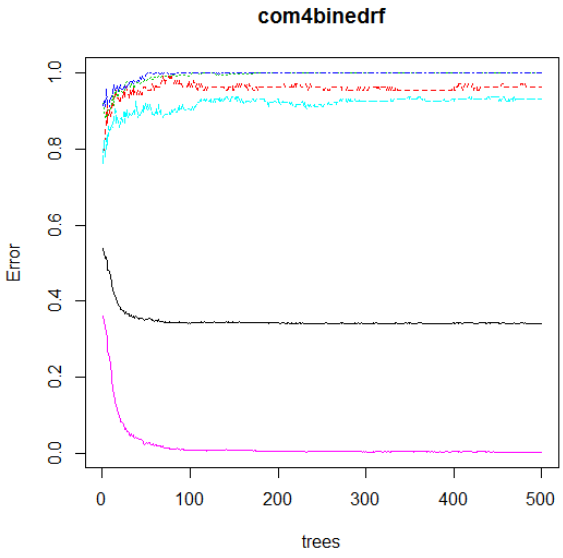


Figure 20: Imputed data 4

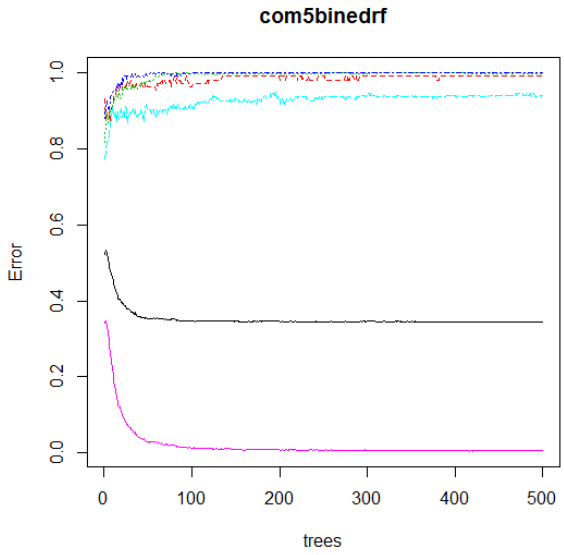


Figure 21: Imputed data 5

Significant Clinical predictors diagnostic accuracy·

Table 14: Parameter estimates and standard errors for the Logistic Regression model, effect on log of odds for influenza when reference is pneumonia.

Parameter	Estimate	Standard error	p_{value}
Intercept	4.722	1.773	0.008
Dur.pr.illness	-0.415	0.079	$< .001$
CRP	-0.019	0.004	$< .001$
Gen.unwell.yn2	-2.556	0.580	$< .001$
Dim.vesic. breath.yn1	-1.931	0.544	$< .001$
Crackles.yn1	-1.821	0.526	$< .001$
Runn.nose.yn2	-1.096	0.414	0.008
Phlegm.yn1	-1.519	0.488	0.002
age	-0.036	0.012	0.008
lung.other.yn2	3.522	1.392	0.011
smoke3	-1.414	0.512	0.006
Wheeze.sev2	2.193	0.793	0.005
Wheeze.sev3	3.915	1.037	$< .001$
Wheeze.sev4	4.75	1.272	$< .001$
Breath.sev2	-1.937	0.892	0.03
Breath.sev3	-2.104	0.994	0.034
Breath.sev4	-3.671	1.178	0.002

Table 15: Parameter estimates and standard errors for the Logistic Regression model, effect on log of odds for asthma when reference is pneumonia

Parameter	Estimate	Standard error	p_{value}
CRP	-0.009	0.003	$< .001$
Fever.yn2	1.323	0.301	$< .001$
age	-0.04	0.010	$< .001$
Syst.BP	0.028	0.008	$< .001$
Runn.nose.yn2	-0.997	0.298	$< .001$
recurrcough1	1.117	0.467	0.017
Dim.vesic. breath.yn1	-1.049	0.375	0.005
Phlegm.yn1	-0.837	0.397	0.035
Crackles3.yn1	-1.074	0.392	0.006
Gen.unwell.yn2	-0.680	0.342	0.047
gender1	0.611	0.306	0.046
smoke2	-0.906	0.364	0.013
Wheeze.sev3	1.248	0.607	0.05
Wheeze.sev4	2.108	0.825	0.010

Table 16: Parameter estimates and standard errors for the Logistic Regression model, effect on log of odds for COPD when reference is pneumonia

Parameter	Estimate	Standard error	p_{value}
Intercept	-1.870	0.915	0.041
Fever.yn2	1.100	0.285	$< .001$
CRP	-0.007	0.002	$< .001$
Crackles.yn1	-1.082	0.355	0.002
Syst.BP	0.028	0.007	0.002
smoke2	-0.71	0.346	0.040
Runn.nose.yn2	-0.618	0.283	0.029
Dim.vesic. breath.yn1	-0.832	0.348	0.017
day.unwell	-0.059	0.027	0.029

Table 17: Parameter estimates and standard errors for the Logistic Regression model, effect on log of odds for influenza when reference is asthma

Parameter	Estimate	Standard error	p_{value}
Intercept	2.912	0.362	$< .001$
Dur.pr.illness	-0.311	0.044	$< .001$
Fever.yn2	-1.525	0.230	$< .001$
smoke3	-0.580	0.268	0.031
Phlegm.yn1	-0.561	0.259	0.030
recurrcough1	-0.994	0.352	0.005
gender1	-0.580	0.236	0.014

		test value: p	test value: n	total
actual value	p′	tp= 7	fn= 317	P′=324
	n′	fp= 8	tn= 1674	N′ =1682
	total	P=15	N=1991	

Table 18: Contingency 2×2 table, asthma when reference is other diagnosis, tp: true positive, fp: false positive, fn : false negative, tn: true negative

.

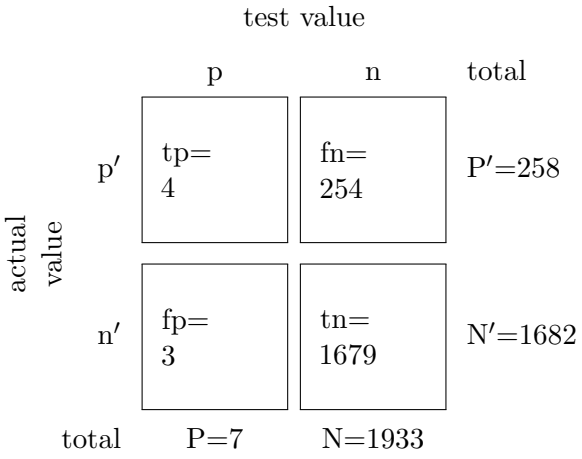


Table 19: Contingency 2×2 table, copd when reference is other diagnosis, tp: true positive, fp: false positive, fn : false negative, tn: true negative

.

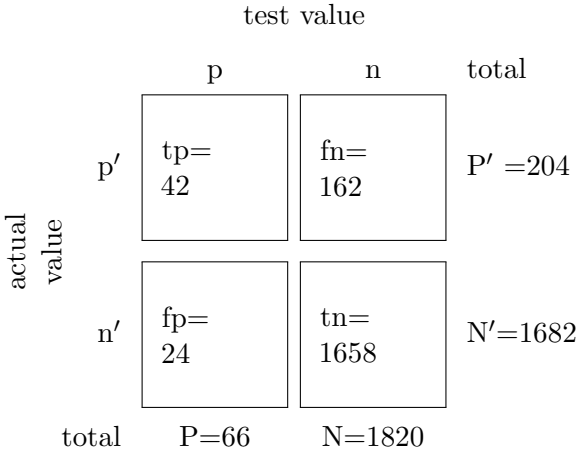


Table 20: Contingency 2×2 table, influenza when reference is other diagnosis, tp: true positive, fp: false positive, fn : false negative, tn: true negative

.

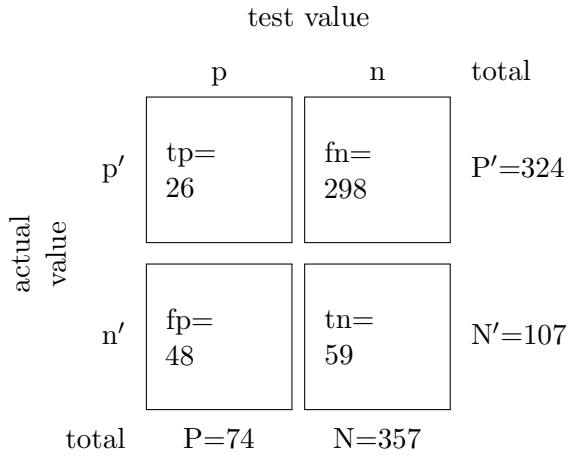


Table 21: Contingency 2×2 table, asthma when reference is pneumonia, tp: true positive, fp: false positive, fn : false negative, tn: true negative

.

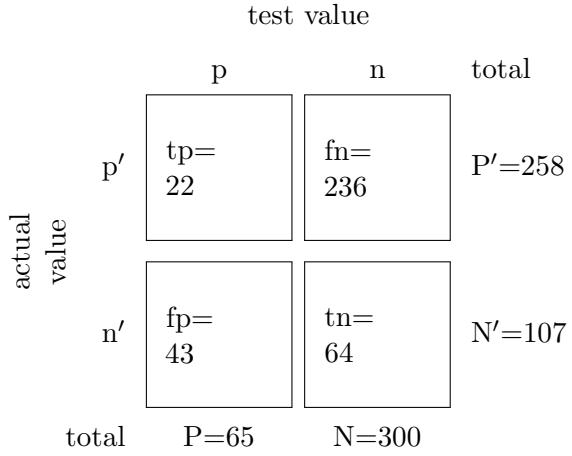


Table 22: Contingency 2×2 table, copd when reference is other diagnosis, tp: true positive, fp: false positive, fn : false negative, tn: true negative

.

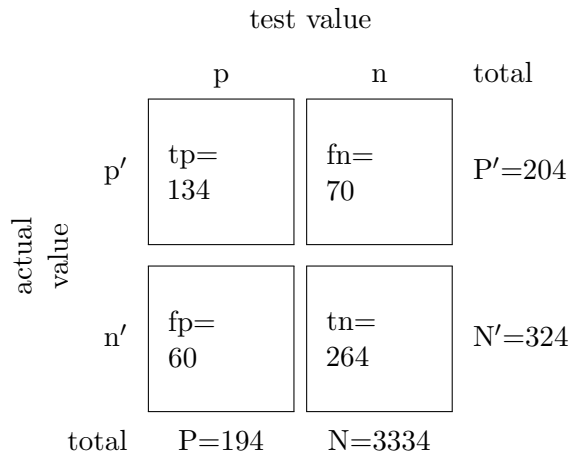


Table 23: Contingency 2×2 table, asthma when reference is influenza, tp: true positive, fp: false positive, fn : false negative, tn: true negative

.

Examine how the scores are calibrated as probability estimates(Calibration Plot).

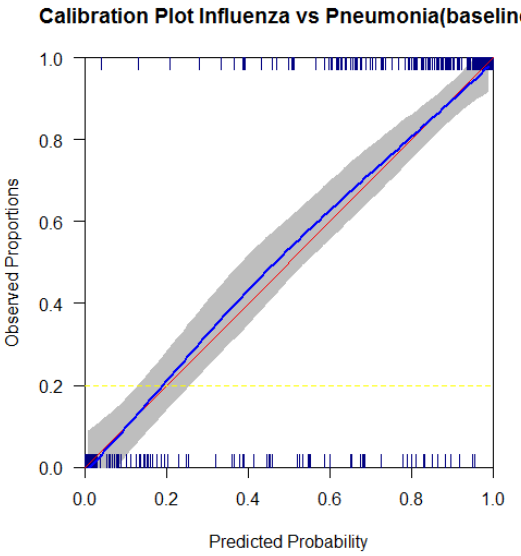


Figure 22: Calibrattion Plot Influenza

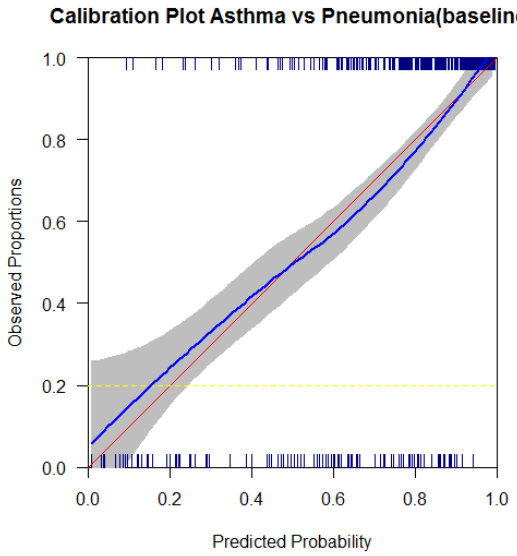


Figure 23: Calibration Plot Asthma

B R code

```
####All codes were repeated for all five
imputation and where necessary estimates
were pooled,for illustration purposes
we show the codes of 1st imputed
dataset.
For diagnostic accuracy part,
for the same
reason we illustrate the procedure for
categories influenza and asthma####
########Section 4.1##########
############################
#####MULTIPLE IMPUTATION ####
############################
library(mice)

mydata <- read.table("C:/New folder/
Thesis/Thesis Last Data/20072015/
multinomialsev2107.txt", header=TRUE,
sep="\t" ,fill = TRUE)



mydata$class <- as.factor(mydata$class)
mydata$pneumonia<-
as.factor(mydata$pneumonia)
mydata$diagASTMA<-as.factor
(mydata$diagASTMA)
mydata$COPD<-as.factor(mydata$COPD)
mydata$Influenza_A_B_interpretation
<-as.factor
(mydata$Influenza_A_B_interpretation)
mydata$other<-as.factor(mydata$other)
mydata$Dur_pr_cough<-
as.numeric(mydata$Dur_pr_cough)
mydata$Dur_pr_illness<-as.numeric
(mydata$Dur_pr_illness)
mydata$age<-as.numeric(mydata$age)
mydata$chest_wheez<-
as.factor(mydata$chest_wheez )
mydata$gender<-as.factor(mydata$gender)
mydata$asthma_family<-
as.factor(mydata$asthma_family)
mydata$smoke<-as.factor(mydata$smoke)
mydata$chest_tightness<-
as.factor(mydata$chest_tightness)
mydata$breaths_min<-
as.numeric(mydata$breaths_min)
mydata$Cough_sev<-
as.factor(mydata$Cough_sev)
mydata$smoke_pack_years<-
as.numeric(mydata$smoke_pack_years)
mydata$Phlegm_yn<-
as.factor(mydata$Phlegm_yn )
mydata$longterm_illness_yn<-
as.factor(mydata$longterm_illness_yn)
mydata$Breath_yn<-as.factor(mydata$Breath_yn)
mydata$asthma_yn<-as.factor(mydata$asthma_yn )
mydata$COPD_yn<-as.factor(mydata$COPD_yn)
mydata$Breath_sev<-as.factor(mydata$Breath_sev)
mydata$other_heart_yn<-
as.factor(mydata$other_heart_yn)
mydata$Runn_nose_yn <-
as.factor(mydata$Runn_nose_yn)
mydata$diabetes_yn <-
as.factor( mydata$diabetes_yn)
mydata$Cough_yn<-as.factor(mydata$Cough_yn )
mydata$Fever_yn<-as.factor(mydata$Fever_yn)
mydata$cough_2wksplus_ter<-
as.factor(mydata$cough_2wksplus_ter)
mydata$Chest_pain_yn<-
as.factor(mydata$Chest_pain_yn )
mydata$Chest_pain_sev
<-as.factor(mydata$Chest_pain_sev)
mydata$Diarhh_yn<-as.factor(mydata$Diarhh_yn)
mydata$Syst_BP<-as.numeric(mydata$Syst_BP)
mydata$Diast_BP<-as.numeric(mydata$Diast_BP )
mydata$Wheeze_yn<-as.factor( mydata$Wheeze_yn )
mydata$Oral_temp<-as.numeric(mydata$ Oral_temp )
mydata$Wheeze_sev<-as.factor(mydata$Wheeze_sev )
mydata$Interf_act_yn<-
as.factor(mydata$Interf_act_yn )
mydata$Muscle_ach_yn<-
as.factor(mydata$Muscle_ach_yn)
mydata$Headache_yn<-as.factor(mydata$Headache_yn)
mydata$Gen_unwell_yn<-
as.factor(mydata$Gen_unwell_yn)
mydata$FEV1<-as.numeric(mydata$FEV1)
mydata$Suspected_pneumonia<-
as.factor(mydata$Suspected_pneumonia)
mydata$ratio_70<-
as.factor(mydata$ratio_70)
mydata$Beats_min<-
as.numeric(mydata$Beats_min)
```

```
mydata$examabnormal<-
as.factor(mydata$examabnormal)
mydata$FV<-as.numeric(mydata$ FVC)
mydata$Gen_tox_yn<-
as.factor(mydata$Gen_tox_yn)
mydata$Dim_vesic_breath_yn<-
as.factor(mydata$Dim_vesic_breath_yn)
mydata$crpgp<-as.factor(mydata$crpgp)
mydata$crp<-as.numeric(mydata$crp)
mydata$Tachypnoea<-
ifelse(comb$breaths_min > 24,
"Tachypnoea","Normal breaths")
comb$Temperature <-
ifelse(comb$Oral_tem >37.8 ,
"Fever","No Fever")
mydata$Tachypnoea<-
as.factor(comb$Tachypnoea)
comb $ Temperature <-
as.factor(comb$Temperature)
mydata$prescr_med_yn<-
as.factor(mydata$prescr_med_yn )
mydata$infl_vacc_yn<-
as.factor(mydata$infl_vacc_yn)
mydata$CRACKLES3_YN<-
as.factor(mydata$CRACKLES3_YN)
mydata$lung_other_yn <-
as.factor(mydata$lung_other_yn )
mydata$allergic_disease_yn
<-as.factor
(mydata$allergic_disease_yn )
mydata$ AB_treat_yn<-
as.factor(mydata$AB_treat_yn)
mydata $ recurrcough<-
as.factor(mydata$recurrcough)
mydata$day_recovery<-
as.numeric(mydata$day_recovery)
mydata$phlegmanycolour
<-as.factor(mydata$phlegmanycolour)
mydata$pulsehigh<-
as.factor(mydata$pulsehigh)
mydata$day_unwell <-
as.numeric(mydata$day_unwell)
mydata $ lung_other_yn <-
as.factor(mydata$lung_other_yn)

##################################
##################################
##################################
remove variables not used
in the papers
##################################
##################################

mydata3<-mydata[,c(1:21,23,26:69,71)]

mydata3pat<-
mydata3[,c(1:8,10:12,14:16,18:21,23:26
,28,31:34,36:41,43,45,46,48,49:57,59,
60,64:67)]
str(mydata3pat)
#####################################
library(data.table)
setDT(mydata3pat)[PCN_name ==
"Brastislava",
Country := "Slovakia"]
setDT(mydata3pat)[PCN_name ==
"Cardiff",
Country := "Wales"]
setDT(mydata3pat)[PCN_name ==
"Southampton",
Country := "England"]
mydata3pat$Country<-
factor(mydata3pat$Country)
table(mydata3pat$Country)
#####################################
#######explore missingness patterns##
############
missi<-md.pattern(mydata3pat)
library(VIM)
marginplot(mydata[, c("recurrcough", "crp")],
col = mdc(1:2), cex = 1.2,
cex.lab = 1.2, cex.numbers = 1.3, pch = 19)
########missing values per variable##
for (Var in names(mydata)) {
    missing <- sum(is.na(mydata[,Var]))
    if (missing > 0) {
        print(c(Var,missing))
    }}
#####################################
Imputation, which variables to use
as predictors and imputation method
for each variable
pmm:predictive mean matching
logreg:logistic regression model
polyreg:Multinomial logit model
#####################################
ini <- mice(mydata3pat,maxit=0,pri=F)
mydata3pat<-as.data.frame(mydata3pat)
```

```
pred<-ini$pred
pred[,c("SubjectCode","ResponsDate",
"pneumonia","class","diagASTMA","COPD",
"Influenza_A_B_interpretation","other",
"smoke_pack_years","PCN_name","Country",
"Diagnosis","day_unwell")] <- 0
meth <- ini$meth

meth[c("age","gender","smoke",
"Dur_pr_illness","crpgp","Cough_sev",
"Phlegm_yn","Breath_sev","Runn_nose_yn",
"Fever_yn","Chest_pain_sev","Diarhh_yn",
"other_heart_yn","Gen_tox_yn",
"Dim_vesic_breath_yn","pulsehigh",
"breaths_min","Syst_BP","Diast_BP",
"Oral_temp","crp","infl_vacc_yn",
"lung_other_yn","diabetes_yn",
"Muscle_ach_yn","Headache_yn",
"Gen_unwell_yn","Interf_act_yn",
"examabnormal","Wheeze_sev",
"allergic_disease_yn","recurrcough",
"phlegmanycolour","asthma_family")]
<-c("pmm","logreg","polyreg","pmm",
"polyreg","polyreg","logreg","polyreg",
"logreg","logreg","polyreg","logreg",
"logreg","logreg","logreg","logreg",
"pmm","pmm","pmm","pmm","pmm","logreg",
"logreg","logreg","logreg","logreg",
"logreg","logreg","logreg","polyreg",
"logreg","logreg","logreg","polyreg")

imp <- mice(mydata3pat,m=5,
maxit=10, printFlag=TRUE,pred=pred,
meth=meth,seed=10082015)
str(imp)
com <- complete(imp, 1)
com2 <- complete(imp, 2)
com3 <- complete(imp, 3)
com4 <- complete(imp, 4)
com5 <- complete(imp, 5)

########Section 4.3##########
#############################
#####RANDOM FORESTS##########
#############################

############Random Forests#########
library(randomForest)
library(lattice)
```

```
myvars <-
c("age","gender","smoke",
"Dur_pr_illness","Cough_sev","Phlegm_yn",
"Breath_sev","Runn_nose_yn","Fever_yn",
"Chest_pain_sev","Diarhh_yn",
"other_heart_yn","Gen_tox_yn",
"Dim_vesic_breath_yn","Beats_min",
"breaths_min","Syst_BP","Diast_BP",
"Oral_temp","crp","CRACKLES3_YN",
"infl_vacc_yn","lung_other_yn",
"diabetes_yn","day_unwell",
"Muscle_ach_yn","Headache_yn",
"Gen_unwell_yn","Interf_act_yn",
"examabnormal","Tachypnoea","pulsehigh",
"Temperature","asthma_family",
"Wheeze_sev",
"Chest_pain_sev","allergic_disease_yn",
"recurrcough","phlegmanycolour","crpgp",
"class")

newdata <- com[myvars]
###### classification trees##########
combinedrf<-randomForest
(class ~age+gender+
Temperature+smoke+Dur_pr_illness
+Cough_sev+Phlegm_yn+Breath_sev
+Runn_nose_yn+Fever_yn+Chest_pain_sev
+Diarhh_yn+other_heart_yn+
Gen_tox_yn+Dim_vesic_breath_yn+
Beats_min+
breaths_min+Syst_BP+Diast_BP+Oral_temp+
crp+CRACKLES3_YN+infl_vacc_yn+
lung_other_yn
+diabetes_yn+day_unwell+Muscle_ach_yn+
Headache_yn+Gen_unwell_yn+Interf_act_yn+
examabnormal+Tachypnoea+pulsehigh+
Temperature+asthma_family+Wheeze_sev+
Chest_pain_sev+allergic_disease_yn+
recurrcough+phlegmanycolour,data=newdata,
ntree=500, mtry=12,importance=TRUE,
nodeSize = 1000,norm.votes=TRUE,
jclasswt =0)

#########Mean decrease in accuracy

imp=as.data.frame(importance(combinedrf,
type=1))

library(lattice)
```

```
data <-as.data.frame(cbind(rownames(imp),
round(imp[,"MeanDecreaseAccuracy"],3)))
colnames(data)<-c("Parameters",
"MeanDecreaseAccuracy")
data$MeanDecreaseAccuracy <-
as.numeric(as.character
(data$MeanDecreaseAccuracy))

data$Parameters <-reorder
(data$Parameters,
data$MeanDecreaseAccuracy)
dotplot(Parameters ~
MeanDecreaseAccuracy,
data = data,aspect = 1.5,
xlab = "Mean Decrease Accuracy",
scales = list(cex = .6),
panel = function (x, y) {
panel.dotplot(x, y, col="black",lty = 2)
panel.abline( h=y[ which(x==0.300
lty = "dotted", col = "red")})

#######tuning to identify best number
of variables mtry###################

tuning <-tuneRF(newdata2 [ ,-41],
newdata2[ ,41], improve=.10,
ntreeTry=500,
mtryStart = 6, stepFactor = 0.5)
tuning
print(tuning)

########Section 4.4##########
#############################
#############################
#############################
#######Multinomial LR########
#############################

myvars <-c("SubjectCode","class",
"Country","Dur_pr_illness","PCN_name"
,"day_unwell","Wheeze_sev","Syst_BP",
"Fever_yn","age",
"smoke","Temperature","crp","Diast_BP",
"Dim_vesic_breath_yn","Muscle_ach_yn",
"pulsehigh","examabnormal",
"Chest_pain_sev",
"Tachypnoea","Phlegm_yn",
"Gen_tox_yn",
"asthma_family","recurrcough",
"Runn_nose_yn",
"Interf_act_yn","CRACKLES3_YN",
"Breath_sev","Gen_unwell_yn",
"Headache_yn","other_heart_yn",
"gender","infl_vacc_yn",
"phlegmanycolour","lung_other_yn")

finaldata<- comb[myvars]

library(mlogit)

long0=mlogit.data
(finaldata,shape="wide",
chid.var = "SubjectCode",
choice="class")

TM <- mlogit.data(long0,
choice = "class", shape = "long",
alt.levels = c("Pneumonia","Asthma",
"COPD","Influenza","Other"),
chid.var = "SubjectCode",
drop.index=TRUE)

simple0=mlogit(class ~ 1 |
Dur_pr_illness + day_unwell +
Wheeze_sev + Syst_BP +Fever_yn +
 age+smoke+crp+Diast_BP+
Muscle_ach_yn+examabnormal
+Chest_pain_sev+Tachypnoea+
Phlegm_yn+Gen_tox_yn+asthma_family+
recurrcough+Runn_nose_yn+Interf_act_yn+
CRACKLES3_YN+Breath_sev+
Dim_vesic_breath_yn+Gen_unwell_yn+
Headache_yn+other_heart_yn+
gender+infl_vacc_yn+phlegmanycolour+
lung_other_yn,TM,reflevel ="Other")

########Section 4.4###############
#################################
#################################
#################################
####cLUSTER ROBUST INFERENCE#####
#################################
#################################
Adjust the standard errors
to account for clustering
/ function#####################
#################################
finaldata$Country<-as.numeric
```

```
(finaldata$Country)

cl.mlogit<-function(logitmodel,cluster){
require(sandwich, quietly = TRUE)
require(lmtest, quietly = TRUE)
M <- length(unique(cluster))
N <- length(cluster)
K <- length(coefficients(logitmodel))
# dfc <- (M/(M-1))*((N-1)/(N-K))
dfc <- (M/(M-1))
uj  <- apply(estfun(logitmodel),2,
function(x) tapply(x, cluster,sum));
vcovCL <- dfc*sandwich(logitmodel,
meat.=crossprod(uj)/N)
coeftest(logitmodel, vcovCL) }


cl.mlogit(simpleb,finaldata$Country)


##########Interaction all with age
and crp*Dur_pr_illness->
none significant
Phlegm_yn*age +
CRACKLES3_YN*age +
Interf_act_yn*age
+Dur_pr_illness*crp +
+Breath_sev*age
p.value:0.06 Phglegm_yn * age
p.value:0.892 Breath_sev4*age
################################

########Section 4.5##########
#############################
#############################
#############################
#####Diagnostic Accuracy#####
#############################
#############################

x.sub1 <-com[com$class %in% c("4","2"),]

x.sub1$class<-relevel(x.sub1$class,
ref="4")m1resfinal<-glm(formula=class ~
Dur_pr_illness +day_unwell +
Wheeze_sev + Syst_BP + Fever_yn +
crp +age+smoke+Dim_vesic_breath_yn+
Muscle_ach_yn+Chest_pain_sev+Phlegm_yn+
Gen_tox_yn+recurrcough+Runn_nose_yn+
CRACKLES3_YN+Breath_sev+
crp+Gen_unwell_yn+other_heart_yn+
gender+lung_other_yn+ crp*Dur_pr_illness,
family = binomial(link = "logit")

multi.grace09intercept <-
glm(formula = class ~ +1 ,
family = binomial(link ="logit"),
data = x.sub1)
########forward selection Logistic
Regression
result=step(multin.grace09intercept,
scope=list(lower=multin.grace09intercept,
upper=m1resfinal), direction="forward")
################################
final model####################
m1resfinalb<-glm(formula = class ~
Dur_pr_illness
+ Fever_yn + smoke + Phlegm_yn +Syst_BP +
recurrcough + other_heart_yn +
day_unwell + gender,
family = binomial(link = "logit"),
data = x.sub1)
########## AUC##############
labels=x.sub1$class
scores=predict(m1resfinalb,
newdata=NULL, type="response")
scores=as.vector(scores)

########procedure for AUC
was repeated for
dataset without the duplicated
patients
test<-comb[!(duplicated
(comb$SubjectCode) |
duplicated(comb$SubjectCode,
fromLast = TRUE)), ]#######

library(ROCR)

pred <- prediction(scores, labels)
perf <- performance( pred, "tpr", "fpr")
auc.tmp <- performance(pred,"auc")
auc <- as.numeric(auc.tmp@y.values)
plot(perf, colorize=T)
abline(a=0, b= 1)
text(0.8,0.2, paste("AUC = ",
format(auc, digits=3)))
title("ROC curve Asthma
```

```
when reference is Influenza")

#####################
Cross Validation - 3 fold
#####################

library(DAAG)


val.daag <-CVbinary(m1resfinalb,
rand=NULL,
nfolds=3, print.details=TRUE)

my.cvfunc<-function (obj =
frogs.glm, rand = NULL, nfolds = 3,
print.details = TRUE, seed=NULL)
{
data <- obj$data
m <- dim(data)[1]
if (is.null(seed))
{
if (is.null(rand))
rand <- sample(nfolds, m,
replace=TRUE)
}
else {
set.seed(seed)
if (is.null(rand))
rand <- sample(nfolds,m,replace = TRUE)
}
form <- formula(obj)
yvar <- all.vars(form)[1]
obs <- data[, yvar]
ival <- unique(rand)
fam <- obj$family$family
hat <- predict(glm(form, data,
family = fam), type = "response")
cvhat <- rep(0, length(rand))
if (print.details)
cat("\nFold: ")
my.out<-vector("list",nfolds)
my.y<-vector("list",nfolds)
for (i in ival) {
if (print.details)
cat("", i)
if (i%%20 == 0)
cat("\n")
here <- i != rand
i.glm <-glm(form, data = data[here, ],
family = fam)
cvhat[!here] <-
predict(i.glm,newdata=data[!here,
], family = fam, type ="response")
my.out[[i]]<-cvhat[!here]
my.y[[i]]<-data[!here,yvar]
}
list(my.out,my.y)
}

testit<-my.cvfunc(m1resfinalb
,seed=10)


pred<-prediction(testit[[1]],
testit[[2]])

perf<-performance(pred,"tpr", "fpr")
plot(perf,col="grey82",lty=3)
plot(perf,lwd=1,avg="vertical",
spread.estimate="boxplot", add=TRUE)
abline(0,1,col=4,lty=3)
mtext('Model Influenza when reference
is Asthma',side=3,line=.5,
font=2)
##########Contigency table for Pneumonia
when reference is Other diagnoses
######################
Yhat<- predict(m1resfinal,type="response")
thresh  <- 0.5  # threshold for
dichotomizing according to
predicted probability
YhatFac <- cut(Yhat, breaks=c(-Inf,
thresh, Inf),
labels=c("5", "1"))
###############contingency table
cTab    <- table(x.sub1$class, YhatFac)
addmargins(cTab)# marginal sums
###############percentage
correct for data
sum(diag(cTab)) / sum(cTab)


#########################
##########AUC difference
test, model with CRP (pred1,lab1),
model without CRP (pred2,lab2)
pred: predicted prob.
lab: factor for pred.
```

```
##########################
##########################
AUC.test(pred1, lab1, pred2,
lab2, conf.level = 0.95,
paired = FALSE)

#####################
Within and between
variable imputation var
#####################
######u:se of each imp
u1<-0.003595
u2<-0.003647
u3<-0.003215
u4<-0.0054568
u5<-0.00327
U= 1/5*(u1+u2+u3+u4+u5)
######g:coef of each imp
g1<- -0.01489
g2<- -0.018617
g3<-  -0.011264
g4<- -0.4963507
g5<--0.01465
qbar<-1/5*(g1+g2+g3+g4+g5)
B<-1/4*((g1-qbar)+(g2-qbar)
+(g3-qbar)+(g4-qbar)+
(g5-qbar))^2
T=U+(1+(1/5))*B
m=5
m1=m-1=4
df=(m1)*(1+((m*U)/
((m+1)*B)))^2
t=qbar/sqrt(T)
2*pt(-abs(t),df=df)



#######Calibration plot
#######observed versus
predicted
####p:scores/estimated/
predicted probabilities
####y:response 0:1
calibration<- function (y, p,
main="Title"){
newp <- seq(0, 1, length=100)
yy <- predict(loess(y ~ p, span=1),
newp, se=T)

yy.ok <- !is.na(yy$fit)
yy$fit <- yy$fit[yy.ok]
yy$se.fit <- yy$se.fit[yy.ok]
newp <- newp[yy.ok]
se.lower <- yy$fit - 2 * yy$se.fit
se.upper <- yy$fit + 2 * yy$se.fit
par(pty="s")
plot(c(0,1), c(0,1), type="n",
xlab="Predicted Probability",
ylab="Observed Proportions",
xaxs="i", yaxs="i", las=1, main=main)
polygon(c(newp, rev(newp), newp[1]),
c(se.lower, rev(se.upper), se.lower[1]),
col = "gray", border = NA)
rug(p[y == 0], side=1, col="navy")
rug(p[y == 1], side=3, col="navy")
abline(0, 1, col="red")
abline(h=0.2, col="yellow", lty=2)
lines(newp, yy$fit, lwd=2, col="blue")
par(pty="m")
 }

####Three way HUM########
########################
Code by Li. J.##########
########################
########################
available at:
http://www.stat.nus.edu.sg/~stalj/
########################
########################

ThreeHUM=function(y,d){

#y is the tri-nomial response,
i.e., a single vector taking
three distinct values,
can be nominal or numerical
#d is the continuous marker

#x1 is position of
observations from the 1st category
#x2 is position
of observations
from the 2nd category
#x3 is position
of observations
from the 3rd category
x1=which(y==1)
```

```
x2=which(y==4)
x3=which(y==5)
n=length(y)

#n is the sample size
a=matrix(0,n,3);
one1=a;
one1[,1]=1;
one2=a;
one2[,2]=1;
one3=a;
one3[,3]=1;

library(nnet)

fm=multinom(y~d)

#extract the probablity
assessment vector
pp=fm$fitted;

dd1=pp-one1;
dd2=pp-one2;
dd3=pp-one3;

jd1=sqrt(dd1[,1]^2+
dd1[,2]^2+dd1[,3]^2);
jd2=sqrt(dd2[,1]^2+
dd2[,2]^2+dd2[,3]^2);
jd3=sqrt(dd3[,1]^2+
dd3[,2]^2+dd3[,3]^2);
jd1=exp(jd1);
jd2=exp(jd2);
jd3=exp(jd3);

mt1=kronecker
(jd1[x1]%*%t(jd2[x2]),jd3[x3]);
mt2=kronecker
(jd1[x1]%*%t(jd3[x2]),jd2[x3]);
mt3=kronecker
(jd2[x1]%*%t(jd1[x2]),jd3[x3]);
mt4=kronecker
(jd2[x1]%*%t(jd3[x2]),jd1[x3]);
mt5=kronecker
(jd3[x1]%*%t(jd2[x2]),jd1[x3]);
mt6=kronecker
(jd3[x1]%*%t(jd1[x2]),jd2[x3]);

cr=sum(mt1==pmin(pmin(
pmin(pmin(pmin(mt1,
mt2), mt3), mt4), mt5), mt6));

#hypervolume under
ROC manifold value
hum=cr/(length(x1)*
length(x2)*length(x3));

return(hum)


}
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Diagnosing pneumonia, influenza and obstructive pulmonary diseases in adult patients presenting to primary care with acute cough: a multinomial logistic regression analysis

Richting: **Master of Statistics-Biostatistics**
Jaar: **2015**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Mesiri, Pavlina

Datum: **8/09/2015**