Master of Statistics: Biostatistics

ASSESSING THE PERFORMANCE OF CASE CLASSIFICATION AND VALIDATION ALGORITHMS; AND SAMPLE SIZE ESTIMATION FOR CLASSIFICATION ALGORITHMS' VALIDATION

By

Forsi Nwebim Boeyeo

Internal Supervisor: Prof. dr. Ziv SHKEDY

External Supervisor: Dr Dominique ROSILLON

Submitted in partial fulfilment of the requirements for the degree of Masters of Statistics: Biostatistics

January 30, 2015

DECLARATION

The thesis titled “*Assessing the Performance of Case Classification and Validation Algorithms; and Sample Size Estimation for Classification Algorithms’ validation”* is the research work of *Forsi Nwebim Boeyeo.* He is solely responsible for all errors, omissions and misinterpretation of facts. The contributions of others are mentioned in the references.

Forsi Nwebim Boeyeo

…………………….
(Student)

Prof. dr. Ziv SHKEDY	Dr Dominique Rosillon
…………………….	…………………….
(Internal Supervisor)	(External Supervisor)

ACKNOWLEDGEMENT

I will always remain heavily indebted to the Almighty God for sailing me safely throughout my studies in Hasselt University. It goes without saying that I appreciate Prof. dr. Ziv SHKEDY and Dr Dominique ROSILLON for dedicating their precious time and energy to correcting my work and guiding me throughout this thesis.

The immense effort of the entire staff of CENSTAT Hasselt University is highly appreciated. I won’t hesitate to extend my sincere appreciations to GSK Vaccines for giving me the opportunity to conduct my thesis in their institution. Special thanks to Martine MACHIELS, Leen JORISSEN, Inge JACOBS, Tatsiana KHAMIAKOVA and Germano FERREIRA for always being around when I am stuck.

I am grateful for my family for their huge support in making sure I am successful in my studies. Special thanks to my parents, Lum Julia Etima, Ngum Beybunui Boeyeo, Germain Etima and Akuo Cheenuwenui Boeyeo.

To all the good people who have always been there to make sure I succeed; Jennifer Diony, Ngu Blessing, Kurnia Wahyudi, Ketunze Brain, Silenou Bernard, Adelino Martins, Abang Felix, Belay Belete and you, I say thank you. Special appreciation goes to Wandji Nadege for always encouraging me when I am down.

ABSTRACT

This report describes the application of classification techniques for classifying adverse pregnancy outcomes (spontaneous abortion) of pregnant women exposed to the Human Papillomavirus (HPV) vaccine based on data obtained from the Clinical Practice Research Database (CPRD) and also to derive a statistical method to estimate the sample size required for the validation of classification algorithms. 782 subjects were extracted and randomly divided into approximately five equal-sized parts making sure only one part is used for testing and the remainder for training. Logistic regression models and a random forest model were built to predict the binary outcome variable; spontaneous abortion. To estimate the sample size required for the validation of classification algorithms, the prevalence, sensitivity and specificity were taken account of in the sample size formula. A predefined algorithm had previously been used to classify the subjects and resulted in a sensitivity of 74.2% with a 95% confidence interval of 64.4% - 82.6% and a specificity of 79.7% with a 95% confidence interval of 76.5% - 82.7%. Also, this algorithm had a correct classification rate of 79.0%. By taking into account other diagnostic information from the database, classification models were constructed. The best model was the model with predefined diagnosis, induced abortion and fullterm delivery as inputs and it had an improved sensitivity of 95.0% and specificity of 99.3% with a corresponding 95% confidence interval of 75.1% – 99.8% and 96.0% - 99.9% respectively, with a correct classification rate of 98.7%. Also, this model notably had a low false negative rate which was of interest due to the low prevalence of spontaneous abortion in the study population. The classification models generally had high predictive values. The confidence interval for the difference in the correct classification rates was 16.3% - 23.1%, suggesting that more information from the database improved the classification performance. Furthermore, it was estimated that 589 subjects were required to obtain the positive predictive value of 95.0% with a clinically acceptable precision of 5%.

Keywords: Spontaneous abortion, Logistic regression, Random forest, Sensitivity, Specificity, Predictive values, Sample size.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 Background

Classification is the task of assigning objects into predefined classes or groups (Fielding, 2007). This process involves the use of techniques designed to find models able to recognize the membership of each record to its proper class on the basis of a set of attributes and once a classification model has been obtained, the membership of unknown objects to one of the defined classes can be predicted (Ballabio & Todeschini, 2009). In this sense, a classification model or a classifier can be seen as a black box that automatically assigns a class label when presented with the attribute set of an unknown record. Classification techniques are most suited for predicting or describing data sets with binary or nominal categories as they do not consider the implicit order among the categories (Tan *et al*., 2005). They make use of a learning algorithm to identify a model that best fits the relationship between the attribute set and the class label of the input data. Usually, the data is divided into two sets; one used to train the classifier or learn the model known as the training data and the other used to assess the predictive strength of the chosen model known as the test set (Tan *et al*., 2005).

A Classifier is judged by its ability to make accurate future predictions. There is nothing to gain if a classifier's predictive ability is restricted to the data that were used to construct it. In this sense, a good classifier or classification model should be able to learn the general features of the training data so that it can make accurate predictions when presented with novel cases. Assessment of the performance of a classifier is of great importance in practice as it guides us in choosing a learning method and gives us an idea of the quality of our chosen classification method. In the medical sciences, classification can be seen as making a diagnosis. The evaluation of the performance of a classification model, in other words diagnostic test, is based on the number of cases correctly and incorrectly predicted or diagnosed. This is usually tabulated in a confusion matrix from which accuracy measures such as misclassification rates, sensitivity, specificity, positive predictive and negative predictive values are computed.

In the light of epidemiology, sensitivity can be defined as the probability that a diagnostic test is positive given a subject has the disease while specificity is the probability that a diagnostic test is negative given a subject has no disease (Agresti, 2007; Akobeng, 2007). A highly sensitive diagnostic tool means many non cases will be misclassified as cases (high false positive rate) ,while a highly specific diagnostic tool will misclassify many cases as non cases (false negatives). Achieving an optimal sensitivity and specificity for a diagnostic tool necessitates a trade off between the sensitivity and specificity. This trade off depends on the consequences of having false positive or false negative cases and this has generated a lot of debate in the field of epidemiology. If getting a false positive is tolerable and without major consequences, but a false negative has detrimental consequences, then one would be contented with a diagnostic tool that has a high false positive rate. Conversely, if getting a false negative has no major consequences, but a false positive has detrimental consequences, then one would prefer to have a high false negative rate (Ioannidis *et al*., 2011).In a broader sense, tolerating high false positive rates or high false negative rates requires a systematic appraisal on a case by case basis and this often extends beyond epidemiology, taking into account public health policies and political considerations(Ioannidis *et al*., 2011). A more important measure of performance of a diagnostic tool is the positive predictive value which is the probability that a disease is present given that a diagnostic tool is positive or the probability of a case given a classifier predicts a case (Agresti, 2007). Also, we desire a classifier with a high positive predictive value. The positive predictive value is directly related to the number of people with the disease in question (cases) and will increase with increasing prevalence (Parikh *et al*., 2008).

Classification methods are increasingly used in several fields including the social, economic, medical and pharmaceutical sciences and they have become very popular in various automatic medical diagnostic tools. It has been reported that, using multiple diagnostic codes in combination with medication and other information may improve the accuracy of case classification (Cooke *et al*., 2011, Quint *et al*., 2014) . In this report, we aim to assess and try to improve the performance of some diagnostic tools or classifiers used for detecting adverse pregnancy outcomes in pregnant women exposed to the Human Papillomavirus(HPV) vaccine based on multiple diagnostic and medication codes obtained from the Clinical Practice Research Database (CPRD) and also in determining an appropriate sample size for validating these diagnostic tools.

1.2 Objectives

Several techniques could be used to diagnose or classify subjects as having a spontaneous abortion or not and these techniques usually have varying degrees of performances. This variation in performance can be seen as a reflection of the shortcomings of one technique or the other. Given that there is an inherent shortcoming in every classification technique, in order to choose a more predictive one, it makes sense to use different classification techniques to construct algorithms and compare their performances. In this study, we sought to develop improved algorithms for a more accurate identification of adverse pregnancy outcomes as compared to that of a predefined algorithm.

The main objectives of this study was to assess the performance of some case classification algorithms in classifying adverse pregnancy outcomes (spontaneous abortion) of pregnant women exposed to the Human Papillomavirus (HPV) vaccine, based on data obtained from the CPRD database and also to derive a statistical method to estimate the sample size required for the validation of classification algorithms.

2.0 METHODOLOGY

2.1 Data

The dataset used in this study was extracted from the Clinical Practice Research Database (CPRD) in the UK. The CPRD offers unique opportunities for health researchers to draw on the power of large multi-linked observational datasets on a previously unprecedented scale (CPRD, 2014). It is the world's largest validated computerised database of anonymised longitudinal medical records for primary care (Williams *et al.*, 2012). Records are derived from a widely used general practitioner software system and contain complete prescribing and coded diagnostic and clinical information as well as information on tests requested, laboratory results and referrals made at or following on from each consultation(Tate *et al*., 2013). The CPRD provides an excellent resource in which to conduct epidemiological studies as they offer a large sample size, the presence of disease severity indicators and long-term follow-up information on a patient's integrated medical history (Quint *et al*., 2014).

In this study, 1046 pregnant women aged between 15 and 25 years residing in the United Kingdom and who reported their last menstrual period between September 2008 and December 2010 were extracted from the CPRD. For the analysis set, 782 (97 with and 685 without spontaneous abortion) pregnant women who had reference labels of the outcome were considered, because to train and validate predictive models, a source of "gold standard" cases with known status is needed (Slipchenko *et al*., 2008). The women had been classified using a predefined algorithm. The classifications were reviewed by two independent experts and by a review of patient profiles and free texts by an independent organisation, PALLAS.

Table 1: *Attributes extracted from the dataset.*

	Attribute	**Type**
1.	Patient identity	Character
2.	Induced abortion	Categorical
3.	Other abortion	Categorical
4.	Stillbirth	Categorical
5.	Preterm delivery	Categorical
6.	Postterm delivery	Categorical
7.	Full term delivery	Categorical
8.	Predefined diagnosis of spontaneous abortion	Categorical
9.	Contraindicated drug	Categorical
10.	Drugs during 1st trimester	Categorical
11.	Ultrasound test at week 12 of gestation	Categorical
12.	Other vaccines received	Categorical
13.	Alcohol consumption	Categorical
14.	Smoking	Categorical

The response variable(Y_i), was the diagnosis of spontaneous abortion in a pregnant woman who had been exposed to the Human Papillomavirus vaccine.

$$Y_i = \begin{cases} \text{Yes}, & \text{if } \text{i}^{\text{th}} \text{ woman had spontaneous abortion} \\ \text{No}, & \text{if } \text{i}^{\text{th}} \text{ woman did not have spontaneous abortion} \end{cases}$$

The gold standard was the classification by PALLAS and the experts.The attribute set considered in this study are presented in Table 1. Approximately 80 % of the data rows had incomplete observations in the attributes. Missing values in the attribute set were imputed using the R function “rfImpute” in the random forest package (Liaw & Weiner, 2014).

2.2 Method of statistical analysis

2.2.1 Characteristics of diagnostic algorithms

The characteristics of a diagnostic algorithm are a reflection of their performance and a choice of which algorithm to use is always based on such characteristics. Common characteristics of classification algorithms include sensitivity, specificity, positive predictive value, negative predictive value and misclassification rate. The sensitivity of a diagnostic algorithm is the probability that the algorithm predicts a spontaneous abortion given a true case of a spontaneous abortion while the specificity is the probability that the algorithm predicts a non-spontaneous abortion given a true non-spontaneous abortion (Agresti, 2007). On the other hand, the positive predictive value which is a more important characteristic to clinicians, is the proportion of cases of spontaneous abortion predicted that are truly spontaneous abortion cases while the negative predictive value is the proportion of non-spontaneous abortion cases that are actually not cases. Like all other diagnostic tools, diagnostic algorithms are not error-free. The false positive rate which is the proportion of true non-spontaneous abortion cases classified as spontaneous abortion and false negative which is the proportion of true spontaneous abortion cases classified as non-spontaneous abortion by the diagnostic algorithms (Slipchenko *et al.*, 2008), were used as a measure of the errors committed.

The sensitivity, specificity, positive predictive value, negative predictive value, false positive rate and false negative rate with their respective 95% confidence intervals were computed for the classification of pregnant women into a spontaneous abortion (Yes) or non-spontaneous abortion (No) class by the pre-defined algorithm. Using logistic regression and random forest techniques, algorithms/classifiers were constructed using the attributes in the dataset with the aim of improving the diagnosis of the predefined algorithm.

	Gold Standard(Spontaneous abortion)	
Algorithm's Diagnosis	Yes	No
Yes	Correct prediction (True positive) A	Incorrect prediction (False positive) B
No	Incorrect prediction (False negative) C	Correct prediction (True negative) D

Figure 1 : *Confusion Matrix of Algorithm's Prediction of Spontaneous Abortion*

Table 2: *Confusion matrix-derived characteristics of diagnostic algorithms*

Characteristic	Formula
Sensitivity	$\frac{A}{(A+C)}$
Specificity	$\frac{D}{(B+D)}$
Positive predictive value	$\frac{A}{(A+B)}$
Negative predictive value	$\frac{D}{(C+D)}$
False positive rate	$\frac{B}{(B+D)}$
False negative rate	$\frac{C}{(A+C)}$

N = *A*+*B*+*C*+*D*

The algorithms were used to classify the subjects into a “yes” and a “No” class and using the gold standard, the counts of “Yes” and “No” correctly and incorrectly predicted by the algorithms were summarised in a confusion or error matrix that cross-tabulates the actual and predicted response patterns as shown in Figure 1 (Fielding, 2007). The different algorithms were compared in terms of their ability to generalize. The sensitivity, specificity, positive predictive value, negative predictive value, false negative and false positive rates were also computed with their corresponding 95% confidence intervals and the formulae used for their computations are presented in Table 2.

2.2.2. Classification Techniques

There are many different classification techniques available for predicting categorical responses. In this report, logistic regression which is a common technique for predicting binary responses was applied for the prediction of subjects as having spontaneous abortion or not. Also, a popular computer-intensive tree-based method, the random forests, which is well known for its ability not to over fit, was used in this report. Since a good classifier requires a relatively low bias and low variance, a trade off is required between a complex model with low bias but high variance and a simple model with high bias but low variance. This can be achieved through variable selection. A basic plan is needed for selecting the variables for the models and methods for assessing the adequacy of the model both in terms of its individual variables and its overall performance (Hosmer *et al*., 2013). Since the Random Forest provides information on variable importance, it was used to complement variable selection (Calle & Urrea, 2010). Univariate logistic regression models were fitted to get the association of each variable with the response. A forward selection was performed by adding the variables in turn starting with the most significant. The dataset was randomly divided into a training set which was used to build the models and a test set, that was used for validation (Hastie *et al.*, 2001; James *et al.*, 2013). The proportion of data reserved for training is usually at the discretion of the analyst and in this report one-third was reserved for testing and the rest for training. Five-fold cross-validation technique was applied to obtain more reliable predictions and generalization errors by dividing the data into approximately five equal-sized parts making sure only one part is used for testing and the remainder for training (Tan *et al*., 2005 ; Hastie *et al*., 2001).

2.2.2.1 Random Forests

Classification trees can easily handle qualitative and quantitative variables but unfortunately, they are unstable (Breiman, 1996). However, by aggregating many classification trees using an ensemble method like random forests, the predictive performance can be substantially improved (James *et al.*, 2013).

Random forest was fitted by varying the number of predictors considered for each split (m) , choosing the one with the smallest out-of-bag error rate (Hastie *et al*., 2009) and fitting 1000 different trees. Assignment into the class of spontaneous abortion (Yes) and non-spontaneous abortion (No) by the random forest was accomplished by voting of the individual random-forest trees. A subject with the highest number of votes for a given response class (Yes or No) was assigned into that class and the number of class assignments were counted and tabulated on a confusion matrix from which the error rates, sensitivity, specificity and predictive values were computed.

Let $\hat{Y}_{i_b}(\mathrm{x})$ be the spontaneous abortion class prediction for the i^{th} pregnant woman by the b^{th} random-forest tree. Then the random forest classification rule is of the form:

$$\hat{Y}_{i_{rf}}^{B}(x) \ = majority\ vote\ \{\ \hat{Y}_{i_b}(\mathrm{x})\}_1^B \tag{1}$$

(Hastie *et al*., 2009)

Where

B is the number of trees, $\hat{Y}_{i_b}$(x) is the spontaneous abortion class prediction of the i^{th} pregnant woman by the bth tree and $\hat{Y}_{i_{rf}}^{B}(x)$ is the spontaneous abortion class prediction of the i^{th} pregnant woman by Random Forest.

Random Forest was used for classification in this report because they have been shown to be efficient on large databases, naturally handle categorical predictors, can handle large numbers of predictors without selection routines and the last but not the least; it provides information on the importance of variables used (Fielding, 2007). Furthermore, random forests are an effective tool in prediction (Breiman, 2001 ; Lehmann *et al*., 2007). Despite its numerous advantages, random forests could be difficult to interpret.

2.2.2.2 Logistic Regression

Logistic regression models were used for predicting the probability of having a spontaneous abortion given the inputs. The parameter estimates of the logistic regression are unknown and were estimated based on the training data. The logistic regression model can be written as:

$$log\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad (2)$$

Where

$\boldsymbol{X} = (X_1, \dots, X_p)$ are the p inputs in the model, $\beta_0, \beta_1, \dots, \beta_p$ are unknown parameters and $p(\boldsymbol{X}) = Pr(Y = \text{Yes}/X)$ is the probability of having a spontaneous abortion given the inputs.

The maximum likelihood method was used to estimate the parameters of the models such that the predicted probabilities of subject having a spontaneous abortion was as close as possible to their true probabilities of having a spontaneous abortion (Hastie *et al*., 2001; Kutner *et al*., 2005). The likelihood function used for the estimation of the parameters of the fitted logistic regression models is of the form:

$$L(\beta) = \prod_{i=1}^{n} p(X_i)^{Y_i}[1 - p(X_i)]^{1-Y_i} \qquad (3)$$

Where

$L(\beta)$ is the likelihood of having a spontaneous abortion, and $p(X_i) = Pr(Y_i = Yes/\, X_i)$ is the probability of the i^{th} pregnant woman having a spontaneous abortion given the inputs.

The Logistic regression classification rule is of the form:

$$\hat{Y}_i = \begin{cases} \text{Yes} & \text{if } p(X_i) > 0.5 \\ \text{No} & \text{Otherwise} \end{cases}$$

Where

$\hat{Y}_i$ is the spontaneous abortion class prediction of the i^{th} pregnant woman.

If the predicted probability, $p(X_i) > 0.5$, the subject was classified in the class of spontaneous abortion (Yes) otherwise she was classified as not (No). The class predictions were cross-tabulated with the gold standard in a confusion matrix from which the false positive rate, false negative rate, sensitivity, specificity, and predictive values were computed. This was done for all possible models. The advantage of using logistic regression is that, logistic regression easily accommodates the binary nature of the response (spontaneous abortion(Y_i) = "Yes" or "No") and also, it uses the method of maximum likelihood for parameter estimation which has better statistical properties (James *et al*., 2013).

2.2.3 Sample Size Computation

The aim of the sample size calculation was to determine an adequate number of subjects required to validate a pre-specified estimate of positive predictive value with a specified precision (Naing *et al*., 2006). The positive predictive value depends on the characteristics of the diagnostic test (sensitivity and specificity) and also on the prevalence of the disease under study (Akobeng, 2007) as shown in equation (4).

$$PPV = \frac{Sn(P)}{Sn(P) + (1 - Sp)(1 - P)} \tag{4}$$

In the light of this interrelationship, the estimation of the required sample size to obtain a predefined predictive value with a given precision was done taking into account the sensitivity, specificity and prevalence. Using a predefined specificity and sensitivity of a diagnostic test, the sample sizes for the sensitivity and specificity were computed using equations (5) and (6). The maximum of the sample sizes for the specificity and sensitivity was considered as the sample size for approximating the corresponding positive predictive value. The sample size formulae used to estimate the predefined sensitivity and specificity is of the form:

$$N_{Sn} = \frac{Z^2_{\alpha/2} Sn(1-Sn)}{PL^2} \tag{5}$$

$$N_{Sp} = \frac{Z^2_{\alpha/2} Sp(1-Sp)}{(1-P)L^2} \tag{6}$$

(Buderer, 1996 ; Jones *et al*., 2003 & Malhotra, 2010)

Where

$N_{Sn}(N_{Sp})$ is the expected number of subjects to validate a specified sensitivity (specificity),

$Sn(Sp)$ is the specified sensitivity (specificity)

α is the significance level fixed at 5 %

L is the precision (width of confidence interval)

P is the prevalence

PPV is the positive predictive value

Since obtaining data is very expensive, by estimating the sample size adequately, the unnecessary expenditure in obtaining excess data will be reduced. On the other hand, if the data is less than the required number to validate the specified positive predictive value, it would result in a waste of time, resources and an unreliable estimate of interest (Hajian-Tilaki, 2011).

In this report, all statistical analyses were carried out using SAS 9.2 and R version 3.1.2.

3.0 RESULTS

3.1 Predefined Algorithm

Table 3: *Confusion Matrix derived from predefined algorithm's prediction of spontaneous abortion*

Predicted diagnosis	**Reference diagnosis**		
	Yes	No	Total
Yes	72	5	77
No	4	546	550
Unknown	21	134	155
Total	97	685	782

Table 4: *Performance characteristics of the predefined algorithm*

Algorithm	**Se%(CI)**	**Sp%(CI)**	**PPV%(CI)**	**NPV%(CI)**	**FP%(CI)**	**FN%(CI)**
Predefined	74.2 (64.4,82.6)	79.7 (76.5,82.7)	93.5 (85.5,97.9)	99.3 (98.2,99.8)	0.7 (0.2,1.7)	4.1 (1.1,10.2)

Se = sensitivity, Sp = specificity, PPV = positive predictive value, NPV = negative predictive value, FP = false positive rate and FN = false negative rate, CI = 95% confidence interval.

Tables 3 and 4 present the performance of the predefined algorithm's diagnosis of spontaneous abortion compared to the gold standard. The predefined algorithm rightly predicted 72 cases of spontaneous abortion and 546 cases without spontaneous abortion giving a sensitivity of 74.2% with a 95% confidence interval of 64.4% - 82.6% and a specificity of 79.7% with a 95% confidence interval of 76.5% - 82.7%. Five pregnant women without spontaneous abortion were misclassified as having spontaneous abortion while 4 with spontaneous abortion were misclassified as not having a spontaneous abortion giving a false negative and positive rate with their corresponding 95% confidence intervals of 4.1 % (1.1 - 10.2) and 0.7% (0.2 - 1.7) respectively. Furthermore, despite the inability of the algorithm to allocate 155 pregnant women to either the class of pregnant women who had a spontaneous abortion or not, it had a high positive predictive value of 93.5% with a corresponding 95% confidence interval of 85.5 – 97.9 and a corresponding high negative predictive value of 99.3% (98.2 - 99.8).

3.2 Random Forests

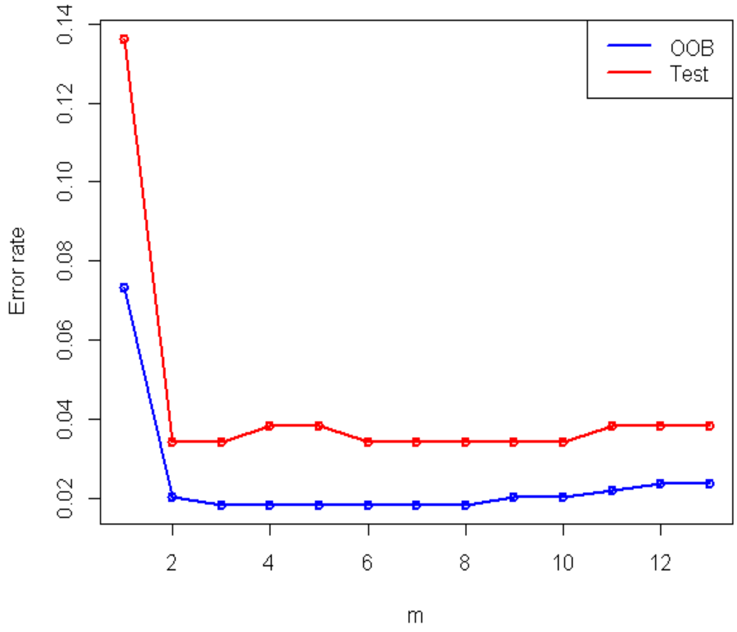

Figure 2 : *Using Out-of-bag and test errors to choose the best m*

The choice of the number of variables to be used for splitting the subjects into the class of spontaneous abortion and non-spontaneous abortion in the random forest (m) was based on the out-of-bag error and test error as presented in Figure 2. From the figure, the out-of-bag error and test error were high when only one variable was considered for each split($m = 1$), but decreased sharply as the number of variables to be used for each split increased to 2. The out-of-bag error rate reached its lowest point when three variables were considered for each split ($m = 3$) and at the same point, the test error was considerably low before slightly increasing. Furthermore, by using all the variables for splitting the subjects, the out-of-bag error and test error rates were higher than using only three variables.

The variable importance plot is a very important plot obtained from the Random Forest. Figure 3 presents the importance of each variable in the random forest algorithm based on mean decrease in accuracy of predictions when a given variable is excluded from the algorithm(mean decrease in accuracy) and the total decrease in node impurity that results from splits over that variable, averaged over all trees in the forest(mean decrease in Gini). The variables are ordered as most important to least important from top to bottom. Based on the mean decrease in Gini, among all the variables considered in the random forest, the variable Predefined diagnosis was indisputably the most important variable with a mean decrease in Gini of approximately 60, followed by the variables Fullterm delivery and

Induced abortion. There were some discrepancies between the two plots but based on variable significance in logistic regression (Table A3 and A4) and mean decrease in Gini from the variable importance plot, Predefined diagnosis, Fullterm delivery and Induced abortion were the most significant variables and were considered for further analyses.

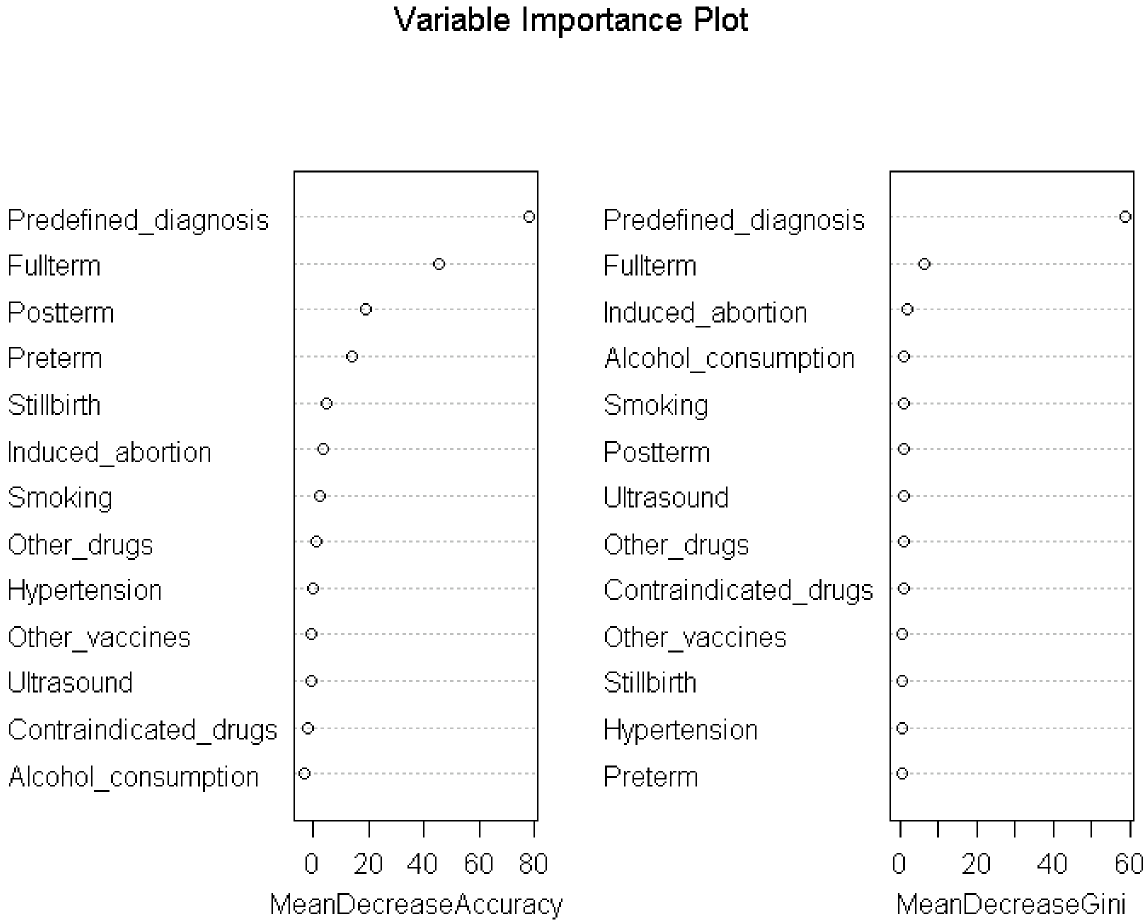

Figure 3 : *Variable Importance Plot for a RandomForest Algorithm*

Table 5 : *Confusion Matrix derived from RandomForest Algorithm (Classification of spontaneous abortion)*

Predicted	**Reference diagnosis**		
diagnosis	Yes	No	Total
Yes	19	1	20
No	1	136	137
Total	20	137	157

Table 6: *Performance characteristics of the Random Forest Algorithm*

Algorithm	Se%(CI)	Sp%(CI)	PPV%(CI)	NPV%(CI)	FP%(CI)	FN%(CI)
Random forest	95.0 (75.1, 99.8)	99.3 (96.0, 99.9)	95.0 (75.1,99.8)	99.3 (96.0,99.9)	0.7 (0.02,4.0)	5.0 (0.1,24.8)

Se = sensitivity, Sp = specificity, PPV = positive predictive value, NPV = negative predictive value, FP = false positive rate and FN = false negative rate, CI = 95% confidence interval

Classification of spontaneous and non-spontaneous abortions in pregnant women by the random forest is presented in Table 5 from which the characteristics of interest were computed as presented in Table 6. Based on the average of the 5-fold cross-validations, the random forest rightly predicted 19 cases of spontaneous abortion and 136 cases without spontaneous abortion giving a sensitivity of 95.0% with a corresponding 95% confidence interval of 75.1% - 99.8% and a specificity of 99.3% with a 95% confidence interval of 96.0% - 99.9%. Only 1 pregnant woman with true spontaneous abortion was misclassified as not having a spontaneous abortion and 1 without spontaneous abortion misclassified as having spontaneous abortion resulting in false negative and positive rates with their corresponding 95% confidence intervals of 5.0 % (0.1 – 24.8) and 0.7% (0.02 – 4.0) respectively. Furthermore, the random forest had high values of the positive predictive and negative predictive values of 95.0 %(75.1 - 99.8) and 99.3 % (96.0 – 99.9) respectively.

3.3 Logistic regression

Figure 4 presents the boxplots of predicted versus actual spontaneous abortion status for the logistic regression models with Predefined diagnosis only and a logistic regression model with Predefined diagnosis, Induced abortion and Fullterm delivery. The model with predefined diagnosis had higher predicted probabilities for spontaneous abortion than non-spontaneous abortion even though with a wide range, thus capable of discriminating between the two classes of spontaneous abortion as expected. Furthermore, when the other significant variables (Fullterm delivery and Induced abortion) were added to this model, it became even more discriminatory with higher predicted probabilities for spontaneous abortion cases.

The actual predictions of spontaneous abortion cases depend on the threshold probability for separating cases from non-cases. Based on the model with the most predictive variable (Predefined diagnosis) and a model with all the significant variables (Final model), a probability threshold for classification was derived.

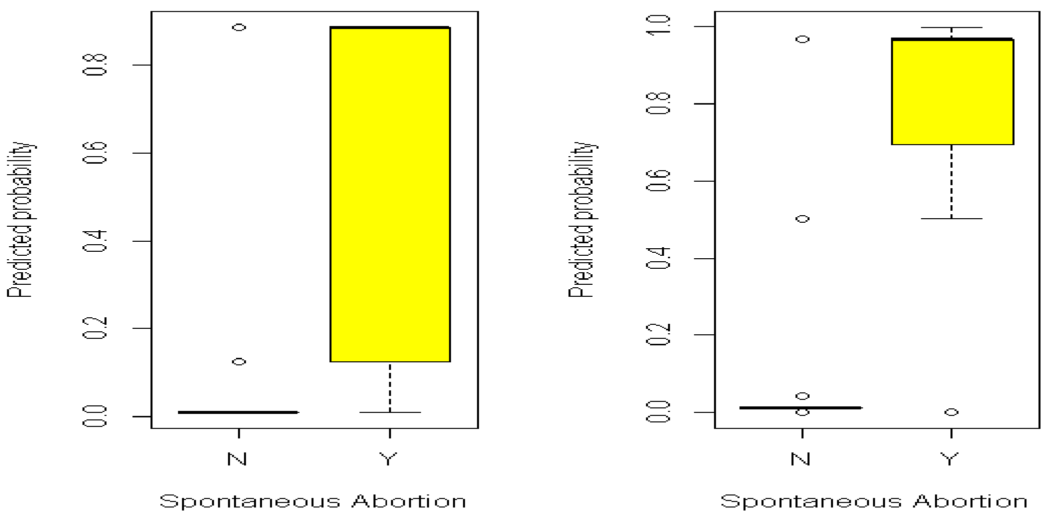

Figure 4: *Boxplot of predicted probabilities versus actual status of Spontaneous abortion for model with Predefined diagnosis only (left) and model with Predefined diagnosis, Induced abortion and Fullterm delivery (right)*

Table 7: *Performance Characteristics for different probability thresholds*

Model	**Probability threshold**	**Se (%)**	**Sp (%)**	**MR (%)**
Predefined diagnosis only	0.1	100	77.8	18.3
	0.2	80.5	100	3.4
	0.5	80.5	100	3.4
Final (All significant variables)	0.1	95.1	100	0.9
	0.2	95.1	100	0.9
	0.5	95.1	100	0.9

Se = sensitivity, Sp = specificity, MR = Misclassification rate

Table 7 presents a range of possible probability thresholds and their corresponding performance characteristics. Using a reasonable conservative probability threshold of 0.1 resulted in a misclassification rate of 18.3% in the model with predefined diagnosis only and 0.9% in the final model (model with all significant variables). The misclassification rate was stable from a probability threshold of 0.2 upwards, for this reason, the default threshold of 0.5 was retained and used for all analyses.

The results of prediction and performance characteristics from the logistic regression model including all three significant attributes (final model); Predefined diagnosis, Induced abortion, and Fullterm delivery, the models with Predefined diagnosis only, Induced abortion only and Fullterm delivery only are presented in Tables 8, 9, 10, A1 and A2. The model with all three significant variables (Table 8) correctly predicted 19 out of 20 cases of true spontaneous abortion and 136 out of 137 cases of non-spontaneous abortion resulting in an acceptable false negative rate of 5.0 % with a corresponding 95% confidence intervals of (0.1 – 24.8) and a false positive rate of 0.7% (0.02 – 4.0) (Table 10). This model had a high specificity of 99.3% (96.0 - 99.9) and a high sensitivity of 95.0% (75.1 – 99.8). Also, this model produced high positive and negative predictive values of 95.0% (75.1 – 99.8) and 99.3% (96.0 - 99.9) respectively. On the other hand, the model with Predefined diagnosis only (Table 9) correctly predicted 15 out of 20 cases of true spontaneous abortion and 136 out of 137 cases of non-spontaneous abortion resulting in a false positive and false negative rates with their corresponding 95% confidence intervals of 0.7% (0.02 – 4.0) and 25.0% (8.7 – 49.1) respectively. Also, this model had a relatively high sensitivity of 75.0% (51.0 – 91.3) and high specificity of 99.3% (96.0, 99.9). The model with Induced abortion only (Table A1) correctly predicted only 2 out of 20 cases of true spontaneous abortion but 135 out of 137 cases of non-spontaneous abortion resulting in an unacceptable false negative rate of 90.0 % with a corresponding 95% confidence intervals of (68.3 – 98.8) and a false positive rate of 1.5% (0.2 – 5.2). This model had a very poor sensitivity of 10.0 %(1.2, 31.7) and positive predictive value of 50.0% (6.8, 93.2). Despite the significance of the variable Fullterm delivery, the model with Fullterm delivery could not differentiate spontaneous abortion cases from non cases and thus classified all subjects as not having a spontaneous abortion (Table A2). This model had an unacceptable sensitivity of 0% and specificity of 100% (97.3 - 100) with an unacceptable false negative rate of 100% (83.2 - 100). Despite the high specificity, this model was undoubtedly useless as it was unable to discriminate between spontaneous abortion and non-spontaneous abortion cases.

Table 8: *Confusion Matrix derived from Logistic Regression Model with Predefined diagnosis, Induced abortion and Fullterm delivery (Final Model)*

Predicted	**Reference diagnosis**		
diagnosis	Yes	No	Total
Yes	19	1	20
No	1	136	137
Total	20	137	157

Table 9: *Confusion Matrix derived from Logistic Regression Model with Predefined diagnosis only*

Predicted diagnosis	**Reference diagnosis**		
	Yes	No	Total
Yes	15	1	16
No	5	136	141
Total	20	137	157

Table 10: *Performance Characteristics of the different Logistic Regression Models*

Model	**Se%(CI)**	**Sp%(CI)**	**PPV%(CI)**	**NPV%(CI)**	**FP%(CI)**	**FN%(CI)**
LR_{FM}	95.0 (75.1, 99.8)	99.3 (96.0, 99.9)	95.0 (75.1,99.8)	99.3 (96.0,99.9)	0.7 (0.02,4.0)	5.0 (0.1,24.8)
$LR_{Predefined}$	75.0 (51.0, 91.3)	99.3 (96.0, 99.9)	93.8 (69.7,99.8)	96.5 (91.9,98.8)	0.7 (0.02,4.0)	25.0 (8.7,49.1)
$LR_{Induced}$	10.0 (1.2, 31.7)	98.5 (94.8, 99.8)	50.0 (6.8,93.2)	88.2 (82.1,92.8)	1.5 (0.2, 5.2)	90 (68.3,98.8)
$LR_{Fullterm}$	0 --	100 (97.3, 100)	0 --	87.3 (81.0, 92.0)	0 --	100 (83.2, 100)

Se = sensitivity, Sp = specificity, PPV = positive predictive value, NPV = negative predictive value, FP = false positive rate, FN = false negative rate, LR_{FM} = logistic regression model with Predefined diagnosis, Induced abortion and Fullterm delivery, $LR_{Induced}$ = logistic regression model with Induced abortion only, $LR_{Predefined}$= logistic regression model with Predefined diagnosis only, $LR_{Fullterm}$= logistic regression model with Fullterm delivery only, CI = 95% Confidence interval.

3.4 Comparison of Algorithms

A ROC plot comparing the performance of the different models used in this report is presented in Figure 5. Models close to the upper left corner of the ROC plot were the most accurate. Based on the plot, it’s pretty clear that in this population of pregnant women, the Random forest model and Logistic regression model with Predefined diagnosis, Induced abortion and Fullterm delivery outperformed the other models. As was expected, the model with Predefined diagnosis (Prediagnosis) had a fairly good performance. The model with only Induced abortion and the model with only Fullterm delivery performed poorly as they were not much different from random guessing.

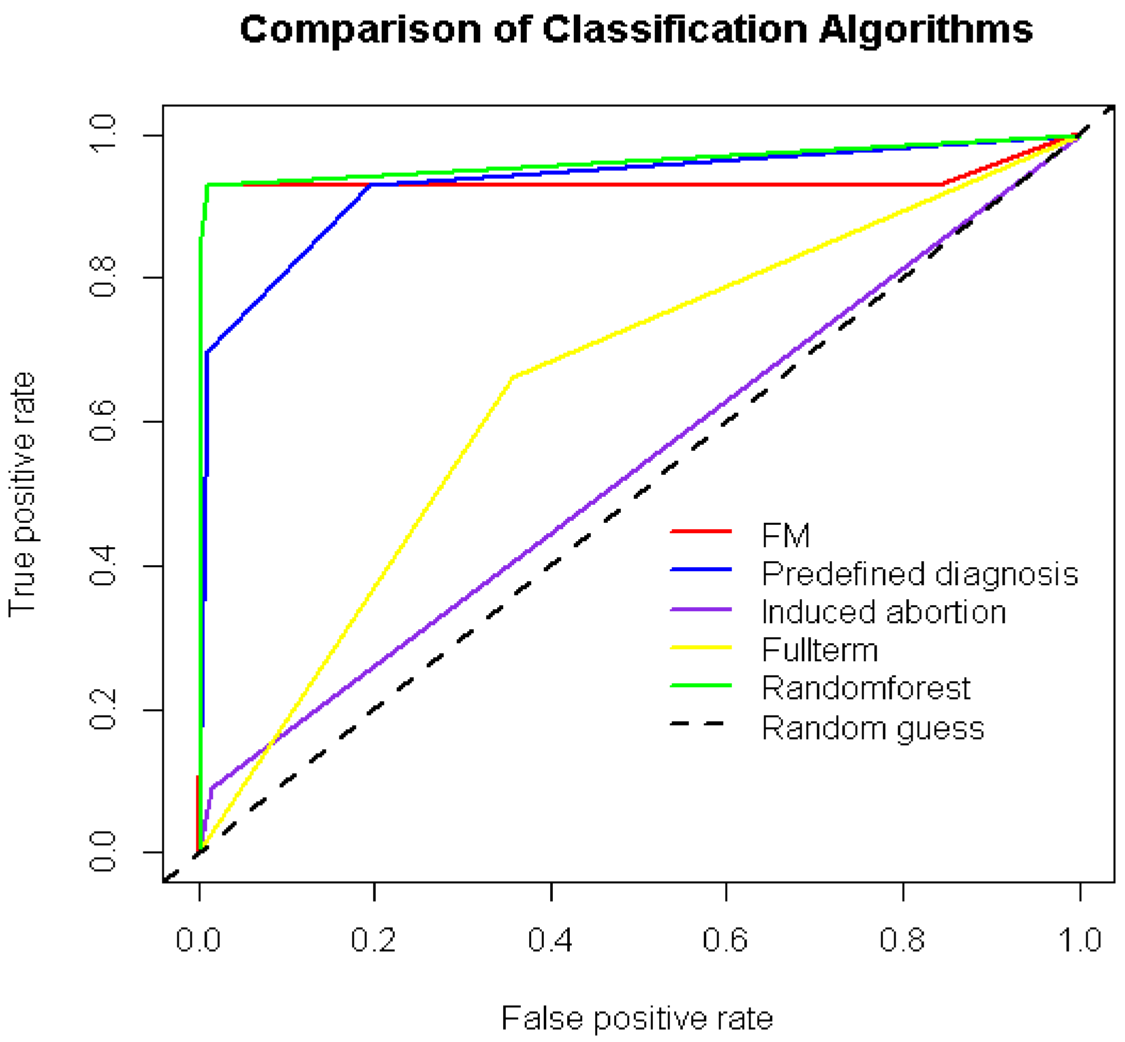

Figure 5 : *ROC curve comparing the performances of the Random forest algorithm; Logistic regression model with Predefined diagnosis, Induced abortion, and Fullterm delivery (FM); Logistic regression model with induced abortion only; Logistic regression model with Fullterm delivery only; and Logistic regression model with Predefined diagnosis only.*

Table 11: *Training and test errors (in percentages) for the different algorithms based on 5-fold cross-validations*

Data	**n**	**RandomForest**	**LR_{FM}**	**$LR_{Predefined}$**	**$LR_{Induced}$**	**$LR_{Fullterm}$**
Train set	625	1.12	1.28	3.68	12.16	12.48
Test set	157	1.27	1.27	3.82	12.74	12.74

LR_{FM} = logistic regression with predefined diagnosis, Induced abortion and Fullterm delivery, $LR_{Induced}$ = logistic regression model with Induced abortion only, $LR_{Predefined}$= logistic regression model with predefined diagnosis only, $LR_{Fullterm}$= logistic regression model with Fullterm delivery only.

Table 11 compares the predictive performance of the different models based on the training and generalization errors computed using 5 fold cross-validations. Generally, as was expected, the training errors were smaller than the generalization errors. The random forest model and the logistic regression model with Predefined diagnosis, Induced abortion and Fullterm delivery had the smallest generalization error (1.27%), thus best performance. The model with Predefined diagnosis had a relatively low generalization error of 3.82% even though it was three times greater than the model with Predefined diagnosis, Induced abortion and Fullterm delivery. On the other hand, the model with only Induced abortion and the model with only Fullterm delivery had relatively higher generalization errors of approximately 13.0% showing their poor performance individually.

3.5 Sample Size Computation

The proportion of pregnant women with spontaneous abortion in this study population was estimated at 12.4%. The classifier with Predefined diagnosis, Induced abortion and Fullterm delivery scored a sensitivity of 95.0% , specificity of 99.3% and a corresponding estimated positive predictive value of 95.0%. We would require a sample size of approximately 589 subjects to obtain the sensitivity of 95.0% at a clinically acceptable precision of 5% (Figure 6). Similarly, for the specificity of 99.3%, we would require a sample size of approximately 13 subjects (Figure 7). The sample size required to estimate the corresponding positive predictive value with the same precision is the maximum of the sample sizes for the sensitivity and specificity which was approximated at 589 subjects.

Figures 6 and 7 present a plot of sensitivity and specificity versus sample size for different prevalence values. From the figures, it is clear how the sample sizes for sensitivity and specificity vary with prevalence.The sample size for sensitivity increases as the prevalence decreases while the sample size of specificity increases with increasing prevalence. By exploiting the relationship between sensitivity, specificity and prevalence, the sample size for the positive predictive value can be obtained from figures 6 and 7 by looking for the sample sizes of the corresponding sensitivity and specificity and choosing the maximum.

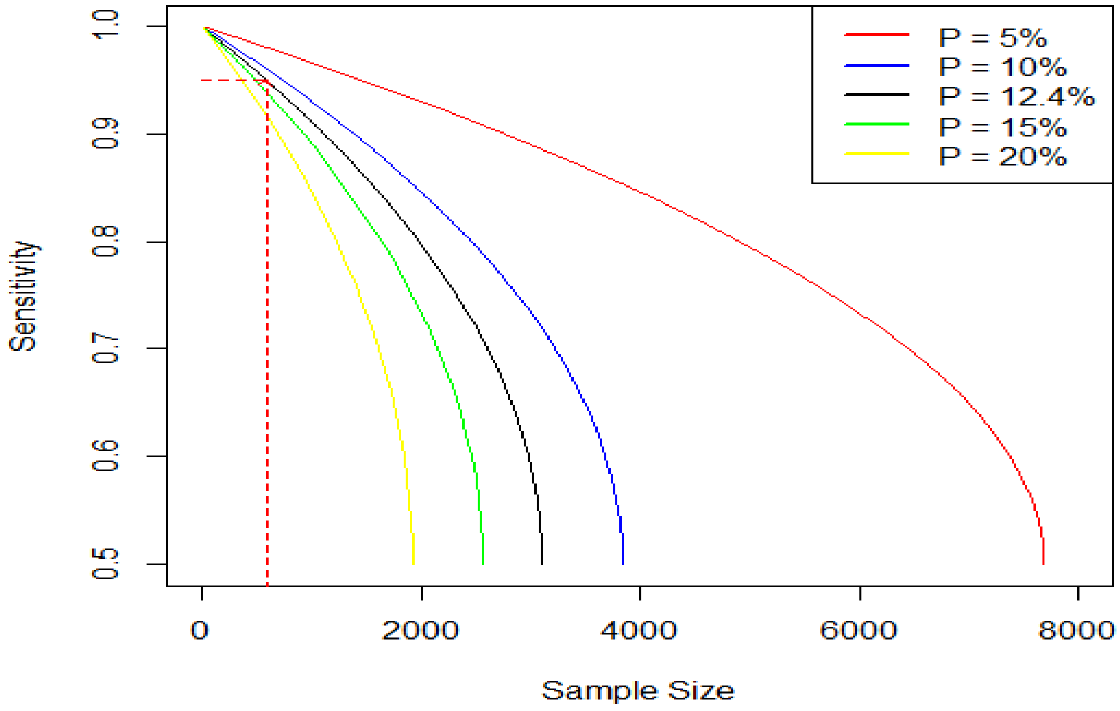

Figure 6: *Sample Size for Sensitivity for different Prevalence values (P) based on a precision of 5%.*

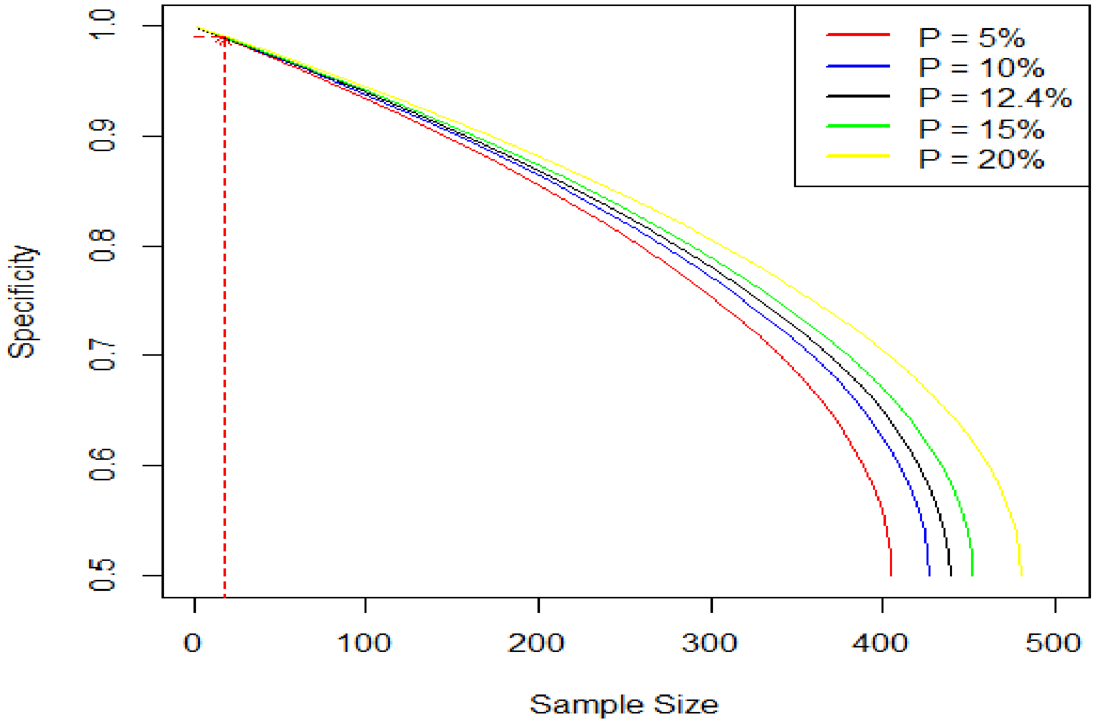

Figure 7: *Sample Size for Specificity for different Prevalence values (P) based on a precision of 5%.*

4.0 DISCUSSION AND CONCLUSION

There is an inherent shortcoming in every diagnostic test. In this study, we sought to develop improved algorithms for a more accurate identification of adverse pregnancy outcomes as compared to a predefined algorithm. The main objectives of this study was to assess the performance of some case classification algorithms in classifying pregnant women as having a spontaneous abortion or not after exposure to the Human Papillomavirus vaccine and also to derive a statistical method to estimate the sample size required for validation of diagnostic algorithms.

A predefined algorithm was able to correctly classify 618 out of 782 pregnant women into the classes of spontaneous abortion and non-spontaneous abortion resulting in a sensitivity of 74.2% and a positive predictive value of 93.5%. However, 155 subjects could neither be classified into any of the two classes. To improve the performance of the predefined algorithm, variables were extracted from the CPRD database and classifiers (models) were constructed. The sensitivity, specificity, predictive values, false negative and false positive rates were computed for the different classifiers.

The optimal complexity of a classifier is the one which produces the least generalization error (Hastie *et al*., 2001). In combination with forward variable selection, the Mean Decrease in Gini was used for obtaining an appropriate complexity. This is because the random forest Mean Decrease in Gini provides more robust results compared to Mean Decrease in Accuracy (Calle & Urrea, 2010). The models with the attributes Predefined diagnosis, Induced abortion and Fullterm delivery yielded the least generalization (test) errors. The model with induced abortion was too simple and did not fit the data well resulting in higher generalization errors and thus poorer performance meanwhile the model with only Fullterm delivery apart from being simple was useless as it was unable to differentiate between spontaneous abortion and non-spontaneous abortion on its own, thus classifying everyone in the non-spontaneous abortion class.

It has been reported that, using multiple diagnostic codes and other information from large databases may improve the accuracy of case classification (Cooke *et al*., 2011, Quint *et al*., 2014). The diagnosis of the predefined algorithm (predefined diagnosis) was incorporated into the Random forest model and into a logistic regression model with Induced abortion and Fullterm delivery and they yielded much improvement in performance. Both models scored a sensitivity of 95.0%(75.1,99.8), specificity of 99.3%(96.0, 99.9), positive predictive value of 95.0%(75.1, 99.8) and a false negative rate of 5.0%(0.13,24.8). Since spontaneous abortion in the study population is rare (Prevalence of 12.4%), it was of interest to minimize the false negative rate. The model with Predefined diagnosis, Induced abortion and Fullterm delivery was chosen as the best classifier as it produced optimal performance. The null hypothesis of no difference in the correct classification rate of the predefined algorithm and the best classifier was rejected (confidence interval of difference in correct classification rate: 16.3% - 23.1%) indicating an improvement in classification. Also, by using the model with Predefined diagnosis, Induced abortion and Fullterm delivery, all subjects were classified in contrast to the predefined algorithm in which 155 subjects could not be classified. Similar studies were not available for comparison of results and also, the results could not be compared to similar studies based on a different population. A major limitation is the small sample size used for validation leading to estimates with wide confidence intervals.

Furthermore, to obtain the same positive predictive value (95.0%) as the best model with a clinically acceptable precision of 5%, the sample size required for obtaining the corresponding sensitivity of 95.0% and specificity of 99.3% were estimated at 589 and 13 subjects respectively. The required sample size for obtaining the estimated positive predictive value (95.0%) with a precision of 5% is the maximum of the sample size for sensitivity and specificity. Thus we require a validation set of 589 subjects to obtain a positive predictive value of 0.95±0.05. The limitation of this approach is that, the pre-specified precision (width of confidence interval) is not necessarily observed.

5.0 REFERENCES

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, Wiley.

Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatrica*, 96,338-341.

Beiman, L. (1996). Bagging Predictors. *Machine Learning,* 24**,** 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning,* 45**,** 5-32.

Buderer, N. M. F. (1996). Statistical Methodology: I. Incorporating the Prevalence of Disease into the Sample Size Calculation for Sensitivity and Specificity. *Academic Emergency Medicine,* 3: 895–900

Calle, M. L. & Urrea, V. (2010). Letter to the Editor: Stability of Random Forest importance measures. *Briefings in Bioinformatics.*

Cooke, C., Joo, M., Anderson, S., Lee, T., Udris, E., Johnson, E. & Au, D.(2011). The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. *BMC Health Services Research,* 11**,** 37.

CPRD (2014). Available at http://www.cprd.com/ObservationalData/

Fielding, A. H. (2007). Cluster and Classification Techniques for the Biosciences. Cambridge University press.

Hajian-Tilaki, K. (2011). Sample size estimation in epidemiologic studies. *Caspian Journal of Internal Medicine,* 2**,** 289-298.

Hastie, T., Tibshirani, R., Friedman, J.(2009). *The elements of statistical learning,* Springer New York.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). The Elements of Statistical Learning. Vol. 1. Np. Springer New York.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied Logistic Regression,* Wiley.

Ioannidis, J. P. A., Tarone, R. & Mclaughlin, J. K. (2011). The False-positive to False-negative Ratio in Epidemiologic Studies. *Epidemiology,* 22**,** 450-456 .

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning,* Springer.

Jones, S.R., Carley, S. & Harrison, M. (2003). An introduction to power and sample size estimation. *Emerg Med J* ;20:453–458.

Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W. (2005). Applied Linear Statistical

Models (Fifth Edition). McGraw-Hill.

Lehmann, C., Koenig, T., Jelic, V., Pricehep, L., John, R. E., Wahlund, L.-O., Dodge, Y. & Dierks, T. (2007). Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *Journal of Neuroscience Methods,* 161**,** 342-350.

Liaw, A. & Weiner, M. (2014). Breiman and Cutler's random forests for classification and regression.

Malhotra, R. K., & Indrayan, A. (2010). A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. *Indian Journal of Ophthalmology*, *58*(6), 519–522.

Naing, L., Winn, T. & Rusli, B. (2006). Practical issues in calculating the sample size for prevalence studies. *Archives of Orofacial Sciences,* 1**,** 9-14.

Parikh, R., Mathai, A., Parikh, S., Chandra -Sekhar, G. & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology,* 56**,** 45-50.

Quint, J. K., Müllerova, H., Disantostefano, R. L., Forbes, H., Eaton, S., Hurst, J. R., Davis, K. & Smeeth, L. (2014). Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ Open,* 4.

Slipchenko, T., Bowman, C., Chen, Y., Sugar, C., Gifford, A. (2008). In Search for a Golden Algorithm. SAS Institute Inc. [Available at: http://www2.sas.com/proceedings/forum2008/144-2008.pdf]

Tan, P.-N., Steinbach, M. & Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*, Addison-Wesley Longman Publishing Co., Inc.

Tate, A. R., Beloff, N., Al-Radwan, B., Wickson, J., Puri, S., Williams, T., Van staa, T. & Bleach, A. (2013). Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface. *Journal of the American Medical Informatics Association*.

Thomas, S. L., Edwards, C. J., Smeeth, L., Cooper, C. & Hall, A. J. (2008). How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the general practice research database? *Arthritis Care & Research,* 59**,** 1314-1321.

Williams, T., Vanstaa, T., Puri, S. & Eaton, S. (2012). Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Therapeutic Advances in Drug Safety*.

6.0 APPENDIX

Table A1: *Confusion Matrix derived from Logistic Regression model with Induced abortion only*

Predicted diagnosis	**Reference diagnosis**		
	Yes	No	Total
Yes	2	2	4
No	18	135	153
Total	20	137	157

Table A2: *Confusion Matrix derived from Logistic Regression model with Fullterm delivery only*

Predicted diagnosis	**Reference diagnosis**		
	Yes	No	Total
Yes	0	0	0
No	20	137	157
Total	20	137	157

Table A3: *Variable significance from univariate logistic regression*

Variable	**Coefficient**	**Standard error**	**P-value**
Predefined diagnosis U	2.647	0.584	< 0.0001*
Y	6.647	0.692	< 0.0001*
Induced abortion	1.778	0.589	0.002*
Stillbirth	1.796	0.924	0.05
Fullterm delivery	-1.257	0.297	< 0.0001*
Ultrasound	0.038	0.368	0.92
Smoking	0.169	0.288	0.55
Hypertension	-0.393	0.617	0.52
Alcohol consumption	0.032	0.368	0.93
Other vaccines	-0.196	0.543	0.72
Contraindicated drugs	0.039	0.295	0.89
Other drugs	0.026	0.318	0.94
Preterm delivery	-15.452	761.367	0.98
Postterm delivery	-1.557	1.023	0.12

**Statistically significant at 5% significance level, U=unknown, Y=yes*

Table A4: Logistic regression - *Forward selection*

Variable	**Coefficient**	**Standard error**	**P-value**
Predefined diagnosis U	5.188	0.779	< 0.0001*
Y	11.650	1.374	< 0.0001*
Fullterm delivery	3.917	0.782	< 0.0001*
Induced abortion	-3.361	1.236	0.006*

**Statistically significant at 5% significance level, U=unknown, Y=yes*

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Assessing the performance of case classification and validation algorithms; and sample size estimation for classification algorithms' validation<br />

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,

Boeyeo, Forsi Nwebim

Datum: **28/01/2015**