J.Natn.Sci.Foundation Sri Lanka 2012 ***40****(4):321-331*

RESEARCH ARTICLE

Determining the sensitivity and specificity of a substitute test as a diagnostic for its gold standard in the presence of severe missingness

D. M. Senaratna and M.R. Sooriyarachchi*
Department of Statistics, Faculty of Science, University of Colombo, Colombo 3.

Revised: 10 October 2012 ; Accepted: 19 November 2012

Abstract: When the need arises to identify a disease, substitute tests or screening tests are commonly used to recommend patients for its respective "Gold Standard". Since it is seldom that these gold standards are carried out for those who pass the substitute tests, calculating the sensitivity and specificity of the substitute test has become a near impossible task using conventional methods. However, due to the life threatening nature of certain diseases such as coronary artery disease (CAD), understanding the effectiveness of these substitute tests in detecting the disease for sub-regions of the world is of utmost importance. Therefore, the primary objective of this study was to develop a theoretical framework to determine the sensitivity and specificity of a diagnostic test in the presence of severe missingness in the results of its gold standard.

The methodology involves missing value imputation for the missing response, which is the result of the gold standard for those who have passed the substitute test. Logistic models were used to predict the existence of the disease using pre-defined risk factors. Subsequently, receiver operator characteristic (ROC) curves were used to confirm the existing cut-off for the substitute test.

This procedure is illustrated on data from a retrospective study carried out in a General Hospital in Sri Lanka. The ROC curve analysis verified the existing Bruce protocol method cut-off as being the best to classify the existence of CAD. The study confirms that the results conform to world standards.

Keywords: Angiogram, diagnostic tests, gold standard, missing value analysis, multiple imputation, ROC curve analysis, substitute test.

INTRODUCTION

Background

Screening tests or substitute tests are commonly used in the medical field, either, in order to select individuals before recommending them to more conclusive tests, or to identify diseases prior to conspicuous symptoms being noticed. Primarily due to high costs and administrative difficulties, many tests have been developed to act as diagnostic for conclusive tests, which are commonly referred to as the "Gold Standard". In the domain of heart disease, cardiac stress tests (CST), echo - cardiography, and baseline electrocardiography are a few of the screening tests used to identify coronary artery disease, whilst the angiogram is considered as its ultimate Gold Standard (Greenland *et al.*, 2010). In the domain of diabetes mellitus, fasting plasma glucose and 2-hour plasma glucose during an oral glucose tolerance test are commonly used as substitute tests to identify Type-2 Diabetes whilst its controversial gold standard is the level of glycosylated haemoglobin (American Diabetes Association, 2004).

Medical consultants use many factors such as gender, medical history, and social habits along with the results of such screening test results as predictors of a disease. However, the predictive capabilities of screening tests such as the CST have shown to differ from region to region, resulting in controversies (Bokhari *et al.*, 2008).

* Corresponding author (roshini@mail.cmb.ac.lk)

The appropriateness of diagnostic methods is rarely tested in developing nations. The primary reason being problems regarding funding and other resources in such study domains. Here, a complex procedure is needed to identify the actual disease condition of individuals. The disease condition as indicated by a gold standard or a similar device for individuals who pass these screening tests are seldom available as doctors in developing countries rarely advice the gold standard for those passing the screening test (Atukorale, 2005; Wake & Yoshiyama, 2009); hence resulting in 'missingness' with regard to vital data. Therefore the motivation for this paper arose, since perhaps the only manner in which at least an approximate understanding on the precision of the stress test could be obtained, through statistical techniques using missing data imputation.

Review of literature

In order to unravel the problem of the sensitivity and specificity of the substitute test, four primary statistical methods are needed. Namely, a sample size calculation and design of the study, missing value analysis, statistical modeling and ROC curve analysis.

Approaches on design reviews can be seen in studies such as that has been carried out by Bolland *et al.* (1998). Royston and Barbiker (2002) recommended the relatively recent review on sample size calculation methods by Sahai and Khurshid (1996), along with the work carried out by Ury and Fleiss (1980) and Lachin (1977), for an extensive understanding on two group binary outcome studies.

In the area of missing value analysis, the literature points to a vast range of techniques ranging from crude methods such as mean substitution, to approximate bayesian bootstrap methods, EM algorithms, and to even non-parametric decision-tree methods, to impute datum. Imputation methods can be further divided into single value imputation (SI) and multiple-value imputation (MI) (Rubin, 1987). In Van-Leeuwen *et al*. (2007)'s study, imputation was repeated 100 times to classify unverified women as having gestational diabetes mellitus or not.

For the purpose of summarizing the predictive power of a binary outcome situation, when test data do not fall into two obviously defined categories, Agresti (2007) recommends the use of ROC curves. The area under the curve (AUC) is one of the most commonly used methods as a summary measure of the ROC curve to compare classification capabilities (Vergara *et al.*, 2008). Parametric, semi-parametric or non-parametric estimation methods can be used to estimate the AUC of a ROC (Vergara *et al.*, 2008). Hanley and McNeil (1983) recommend a test for comparing two AUCs, pair-wise.

Objectives of the study

The primary objective of this study was to recommend a methodology to identify the sensitivity and specificity of a substitute test in the presence of missingness. Further, it was also intended to provide advice on the identification of the significance of the substitute test as a diagnostic for its gold standard, and to verify the existing cut-off for using an example set of data.

The data

The methodology was illustrated using data from the Cardiology Unit of the Sri Jayawardenapura Hospital, in Sri Lanka, since it had a relatively well organized records room, owing to being a general hospital that had patients from all over the country, and easy administrative conveniences in obtaining large numbers of records. Five hundred patient records details were collected from bed-head tickets (BHT) from January 2008 to October 2009. A further set of 50 BHTs were also obtained for the purpose of validating the study findings from the year 2010. Table 1 includes a description of the variables collected for the study.

Brief description of the methodology

The theoretical framework used a sample size design, missing value analysis, statistical modeling and ROC curve analysis.

The sample size formula by Lachin (1977) for the comparison of more than two groups with dichotomous outcomes in an r x c contingency table was adopted for the sampling design. The logistic regression method, an extension of the regression method, was used to multiple impute the missing responses of the gold standard (Yuan, 2001). The missing values were sampled from the posterior distribution of the responses using Monte-Carlo simulation (Tan *et al.*, 2010). Logistic models, as predominantly used in the medical field to model a disease status, were used as the underlying statistical model (Agresti, 2007). Finally, the area under the receiver operating characteristic curve (AUC) was used to identify the best cut-off for the substitute test, using the Dorfman and Alf maximum likelihood estimation approach (Hanley & McNeil, 1983).

Table 1: Variables collected for the purpose of the study

	Variable	Type	Levels	Description
1.0	Angiogram results	Nominal	99 – Not carried out 0 – No disease 1 – Single vessel disease 2 – Double vessel disease 3 – Triple vessel disease	Disease status identified through the angiogram
2.0	Cardiac stress test result	Nominal / ordinal	99 – Not carried out 1 – Stage 1 difficulty 2 – Stage 2 difficulty 3 – Stage 3 or higher difficulty or minor difficulties 4 – Completed the CST or patient was diagnosed as adequately stressed.	Performance with respect to the CST
3.0	Strong Indication of the disease expressed by medical consultants	Binary	0 – No 1 – Yes	
4.0	Age	Continuous		Patient's age
5.0	Gender	Binary	0 – Female 1– Male	
6.0	Hypertension			
6.1	Hypertension – history	Binary	0 – No 1 – Yes	
6.2	Hypertension – pressure	Continuous		Pressure on admission
6.3	Hypertension – pulse	Continuous		Pulse on admission
7.0	Cholesterol			
7.1	Cholesterol – LDL	Continuous		
7.2	Cholesterol – triglyceride	Continuous		
7.3	Cholesterol – HDL	Continuous		
7.4	Cholesterol – total	Continuous		
8.0	Diabetes mellitus	Binary	0 – No 1 – Yes	
9.0	Family history of disease or known related factors	Binary	0 – No 1 – Yes	
10.0	Cigarette consumption	Binary	0 – No 1 – Yes	
11.0	Alcohol consumption	Binary	0 – No 1 – Yes	
12.0	Marital status	Binary	0 – No 1 – Yes	
13.0	Date	Date		

METHODOLOGY

The design review

Following the initial data collection process (internal pilot study), a design review for the sample size calculation is usually conducted. Random Sampling methods (Kish, 1995) can be used as a sampling method. In study domains that are either possibly the first of its kind or where there is a lack of prior information, a crude guess ($n_{Preliminary}$) for an initial sample size is decided based on past literature and the data collection carried out. Then, after gathering information about the study parameters using the intial observations, the sample size is re-estimated and a final sample size is fixed (n_{New}). Bolland *et al*. (1998) recommend an upper bound for the new sample size. The reason for an upper bound ($n_{upperbound}$) is in the instance where collecting a large sample size is infeasible.

Sample size calculation

Sample size calculation 1: two groups with dichotomous outcomes

The formula for the comparison of two groups with dichotomous outcomes (that is, having proportions P_1 and P_2) is given by Ury and Fleiss (1980) for equal groups as well as for the comparison of unequal groups with dichotomous outcomes.

Sample size calculation 2: more than two groups with dichotomous outcomes

Royston and Barbiker (2002) recommended the sample size formula for the comparison of more than two groups with dichotomous outcomes by Lachin (1977) for determining r x c contingency tables. This procedure can easily be extended for the case where c > 2, by simply substituting for 'j', the number of possible outcomes for the response. The methodology for c = 2 has been incorporated in the ART module of Stata (Royston & Barbiker, 2002) in which the user can obtain the required sample size for six or less treatment groups (i={2,3,4,5,6}).

Missing value imputation

The literature mentions three types of missingness, namely, missing completely at random (MCAR), missing at random (MAR) and non-ignorable missingness or missing not at random (MNAR) (Acock, 2005). The definitions of missingness are explained by Tan *et al.* (2010).

Posterior distribution (Tan et al., 2010)

The Bayesian approach to missing value imputation consists of three steps (Gelman *et al*., 1995)

1. Constructing a full probability model summarized by a joint distribution for all observable and unobservable quantities
2. Summarizing the findings for observed quantities of interest based on the derived conditional distributions of these quantities given the observed data
3. Assessing model adequacy

The joint posterior distribution of Y_{com} and θ :

$$f(Y_{com},\theta) = f(Y_{com}|\theta)\pi(\theta) \quad ...(1)$$

where $f(Y_{com},\theta)$ is the sampling distribution for the missing values. When $f(Y_{com},\theta)$ is regarded as a function of θ with fixed Y_{com}, it is the familiar likelihood function denoted by $L(Y_{com}|\theta)$.

Conditional distributions of these quantities are obtained by the Bayes theorem:

$$f(\theta|Y_{com}) = \frac{f(Y_{com}|\theta)\pi(\theta)}{f(Y_{com})} \alpha\, f(Y_{com}|\theta)\pi(\theta) \quad ...(2)$$

Where

$$f(Y_{com}) = \int f(Y_{com},\theta).d\theta = \int f(Y_{com},\theta)\pi(\theta).d\theta \quad ...(3)$$

Is the normalizing constant of $f(\theta|Y_{com})$.

After $Y_{com}$is observed, one can predict or forecast the future observation, denoted by $\tilde{y}$. The posterior predictive distribution of $\tilde{y}$ given the data Y_{com} is defined as:

$$f(\tilde{y}|Y_{com}) = \int f(\tilde{y},\theta|Y_{com})d\theta$$

$$= \int f(\tilde{y}|Y_{com},\theta)f(\theta|Y_{com})d\theta \quad ...(4)$$

Most frequently, the future observation $\tilde{y}$ and Y_{com} are conditionally independent given θ . In this case we have

$$f(\tilde{y}|Y_{com},\theta) = f(\tilde{y}|\theta) \quad ...(5)$$

Regression method for imputing (Yuan, 2001)

The methodology for imputing using the regression technique is as follows (Yuan, 2001).

The imputation model for the standard regression model, is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \quad ...(6)$$

where, Y_i = the variable inclusive of missing values given the covariate variables, X_1, X_2, ... , X_n. The fitted model includes the regression parameters β_0, β_1, ... ,β_n and the associated covariance matrix $\hat{\sigma}_i^2 V_i$ where V_j is the usual $(X'X)^{-1}$ matrix derived from the intercept and covariates X_1, X_2, ... , X_n.

The following steps are used to generate imputed values for each imputation.

New parameters β_*= β_{*0}, β_{*1}, ..., β_{*n} and $\hat{\sigma}^2_{*j}$ are drawn from the posterior distribution of the parameters. That is they are simulated from estimates, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_n$, $\hat{\sigma}^2_{*j}$ and V_j. The variance is drawn as $\sigma^2_{*j} = \hat{\sigma}^2_j (n_j - n - 1)/g$ where g is a $\chi^2_{n_j - n - 1}$ random variable and n_j is the number of non-missing observations for Y_j. The regression coefficients are drawn as $\beta_* = \hat{\beta} + \hat{\sigma}_{*j} V^*_{hj} Z$ where V^*_{hj} is the upper triangular matrix in the Cholesky decomposition, $V_j = V^*_{hj} V_{hj}$ and Z is a vector of n+1 independent random normal variates. The missing values are then replaced by $\beta_{*0} + \beta_{*1} x_1 + \beta_{*2} x_2 + ... + \beta_{*n} x_n + z_i \sigma_{*j}$, where x_1, x_2,..., x_n are the values of the covariates and z_i is a simulated normal deviate. The logistic regression method is an extension of the regression method and is defined by $Logit(P_i) = Log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{2,i} + ... + \beta_n X_{n,i}$ where, P_i is the probability of disease for treatment '*i*' given the explanatory variables, X_1, X_2,..., X_n, are fitted, and missing P_i values are imputed using the same procedure stated above.

The conventional method to obtain the final imputed values, as carried out in many studies, is to multiple impute data sets and carry out a statistical analysis for each of these and then combine the results using Rubin's rule (Mehta *et al.*, 2007). Another approach as used by Van-Leeuwen *et al.* (2007) is to multiple impute many data sets and obtain the average for each observation. In this study 100 such imputations were averaged out. The averaged out observation were grouped as 1 if $\hat{P}_i > 0.5$ or else grouped as 0. If a different threshold can be reasoned to be more appropriate, then the same threshold can be used.

Monte-Carlo simulation: the inversion method (Tan et al., 2010)

Let X be a random variable with cumulative distribution function F. Since F is a non-decreasing function, the inverse of the function F^{-1} may be defined by

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}, \; u \in (0,1) \qquad ...(7)$$

If $U \sim U(0,1)$ then $F(X) \sim U(0,1)$ or equivalently

$$X = F^{-1}(U) \qquad ...(8)$$

has the cumulative distribution function F. Hence, in order to generate one sample, say x, from random variable $X \sim F$, we first draw from $U \sim U(0,1)$, then compute $F^{-1}(U)$ and set it equal to x. Hence the steps can be stated as, first draw U from $U(0,1)$ and then return $X = F^{-1}(U)$

Verification bias in the response variable

As mentioned previously in this study, results for the diagnostic gold standard (CAD) are available primarily for patients who are positive for the test under investigation (CST). When this type of missingness is present, data from such studies are subject to what has been termed "verification bias". There are several ways to adjust for verification bias using statistical correction methods (Laurer *et al*., 2007; Cronin & Vickers, 2008).

Another approach for correcting verification bias under the assumption that the data are missing at random (MAR), the response variable is binary and the number of covariates is relatively large, requiring parametric models for the probability of verification, is multiple imputation (Harel & Zhou, 2006 ; Hua, 2009). Here, multiple imputation based on data augmentation has been used to correct for verification bias. Using simulation, Harel and Zhou (2006) show that imputation methods are better than the existing methods with regard to nominal coverage and confidence interval length for the sensitivity and specificity of the test. Harel and Zhou (2006) also go on to show that for a sample as large as in this study (greater than 200 observations), the biases of sensitivity and specificity from multiple imputation procedures are only marginally higher than from the existing methods.

These findings support our use of multiple imputation and indicate that there is no use of making further verification bias corrections.

Model building and ROC

The theory behind logistic models is well established and has been described by many authors such as Agresti (2007), Collett (1991) and Hosmer and Lemeshow (2000). Agresti (2007) states that in the use of most diagnostic tests when test data do not fall into two obviously defined categories, the area under the curve (AUC) of a receiver operating characteristic curve (ROC curves) is one of the most reliable measures of the logistic models classification capabilities.

ROC Curves are plots of sensitivity as a function of 1-specificity, and are calculated using all possible cut-offs (Agresti, 2007; Vergara *et al.*, 2008). Using the obtained model, $Y = Log\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta_1 X_1 + ... + \beta_n X_n$, predicted values, $\hat{Y}$, for the existing data set can be obtained from $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + ... + \hat{\beta}_n X$, where $\hat{\alpha}, \hat{\beta}_1, ..., \hat{\beta}_n$ are the estimates of the unknown parameters α, β_1,...,

β_n respectively. That is, if $\hat{Y}$ > Threshold (k) the predicted outcome is positive, or else it is categorized as negative. Using k values ranging from 0 to 1, the sensitivity and specificity were calculated for 'each' of these thresholds.

That is, $\hat{Y} = \begin{cases} 1 \text{if } Y > k \\ 0 \text{ otherwise} \end{cases}$ and $k \in (0,1)$

Estimates for ROCs

In order to calculate the AUC, both parametric or semi-parametric estimation methods give a smooth ROC curve and more importantly, as a result of their distributional assumptions, statistical inferences such as hypothesis testing and confidence intervals can be very easily achieved (Vergara *et al.*, 2008). Researchers like Hanley and McNeil (1983) have shown a preference towards using the Dorfman and Alf (1969) maximum likelihood estimation approach. This is the same approach used in the software ROCKIT (Metz *et al*., 1998). In the terminology, negative cases are patients who actually do not have a disease or given condition, and positive cases are patients who actually do have a disease or a given condition.

Comparing AUCs

Once the estimates for the AUC, its variance and standard errors are obtained, a pair-wise comparison can be made following Hanley and McNeal (1983). It is generally accepted that for a sufficiently large dataset the AUC estimate approximates a normal random variable. Hence the test statistic for the difference between two AUC's would be:

$$Z = \frac{AUC_1 - AUC_2}{\sqrt{\left(\hat{VAR}(AUC_1)\right) + \left(\hat{VAR}(AUC_2)\right) - 2\left(\hat{COV}(AUC_1, AUC_2)\right)}} \quad ...(9)$$

APPLICATION

Design review and sample size calculation

After planning out the data collection process, a design review for the sample size calculation was conducted based on the study by Bolland *et al*. (1998). A crude guess of 250 observations was first decided upon and an upper bound of 500 data points was set due to the difficulty in obtaining records. This guess of 250 observations was calculated using the data in the study by McNeer *et al.* (1978).

Gender, age (categorized into 3 levels), hypertension status and diabetes mellitus status were also selected for the sample size calculation, owing to their impact on the disease (Wilson *et al.*, 1998). The required sample size for each of the selected factors was calculated using the methodology in the ART menu and described under the design review. The significance level for this study was fixed at 0.05 for a two tailed test, and the power was decided upon as 0.80. The reason for choosing such values was due to the infeasibility in collecting large samples due to both inherent missing values and administrative inconveniences.

Table 2 depicts the sample sizes obtained for the respective factors. As can be seen from the results, the maximum sample size was 295. Therefore, the required sample size was set at 300 with the test having a power of 80 % with a type I error of 5 %.

Table 2: Required sample sizes by variable

Description	Sample size required
Cardiac stress test (6 levels)	99
Gender (2 levels)	87
Age (3 levels)	295
Hypertension (2 levels)	172
Diabetes mellitus (2 levels)	72

Missing value imputation

The primary plausible reason for the missingness of most of the data points on covariates, apart from that of the angiogram status, is due to medical staff not being able to complete records. However, the missingness on the angiogram response status was mainly due to the fact that those individuals who passed the CST were not subjected to an angiogram. Therefore, making it impossible to calculate the sensitivity or specificity of the CST since this number was extremely small and as good as non-existent, though a comparison of CST levels only, similar to the work done by McNeer *et al*. (1978), could have been conducted.

The opinion of the medical doctors involved in the study regarding the missingness of the covariates was that their results could be biased and those missing values were not conditional on another variable, and hence according to the discussion made by Acock (2005) were not missing at random (NMAR). On the other hand, the missingness of the response variable (CAD) was entirely dependent on another variable, namely, the CST and thus,

the values of the missingness of the CAD falls under the preview of missing at random (MAR). Current research indicates that while using imputed missing values that are missing at random or missing completely at random (MAR or MCAR) does not bias the results the same, is not the case for missing values, which are not missing at random (NMAR).

It must be noted that though angiograms have been in use in Sri Lanka for well over 15 years, it is surprising that a study concerning the sensitivity and specificity levels of the CST has not been published possibly due to this reason. Therefore, instead of confining this study to a comparison of the CST levels, it was thought as necessary to impute these missing values.

Sterne *et al.* (2009) stated that the "missing at random (MAR) assumption may be reasonable if a variable that is predictive of missing data in a covariate of interest is included in the imputation model". Following from this definition, since the variable needed to be imputed is that of the response, the MAR assumption was valid. Further, since a logistic model was to be used in the final analysis, it was considered best to use this method opposed to mean imputation or hot deck methods. It must also be noted that imputing missing values for the response or dependant variable is seldom carried out when explanatory variables are not missing or imputed, since "in this case MI is the same as list-wise deletion and such imputation only increases sampling variability" (Allison, 2004). However, in this study, as explained above, since if these particular values were not imputed, it would be impossible to find out the sensitivity and specificity of the CST, hence this procedure was carried out.

Two approaches can be used to impute data. The first method is to compute multiple imputations, generally around 5, analyze those multiple imputations individually using conventional statistical methods, and, finally, to combine the results using Rubin's rule (Rosenbaum & Rubin, 1983). Method two, however, involves creating many multiple imputations, around 100 and averaging the results (Van-Leeuwen *et al.*, 2007). Though the first method is more popular and perhaps better validated in many studies due to computational ease and sound methodology, the second method was adopted. These values were included into the original data set. The variables used for imputing include age, gender, hypertension, diabetes mellitus, cigarette and alcohol consumption, systolic and diastolic blood pressures, marital status and CST status. No interaction terms were included.

After the imputation, logistic regression models were used on the imputed data set to determine important covariates. Variables that were considered as insignificant remained to be so and those that were significant remained to be significant apart from the variable family history. Yet, even this variable is more significant than the other variables, as was the case before imputation. Perhaps due to the large increase in power, as a result of the increase in sample size, the significance levels of these variables appear to have increased vastly.

Logistic models

Both a forward selection and backward elimination procedure were carried out, considering up to two interaction terms only. The final model obtained using backward elimination process for the total set of observations including imputed observations is as follows:

$$\log it(P_{iiklmn}) = const + \beta_1(Age) + \beta_i^{HT} + \beta_i^{FH} + \beta_k^{DM} + \beta_l^{CST} + \beta_m^{Alc} + \beta_n^{Sex} \quad ...(10)$$

After building a model, it was clearly observed how the odds of getting CAD decreased as an individual's ability to withstand a CST stage level increased out of those who failed the stress test. It was also observed that those who passed the CST had the smallest odds of getting CAD. Following from this observation, the final objective of this study was to identify the best cut-off for the CST. That is, to identify if instead of using the conventional Bruce-protocol method to pass and fail individuals, if having a stage as a cut-off gives a significantly better or even similar classification capability. For this purpose, the CST variable was grouped into three categories as given below:

1. Group 1: Those who failed in stage 1 versus the rest (those who passed up to a stage ≥ 1). The corresponding model based on backward elimination is

$$\log it(P_{ijklmn}) = const + \beta_1(Age) + \beta_i^{HT} + \beta_i^{FH} + \beta_k^{DM} + \beta_l^{Alc} + \beta_m^{Sex} \quad ... (11)$$

2. Group 2: Those who failed in stages 1 or 2 versus the rest (those who passed up to a stage ≥ 2). The corresponding model based on backward elimination is:

$$\log it(P_{ijklmn}) = const + \beta_1(Age) + \beta_i^{HT} + \beta_j^{FH} + \beta_k^{DM} + \beta_l^{CST} + \beta_m^{Alc} + \beta_n^{Sex} \quad ...(12)$$

3. Group 3: Those who failed the test (all stages 1 and above) versus those who passed the CST. The corresponding model based on backward elimination is:

$$\log it(P_{ijklmn}) = const + \beta_1(Age) + \beta_i^{HT} + \beta_j^{FH} + \beta_k^{DM} + \beta_l^{CST} + \beta_m^{Alc} + \beta_n^{Sex} + \beta_2(Age \times CST) + \beta_{\ln}^{CST \times Sex} + \beta_{kl}^{DM \times CST} \quad ...(13)$$

ROC curve analysis

Three ROC curves were constructed for the three cut-off models. In order to obtain a graphical overall look at the ROC curves, the ROC curves corresponding to the three models were plotted and are given in Figure 1.

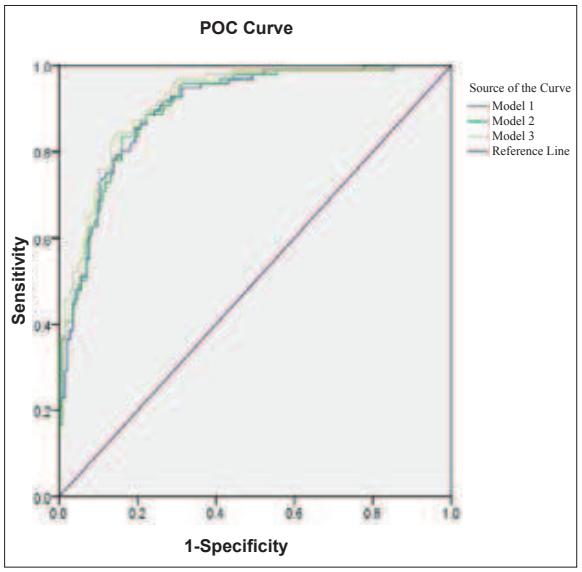

Figure 1: ROC curves for cut-off models

It can be seen that the first two CST groupings appear to be similar, since their respective lines overlap somewhat. However, in contrast to the first two groupings, the last grouping (colour-coded in off-white), which is the Bruce protocol cut-offs of pass and fail as currently used by medical practitioners in Sri Lanka, appears to have its ROC curve line almost always above the other two groupings, though quite close. In other words, the third grouping not only has the most significant CST grouping with respect to its logistic model but it also appears to have better discrimination power than the other two based on the observations of the ROC curve. However in order to test if this difference is significant, a statistical pair-wise comparison test was carried out.

Using the above data obtained through ROCKIT, and under the assumption that for a sufficiently large data set the AUC is distributed normally, the following hypothesis was tested for the three possible comparisons.

$$H_0 : (AUC)_i = (AUC)_j \quad \text{where, (i,j)=\{(1,2) (1,3) (2,3)\}}$$
$$H_1 : (AUC)_i \neq (AUC)_j$$

Using the explained test statistic and the calculated estimates as given in Table 3, the outcome as given in Table 4 was obtained. It can clearly be seen that the AUCs for the first two groupings were not significantly different. However, the AUC of the third grouping was significantly different from both the first and second grouping at 5 % significance level. Further, since the Z statistic value is positive, it implies that the AUC of grouping 3 is significantly 'larger' than that of the other two at significant level even smaller than 5 %.

Table 3: The estimates for the AUCs and their corresponding standard errors for three CST groupings

CST grouping	AUC	Standard error of AUC
1	0.8989	0.0197
2	0.9025	0.0193
3	0.9225	0.0171

Table 4: Pair-wise comparison between the three groupings.

Comparison	Mean (Difference)	VAR (Difference)	Z	p - value (Two-tailed)
Group 1 vs 2	0.0019	0.000037	0.3121	0.7530
Group 1 vs 3	0.0245	0.000110	2.3370	0.0194
Group 2 vs 3	0.0212	0.000102	2.0970	0.0351

Validation of the model

In studies related to the medical research field, many practical dilemmas could occur and hence bias the results either knowingly or unknowingly to the researcher. Some of these drawbacks include: possible lack of representativeness in the sample due to administrative issues or missing data; inadequate sample size; omission of important confounders due to issues ranging from lack of knowledge in the subject area to the inability to measure the existence of a confounder due to it being

controlled as a precaution, an example being the problem encountered with cholesterol levels in this study. Though such drawbacks are common in medical related studies, it is, however, very important that the inferences or results obtained are accurate and precise enough, due to the possibly life threatening nature of the disease. Therefore, it is important to verify even a small doubt. Incumbent validation procedures include methods such as bootstrapping and independent test case validation. For the purpose of validation, it was considered best to use an independent test case since it would verify the validity of the inferences obtained with a dataset completely unrelated with the first, and also, due to the accepted nature and simplicity of this method. Data collection was a problematic issue throughout this study, and therefore, it was not possible to obtain a large test dataset. However, the objective of validating using a test dataset was not to make inferences concerning the study hypotheses, but instead to observe if the data agreed in general with the previously obtained models. Another unrelated set of 50 BHTs and CST records were once again requested from the Sri Jayawardenapura hospital. Firstly, the dataset was cleaned, and as before, a missing data imputation procedure was carried out. Then using the forecasts and the actual results, from the modeling procedure, the false positive rate (FPR), false negative rate (FNR), true positive rate (TPR), and true negative rate (TNR) were tabulated.

Further, ROC curves were constructed for the test cases as well. The above mentioned calculations were carried out after deleting the individuals who did not do the CST as they were not needed. This final test dataset had just 34 observations. Therefore, in drawing conclusions from this small dataset, two things should be kept in mind. Firstly, the AUCs could be imprecisely estimated, and secondly, the estimated AUC may not have a normal distribution. Thus results should be carefully interpreted.

Though the overall model has a high TPR of 82 % its TNR is only 67 %. In the case of the model chosen to find the best cut-off, the TPR was lesser with its value being just 73 %, with however the same TNR value observed for the first model, that is 67 %. In general, the overall model correctly identified roughly 76 % of the cases while the other correctly classified over 70 % of the cases. In order to obtain an idea for the classification capabilities of the two models, ROC curves were once again constructed but are not presented here. Calculations were carried out for the AUCs and their respective standard errors. The AUCs for the two models were 0.811 and 0.765, respectively. According to Hosmer and Lemeshow (2000), these AUC values would imply that the first AUC has excellent discrimination while the second has acceptable discrimination. Further, it was found out through ROCKIT that these AUCs were not significantly different from each other. Yet due to the small sample size it must be noted that these estimates and inferences may not be very accurate. Therefore, the ROC curves were used only to obtain a graphical view of the classification capability of the two models. It was observed that the AUCs for both models are very much further away from the diagonal of the curve.

DISCUSSION

In this study missing value imputation has been successfully used for determining the values of sensitivity and specificity and thereby determining the diagnostic capabilities of the CST as a substitute for the angiogram. This technique can be similarly used in cases where passing the substitute test results in no gold standard test being done.

The main finding of this study was that the Bruce protocol cut-off was the best classifier of the CAD. Also the results obtained for the example dataset are conformed to the world standards. The sensitivity and specificity values obtained in this study with the aid of missing value imputation were, for the Bruce protocol method, a sensitivity of 87 % and specificity of 77 % after adjusting for the other confounding variables. Similarly, for the overall model a sensitivity of 85 % and specificity of 79 % can be observed. In both these cases, we can observe that the sensitivity is slightly higher than the specificity. The sensitivity and specificity values obtained for the above situations may however have been enhanced by the confounders' predictive capabilities.

The American Heart Association guidelines state a risk factor-unadjusted "sensitivity and specificity of 68 % and 77 % for detecting significant coronary disease at angiography" whilst Hill and Timmis (2002) have concluded in their study as this test having a risk factor-unadjusted sensitivity of 78 % and a specificity of 70 % in detecting coronary artery disease. Fuster *et al.* (2004) state in their study that, "the true diagnostic value of the exercise ECG relates to its relatively high specificity". As can be seen, these values seem to change somewhat and to quote Bokhari *et al.* (2008) "wide variations in the sensitivity and specificity of the exercise ECG for the diagnosis of coronary artery disease (CAD) have been reported". It is interesting to note that the sensitivity and specificity values of these unadjusted studies are low, relative to those obtained in

this study, which was adjusted for risk factors. However, Koide *et al.*'s study (2001), which adjusted for some risk factors give sensitivity and specificity values of 84 % and 90 %, respectively indicating that our values are somewhat higher than usually reported values due to adjustment for risk factors. In general, it can be observed that the sensitivity and specificity values obtained after using missing value imputation, are similar to world wide standards. Though the Bruce protocol method gave reasonable values for the sensitivity and specificity of the CST, that does not, of course, rule out the fact that other methods are better or worse.

A very interesting sub-finding of this study was the interaction terms obtained in the final cut-off model. It is generally accepted that gender and age (Roger *et al*., 1998) can have a marked impact on the CST results. So much so, that the Bruce protocol cut-offs are adjusted for age.

The reason for the odds ratio of getting CAD for male *versus* that of female to increase dramatically for those who had passed the CST could be due to the fact that women who had CAD were more sensitive to the stress test than were the males. That is, if we take the group of individuals who passed the stress test, we can assume there to be a very few females with CAD, as opposed to males who may still have CAD but managed to pass the CST due to better fitness rates. This could explain why there was a positive interaction for gender with CST. Another interesting finding was the positive interaction term for age with the level passed in the CST. This implies that as a person gets older, the impact of the CST lessens. Medical practitioners in Sri Lanka also state that if a younger person fails the CST that would imply that the patient has a higher chance of having CAD than an older individual. This interaction term appears to agree with this hypothesis. Yet the interaction with diabetes cannot be explained by the above arguments for gender or age. The overall reasoning behind why gender, age and diabetes had positive interactions with CST could be due to the fact that when an individual has CAD, the CST predicts it well, hence obscuring the impact of the other risk factors as opposed to the case where they failed it. Further, the small counts observed may have exaggerated the actual estimates and also resulting in the large confidence intervals obtained.

It can, however, be argued that this observation comes as a result of the imputation procedure. But since the imputation was carried out using many other variables such as systolic and diastolic blood pressure, highly correlated variables such as alcohol and cigarette consumption and even variables such as marital status, imputation appears to be an unlikely cause. That is, due to the inclusion of a large number of other variables in the imputation procedure, which were independent yet highly correlated with CAD, it would be expected that the impact of a few of these variables to be lessened and not strengthened. This would be an interesting topic for further research.

REFERENCES

1. Acock A. C. (2005). Working with missing values. *Journal of Marriage and Family* **67**(4): 1012–1028.
2. Agresti A. (2007). *An introduction to Categorical Data Analysis*. Wiley- Interscience, New Jersey, USA.
3. Allison P. D. (2005). Imputation of categorical variables with PROC MI. 30-113. *Proceedings of the 30th Annual SAS® Users Group International Conference*, April 10 -13, Philadelphia, Pennsylvania, pp. 30.
4. American Diabetes Association. Screening for type 2 diabetes (2004). *http://www.guideline.gov*.
5. Atukorale D.P. (2005). Is coronary angiography indicated in every acute chest pain? *The Island, http://www.island.lk/2005/07/23/features7.html.*
6. Bokhari S., Shahzad A. & Bergmann S.R. (2008). Superiority of exercise myocardial perfusion imaging compared with the exercise ECG in the diagnosis of coronary artery disease. *Coronary Artery Disease* **19**(6): 399 − 404.
7. Bolland K.M., Sooriyarachchi M.R. & Whitehead J. (1998). Sample size review in head injury trial with ordered categorical responses. *Statistics in Medicine* **17**(24): 2835−2847.
8. Collett D. (1991). *Modelling Binary Data*. Chapman & Hall, London,UK.
9. Cronin A.M. & Vickers A.J. (2008). Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are a few false negatives : a simulation study. *BMC Medical Research Methodology* **8**: 75.
10. Dorfman D.D. & Alf E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology* **6**(3): 487−496.
11. Fuster V., Alexander R.W., O'Rourke R.A., Roberts R., King S.B. & Prystowsky E.N. (2001). *Hurst's the Heart*. McGraw-Hill Professional, Columbus, USA.
12. Greenland P. *et al*. (17 authors) (2010). ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation* **122**(25): e584−e636.
13. Hanley J.A. & McNeil B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**(3): 839 − 843.

14. Harel O. & Zhou X.H. (2006). Multiple imputation for correcting verification bias. *Statistics in Medicine* **25**(22): 3769−3786.
15. Hill J. & Timmis A. (2002). ABC of clinical electrocardiography Exercise tolerance testing. *British Medical Journal* **324**: 1084
16. Hosmer D.W. & Lemeshow S. (2000). *Applied Logistic Regression*. John Wiley & Sons, Inc., Columbus, USA.
17. Hua H. (2009). Correcting verification bias in the assessment of the accuracy of diagnostic tests. *Ph.D. thesis*, University of Rochester, New York, USA.
18. Kish L. (1995). *Survey Sampling*. Wiley-Interscience, Columbus, USA.
19. Koide Y., Yotsukura M., Yoshino H. & Ishikawa K. (2001). A new coronary artery disease index of treadmill exercise electrocardiograms based on the step-up diagnostic method. *American Journal of Cardiology* **87**(2):142−147.
20. Lachin J. M. (1977). Sample size determinations for r x c comparative trials. *Biometrics* **33**(2): 315−324.
21. Lauer M.S., Murthy S.C., Blackstone E.H., Okereke I.C. & Rice T.W. (2007) Fluorodeoxyglucose uptake by position emission tomography for diagnosis of suspected lung cancer. *Archives of Internal Medicine* **167**: 161−165.
22. Marshall A., Altman D.G., Holder R.L. & Royston P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology* **9**:57.
23. Mcneer J.F., Margolis J.R., Lee K.L., Kisslo J.A., Peter R.H., Kong Y., Behar V.S., Wallace A.G., McCants C.B. & Rosati R.A. (1978). The role of the exercise test in the evaluation of patients for ischemic heart disease. *Circulation* **57**(1): 64 –70.
24. Mehta K., Rustagi M., Kohli S. & Tiwari S. (2007). Implementing multiple imputation in an automatic variable selection. *Proceedings of the NESUG Conference*, November, Baltimore, Maryland, USA.
25. Metz C. E., Herman B. A. & Roe C. A. (1998). Statistical comparison of two ROC- curve estimates obtained from partially-paired datasets. *Medical Decision Making* **18**(1): 110 –121.
26. Roger V.L., Jacobsen S.J., Pellikka P.A., Miller T.D., Bailey K.R. & Gersh B.J. (1998). Gender differences in use of stress testing and Coronary heart disease mortality : a population-based study in Olmsted County, Minnesota. *Journal of the American College of Cardiology* **32**(2): 345−352.
27. Rosenbaum P.R. & Rubin D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1): 41− 55.
28. Royston P. (2005). Multiple imputation of missing values: update of ice. *The Stata Journal* **5**(4): 527 − 536.
29. Royston P. & Babiker A. (2002). A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome. *The Stata Journal* **2**(2): 151–163.
30. Rubin D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc., Columbus, USA.
31. Sahai H. & Khurshid A. (1996). Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two sample design: a review. *Statistics in Medicine* **15**(1): 1–21.
32. Schafer J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London, UK.
33. Sterne J., White I., Carlin J. & Carpenter J. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal* **338**: 2393
34. Tan M.T., Tian G. & Ng K.W. (2010). *Bayesian Missing Data Problems*. Chapman & Hall, London, UK.
35. Ury H.K. & Fleiss J.L. (1980). On approximate sample sizes for comparing two independent proportions with the use of Yates correction. *Biometrics* **36**(2): 347 − 351.
36. Van-Leeuwen M.V., Zweers E.J.K., Opmeer B.C., van Ballegooie E., ter Brugge H.G., de Valk H.W., Mol B.W. & Visser G.H. (2007). Comparison of accuracy measures of two screening tests for gestational diabetes mellitus. *Diabetes Care* **30**(11): 2779 −2784.
37. Vergara I.A., Norambuena T., Ferrada E., Slater A.W. & Melo F. (2008). StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatics* **9**:265.
38. Wake R. & Yoshiyama M. (2009). Gender differences in ischemic heart disease. *Recent Patents on Cardiovascular Drug Discovery* **4**(3): 234 − 240.
39. Wilson P.W.F., D'Agostino R.B., Levy D., Belanger A.M., Silbershatz H. & Kannel W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation* **97**:1837 −1847.
40. Yuan Y.C. (2001). Multiple imputation for missing data: concepts and new development SAS/STAT 8.2. *SAS Institute Inc. Cary, NC*. Available at *http://www.sas.com/statistics*.