Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2022, 2022:21

https://doi.org/10.28919/cmbn/7089

ISSN: 2052-2541

A MULTIPLE IMPUTATION APPROACH TO EVALUATE THE ACCURACY OF DIAGNOSTIC TESTS IN PRESENCE OF MISSING VALUES

AHMED M. GAD[1,*], ASMAA A. M. ALI[2], RAMADAN H. MOHAMED[2]

[1]Business Administration Department, Faculty of Business Administration, Economics and Political Science

The British University in Egypt (BUE) El Sherouk City, Cairo, Egypt

[2] Statistics Department, Faculty of Economics and Political Science, Cairo University, Egypt

Abstract: Diagnostic tests are used to determine the presence or absence of a disease. Diagnostic accuracy is the main tool to evaluate a test. Four accuracy measures are used to evaluate how well the results of the test under evaluation (index test) agree with the outcome of the reference test (gold standard). These measures are sensitivity, specificity, positive predictive value and negative predictive value. Some subjects are only measured by a subset of tests which result in missing values. This leads to biased results. The mechanism of missing data could be missing completely at random (MCA), missing at random (MAR), or missing not at random (MNAR). Various methods such as the complete-case analysis (CCA) and the maximum likelihood (ML) method are used to handle missing data. Also, imputation methods could be used. The article aims to use a multiple imputation approach to evaluate binary diagnostic tests with missing data under the MCAR mechanism. The proposed approach is applied to a real data set. Also, a simulation study is conducted to evaluate the performance of the proposed approach.

*Corresponding author

E-mail address: ahmed.gad@feps.edu.eg

Received December 16, 2021

Keywords: diagnostic accuracy measures; diagnostic tests; missing completely at random; multiple imputation by chained equations approach.

2010 AMS Subject Classification: 62-07, 62P10.

1. INTRODUCTION

Clinicians use diagnostic tests to determine the presence or absence of a disease. Accurate diagnosis of a disease is often the first step toward its treatment and prevention. The aim of diagnostic accuracy studies is usually to find out the ability of a test to differentiate between patients with and without a disease. The presence or absence of a disease is determined by a gold standard test [3]. The gold standard is the best available test with known results. Usually, the gold standard test is often invasive or expensive. The results of a new non-invasive test (the index test) are compared with the results of the gold test. The basic structure of all diagnostic test studies is to select a series of patients to receive the index test(s) then followed by the gold standard test. Finally, the results of the index test and the gold standard are used to estimate the accuracy parameters. The accuracy measures express how well the results of the test under evaluation (index test) agree with the outcome of the gold standard test. These measures are the sensitivity, specificity, negative predictive value and positive predictive value Sensitivity of a test is the probability of testing positive given the presence of disease. Specificity of a test refers to the probability of testing negative given the absence of the disease. Positive predictive value of a test is the probability of disease given testing positive. Negative predictive value of a test is the probability of no disease given testing negative [9].

Missing values are very common in medical studies and in diagnostic tests. Missing data can be caused by several mechanisms. Missing data mechanism is called missing completely at random (MCAR) if the probability that an observation is missing is not related to any other patient characteristics. A mechanism is said to be missing at random (MAR) when the reason for the missingness is based on other observed patient characteristics. If the probability that an

observation is missing depends on information that is not observed, the missing data are called missing not at random (MNAR) [10].

There are various approaches that are used to handle missing data. These methods ranges from the complete case analysis (CCA) to the imputation techniques. The multiple imputation (MI) is very common as imputation method. The MI method consists of three steps. In the first step an M $(M>1)$ complete (imputed) data sets are obtained by filling each missing value M times using a convenient imputation model. In the second step the analysis of the M data sets is conducted using standard complete-data techniques. In the third step the results from the M imputed complete data sets are combined in an appropriate way to obtain the estimates [10].

A well-known variant of multiple imputation technique is the multiple imputation using the chained equations (MICE) Approach. The MICE is a practical approach for generating imputation based on a set of imputation models, given that there is one model for each variable with missing values. Consider a set of variables $Y_1 \ldots\ldots Y_k$, where some or all have missing values. If Y_1 has missing values, it will be regressed on the other variables Y_2 *to* Y_k. The missing values of Y_1 are simulated from $P(Y_1 \mid Y_2^t, Y_3^t, \ldots\ldots, Y_k^t)$, where, t is an iteration counter and the estimation is thus restricted to individuals with observed Y_1. If Y_2 has missing values, Y_2 is regressed on all the other variables Y_1^{t+1}, Y_3 to Y_k. The missing values of Y_2 are simulated from $P(Y_2 \mid Y_1^{t+1}, Y_3^t, \ldots\ldots, Y_k^t)$, the estimation is thus restricted to individuals with observed Y_2. The process is repeated for all other variables with missing values in a cycle. In order to stabilize the results, the procedure is generally repeated for several replications (e.g. 10 or 20) to produce a single imputed data set, and the whole procedure is repeated M times to give M imputed data sets [2].

The aim of this article is to use a multiple imputation approach to evaluate binary diagnostic tests in the presence of missing values under the MCAR assumption. Multiple imputation by chained equations (MICE) is suggested for the evaluation of binary diagnostic tests with missing values.

The rest of the article is organized as follows. In Section 2, different methods to evaluate the accuracy measures are presented. In Section 3, literature review of diagnostic test with missing

values issue is presented. The proposed MI technique is described in Section 4. Section 5 is devoted to apply the proposed techniques to a dataset described in [9]. In Section 6, the performance of MICE in evaluating diagnostic tests is evaluated using a simulation study. Finally, discussion and conclusion are presented in Section 7.

2. Evaluation of the Accuracy Measures

There are four methods to evaluate the four accuracy measures. They are the simple proportion method, the logistic modelling method and the GEE method. The simple proportion method is the simplest one. It ignores the correlation among tests. The four diagnostic measures of the test and their standard errors are obtained as follows. The sensitivity of a test and its standard error are:

$$\text{Sens (test)} = P(test+ \mid disease+) = \frac{TP}{TP+FN} \quad \text{and} \quad \text{SE (Sens)} = \sqrt{\frac{Sens(1-Sens)}{TP+FN}}.$$

The specificity of a test and its standard error are:

$$\text{Spec (test)} = P(test- \mid disease-) = \frac{TN}{FP+TN} \text{ and SE (Spec)} = \sqrt{\frac{Spec(1-Spec)}{FP+TN}}.$$

The positive predictive value of a test and its standard error are:

$$\text{PPV (test)} = P(disease+ \mid test+) = \frac{TP}{TP+FP} \text{ and SE (PPV)} = \sqrt{\frac{PPV(1-PPV)}{TP+FP}}.$$

The negative predictive value of a test and its standard error are:

$$\text{NPV (test)} = P(disease- \mid test-) = \frac{TN}{FN+TN} \text{ and SE (NPV)} = \sqrt{\frac{NPV(1-NPV)}{FN+TN}},$$

where TP is the true Positive, FP is the false positive, FN is the false negative and TN is the true negative [6].

The logistic regression model may be used to estimate the accuracy parameters. The dependent variable (Y) is defined as the dichotomous results of the test. The presence or absence of disease, as defined by the "gold standard", is included as a binary explanatory variable (D) [5] as follows:

$$Logit\ (p) = ln\left(\frac{P}{1-P}\right) = \beta_0 + \alpha D + \beta_1 X_1 + \cdots + \beta_k X_k,$$

where $p = P(Y=1 \mid D, X_1, \ldots\ldots, X_k)$, D represents the disease status and $X_1, \ldots\ldots, X_k$ are other covariates in the model. Because $p = P(Y=1 \mid D, X_1, \ldots\ldots, X_k) = \frac{e^{\beta_0 + \alpha D + \beta_1 X_1 + \cdots + \beta_k X_k}}{1+e^{\beta_0 + \alpha D + \beta_1 X_1 + \cdots + \beta_k X_k}}$, then,

Sens (test) $=P(\text{test} + \mid \text{disease}+) = P\ (Y=1 \mid D=1,\ X_1, \ldots\ldots,\ X_k) = \frac{e^{\beta_0 + \alpha + \beta_1 X_1 + \cdots + \beta_k X_k}}{1+e^{\beta_0 + \alpha + \beta_1 X_1 + \cdots + \beta_k X_k}}$,

Spec (test) $= P\ (\text{test} - \mid \text{diease}-) = P\ (Y=0 \mid D=0,\ X_1, \ldots\ldots,\ X_k\ ) = 1 - \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k}}{1+e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k}}$. Also, PPV and NPV can be estimated using the logistic regression model by defining the dependent variable (Y) to be the presence or absence of disease and the diagnostic test result is included as a binary explanatory variable in the model. PPV (test) is obtained the same as Sens(test) above. NPV (test) is obtained the same as Spec (test) above.

Leisenberg et al. [8] demonstrated how to estimate sensitivity and specificity for two and three diagnostic tests, respectively, using the generalized estimating equation (GEE) approach. Sternberg and Hadgu [11] generalized their formulation to estimate the sensitivity and specificity when J diagnostic tests are applied to each subject. This is done by assume arbitrary number of covariates measured on each subject. These include R_i is an ordered set of indexes that correspond to the diagnostic tests applied or observed on the ith individual, n_i is the size of the set R_i and let j_k be the kth element of the set R_i. Also, Y_{ij_k} is the outcome of diagnostic test j_k performed for individual i, where $k = 1, \ldots\ldots, n_i$, such that

$$Y_{ij_k} = \begin{cases} 1 & \text{positive} \\ 0 & \text{negative} \end{cases} \quad j_k \in R_i.$$

Let D_{ij_k} be the result of the perfect gold standard, which indicates disease status, such that

$$D_{ij_k} = \begin{cases} 1 & \text{disease} \\ 0 & \text{no disease} \end{cases} \quad j_k \in R_i.$$

For the J diagnostic tests, we create a set of $J-1$ indicator variables, and arbitrarily assume that the test 1 be the reference test and let

$$T_{ij_km} = \begin{cases} 1 & \text{if } j_k = m \\ 0 & \text{otherwise} \end{cases} \quad j_k \in R_i \quad \text{and } m=2,\ldots., J\ .$$

Let $X_{ij_k}^T = (X_{ij_k1}, X_{ij_k2}, \ldots\ldots\ldots, X_{ij_kp})$ be the $1\times P$ vector of covariates corresponding to Y_{ij_k}. A reasonable model to estimate sensitivity and specificity for an arbitrary test j, for a given set of covariates X_{ij} is

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 D_{ij} + \beta_2 T_{ij2} + \cdots + \beta_J T_{ijJ} + \beta_{J+1} D_{ij} T_{ij2} + \cdots + \beta_{2J-1} D_{ij} T_{ijJ} + \beta_{2J} X_{ij1} + \beta_{2J+1} X_{ij2} + \cdots + \beta_{2J+p-1} X_{ijp}.$$

For j =1 Sens (test) = $\frac{e^{\beta_0+\beta_1+\beta_*' x_{i1}}}{1+e^{\beta_0+\beta_1+\beta_*' x_{i1}}}$ and Spec (test) = $\frac{1}{1+e^{\beta_0+\beta_*' x_{i1}}}$. For j = 2, 3,….., J Sens (test) = $\frac{e^{\beta_0+\beta_1+\beta_j+\beta_{j+J-1}+\beta_*' x_{ij}}}{1+e^{\beta_0+\beta_1+\beta_j+\beta_{j+J-1}+\beta_*' x_{ij}}}$ and Spec (test) = $\frac{1}{1+e^{\beta_0+\beta_j+\beta_*' x_{ij}}}$, where, β_*' is a $P \times 1$ vector of regression coefficient for all covariates (X) in the model.

3. DIAGNOSTIC TESTS WITH MISSING DATA

Barnhart and Kosinski [1] studied the use of subunit-level sensitivities and specificities for assessing the performance of a diagnostic test performed at the subunit level. They obtained an adjusted formula for estimates of the subunit sensitivities and specificities under the assumption that the subunit disease status is missing at random. They introduced a WLS approach for inference concerning correlated subunit-level sensitivities and specificities, especially for testing the equality of subunit-level sensitivities and the equality of subunit-level specificities.

Poleto et al. [9] presented data extracted from an observational study to diagnose endometriosis (D) by a laparoscopy procedure (gold standard) versus three diagnostic tests; Ultrasonography (US), Magnetic Resonance (MR) and Echo-Colonoscopy (EC). They considered models that ignore the missing data mechanism such as the complete case analysis (CCA) method. Also, they considered models that include the missing data mechanism, such as the maximum likelihood methods (ML). The ML method showed better performance comparable to the CCA under missing completely at random (MCAR) and with high rates of missingness.

Zhang et al. [16] developed an EM algorithm-based approach to evaluate the diagnostic accuracy of multiple imperfect tests in the absence of a gold standard under either an MAR assumption or an MNAR mechanism. The tests are assumed to be independent conditional on the true disease status. They applied the proposed methods to a real data set from the National Cancer Institute (NCI) colon cancer family registry on diagnosing microsatellite instability for hereditary non-polyposis colorectal cancer.

Jimenez and Nofuentes [7] introduced a global hypothesis test to simultaneously compare the positive predictive values of two or more binary diagnostic tests when the disease status (either present or absent) is unknown for a subset of individuals. The global hypothesis test is based on the chi-squared distribution and can be solved through the method of maximum likelihood and the delta method.

4. THE PROPOSED APPROACH

Multiple imputation is achieved by three steps. First, generating M ($M > 1$) complete (imputed) data sets by filling each missing value M times. Second, analyzing the M imputed complete data sets using standard complete-data technique. This step includes estimation of the four parameters; sensitivity, specificity, positive predictive value and negative predictive value. There are different methods for estimating sensitivity, specificity, positive predictive value and negative predictive value. Finally (the third step), combining the analysis results from the M imputed complete data sets in appropriate way. The evaluation of diagnostic tests will be carried out based on the combined estimates of the parameters from the third step.

The adopted approach is the multiple imputation by chained equations (MICE). The steps of the MICE to evaluate binary diagnostic tests in the presence of missing values can be conducted as follows. The MICE needs an imputation model and an analysis model. These models need to be compatible. This means that all the variables in the analysis model need to be included in the imputation model (regardless of whether they contain missing values or not) including the dependent variable. If the analysis model contains interactions and the imputation is performed through standard MICE, this must be specified in the imputation model as well to avoid bias in the analysis [14]. The interaction terms will be incomplete if the variables that make up the interaction are incomplete. A reasonable imputation method is considered to take into account the interaction during the imputation process [13]. The interactions are included in the third analysis method: the GEE method [11].

Let $\widehat{Q}_i$ and $\widehat{U}_i$ be the point and variance estimates of the needed parameter from the ith imputed data set, *i*=1, 2, ..., *M*. Then the point estimate for Q from multiple imputations is the average of the *M* complete-data estimates [15]:

$$\bar{Q} = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \widehat{Q}_{\mathrm{i}} \ .$$

Let $\overline{\mathrm{U}}$ be the within-imputation variance, which is the average of the *M* complete-data estimates $\bar{U} = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \widehat{U}_{\mathrm{i}}$ and B be the between-imputation variance $B = \frac{1}{m-1} \sum_{\mathrm{i}=1}^{\mathrm{m}} (\widehat{Q}_i - \bar{Q})^2$

Then the variance estimate associated with $\overline{\mathrm{Q}}$ is the total variance

$$T = \bar{U} + (1 + \frac{1}{m}) B$$

5. Application: Comparing the Accuracy of Such Non-Invasive Tests (US, MR and EC) with Missing Values

We apply the proposed technique to the data presented in Poleto et al. [9]. The data consider 219 patients submitted to a laparoscopy procedure (gold standard) to diagnose endometriosis (D) were also evaluated with one or more non-invasive methods (ultrasonography (US), magnetic resonance (MR) and echocolonoscopy (EC)). The true status of the patients is determined by laparoscopically (gold standard). The frequencies of patients with positive (+) and negative (−) results are presented in Table 1.

All the 219 patients were evaluated via US. Out of these 219 patients, 91 had additional MR measurements only, 17 were also evaluated via EC, 13 had both (MR and EC) measurements, and for 98 patients, neither MR nor EC measurements were available.

The main reason for the missing data was the unavailability of the corresponding equipment at the occasion of evaluation; in some cases, diagnostic tests results for retrocervical endometriosis were missing because some patients were only evaluated for endometriosis occurring at other sites.

Table 1: Observed frequency of patients

			D	
US	MR	EC	-	+
		-	6	0
	-	+	1	0
		Missing	51	1
		-	0	0
-	+	+	0	0
		Missing	4	1
		-	3	1
	Missing	+	3	1
		Missing	51	2
		-	0	1
	-	+	0	2
		Missing	0	21
		-	0	1
+	+	+	0	2
		Missing	1	12
		-	0	4
	Missing	+	0	5
		Missing	2	43

Analyses that depend on all the observed data are not generally implemented in statistical packages, so many users pragmatically decide to consider only some subset of the data which can be analyzed using the available software.

Therefore, MICE is considered a suitable choice for comparing the three tests (US, MR and EC) because multiple imputation generally assumes that the data are, at the least, MAR. This

approach can also be used on data that are MCAR [2]. Since the partial response rate is above 10%, this means that using an MICE framework to handle missing data is appropriate [14]. After generating the imputations, we have 35 complete datasets for the simple proportion method, for the LR model (because the percent of missing values = 35%) and 5 for the GEE method. We estimated the four parameters; the sensitivity (Sens), the specificity (Spec), the positive predictive value (PPV) and the negative predictive value (NPV) for each diagnostic test (EC, MR and US) on each complete dataset.

We employed the multiple imputation by chained equations (MICE) approach. There are three estimation methods included in the second step of multiple imputation. The three estimation methods are: the simple proportion method, the logistic regression approach and the GEE approach. In our data, $\hat{Q}_i$ represents the estimate for each parameter of the four parameters (Sens, Spec, PPV and NPV) for EC, MR and US on each complete dataset. $\widehat{U}_{\mathrm{i}}$ is the variance estimate for each parameter of the four parameters (Sens, Spec, PPV and NPV) for EC, MR and US on each complete dataset. Then we can calculate $\overline{Q}$,$\overline{U}$, B and T in straightforward way.

The results are presented in Table (2). The results show that the estimates of sensitivity, specificity, positive predictive value and negative predictive values using the simple proportion method are consistent with those of the LR model. The GEE approach showed an improvement in the estimates of the $\mathrm{Sens}_{(\mathrm{MR})}$, $\mathrm{Sens}_{(\mathrm{EC})}$, $\mathrm{NPV}_{(\mathrm{MR})}$ and $\mathrm{NPV}_{(\mathrm{EC})}$. The $\widehat{\mathrm{Sens}_{(\mathrm{MR})}}$ increased from 0.4071 to 0.4385. The $\widehat{\mathrm{Sens}_{(\mathrm{EC})}}$ increased from 0.5997 to 0.6652. The $\widehat{\mathrm{NPV}_{(\mathrm{MR})}}$ increased from 0.657 to 0.6614. The $\widehat{\mathrm{NPV}_{(\mathrm{EC})}}$ increased from 0.6677 to 0.6838. In contrast, the estimates of the $\mathrm{Spec}_{(\mathrm{MR})}$, $\mathrm{Spec}_{(\mathrm{EC})}$, $\mathrm{PPV}_{(\mathrm{MR})}$ and $\mathrm{PPV}_{(\mathrm{EC})}$ that are obtained under the GEE method became lower. The $\widehat{\mathrm{Spec}_{(\mathrm{MR})}}$ decreased from (0.9009) to 0.8703. The $\widehat{\mathrm{Spec}_{(\mathrm{EC})}}$ decreased from 0.6478 to 0.5548. The $\widehat{\mathrm{PPV}_{(\mathrm{MR})}}$ decreased from 0.7659 to 0.7285. The $\widehat{\mathrm{PPV}_{(\mathrm{EC})}}$ decreased from 0.5837 to 0.5338. According to the three estimation methods, the (US) test has the highest estimates of Sens is 0.9381, Spec is 0.9754, PPV is 0.9681 and NPV is 0.95). The MR test has higher estimates of

Spec and PPV than the estimates of Spec and PPV of the EC test. The EC test has higher estimates of Sens and NPV than the estimates of Sens and NPV of the MR test.

Table (2): Analysis of MICE (through the three estimation methods) under MCAR (MR × EC × US) for the four diagnostic measures.

	Simple proportion		LR		GEE	
Parameter	**Estimate**	**Std. error**	**Estimate**	**Std. error**	**Estimate**	**Std. error**
$Sens_{(MR)}$	0.4071	0.0712	0.4064	0.0718	0.4385	0.0934
$Sens_{(EC)}$	0.5997	0.0944	0.6004	0.0975	0.6652	0.1953
$Sens_{(US)}$	0.9381	0.0245	0.9381	0.0245	0.9381	0.0245
$Spec_{(MR)}$	0.9009	0.0388	0.9044	0.0392	0.8703	0.0563
$Spec_{(EC)}$	0.6478	0.1257	0.6571	0.1314	0.5548	0.0934
$Spec_{(US)}$	0.9754	0.014	0.9754	0.014	0.9754	0.014
$PPV_{(MR)}$	0.7659	0.0856	0.7715	0.0868	0.7285	0.0972
$PPV_{(EC)}$	0.5837	0.1009	0.5863	0.1037	0.5338	0.0828
$PPV_{(US)}$	0.9681	0.0181	0.9681	0.0181	0.9681	0.0181
$NPV_{(MR)}$	0.657	0.0432	0.6573	0.0434	0.6614	0.0521
$NPV_{(EC)}$	0.6677	0.0754	0.6702	0.0752	0.6838	0.1548
$NPV_{(US)}$	0.952	0.0191	0.952	0.0191	0.952	0.0191

Ranking of the three tests US, EC & MR according to the three estimation methods are presented in Table 3. The results show that the US test has the first rank of the estimates (Sens, Spec, PPV and NPV). The MR test has the second rank of the estimates (Spec and PPV). The EC test has second rank of the estimates (Sens and NPV).

Table 3: Ranking of the three tests (US, EC & MR) according to the three estimation methods

parameter / test	US	MR	EC
Sens	1	3	2
Spec	1	2	3
PPV	1	2	3
NPV	1	3	2

The *p*-values are presented in Table 4. The results show that the P-values obtained under the simple proportion method are consistent with those obtained under the LR model.

Table 4: The P-values of the three methods

	Simple proportion		**LR**		**GEE**	
	$Sens_{(US)}$	$Sens_{(EC)}$	$Sens_{(US)}$	$Sens_{(EC)}$	$Sens_{(US)}$	$Sens_{(EC)}$
$Sens_{(MR)}$	0.00012	1.24052	0.00012	1.31119	0.00012	3.67002
$Sens_{(EC)}$	0.00626		0.00938		1.96416	
	$Spec_{(US)}$	$Spec_{(EC)}$	$Spec_{(US)}$	$Spec_{(EC)}$	$Spec_{(US)}$	$Spec_{(EC)}$
$Spec_{(MR)}$	0.85477	0.65075	1.05632	0.85854	0.73448	0.04147
$Spec_{(EC)}$	0.11471		0.1927		0.00012	
	$PPV_{(US)}$	$PPV_{(EC)}$	$PPV_{(US)}$	$PPV_{(EC)}$	$PPV_{(US)}$	$PPV_{(EC)}$
$PPV_{(MR)}$	0.24966	2.02214	0.32038	2.05199	0.14731	0.84169
$PPV_{(EC)}$	0.00211		0.00347		0.00012	
	$NPV_{(US)}$	$NPV_{(EC)}$	$NPV_{(US)}$	$NPV_{(EC)}$	$NPV_{(US)}$	$NPV_{(EC)}$
$NPV_{(MR)}$	0.00012	10.8215	0.00012	10.5862	0.00012	10.6033
$NPV_{(EC)}$	0.00306		0.00336		0.98548	

The results obtained using the ML approach under MCAR mechanism in Poleto et al. [9] is presented in Table 5 for the sake of comparison with our results. The results show that the parameters estimate that are obtained using the ML approach are so close to the parameters estimates that are obtained under the simple proportion method and the LR model. The estimates of the $\text{Sens}_{(\text{MR})}$, $\text{Sens}_{(\text{EC})}$, $\text{NPV}_{(\text{MR})}$ and $\text{NPV}_{(\text{EC})}$ that are obtained under the GEE approach are higher than those obtained by the ML approach. The $\widehat{\text{Sens}_{(\text{MR})}}$ increased from 0.391 to 0.4385. The $\widehat{\text{Sens}_{(\text{EC})}}$ increased from 0.592 to 0.6652. The $\widehat{\text{NPV}_{(\text{MR})}}$ increased from 0.653 to 0.6614. The $\widehat{\text{NPV}_{(\text{EC})}}$ increased from 0.675 to 0.6838. In contrast, the GEE approach has lower estimates of the $\text{Spec}_{(\text{MR})}$, $\text{Spec}_{(\text{EC})}$, $\text{PPV}_{(\text{MR})}$ and $\text{PPV}_{(\text{EC})}$ than the ML approach. The $\widehat{\text{Spec}_{(\text{MR})}}$ decreased from 0.909 to 0.8703. The $\widehat{\text{Spec}_{(\text{EC})}}$ decreased from 0.674 to 0.5548. The $\widehat{\text{PPV}_{(\text{MR})}}$ decreased from 0.774 to 0.7285. The $\widehat{\text{PPV}_{(\text{EC})}}$ decreased from 0.591 to 0.5338.

Table 5: Analysis of MICE (through the three estimation methods) and the ML approach under MCAR (MR × EC × US) for the four diagnostic measures.

	ML		Simple proportion		LR		GEE	
Para.	**Est.**	**Std. error**	**Est.**	**Std. error**	**Est.**	**Std. error**	**Est.**	**Std. error**
$\text{Sens}_{(\text{MR})}$	0.391	0.072	0.4071	0.0712	0.4064	0.0718	0.4385	0.09344
$\text{Sens}_{(\text{EC})}$	0.592	0.135	0.5997	0.0944	0.6004	0.0975	0.6652	0.19532
$\text{Sens}_{(\text{US})}$	0.938	0.024	0.9381	0.0245	0.9381	0.0245	0.9381	0.02446

$Spec_{(MR)}$	0.909	0.035	0.9009	0.0388	0.9044	0.0392	0.8703	0.05626
$Spec_{(EC)}$	0.674	0.111	0.6478	0.1257	0.6571	0.1314	0.5548	0.0934
$Spec_{(US)}$	0.975	0.014	0.9754	0.014	0.9754	0.014	0.9754	0.01402
$PPV_{(MR)}$	0.774	0.078	0.7659	0.0856	0.7715	0.0868	0.7285	0.09725
$PPV_{(EC)}$	0.591	0.104	0.5837	0.1009	0.5863	0.1037	0.5338	0.08278
$PPV_{(US)}$	0.968	0.018	0.9681	0.0181	0.9681	0.0181	0.9681	0.01813
$NPV_{(MR)}$	0.653	0.042	0.657	0.0432	0.6573	0.0434	0.6614	0.05206
$NPV_{(EC)}$	0.675	0.086	0.6677	0.0754	0.6702	0.0752	0.6838	0.15481
$NPV_{(US)}$	0.952	0.019	0.952	0.0191	0.952	0.0191	0.952	0.01912

6. SIMULATION STUDY

The aim of this simulation study is to evaluate the performance of MICE for estimating diagnostic measures using the three proposed estimation methods.

Simulation Setting

A random reference test variable (D) is generated from the Bernoulli distribution with a mean equal to probability (p). Then, three binary diagnostic tests t_1, t_2, t_3 for each subject are generated from the Bernoulli distribution. For the simple proportion method the mean (p_1) = sensitivity (true

value) for the diseased subjects and mean (p_2) = 1-specificity (true value) for the non-diseased subjects. For the logistic model the mean *pr_1*, *pr_2* and *pr_3*. *pr_1*, *pr_2* and *pr_3* were determined using the following functions (Carsey and Harden, 2014):

$Logit\ (pr\_1) = b_1 + b_2 * D$, where $0 \leq pr\_1 \leq 1$

$Logit\ (pr\_2) = c_1 + c_2 * D$, where $0 \leq pr\_2 \leq 1$

$Logit\ (pr\_3) = r_1 + r_2 * D$, where $0 \leq pr\_3 \leq 1$

The b_1, c_1 and r_1 are regression intercepts. The b_2, c_2, and r_2 are regression coefficients. For the GEE approach, first, we simulated (D) with a mean equal to probability (p) and the indicator variables were also created. To generate $Y_1, \ldots\ldots, Y_J$ (in our data J=3=the number of the tests) correlated binary outcomes given the gold standard (D) and the other indicator variables, NORTA method (Cario and Nelson, 1997; Touloumis, 2016) was used as follows:

For $Y_j \sim F$ for all $j = 1, \ldots, J$ and a given correlation matrix R_y:

1. Generate a random vector $Z = (Z_1, \ldots, Z_J)'$ from a standard multivariate normal distribution with correlation matrix corr $(Z) = R_Z$. The elements of R_Z are calculated by solving numerically $J(J-1)/2$ equations, such that each equation relates corr $(Z_j, Z_{j'})$ with corr $(Y_j, Y_{j'})$ for all $j < j'$.
2. Apply the transformation $Y_j = F^{-1}[\phi(Z_j)]$ for all j, where ϕ is the cumulative distribution of the standard normal distribution. Since $Y_1, \ldots\ldots, Y_J$ are correlated binary outcomes, then F is the CDF of the standard logistic distribution.
3. Then $R_Z \approx R_y$ due to the well-known approximation $\phi(x) = F(x\pi/3)$ for all $x \in \mathbb{R}$ (Cario and Nelson, 1997).

Once data are simulated the missingness in t_2 and t_3 under MCAR are introduced. The process is replicated 1000 times. Then MICE is applied to calculate the percent relative bias (RB %) and the MSE in all common four measures of diagnostic accuracy (Sens, Spec, PPV and NPV) for the three tests.

Simulation Results

The relative bias (RB%) and the mean square errors (MSE) are presented in Table 6.

Table 6: The RB (%) and the MSE of the three estimation methods under MCAR at 0% with US, 52.5% with MR and 86.3% with EC for the four diagnostic measures at n=219.

n=219									
	Simple prop.	**LR**	**GEE**	**Simple prop.**	**LR**	**GEE**	**Simple prop.**	**LR**	**GEE**
	(US)			**(MR)**			**(EC)**		
Sens $_{original}$	0.94	0.94	0.94	0.41	0.41	0.44	0.60	0.60	0.66
Sens	0.94	0.94	0.94	0.42	0.42	0.44	0.59	0.59	0.67
RB$_{Sens}$	-0.07	-0.02	0.17	2.31	2.39	0.00	-1.96	-0.77	0.55
MSE$_{Sens}$	0.01	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.01
Spec $_{original}$	0.97	0.97	0.97	0.90	0.90	0.87	0.65	0.66	0.55
Spec	0.97	0.97	0.97	0.89	0.89	0.88	0.63	0.65	0.55
RB$_{Spec}$	-0.14	0.03	-0.38	-1.44	-1.03	1.52	-2.58	-1.31	-0.33
MSE$_{Spec}$	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01
PPV $_{original}$	0.97	0.97	0.97	0.77	0.77	0.73	0.58	0.59	0.53
PPV	0.97	0.97	0.97	0.76	0.77	0.73	0.57	0.58	0.53
RB$_{PPV}$	-0.03	0.18	-0.23	-1.17	-0.64	0.65	-1.42	-0.27	0.09
MSE$_{PPV}$	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
NPV $_{original}$	0.95	0.95	0.95	0.66	0.66	0.66	0.67	0.67	0.68
NPV	0.95	0.95	0.95	0.65	0.65	0.66	0.65	0.66	0.68
RB$_{NPV}$	-0.19	-0.16	-0.36	-1.07	-0.94	-0.47	-2.26	-0.87	0.13
MSE$_{NPV}$	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01

The results show that the US test, the percent of the RB of the estimates of the four diagnostic measures for 3 estimation method is less than 1. The MSE of the estimates is the same for the simple proportion method and the LR model, but it is slightly higher for the GEE method.

For the MR test, the performance of the simple proportion method and the LR model is the same for the estimation of the four parameters according to the MSE, but the MSE of the estimates of the parameters is slightly higher for the GEE method except for the estimate of the PPV. The simple proportion method and the LR model have approximately the same percent of the RB for the estimate of the Sens. The GEE method records the least percent of the RB for the estimates of the Sens and NPV. The percent of the RB for the estimate of the Spec goes from 1 to 2 for the three estimation methods. The GEE method and the LR model have approximately the same percent of the RB for the estimate of the PPV, but the simple proportion method has higher percent.

For the EC test, the GEE method showed the best performance in the estimation of the four parameters according to the percent RB and the MSE. The LR model has lower RB% than the simple proportion method in the estimation of the four parameters, but the two methods have the same MSE.

7. Discussion and Conclusion

Accurate diagnosis of a disease or classification of a sub-type of a disease is often the first step toward its treatment and prevention. Missing data are common in diagnostic medical settings where some subjects are only measured by a subset of tests (Zhang et al., 2014). Considerable methods are developed to assess the diagnostic accuracy of the index tests whose performance is under evaluation in the presence of the missing values.

The missing data mechanisms can be classified according to the process causing missingness [10]. The missing data mechanisms are: MCAR, MAR, MNAR. Both MCAR and MAR are considered ‘ignorable’ missing data mechanism, as MCAR is a special case of MAR. MNAR is denoted as non-ignorable missing. Also, an important issue to determine the missing data pattern to select the proper imputation method. The missing data patterns are: univariate, monotone and arbitrary.

In this article we considered the multiple imputation by chained equations (MICE) approach to evaluate binary diagnostic tests with missing data under the MCAR assumption. The MICE approach is achieved through three steps. Creating m imputed data sets in the first step. Analyzing the m imputed data sets using the simple proportion method, the LR model and the GEE method in the second step. Finally, combining the estimates from the second step. The applications are conducted using the data analyzed by Poleto et al. [9].

Poleto et al. [9] introduced an ML approach to evaluate three binary diagnostic tests in the presence of missing values. The results that are obtained by the MICE approach including the simple proportion method are consistent with those obtained through the LR model, also they are so close from those obtained by Poleto et al. [9] using the ML approach. The GEE approach has an improvement in the estimates of the parameters, and it also has a drop.

The simulation results of the current study showed that the MICE approach including the simple proportion method and the MICE approach including the LR model have the same performance according to the MSE. The GEE method showed the best performance in the estimation of the four parameters according to the percent RB and the MSE for one of the 3 tests. The percent relative bias of the estimates of the four parameters of the three tests doesn't exceed 3% for the three estimation methods.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] H.X. Barnhart, A.S. Kosinski, Evaluating medical diagnostic tests at the subunit level in the presence of verification bias, Statist. Med. 22 (2003), 2161–2176.

[2] D.S. Bouhlila, F. Sellaouti, Multiple imputation using chained equations for missing data in TIMSS: a case study, Large-Scale Assess Educ. 1 (2013), 4.

[3] J.R. Cardoso, L.M. Pereira, M.D. Iversen, A.L. Ramos, What is gold standard and what is ground truth? Dental Press J. Orthod. 19 (2014), 27–30.

[4] M. C. Cario and B. L. Nelson, Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix, Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois (1997).

[5] S.S. Coughlin, B. Trock, M.H. Criqui, L.W. Pickle, D. Browner, M.C. Tefft, The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test, J. Clinical Epidemiol. 45 (1992), 1–7.

[6] T.S.S. Genders, S. Spronk, T. Stijnen, E.W. Steyerberg, E. Lesaffre, M.G.M. Hunink, Methods for calculating sensitivity and specificity of clustered data: a tutorial, Radiology. 265 (2012), 910–916.

[7] Jimenez, A. E., and Nofuentes, J. A. Comparison of the predicted values of multiple binary diagnostic tests in the presence of ignorable missing data, REVSTAT – Stat. J. 15 (2017), 45-64.

[8] W. Leisenring, M.S. Pepe, G. Longton, A marginal regression modelling framework for evaluating medical diagnostic tests, Stat. Med. 16 (1997), 1263–1281.

[9] F.Z. Poleto, J.M. Singer, C.D. Paulino, Comparing diagnostic tests with missing data, J. Appl. Stat. 38 (2011), 1207–1222.

[10] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, Wiley, New York, (1987).

[11] M.R. Sternberg, A. Hadgu, A GEE approach to estimating sensitivity and specificity and coverage properties of the confidence intervals, Stat. Med. 20 (2001), 1529–1539.

[12] A. Touloumis, Simulating correlated binary and multinomial responses under marginal model specification: The SimCorMultRes Package, The R Journal, 8 (2016), 79–91.

[13] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, J. Stat. Softw. 45 (2011), 1–67.

[14] J.N. Wulff, L. Ejlskov Multiple imputation by chained equations in praxis: guidelines and review, Electron. J. Bus. Res. Methods, 15 (2017), 41-56.

[15] Yuan, Y. C. Multiple imputation for missing data: concepts and new development, In: Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (2000), 267-25, Cary, North Carolina.

[16] Y. Zhang, H. Chu, D. Zeng, Evaluation of incomplete multiple diagnostic tests, with an application in the colon cancer family registry study, J. Appl. Stat. 41 (2014), 688–700.