Dwivedi, Alok Kumar; Mallawaarachchi, Indika; Figueroa-Casas, Juan B.; Morales, Angel M.; Tarwater, Patrick

Article

MULTINOMIAL LOGISTIC REGRESSION APPROACH FOR THE EVALUATION OF BINARY DIAGNOSTIC TEST IN MEDICAL RESEARCH

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Dwivedi, Alok Kumar; Mallawaarachchi, Indika; Figueroa-Casas, Juan B.; Morales, Angel M.; Tarwater, Patrick (2015) : MULTINOMIAL LOGISTIC REGRESSION APPROACH FOR THE EVALUATION OF BINARY DIAGNOSTIC TEST IN MEDICAL RESEARCH, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, NY, Vol. 16, Iss. 2, pp. 203-222, https://doi.org/10.21307/stattrans-2015-011

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 203–222

MULTINOMIAL LOGISTIC REGRESSION APPROACH FOR THE EVALUATION OF BINARY DIAGNOSTIC TEST IN MEDICAL RESEARCH

Alok Kumar Dwivedi[1], **Indika Mallawaarachchi**[2], **Juan B. Figueroa-Casas**[3], **Angel M. Morales**[4], **Patrick Tarwater**[5]

ABSTRACT

Evaluating the effect of variables on diagnostic measures (sensitivity, specificity, positive, and negative predictive values) is often of interest to clinical researchers. Logistic regression (LR) models can be used to predict diagnostic measures of a screening test. A marginal model framework using generalized estimating equation (GEE) with logit/log link can be used to compare the diagnostic measures between two or more screening tests. These individual modeling approaches to each diagnostic measure ignore the dependency among these measures that might affect the association of covariates with each diagnostic measure. The diagnostic measures are computed using joint distribution of screening test result and reference test result which generates a multinomial response data. Thus, multinomial logistic regression (MLR) is a more appropriate approach to modeling these diagnostic measures. In this study, the validity of LR and GEE approaches as compared to MLR model was assessed for modeling diagnostic measures. All methods provided unbiased estimates of diagnostic measures in the absence of any covariate. LR and GEE methods produced more biased estimates as compared to MLR approach especially for small sample size studies. No bias was obtained in predicting sensitivity measure using MLR method for one screening test. Our proposed MLR method is robust for modeling

[1] Assistant Professor, Division of Biostatistics & Epidemiology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, USA.
E-mail: alok.dwivedi@ttuhsc.edu.

[2] Research Associate, Division of Biostatistics & Epidemiology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, USA.
E-mail: indika.mallawaarachchi@ttuhsc.edu.

[3] Associate Professor, Division of Pulmonary and Critical Care Medicine, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, USA.
E-mail: Juan.Figueroa@ttuhsc.edu.

[4] Assistant Professor, Department of Surgery, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, TX 79905, USA. E-mail: angel.morales@ttuhsc.edu.

[5] Professor, Division of Biostatistics & Epidemiology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, USA.
E-mail: patrick.tarwater@ttuhsc.edu.

diagnostic measures of a screening test as opposed to LR method. MLR method and GEE method produced similar estimates of diagnostic measures for comparing two screening tests in large sample size studies. The proposed MLR model for diagnostic measures is simple, and available in common statistical software. Our study demonstrates that MLR method should be preferred as an alternative for modeling diagnostic measures.

Key words: multinomial logistic regression, predictive values, sensitivity, specificity, acute appendicitis, pulmonary abnormalities, medical diagnostic test.

1. Introduction

Diagnostic tests are an essential component in medical care for confirming or establishing the disease diagnosis, evaluating disease prognosis, stratifying risk of disease, and screening for early detection. Clinical researchers conduct studies about diagnostic tests mainly for the purpose of either estimating the diagnostic accuracy of a test according to different patient or environmental characteristics or comparing diagnostic accuracy of different tests according to different patient or environmental characteristics. Very limited statistical methods are available to evaluate the diagnostic measures in regression framework (Leisenring et al., 1997). Studies are required to develop robust statistical methods to analyze data from diagnostic studies and assess the properties of available statistical methods. In this study, we proposed a statistical regression method to analyze data from diagnostic studies.

In diagnostic studies, an investigational/new test is often referred to as screening/diagnostic test and a definite diagnostic test is referred to as the reference or gold standard test. When the screening test and reference test are measured in a binary outcome then various measures are required to assess the performance of screening tests in relation to the reference test. Most commonly used diagnostic performance measures are sensitivity (P{positive test result|disease}), specificity (P{negative test result|no disease}), positive predictive value (P{disease|positive test result}), and negative predictive value (P{no disease|negative test result}) (Leisenring et al., 2000). Sensitivity and specificity are probabilities of the test result measured through a screening test, conditional on disease status measured through a reference test while a predictive value is the probability of disease conditional on the test result measured through a screening test. Clinical researchers are often interested in evaluating these four diagnostic measures of screening tests according to patient and clinical characteristics. Regression approaches are needed to address such clinical questions.

Application of logistic regression (LR) in predicting common diagnostic measures including sensitivity (Se), specificity (Sp), positive predictive value (PPV), and negative predictive value (NPV) of a screening test according to patient or other environmental covariates was proposed by Coughlin et al. (1992). LR models for Se and Sp include reference test result as an independent variable while modeling PPV and NPV include screening test result as a predictor. We

refer to this modeling approach as adjusted LR models for diagnostic measures. The adjusted LR models have been used in clinical studies for evaluating diagnostic measures (Coughlin et al., 1992; Elie et al., 2008). Another alternative is the use of LR models for Se and Sp by restricting the analysis to disease and non-disease group respectively. Similarly, LR models can be used to model PPV and NPV by restricting the analysis to positive screening test result and negative screening test result respectively. We refer to the modeling approaches restricted to a group of individuals as subgroup LR models. Subgroup LR models have also been used in clinical studies (Carney et al., 2003; Laya et al., 1996). Recently an application of LR model for predicting likelihood ratio was also developed (Janssens et al., 2005). Ordinary LR models are sensitive to small sample size and rare events (Nemes et al., 2009; King and Zeng, 2001). Thus, LR models may produce biased estimates of diagnostic measures. Therefore, we determined the bias in estimating diagnostic measures using adjusted and subgroup LR models in presence of a binary cofactor in various scenarios.

The diagnostic measures depend on the four cell frequencies generated from a 2x2 table of screening test result and reference test result. The most natural way is to model the joint distribution of screening test result and reference test result. Typically, each diagnostic measure is modelled independently using LR as a function of risk factors. Since the diagnostic measures are computed using the joint distribution of screening test result and reference test result thus these measures are dependent. Independent modeling of these measures ignores dependency among these measures and that subsequently might affect the association of cofactors with these measures (Puggioni et al., 2008). Since the joint distribution of screening test result and reference test result follows a multinomial distribution, thus a multinomial logistic regression (MLR) can be used to estimate the diagnostic measures. We compared the performance of LR models and MLR model for estimating the common diagnostic measures using simulation studies and our published study data (Figueroa-Casas et al., 2014). We can easily extend the MLR model for comparing two or more screening tests. However, studies involving two or more screening tests often provide paired structure data since each patient usually undergoes through each screening test. Thus, such studies require accounting for clustering effects in the analysis. Sandwich error estimation is commonly used to analyze clustered data, repeated measures data, and data obtained through clustered randomized design. Such procedure provides robust variance estimation. Robust variance approach appropriately accounts for correlation structure in the dataset (Leisenring et al., 1997; Liu, 1998). We suggest using a robust variance approach while modeling diagnostic measures using MLR method for two or more screening tests.

A marginal model framework using generalized estimating equation (GEE) with logit link approach has been proposed to compare diagnostic measures between two or more screening tests. It has been advocated to use independent working correlation matrix for fitting marginal models for diagnostic measures with robust variance estimates (Leisenring et al., 1997; Leisenring et al., 2000).

As discussed earlier for LR models, adjusted GEE and subgroup GEE models can be fitted to compare diagnostic measures between two or more screening tests. Further, the individual approaches to modeling each diagnostic measure through GEE models do not account for dependency among these measures. We also compared the individual modeling approach using GEE models and a joint modeling approach using MLR models for estimating diagnostic measures with simulation studies and real study dataset.

The aim of this study was to propose an alternative regression approach to evaluating a binary diagnostic test based on joint distribution of a new test result with reference test result. Specifically, we evaluated the validity of the proposed MLR approach in estimating diagnostic measures and compared with subgroup and adjusted LR models of diagnostic measures. In addition, we extended MLR approach to modeling more than one screening test for comparing diagnostic measures between screening tests and compared it with GEE approach to modeling diagnostic measures for more than one screening test. The applications of MLR approach for estimating and comparing diagnostic measures were illustrated using data from medical research studies.

2. Methods

2.1. Estimating diagnostic accuracy using a logistic regression (LR) model

Suppose a diagnostic study involves a screening test (T) and a reference test (D). If both the screening test and reference test provide binary (positive/negative) results then data can be summarized using a 2x2 table as presented in Table 1. Se, Sp, PPV, and NPV can be estimated as a/(a+c), d/(b+d), a/(a+b), and d/(c+d) respectively. We need regression approaches to estimate these diagnostic measures in presence of significant patient characteristics or other clinical covariates. LR models (Coughlin et al., 1992) can be used to predict Se, Sp, PPV, and NPV in relation to cofactors.

2.2. Multinomial logistic regression (MLR) for estimating diagnostic accuracy

The common diagnostic measures are based on the four cell frequencies obtained from Table 1. The probabilities of these four cells follow a multinomial distribution. Thus, MLR can be used for estimating common diagnostic measures in presence of patient and environmental covariates. Data summarized in Table 1 have unobserved probability p_k corresponding to each of the 4 cells, where $\sum_{k=1}^{4} p_k = 1$.

The joint probabilities for a screening test (T) and a reference test (D) would be:

P(T=1 and D=1) = True positive probability = a/(a+b+c+d)

P(T=1 and D =0) = False positive probability = b/(a+b+c+d)
P(T=0 and D =1) = False negative probability = c/(a+b+c+d)
P(T=0 and D =0) = True negative probability = d/(a+b+c+d)

A new outcome variable with four categories needs to be generated for applying MLR model. The four categories of the new outcome variable (Y=1, 2, 3, and 4) will be true positive (Y=1: T=1 and D=1), false positive (Y=2: T=1 and D=0), false negative (Y=3: T=0 and D =1), and true negative (Y=4: T=0 and D =0) as described in Table 1. We can fit MLR models by considering any one category as a reference category. For example, if we consider the false negative (Y=3) as a referent category then it compares the likelihood of true positive over false negative which is equivalent to fitting LR model for Se of screening test T. At the same time this model also provides comparison of true negative over false negative which is equivalent to fitting LR model for NPV of screening test T. Thus, a single MLR model can be used to predict Se, Sp, PPV, and NPV of a screening test. However, at least two LR models (one for Se and one for Sp) are needed to estimate all four diagnostic measures.

2.3. Comparing diagnostic accuracy using generalized estimating equation (GEE) and MLR methods

The data needs to be reorganized for comparing diagnostic measures of two or more screening tests using GEE or MLR methods. Suppose we have "n" subjects who underwent two screening tests, then it will be 2n records in a reorganized dataset. In this approach, an indicator variable (Z) is defined to classify each record for each specific test. In other words, each subject will have two records corresponding to each test. Suppose a subject has data on three variables: D, T_1 (screening test 1), and T_2 (screening test 2) in an original dataset, then that subject will have two records: D, T, Z=1 with $T=T_1$ and D, T, Z=0 with $T=T_2$ in a reorganized dataset. The logit or log link under the GEE framework can be applied in the reorganized dataset to compare the diagnostic measures between two screening tests (Moskowitz and Pepe, 2006). The equations for developing GEE models of diagnostic measures are published (Leisenring et al., 1997; Leisenring et al., 2000). MLR models using a robust variance approach can be used to compare diagnostic measures between two or more screening tests by modeling a new dependent variable Y (as described in section 2.2) in the reorganized dataset. The details of LR and MLR models can be found in the Appendix.

3. Data analysis

3.1. Simulation studies.

The performance of MLR as compared with LR models for estimating diagnostic measures was evaluated using Monte Carlo simulation studies. We first created a unique ID variable and a variable (X) from a Bernoulli distribution. We then created a random reference test variable (D) from the Bernoulli distribution with a mean equal to probability (p)

$\text{Logit}(p)=a_1+a_2*X$, where $0\leq p\leq 1$

After that we randomly created a binary screening test (T) for each subject from the Bernoulli distribution having a mean p'. The p' was determined using the following function:

$\text{Logit}(p')=b_1+ b_2*D - b_3*(1-D)$, where $0\leq p'\leq 1$

where a_1 and b_1 are regression intercepts. The a_2, b_2, and b_3 are regression coefficients.

First, we compared the bias in all common four measures of diagnostic accuracy estimated using LR and MLR models in the absence of any cofactor. Then, we focused only on comparing the bias in the estimate of Se of the screening test T. The true Se for screening test T in relation to reference test D was obtained and compared with Se estimated using adjusted LR, subgroup LR, and MLR approaches.

The comparison of MLR and GEE methods for estimating diagnostic measures of two screening tests was also evaluated in various simulation studies as described for a single screening test. We randomly created two binary screening tests (T_1 and T_2) for each subject from the Bernoulli distributions having mean $p^{\dagger}$ and p^{*} respectively. The $p^{\dagger}$ and p^{*} were determined using the following functions:

$\text{Logit}(p^{\dagger})=c_1+ c_2*D - c_3*(1-D) + u_1$, where $0\leq p^{\dagger}\leq 1$

$\text{Logit}(p^{*})=d_1+ d_2*D - d_3*(1-D)+ u_2$, where $0\leq p^{*}\leq 1$

where c_1 and d_1 are regression intercepts. The c_2, c_3, d_2, and d_3 are regression coefficients. To introduce correlation between two screening tests, a random effect component (u_1, u_2) was included for the outcome of each test. The u_1 and u_2 were drawn from a bivariate normal distribution with a known correlation structure. The true Se for screening test T_1 and T_2 in relation to reference test D were obtained and compared with the estimated Se for each test obtained using adjusted GEE, subgroup GEE, and MLR approaches.

The percent relative bias in the estimate was reported. Each simulation study was conducted for a small sample size (100) as well as a large sample size (500). Each simulation study was also conducted for low prevalence (<10%) and

moderate prevalence (20-30%). The effect of different prevalence of a binary cofactor was also examined. Each simulation study was repeated for 1000 simulations. The percent of relative bias was estimated using average of [(true diagnostic value – estimated diagnostic value)*100/true diagnostic value] from 1,000 random data sets. The choice of regression coefficients in the above models was made according to the simulation study. Statistical package STATA 12.1 was used for data analysis.

3.2. Real data analysis

To demonstrate our proposed strategy, we used data from two studies (study I and study II). In study I (single screening test), we were interested in assessing the accuracy of chest radiographs (chest x-ray) to identify bilateral pulmonary infiltrates consistent with acute respiratory distress syndrome in relation to computed tomography (CT, reference test). We used a subgroup LR model to determine the clinical characteristics associated with diagnostic performance measures of chest radiographs. A total of 90 patients met the inclusion criteria and had near simultaneous chest radiograph and CT results to evaluate for specified pulmonary abnormalities. The prevalence of these pulmonary abnormalities was 74% determined using CT (Figueroa-Casas et al., 2014). In the present study, we compared the results of subgroup LR models with our proposed MLR model to assess factors associated with the diagnostic measures of chest radiograph. For study II (two screening tests), we used our motivating study data on acute appendicitis. In study II, a total of 200 patients were evaluated with computer tomography (CT) for the diagnosis of appendicitis. The prevalence of acute appendicitis was found as 95.5%. The surgery residents and radiologists reviewed independently CT for each patient and made diagnosis for acute appendicitis. For each patient, we have pathological diagnosis for acute appendicitis. In this case, pathological diagnosis was considered as a reference test. The aim of this study was to compare the accuracy of CT readings with surgical residents as compared with radiologists. We compared the Se, Sp, PPV, and NPV of CT reading with surgical residents and radiologists in relation to pathological findings using GEE with logit link and robust variance estimation. MLR was also performed to compare Se, Sp, PPV, and NPV of CT reading with surgical residents with radiologists. The results of subgroup LR, subgroup GEE, and MLR approaches were reported using regression coefficient (RC), standard error (SE), and p-value.

4. Results

We found no bias in estimating Se, Sp, PPV, and NPV using either LR (adjusted or subgroup) or MLR methods in the absence of any cofactors. Table 2 shows the percent bias in estimating Se using subgroup LR, adjusted LR, and MLR methods. Subgroup LR model provided biased estimate of Se in the range of 0.06% to 31% while adjusted LR model provided biased estimate of Se in the

range of 0.68% to 38.6% when sample size was 100. There was less bias in the estimate of Se using subgroup LR and no bias using MLR models when sample size was 500. However, we obtained biased estimates of Se in the range of 1.43% to 15.2% using adjusted LR models for sample size 500. The bias in the estimate of Se using LR model was found larger when the prevalence of disease was not similar between the two levels of a cofactor as compared to when the prevalence of disease was similar between the two levels of a cofactor. There was no bias obtained in estimating Se using MLR in any scenario.

Table 3 demonstrates the percent bias in estimating Se using subgroup LR, adjusted LR, and MLR models when the prevalence of disease was moderate (20-30%) for sample size n=100 and n=500. The bias in the estimate of Se using subgroup LR model was less than 8% when the sample size was small while no bias was obtained when the sample size was high (n=500). The bias in the estimated Se using adjusted LR model was obtained from 1% to 19.8% when the sample size was 100 while the bias in the Se using adjusted LR was obtained from 0.35% to 8.47% when the sample size was 500. No bias in any situations was obtained in estimating Se using MLR model.

In summary, the subgroup LR model always provided less biased estimate of Se as compared to adjusted LR model in any scenario. The bias in the estimate of Se was found to be higher when the prevalence of disease was different in different levels of a cofactor. Further, LR model with low prevalent cofactor provided large bias in the estimate of Se as compared to LR model with high prevalent cofactor. The two methods, subgroup LR and MLR, provided unbiased estimate of Se when the disease prevalence was more than 20% and cofactor prevalence was moderate (50%). MLR method always provided an unbiased estimate of Se in any scenario.

Table 4 illustrates the comparison of subgroup LR model and MLR model to evaluate factors associated with the diagnostic measures of chest x-rays in identifying bilateral pulmonary infiltrates consistent with the diagnosis of acute respiratory distress syndrome. Subgroup LR models provided slightly different estimates of regression coefficients and p-values as compared to MLR models. Slightly lower p-values were obtained in the subgroup LR models as compared to MLR models. We further developed LR and MLR models including only gender variable as a cofactor using study I data. We did not find bias in the estimates of diagnostic measures obtained using subgroup LR and MLR models when we included only gender variable. However, less than 3% bias in the estimates of diagnostic measures was obtained using adjusted LR model in study I dataset.

No bias was obtained in estimating any diagnostic measures using different methods for two screening tests in the absence of cofactors. The absolute percent relative bias in the estimate of Se using GEE and MLR methods for different scenarios is shown in Table 5 when the disease prevalence was low. The bias in the estimate of Se was found to be almost similar with subgroup GEE method and MLR method when disease prevalence was low and cofactor prevalence was 50%. However, slightly lower bias in the estimate of Se was obtained using MLR

method as compared to subgroup GEE method when disease prevalence varied according to covariate strata. Less than 10% bias in the estimate of Se was obtained using subgroup GEE and MLR methods when cofactor prevalence was 50%. Adjusted GEE approach provided large bias in the estimate of Se as compared with subgroup GEE and MLR methods. Similar bias pattern was obtained across different methods for estimating Se when cofactor prevalence was 20%. The bias was found to be larger with each method when cofactor prevalence was low.

The absolute percent relative bias in the estimate of Se using GEE and MLR for different scenarios is shown in Table 6 when the disease prevalence was greater than 20%. The range of bias in the estimate of Se using subgroup GEE model and MLR model was found to be 1.71%-5.81% for small sample size (n=100) and 0.33%-3.76% for large sample size (n=500) when cofactor prevalence was 50%. The bias was less than 9% with subgroup GEE and MLR models in an equal prevalence scenario and when the cofactor prevalence was 20%. The subgroup GEE and MLR models both produced bias in estimate of Se up to 16% when cofactor prevalence was 20% and disease prevalence was not the same in different strata. Adjusted GEE method provided very large bias in the estimate of Se in most of the scenarios.

Table 7 delineates a comparison of subgroup GEE and MLR models for determining the differences in diagnostic performance of radiologist CT reading for the diagnosis of appendicitis as compared to surgical residents after adjusting cofactors. Both approaches showed that CT reading with radiologist for the diagnosis of appendicitis had significantly higher Se and lower Sp than CT reading with surgical residents. The p-values obtained from MLR models were slightly different than the p-values obtained using GEE models. The p-values for comparison of specificity between two screening tests were obtained as 0.02 and 0.04 using MLR model and GEE model respectively after adjusting other cofactors.

5. Discussion

The diagnostic measures of a screening test depend upon (1) the cell frequencies generated from a cross-tabulation of screening test result and reference test result, and (2) the study population and clinical characteristics. We need a regression approach to modeling diagnostic measures that describe joint distribution of screening test result and reference test result. We proposed MLR model as direct modeling approach to modeling each common diagnostic measure. We further extended our approach to comparing diagnostic measures of two or more screening tests. The validity of available regression approaches in estimating the diagnostic measures in different scenarios was also estimated in this study. We found that our proposed MLR approach provides unbiased estimates of diagnostic measures as compared to LR methods. We also found our

proposed MLR approach to be more appropriate for comparing two or more screening tests as opposed to adjusted GEE method.

Adjusted LR method provided bias in the estimate of Se up to 19% for small sample size and 12.4% for large sample size, when disease prevalence was low (10%) and cofactor prevalence was 50%. This bias was increased to 31% when the prevalence of covariate was 20% for a low sample size and a low prevalence study. Coughlin et al. (1992) also found 25% bias in the estimate of Se when the prevalence was unequal across covariate strata. In our study, subgroup LR method provided bias in the estimate of Se up to 8% when the prevalence was unequal across covariate strata and prevalence of covariate was 50%. Coughlin et al. (1992) found 7% bias using subgroup model when the prevalence was unequal across covariate strata. In general, adjusted LR model provided biased estimate of Se in all scenarios. Additionally, subgroup LR model provided bias estimates for large sample size studies when disease prevalence was less than 10% and cofactor prevalence was 20%. Our proposed MLR method produced unbiased estimate of Se in all scenarios.

It has been shown that ordinary LR model produces bias estimates for small sample size studies (Nemes et al., 2009; Bergtold et al., 2011). LR model produces large bias when the sample size is small and the outcomes are rare (King and Zeng, 2001). Thus, obviously utilizing LR models for modeling diagnostic measures in such cases will produce biased estimates. Our study demonstrates that MLR is less sensitive to small sample size as compared with LR models for modeling diagnostic measures. Ye and Lord (2014) also showed that MLR model requires smaller sample size as compared with mixed logit model using crash severity data. Further, modeling diagnostic measures directly through MLR approach avoids the dependency problem that arises through individual modeling of each diagnostic measure using LR approach.

In our real data example of accuracy of chest radiograph for detecting pulmonary abnormalities according to gender status, we found no bias in the estimates of any diagnostic measures using subgroup LR and MLR models while up to 3% bias was observed using adjusted LR model. It was expected to obtain unbiased estimates of diagnostic measures using subgroup LR model because disease prevalence (74%) and male gender prevalence (63%) were very high in the study. Slightly lower P-values were obtained in subgroup LR models as compared to MLR models. It has been observed that binary LR models for each pair of multi-response data underestimate the standard errors of the coefficients as compared with MLR model (Agresti, 2007). In other words, ignoring the dependency among diagnostic measures through individual modeling may provide smaller standard errors for the model parameters. Thus, individual modeling of each diagnostic measure using LR model may provide inappropriate inferences as opposed to our proposed MLR method of modeling diagnostic measures.

Subgroup GEE and MLR approaches provided similar results for modeling diagnostic measures of more than one screening test. Subgroup GEE method

produced slightly higher biased estimates as compared with MLR model especially for studies with low sample size and low disease prevalence. Subgroup GEE and MLR approaches provided bias in the estimate of Se up to 9% when disease prevalence was low and up to 6% when disease prevalence was greater than 20% with cofactor prevalence 50%. This bias increased up to 33% when cofactor prevalence was 20%. This bias can be eliminated by restricting the analysis to the specific cofactor strata in MLR or subgroup GEE models. Adjusted GEE approach produced biased estimate of Se in almost all scenarios.

In our real data example for comparison of two screening tests, the Se of CT reading with radiologists was found larger than CT reading with residents. Another study observed no differences in diagnosing acute appendicitis though CT readings by radiology residents as compared with CT readings by radiology faculty (Albano et al., 2001). There were no differences observed between two approaches to comparing each diagnostic measure in the absence of any cofactors in study II dataset. It was expected that the MLR model for multi-response data is similar to modeling two separate logistic regressions in the absence of predictor and interaction (Fidler and Nagelkerke, 2013). Our simulation studies also confirmed these findings that the modeling diagnostic measures through GEE approach and MLR approach provide the same results in the absence of any covariates. Robust variance estimate and independent working correlation matrix were used in GEE models and robust variance estimate was used in MLR models. Despite this, GEE and MLR approaches produced slightly different estimates and p-values for comparing diagnostic measures between two screening tests. This further confirms that ignoring the dependency among diagnostic measures through individual modeling or choosing binomial marginal distribution for estimating diagnostic measures may provide inaccurate results as opposed to joint modeling of screening test result and reference test result using multinomial distribution. Further, it has been demonstrated that the MLR approach is not equivalent to modeling two separate logistic regressions for multi-response data in the presence of an interaction effect. The MLR approach should be preferred over two separate logistic regressions in the presence of a cofactor (Fidler and Nagelkerke, 2013; Miettinen, 1976).

In simulation studies, we have considered only one binary cofactor for the sake of simplicity. The MLR approach can handle both categorical and continuous cofactors. We demonstrated the MLR modeling of common diagnostic measures in presence of a perfect reference test. This approach can also be used in the absence of a perfect reference test. This application is under investigation by us for a future publication. We have shown bias in the estimate of Se measure using different methods. Similarly, we can demonstrate for other diagnostic measures. This study has not provided an inferential comparison in evaluating the association of a cofactor with diagnostic measures using different methods.

6. Conclusions

In this study, we showed MLR model can be used directly for modeling Se, Sp, PPV, and NPV as a function of covariates. We also demonstrated that MLR model can easily be extended for comparing diagnostic measures between more than one screening test. The correlation involved in multiple screening tests can be handled using robust variance approach available in statistical software. Developing MLR models for diagnostic measures is straightforward, simple, and available in common statistical software. In the absence of cofactors, all methods provided unbiased estimates of diagnostic measures. In general, all approaches provided very consistent results in many conditions. The MLR method always produced unbiased estimate of each diagnostic measure of a screening test. Subgroup LR method also produced unbiased estimate of each diagnostic measure in large sample size studies. The results of subgroup GEE and MLR with robust variance estimate for more than one screening test were found consistent. For small sample sizes, subgroup GEE and MLR approaches can produce bias estimates, especially with low prevalent cofactor. In such cases, a restricted analysis of covariate strata can be performed to correct the bias. Adjusted LR and adjusted GEE models should be avoided for predicting diagnostic measures. Subgroup LR and subgroup GEE models can be utilized for estimating diagnostic measures for large sample size studies. However, these methods may provide inaccurate inferences due to ignoring the dependency among the diagnostic measures. We suggest using MLR as an alternative and more appropriate approach to GEE with logit link and LR models for modeling Se, Sp, PPV and NPV.

REFERENCES

AGRESTI, A., (2007). An Introduction to Categorical Data Analysis. John Wiley & Sons, Inc., Hoboken, New Jersey, p. 174.

ALBANO, M. C., ROSS, G. W., DITCHEK, J. J., DUKE, G. L., TEEGER, S., SOSTMAN, H. D., FLOMENBAUM, N., SEIFERT, C., BRILL, P. W., (2001). Resident Interpretation of Emergency CT Scans in the Evaluation of Acute Appendicitis. Academic Radiology, 8, 915−918.

BERGTOLD, J. S., YEAGER, E. A., FEATHERSTONE, A., (2011). Sample Size and Robustness of Inferences from Logistic Regression in the Presence of Nonlinearity and Multicollinearity. The Annual Meeting of Agricultural and Applied Economics Association.

CARNEY, P. A., MIGLIORETTI, D. L., YANKASKAS, B. C., KERLIKOWSKE, K., ROSENBERG, R., RUTTER, C. M., GELLER, B. M., ABRAHAM, L. A., TAPLIN, S. H., DIGNAN, M., CUTTER, G., BALLARD-BARBASH, R., (2003). Individual and Combined Effects of Age, Breast Density, and Hormone Replacement Therapy Use on the Accuracy of Screening Mammography. Annals of Internal Medicine, 138(3), 168−75.

COUGHLIN, S. S., TROCK, B., CRIQUI, M. H., PICKLE, L. W., BROWNER, D., TEFFT, M. C., (1992). The Logistic Modeling of Sensitivity, Specificity, and Predictive Value of a Diagnostic Test. Journal of Clinical Epidemiology, 45, 1−7.

ELIE, C., COSTE, J., THE FRENCH SOCIETY OF CLINICAL CYTOLOGY STUDY, (2008). A Methodological Framework to Distinguish Spectrum Effects from Spectrum Biases and to Assess Diagnostic and Screening Test Accuracy for Patient Populations: Application to the Papanicolaou Cervical Cancer Smear Test. BMC Medical Research Methodology, 8, 7.

FIDLER, V., NAGELKERKE N., (2013). The Mantel-Haenszel Procedure Revisited: Models and Generalizations. PLoS One, 8(3), e58327.

FIGUEROA-CASAS, J. B., CONNERY, S. M., MONTOYA, R., DWIVEDI, A. K., LEE, S., (2014). Accuracy of Early Prediction of Duration of Mechanical Ventilation by Intensivists. Annals of the American Thoracic Society, 11(2), 182−185.

JANSSENS, A. C., DENG, Y., BORSBOOM, G. J., EIJKEMANS, M. J., HABBEMA, J. D., STEYERBERG, E. W., (2005). A New Logistic Regression Approach for the Evaluation of Diagnostic Test Results. Medical Decision Making, 25(2), 168−177.

KING, G., ZENG, L., (2001). Logistic Regression in Rare Events Data. Political Analysis, 9, 137−163.

LAYA M. B., LARSON E. B., TAPLIN S. H., WHITE E., (1996). Effect of Estrogen Replacement Therapy on the Specificity and Sensitivity of Screening Mammography. Journal of National Cancer Institute, 88(10), 643−649.

LEISENRING, W., PEPE, M. S., LONGTON, G., (1997). A Marginal Regression Modelling Framework for Evaluating Medical Diagnostic Tests. Statistics in Medicine, 16, 1263−1281.

LEISENRING, W., ALONZO, T., PEPE, M. S., (2000). Comparisons of Predictive Values of Binary Medical Diagnostic Tests for Paired Designs. Biometrics, 56, 345−351.

LIU, H., (1998). Robust Standard Error Estimate for Cluster Sampling Data: A SAS/IML Macro Procedure for Logistic Regression with Huberization. In: Proceedings of the Twenty-Third Annual SAS Users Group International.

MIETTINEN, O. S., (1976). Stratification by a Multivariate Confounder Score. American Journal of Epidemiology. 104, 609−620.

MOSKOWITZ, C. S., PEPE, M. S., (2006). Comparing the Predictive Values of Diagnostic Tests: Sample Size and Analysis for Paired Study Designs. Memorial Sloan-Kettering Cancer Center, Department of Epidemiology & Biostatistics Working Paper Series. Working Paper 5.

NEMES, S., JONASSON, J. M., GENELL, A., STEINECK, G., (2009). Bias in Odds Ratios by Logistic Regression Modelling and Sample Size. BMC Medical Research Methodology, 9, 56.

PUGGIONI, G., GELFAND, A. E., ELMORE, J. G., (2008). Joint Modeling of Sensitivity and Specificity. Statistics in Medicine, 27(10), 1745−1761.

YE, F., LORD, D., (2014). Comparing Three Commonly Used Crash Severity Models on Sample Size Requirements: Multinomial Logit, Ordered Probit and Mixed Logit Models. Analytic Methods in Accident Research, 1, 72−85.

APPENDICES

Appendix 1.

Table 1. Cross-tabulation of test results (T) with reference test (D)

Test result[1]	**References test**		**Total**
	Positive	Negative	
Positive	a (True positive)	b (False Positive)	a+ b
Negative	c (False Negative)	d (True Negative)	c+ d
Total	a+ c	b+ d	a+ b+ c+ d

Table 2. The percent relative bias in estimating Se using subgroup LR, adjusted LR, and MLR

Disease prevalence <= 10%		**N=100**			**N=500**		
		Sub-group LR	**Adjusted LR**	**MLR**	**Sub-group LR**	**Adjusted LR**	**MLR**
When x=50%							
Equal prevalence	X=1	2.99	10.27	0.00	0.00	-12.46	0.00
	X=0	0.22	3.11	0.00	0.00	-6.30	0.00
Unequal prevalence	X=1	7.60	19.10	0.00	0.00	-12.41	0.00
	X=0	0.80	6.60	0.00	0.00	-5.71	0.00
When x=20%							
Equal prevalence	X=1	22.85	33.01	0.00	0.46	-1.35	0.00
	X=0	0.00	-1.32	0.00	0.00	-1.20	0.00
Unequal prevalence	X=1	31.00	38.60	0.00	1.37	15.21	0.00
	X=0	0.06	-0.68	0.00	0.00	-1.43	0.00

*Se: 20-30%; Sp: 20-30%; Se: sensitivity; Sp: specificity, LR: logistic regression.

Table 3. The percent relative bias in estimating Se using subgroup LR, adjusted LR, and MLR

Disease prevalence > 20%		N=100			N=500		
		Sub-group LR	Adjusted LR	MLR	Sub-group LR	Adjusted LR	MLR
When x=50%							
Equal prevalence	X=1	0.00	-3.58	0.00	0.00	-1.98	0.00
	X=0	0.00	-6.20	0.00	0.00	-1.59	0.00
Unequal prevalence	X=1	0.19	-1.02	0.00	0.00	-5.23	0.00
	X=0	0.00	-4.34	0.00	0.00	-0.49	0.00
When x=20%							
Equal prevalence	X=1	2.95	14.29	0.00	0.00	-8.47	0.00
	X=0	0.00	-2.70	0.00	0.00	-0.50	0.00
Unequal prevalence	X=1	8.08	19.82	0.00	0.00	-10.69	0.00
	X=0	0.00	-1.58	0.00	0.00	-0.35	0.00

*Se: 20-30%; Sp: 20-30%; Se: sensitivity; Sp: specificity, LR: logistic regression.

Table 4. Models of sensitivity, specificity and predictive values using subgroup LR and MLR approaches

Diagnostic models	Subgroup LR			MLR		
	RC	SE	p-value	RC	SE	p-value
Se						
Female gender	1.722	0.807	0.033	1.683	0.806	0.037
BMI(Kg/m^2)>25	-0.761	0.854	0.372	-0.628	0.850	0.461
Sp						
Female gender	-1.596	1.025	0.120	-1.444	0.985	0.143
BMI (Kg/m^2)>25	-0.860	1.342	0.522	-0.426	1.273	0.738
PPV						
Female gender	-1.294	0.887	0.145	-1.310	0.888	0.140
BMI(Kg/m^2)>25	-0.460	1.156	0.690	-0.522	1.152	0.651
NPV						
Female gender	1.526	0.918	0.096	1.548	0.912	0.090
BMI(Kg/m^2)>25	-0.380	1.050	0.718	-0.532	1.001	0.595

BMI: Body mass index; Se: sensitivity; Sp: specificity; PPV: positive predictive value; NPV: negative predictive value, LR: logistic regression.

Table 5. The absolute percent relative bias in estimating Se using subgroup GEE, adjusted GEE, and MLR

Disease prevalence <= 10%	N=100			N=500		
	Sub-group GEE	**Adjusted GEE**	**MLR**	**Sub-group GEE**	**Adjusted GEE**	**MLR**
When x=50%						
Equal prevalence	1.50-9.07	7.12-22.65	1.63-8.55	2.30-8.23	3.45-36.36	2.24-8.40
Unequal prevalence	2.31-14.79	5.96-28.48	2.19-13.29	2.63-5.91	1.04-31.33	2.61-6.03
When x=20%						
Equal prevalence	0.11-30.05	5.18-45.16	0.38-20.61	0.37-2.54	8.57-21.04	0.53-2.53
Unequal prevalence	0.40-33.27	8.99-50.18	0.54-24.69	0.18-7.17	1.28-20.68	0.35-6.82

*Se: 20-30%; Sp: 20-30%; Se: sensitivity; Sp: specificity, GEE: generalized estimating equation.

Table 6. The absolute percent relative bias in estimating Se using subgroup GEE, adjusted GEE, and MLR

Disease prevalence > 20%	N=100			N=500		
	Sub-group GEE	**Adjusted GEE**	**MLR**	**Sub-group GEE**	**Adjusted GEE**	**MLR**
When x=50%						
Equal prevalence	2.68-5.60	1.35-26.82	2.64-5.81	0.72-2.58	6.14-16.65	0.69-2.62
Unequal prevalence	1.71-4.04	2.51-26.70	1.71-4.37	0.35-3.71	5.18-19.67	0.33-3.76
When x=20%						
Equal prevalence	0.70-8.91	4.45-19.64	0.64-8.31	0.11-9.09	4.01-27.92	0.09-9.14
Unequal prevalence	0.53-15.54	4.17-29.08	0.49-14.11	0.15-11.75	5.88-33.10	0.12-11.77

*Se: 20-30%; Sp: 20-30%; Se: sensitivity; Sp: specificity, GEE: generalized estimating equation.

Table 7. Models of sensitivity, specificity and predictive values of diagnosing acute appendicitis using subgroup GEE and MLR approaches

Diagnostic models	Subgroup GEE			MLR		
	RC	SE	p-value	RC	SE	p-value
Se						
Radiologist*	-1.994	0.543	0.000	-1.993	0.543	0.000
Age (years)	0.001	0.014	0.955	0.001	0.014	0.964
Male gender	-0.022	0.503	0.964	-0.021	0.505	0.967
WBC	0.028	0.059	0.634	0.028	0.060	0.636
Sp						
Radiologist*	1.997	0.965	0.038	1.948	0.848	0.022
Age (years)	0.009	0.045	0.838	0.015	0.046	0.741
Male gender	-0.470	1.746	0.788	-0.349	1.049	0.739
WBC	0.049	0.211	0.816	0.030	0.121	0.807
PPV						
Radiologist*	0.765	0.430	0.075	0.749	0.437	0.086
Age (years)	0.031	0.036	0.384	0.029	0.035	0.406
Male gender	-0.211	0.870	0.808	-0.166	0.894	0.852
WBC	0.196	0.057	0.001	0.192	0.057	0.001
NPV						
Radiologist*	-0.888	0.823	0.281	-0.795	0.789	0.314
Age (years)	-0.005	0.028	0.857	-0.013	0.038	0.732
Male gender	-0.413	0.932	0.658	-0.203	0.928	0.826
WBC	-0.137	0.121	0.258	-0.134	0.141	0.343

WBC: white blood cells;*referent: surgical residents; Se: sensitivity; Sp: specificity; PPV: positive predictive value; NPV: negative predictive value, GEE: generalized estimating equation.

Appendix 2.

Estimating diagnostic accuracy using a logistic regression (LR) model

Suppose a diagnostic study involves a screening test (T) and a reference test (D). LR models can be used to predict Se in relation to cofactors:

$$\text{Logit}(P(T=1|D=1, X)=\alpha'_{1D}+\alpha'_{2D}*X_1+...+\alpha'_{kD}*X_k \quad \text{(1a: sub-group)}$$

$$\text{Logit}(P(T=1|D, X)=\alpha_{1D}+\alpha_{2D}*(D=1)+\alpha_{3D}*X_1+...+\alpha_{kD}*X_k \quad \text{(1b: adjusted)}$$

The equation (1a) is referred to as a subgroup model and the equation (1b) is referred to as an adjusted model. Substituting D=0 in the above equations will provide models for 1-specifcity. Thus, LR models can also be used to predict Sp in the presence of cofactors:

$$\text{Logit}(P(T=0|D=0, X)=\alpha'_{1\bar{D}}+\alpha'_{2\bar{D}}*X_1+...+\alpha'_{k\bar{D}}*X_k \quad \text{(2a: sub-group)}$$

$$\text{Logit}(P(T=0|D, X)=\alpha_{1\bar{D}}+\alpha_{2\bar{D}}*(D=0)+\alpha_{3\bar{D}}*X_1+...+\alpha_{k\bar{D}}*X_k \quad \text{(2b: adjusted)}$$

Possible LR models for predicting PPV and NPV are:

$$\text{Logit}(P(D=1|T=1, X)=\beta'_{1T}+\beta'_{2T}*X_1+...+\beta'_{kT}*X_k \quad \text{(3a: sub-group)}$$

$$\text{Logit}(P(D=1|T, X)=\beta_{1T}+\beta_{2T}*(T=1)+\beta_{3T}*X_1+...+\beta_{kT}*X_k \quad \text{(3b: adjusted)}$$

$$\text{Logit}(P(D=0|T=0, X)=\beta'_{1\bar{T}}+\beta'_{2\bar{T}}*X_1+...+\beta'_{k\bar{T}}*X_k \quad \text{(4a: sub-group)}$$

$$\text{Logit}(P(D=0|T, X)=\beta_{1\bar{T}}+\beta_{2\bar{T}}*(T=0)+\beta_{3\bar{T}}*X_1+...+\beta_{k\bar{T}}*X_k \quad \text{(4b: adjusted)}$$

In the above equations (1, 2, & 3), α'_{1D},α_{1D} $\alpha'_{1\bar{D}}$, $\alpha_{1\bar{D}}$ β'_{1T},β_{1T}, $\beta'_{1\bar{T}}$,and $\beta_{1\bar{T}}$ are the intercepts while α'_{2D}.. α'_{kD},α_{1D}...α_{kD} $\alpha'_{2\bar{D}}$....$\alpha'_{k\bar{D}}$, $\alpha_{2\bar{D}}$...$\alpha_{k\bar{D}}$ β'_{2T}...β'_{kT},β_{2T}...β_{kT}, $\beta'_{2\bar{T}}$...$\beta'_{k\bar{T}}$,and $\beta_{2\bar{T}}$..$\beta_{k\bar{T}}$ are the regression coefficients and X (X_1….. X_k) is the vector of k covariates. D and $\bar{D}$ denote the presence and the absence of disease respectively while T denotes positive test result and $\bar{T}$ denotes negative test result.

Multinomial logistic regression (MLR) for estimating diagnostic accuracy

The MLR models for predicting a new outcome variable Y (1=true positive; 2=false positive; 3= false negative; 4=true negative):

$$\log\left[\frac{P(Y=1|X)}{P(Y=3|X)}\right]=\mu_1+\mu_2*X_1+...+\mu_k*X_k \quad \text{Se model}$$

$$\log\left[\frac{P(Y=4|X)}{P(Y=2|X)}\right]=\pi_1+\pi_2*X_1+...\pi_k*X_k \quad \text{Sp model} \quad (5)$$

$$\log\left[\frac{P(Y=1|X)}{P(Y=2|X)}\right]=\rho_1+\rho_2*X_1+...\rho_k*X_k \quad \text{PPV model}$$

$$\log\left[\frac{P(Y=4|X)}{P(Y=3|X)}\right]=\tau_1+\tau_2*X_1+...\tau_k*X_k \quad \text{NPV model}$$

where μ_1 π_1, ρ_1, and τ_1 are the regression intercepts and μ_2…, μ_k, π_2,…, π_k, ρ_2,…, ρ_k and τ_2,…, τ_k are the regression coefficients and X (X_1….. X_k) is the vector of k covariates.

Comparing diagnostic accuracy using generalized estimating equation (GEE) and MLR methods

The MLR described in the above equation (5) can be extended for two screening tests as:

$$\log\left[\frac{P(Y=1|Z,X)}{P(Y=3|Z,X)}\right]=\mu_1+\mu_2*Z+\mu_3*X_1+...+\mu_k*X_k$$

$$\log\left[\frac{P(Y=4|Z,X)}{P(Y=2|Z,X)}\right]=\pi_1+\pi_2*Z+\pi_3*X_1+...\pi_k*X_k \quad (6)$$

$$\log\left[\frac{P(Y=1|Z,X)}{P(Y=2|Z,X)}\right]=\rho_1+\rho_2*Z+\rho_3*X_1+...\rho_k*X_k$$

$$\log\left[\frac{P(Y=4|Z,X)}{P(Y=3|Z,X)}\right]=\tau_1+\tau_2*Z+\tau_3*X_1+...\tau_k*X_k$$

where μ_1, π_1, ρ_1, and τ_1 are the regression intercepts and μ_2…, μ_k, π_2,…, π_k, ρ_2,…, ρ_k and τ_2,…, τ_k are the regression coefficients and X (X_1….. X_k) is the vector of k covariates in equation (6). The μ_2 and π_2 provide the comparison of sensitivities and specificities between two screening tests respectively whereas ρ_2 and τ_2 provide the comparison of positive predictive values and negative predictive values between the two screening tests respectively.