ISSN 2278-5612
Volume 11 Number 2
International Journal of Management and Information Technology

# Analysis of Healthcare Data of NepalHospital using Multinomial Logistic Regression Model

AmitYadav[1], Li Hui[2], Mohsin Ali[3], Ashis Yadav[4], Maira Anis[5]

[1]PhD Scholar, School of Management & Economics,

University of Electronics Science & Technology of China, Chengdu, 611731, China

E-mail: amitaryan2u@yahoo.com

[2]School of Management & Economics,

University of Electronics Science & Technology of China, Chengdu, 611731, China

E-mail: callmelihui@gmail.com

[3]PhD Scholar, School of Management and Economics,

University of Electronic Science and Technology of China, China

Employee, International Islamic University, Islamabad, Pakistan

E-mail: mohsinali757@gmail.com, mohsinali@iiu.edu.pk

[4]Doctor of Medicine (Radiology),

Kathmandu University, Kathmandu, Nepal

E-mail: ashish2y@gmail.com

[5]PhD Scholar, School of Management and Economics

University of Electronic Science and Technology of China, China

E-mail: maira7pk@hotmail.com

## ABSTRACT

Patient data had been collected from the hospital of Nepal with the help of hospital administration, doctors and patient cooperation. Data scrutiny attempts to shows the significant relationship between disease and factors causal of disease. Research explores the utility of multinomial logistic regression (MLR) technique in health domain and its most beneficial use for categorical data. Paper try to exhibit various factors which results in happening of health disorder and highlight application of data mining technique in healthcare. It is conceived that this work render more accuracy and reliability in detection of factors causal of disease, espial of fraud, helpful for all parties associated with healthcare, reduce cost, lessen time and treatment process.

**Keywords:** Healthcare, Diagnosis, primary data, Multinomial Logistic Regression (MLR)

## INTRODUCTION

The healthcare organizations are defined as institutions and resources that are committed to bring forth health related services whose elementary purpose is to improve health and well-being of all the patients. Health is the major factor for sound and healthy society, which directly or indirectly contributes to a country's economic development and poverty reduction.Choice of healthcare facilities depends on the characteristics of facilities provided such as level of care, area of expertise, quality, cost and characteristics of patients, where theseinclude economic status, health status, education and gender (Yip 1998, Thuan 2008).According to a report published in 2000 by the Institute of Medicine, at least 44,000 and perhaps as many as 98,000 patients die in the hospital each year as a result of medical errors (Iglesias 2003). Processes of patient care are so complex that it is difficult for healthcare workers (i.e. doctors, nurses) to create effective care plans for their patients (Bellika 2005). Disease diagnoses are the identifiable problem, which must be amended through intercession and its ultimate goal is to reach an outcome tailored to the aforesaid diagnoses. Healthcare data holds information related with patients as well as revelries involved in healthcare segment. Such data are stored rapidly and to collect health care data intricacy exist. In order to extract meaningful information from such data, traditional methods is not useful. In such situation, data mining is the beneficial where large amount of healthcare data are existing.

Due to improper management, lack of information, poverty, social belief, lack of government policy, a lot of improvement is necessary to improve healthcare system in a country like Nepal. Recent studies indicate that there is an increase in number of private healthcare facilities with respect to the public healthcare in Nepal. Due to lack of proper government policy and mismanagement, public healthcare sector is losing confidence of people and private healthcare is increasing

their trust towards people. The number of private hospitals increased from 69 in 1995 to 147 in 2008; however, the number of public healthcare increased from 78 to 96 during same period. There are almost twice as many hospital beds in the private sector (12310) than in the public sector (6944) in Nepal (RTI International 2010).There is rise in the number of private sectors based on peoples' demands because of the better treatment and care provided to the patients in comparison to the public healthcare organizations in Nepal. We are confronting with difficulties related to healthcare, such as fraud in treatment process, fraud in patient insurance, time taking treatment process, doctor prescribe numerous lab test to recognize disease as well as its causes, high expense bearing treatment and decision making by doctors towards patient disease diagnosis is one of the challenging factor. In several cases even doctor can't recognize the exact cause of disease and patient suffered by this problem. To know the meticulous cause doctors prescribe patient to do several lab test. There are numerous private healthcare to give better treatment than government healthcare but indirectly they used to dupery patient expense. People looking for better treatment, they prefer to go private healthcare but unknown from those fraud. Monopolies by the private healthcare, people are suffering with high price, fraud, misguidance, time, energy etc. Countries like Nepal are facing major problems regarding improved healthcare facilities despite the introduction of free essential healthcare for all citizens in 2008.

Numeral of welfares are provided by data mining, such as significant role in the detection of fraud and abuse, recognition of diseases at early stage, provide better medical treatments at sensible price, brainy healthcare decision support system etc. (Ahmad et al.2015). So, data mining provide improved medical facilities to the patients and helps in numerous medical administration decisions. Koh and Tan (2005) also deliberated about data mining technique advantageous in healthcare such as hospital ranking, number of day's patient stayed in hospital, effective treatment, scam insurance claims, recognizes better treatments process, patient readmission, edifice of effective drug recommendation system etc. Due to above explanation researchers are significantly prejudiced by the capabilities of data mining and for healthcare arena widely used data mining techniques. The ultimate goal is to provide the factors influencing the cause of disease by the age, gender, medical history, family medical history, professional, education, marital status, economic condition, exercise time, alcohol, vegetarian, caste and income. In order to maximize efficiency, minimize fraud, increase care quality and save money as well as time in a healthcare organization.

In health industry, data mining delivers various benefits such as accessibility of medical solution to the patients at lower cost, revealing of fraud in health insurance, causes of diseases and identification of medical treatment methods (Ahmad et al.2015, Tomar and Agarwal 2013, Bellazzi 2008). Koh (2005) stated that these benefits help the healthcare researchers to make efficient healthcare policies, develophealth profiles of individuals, build drug recommendation systems, etc. The healthcare data is always been rigid part for research work because of privacy, hesitation of fraud detection etc. Data always been the significant part for verdicts and without satisfactory data it is tough to make important decision regarding patient health. Healthcare data contains details regarding patients, hospital, medical claims, treatment process, cost, disease diagnosis etc. Therefore, to scrutinize and extract important information from this multifaceted datapowerful tools of data mining plays major role. The results need to be more profoundsince this analysis improves healthcare by enhancing the performance of patient management task.

Quality based on patients' evaluations and their opinions are important deciding factors in selecting a health facility and quality of care.Hence, information collected through various mediums is valuable for the healthcare specialists to find out the causes of diseases and to provide better and cost effective treatment to patients. Data Mining bids information of healthcare which in turn is helpful for making administrative as well as medical decisions likehealth insurance policy, selection of treatments, estimation of medical staff, disease prediction etc. (Silver 2001, Bellazzi 2008). For both analysis and prediction of various diseases, data mining techniques are also used (Kumari 2011, Gupta 2011).There are various data mining techniques such as clustering, classification and association that are used by the healthcare organizations to increase their potential for making better decision regarding patients' health.

Conceiving the discussion referred above, the main goal of this study consist usage of MLR method for diagnosis of disease by showing the relationship with variable which can predict the probability of occurrence i.e. relationship between disease and causal of disease. For this patient data had been collected from the hospital of Nepal with the help of hospital administration, doctors and patient cooperation. MLR analysis result will make convenient for doctor to prescribe medicine and limited lab test. Even private healthcare organization under supervision of government authority can detect fraud by viewing patient detail treatment process. More evidently, this study is to illustrate the applicability of MLR by diagnosis causal of patient disease, usefulness of MLR in healthcare data analysis and its limitation. Which will help the healthcare workers as well as the patients to make medical decisions, accuracy and reliability in detection of fraud, reduce cost and treatment process, selection of treatments and diagnosis of diseases prediction from factors of causes.

## LITERATURE REVIEW

Several research works are going on in the field of healthcare sector worldwide, yet, a lot of researches still needs to be done because of growing population, new diseases, rising of technology, etc. In healthcare, diagnosis of disease is one of the major glitches, which must be amended through intervention. This paperattempts to achieve an outcome by finding the factors thatcauses diseasesusing the data collected for analysis. Researchers have argued that integrate medical information systems are becoming a major part of modern healthcare system and such systems have evolved to an integrated enterprises wide system (Li & Xu 1991, Li et al. 2008, Yoo et al. 2008, Puustjarvi et al. 2010).Some researchers have also suggested that the work in psychiatry and oncology combines or clusters of symptoms are of greater consequence than individual symptoms (Dodd et al. 2001, Gift et al. 2004, Kim et al. 2005). In the nursing research, symptoms cluster as three or more concurrent symptoms that are related to each other but those symptoms may not share the same etiology (Dodd et al. 2001). Exploring multiple symptoms is to create subgroups of patient'sbased ontheir

symptoms and using cluster analysis or various methodologies according to the data. The factors analysis of self-reported symptoms and a cluster analysis of patients with irritable bowel syndrome found that the identification of patient subgroups was more clinically useful (Eslick 2004).

Data mining provides support for constructing a model for managing the healthcare resources, and it is also possible to detect chronic disease and complication of patient so that they can get treatment in timely as well as accurately. Seton Medical Centre used data mining to enhance the healthcare quality by providing various details regarding patient's health and reduced duration of admission of the patients in the hospital (Dakins 2001).Data mining application is used in various sectors of healthcare such as effective management of hospital resources, hospital ranking, better customer relation, hospital infection control, smarter treatment techniques, improved patient care, reduce insurance fraud, recognize high-risk patient and health policy planning (Tomar et al 2013).

Classification divides the data sample into target classes and predicts the target classes for each data pointand it is most widely used method of data mining in the healthcare organization. There are two methods of classificationknown as binary (e.g. high or low) and other is multilevel (e.g. high, medium & low). For analyzing microarray data,different classification methods are used such as decision tree, SVM and ensemble by Hu (2006). Classification methods are also used for anticipating the treatment cost of healthcare services which is increasing with rapid growth and becoming major concern for all (Beller et al. 2008).Linear Discriminate Analysis (LDA) is used to generate early warning for classification of chronic disease, whereK-NN classifier is used for analyzing the patient suffering from heart disease (Shouman et al. 2012).Decision tree is a classifier that use tree life graph and it is most commonly used in theoperation research analysis for calculating conditional probabilities (Goharian et al. 2003).Patients used decision tree for predicting the survivability of breast cancer by Khan (2008) and Chen (2012), who has proposed universal hybrid decision tree classifier for classifying chronic disease patient activity. Similarly, there are various differentmethods for classification such as SVM, Neural Network (NN) and Bayesian Methods, whose merits and demerits have been explained by Tomar (2013).There are several other methods in data mining that are used for multiple purposes, andclustering is one of them which is used for analyzing gene expression data with the help of a new hierarchical clustering approach using genetic algorithm (Tapia 2009) and advantages-disadvantage of various clustering explained by Tomar (2013).

Regression is used to find out functions that explain the correlation among different variables, whereas for statistical modeling two (dependent variable, independent variable) kinds of variables are used. There is always one dependent variable but independent variable may be one or more than that. Regression analysis helps to establish correlation between dependence of one variable upon the others (Fox 1997).In the linear regression, dependent and independent variables are known and target is to berth a line that is correlated between these variables (Fox 1997). Linear regression has a limitation that it can only be used for numerical dataand not forthe categorical data. Moreover, a type of non-linear regression that can accept categorical data and anticipates the probability of occurrence is by using logit function. Binomial and multinomial logistic regression (MLR) is developed to solve problem of categorical data. Here binomial regression predicts the result for a dependent variables having occurrence of only two possible outcomes (e.g. yes or no, code as 0, 1) while multinomial logistic regression solves problems data having three or more outcome or an unordered group of dependent variables (e.g. high, low, middle/ code as 1,2,3). Therefore logistic regression doesn't consider linear relationship between variables (Gutierrez 2011). Healthcare data related with patient history are in categorical form and for such type of data scrutinymultinomial logistic regression is the best methodology for judgments. It can be extensivelyused in medical field for predicting the diseases or survivability of a patient (Ahmad et al.2015, Tomar and Agarwal 2013).In this research paper, we have used multinomial logistic regression for B&B hospital data analysis to show the relation between dependent variables (disease/diagnosis) and independent variables (age, gender, family medical history, medical history, profession, marital status, social status, exercise time, drinking alcohol, vegetarian, caste, income). Gennings (2011) has explained an application of logistic regression for the estimation of relative risk for various medical conditions.

MLR is similar to poly-way eventuality table and log-linear analysis but more intuitive to construe particularly when there are various independent variables being examined with a dependent variable (Tabatchnick 2007). The advantage of MLR is its use of odds ratios as estimators for the predictor variables and both categorical as well as continuous independent variables can be incorporated as predictors. Hossain (2002) had conducted few studies examining the differences in performance for MLR and ordinal regression models compared to linear regression or discriminant analysis. MLR can also be broadened into more advanced statistical analysis that incorporate time as a factor such as competing risks model and survival analysis (Allison 1995). Classification and regression analysis are used for predicting the class or outcome of a function,however, the only difference is the nature of attributes.

The objective of this paper are: 1) find the factors affecting various diseases according to data collected, 2) statically represent the overall data to show the relation between disease and each variable, 3) result will be helpful for all the parties associated with this field, and findings will help related hospital in treatment process as well as improvement of healthcare, and 4) use of R software for analysis shows more convenience, which is very less used for MLR methodology. In the past, mostly SPSS software is used for MLR data analysis. This paper is separated into 6 parts where 1st is introduction of MLR and healthcare, 2nd explains the previous study conducted on various data mining methodology, 3rd gives the brief explanation of MLR process, 4th explain the data analysis steps, 5th shows the results finding using R for MLR, and 6th gives the conclusion of this paper with limitation, simultaneously.

## METHODOLOGY

As mentioned in the previous study, there are various data mining application for data analysis and findings. But to show relation between dependent variables and independent variables, regression is the best method propose by researchers

(Ahmad et al.2015, Tomar 2013). Whereas, linear regression is restricted to numerical variables and cannot scrutinized categorical variables but logistic regression is a type of non-linear regression which can consent categorical data. Logistic regression is alienated into two category, Binomial regression and Multinomial Regression (Multinomial Logistic Regression). Binomial regression envisages result for dependent variable when there is two conceivable outcomes (person does exercise or not) but Multinomial regression can carry analysis when dependent variable had three or more outcome (Ahmad et al.2015, Tomar 2013, Petrucci 2009, Tabatchnick 2007, Hossain 2002, Chan 2005). MLR demand careful consideration of the sample size and scrutinize for outlaying cases. Similar to other data analysis procedures, the initial data analysis includes careful univariate, bivariate, multivariate assessment and multi-collinear which should be evaluated with simple correlations among the independent variable. The sample size guidelines for MLR indicate a minimum of 10 cases per independent variable (Schwab 2002). MLR is often considered as an attractive analysis because it doesn't assume normality, linearity or homoscedasticity, and a power alternative to it is discriminant function analysis which requires that these assumptions are met. For MLR, variable selection or model specification methods are similar to those used with standard multiple regression, e.g. sequential or nested logistic regression analysis. These methods are used when one dependent variable is used as choice on the subsequent dependent variables or criteria for placement.

Logistic regression analysis was used to identify the relationships between dependent and independent variables. The logistic regression model is shown below:

$$\ln(p/1-p) = \beta_0 + \beta_i X_i \quad \text{..... (1)}$$

Where, p = dependent variable probability

Behavior; (p/1-p) = odds of dependent variable

Behavior; $\beta_0$ = constant; Xi = vector of independent variables

$\beta_i$ = parameter estimate for the ith independent variables.

Logistic regression is strong enough in its ability to calculate the individual effects of continuous or categorical independent variables on categorical dependent variables (Wright 1995).

The multinomial logistic regression model is used when there are more than two categorical dependent variables as mentioned in the literature above. The basic idea was extrapolated from binary logistic regression (Aldrich 1984, Hosmer 2000). In MLR model, the estimates of parameter can be identified compared to a benchmark category of dependent variable (Long 1997). This is expressed as below:

$$Log(\pi_i / \pi_1) = \alpha_1 + \beta_i x, i = 1, ....., I-1 \quad \text{..... (2)}$$

The MLR model used benchmark category logits with a predictor x. The MLR model is used in this study to estimates the effect of the individual independent variables on the probability of causes of disease (dependent variable). Several authors provide discussion of binary logistic regression in the context of graduate level textbooks, which provides in-depth view into MLR because of its direct extension (Garson 2011, Mertler 2002, Pedhazur 1997). The steps followed for analysis of data using R for this research paperare mentioned below:

a) The data includes a single categorical dependent variable with five categories [Group1 (Pediatric), Group2 (Gynecology & Obstetric), Group3 (Orthopedics), Group4 (General Medicine) and Group5 (Surgery)] and 13 independent variables (age, gender, medical history, family medical history, professional, education, marital status, economic condition, exercise time, alcohol, vegetarian, caste and income).
b) The data contained enough cases (N= 351 new patient) to satisfy cases to variables.
c) To bring file in R, changed file in 'cvs' formatso that R can read the file and then data using 'foreign' package get summary of data.
d) We need to identify outcome variable as a factor (i.e. categorical). Then load 'mglogit' package (Croissant 2011) in R,which contains the functions for conducting the MLR. Note: The 'mglogit' package requires six other packages.
e) Next, we need to modify the data so that MLR function can process it and expand the outcome variable much like we would do for dummy coding a categorical variable for inclusion in standard multiple regression.
f) Now we can proceed with the MLR analysis using 'mglogit' function and ubiquitous 'summary' function of the result. (Note: - Reference category or Benchmark isspecified and process is continuedchangingthe benchmark).

```
library(nnet)
## Loading required package: nnet
## Warning: package 'nnet' was built under R version 3.1.3
HospitalData <- read.csv("folder path.csv")
summary(HospitalData)
data <-read.csv("file path.csv",head=TRUE)
summary(data)
opar <- par(no.readonly=TRUE)
par(mfrow=c(3,2),mar=c(4,5,2,2))
plot(xlab="Diagnosis or Disease",ylab="Age",data$DiagnosisOrDisease,data$Age)
```

(Note: - Similarly plotted statistical figure for other independent variables relation with dependent variable)

Similarly, we continuously follow the steps using different reference category (benchmark) and find the relation between dependent and independent variables.

```
#benchmark with group1
HosData$Disease2
<- relevel(HospitalData$Disease,ref="Group1(Pediatric)")
mlr<-multinom(Disease2~Age+Gender+MH+FMH+Profession+Edu+Marital+Social Status+ExerciseTime+Alcohol+Vegetarian+Caste+Income,data=HospitalData)
summary(mlogit.disease2)
exp(confint(mlr))
exp(coef(mlr))
```

g) The code “exp(confint(mlr))” give the confidence interval of each variable, if 1 is included in a coefficient’s confidence interval, then coefficient of this independent variable is not significant and remove it. If 1 is not included then those independent variables are significant.

## DATA ANALYSIS

To obtain the eminence and pertinent medical data is one of the most significant challenges of the data mining in healthcare. Healthcare data is multifaceted and not of the same nature because it is collected from various medium, such as discussion with patient, review of physician or medical reports of laboratory. Similarly, for this research paper, data had been collected from various sources such as patients’ interviews, hospital records and doctors diagnoses. There are 14 variables from which one is the dependent variable (Disease/Diagnosis) and the other 13 are independent variables. We have selected disease/diagnosis as the dependent variable because from discussion with hospital doctors and staffs suggestion, the purpose is to find the variables that are the main causes of corresponding disease.Hence, from this result they can distinguishthose variables that cause disease which will be helpful for them in the diagnosis as well as in the treatment process. Researchers need to uphold the quality of data because this data is beneficial to provide cost effective treatments for the patients. Keeping this in mind, we have tried to collect precise and errorless data as much as we can. Some variables were not taken intocontemplation due to inconvenience in collection of data due to privacy issues. Therefore, it is essential to assert the quality and accuracy of data for analysis in order to make effective decision. Due to confidential concerns, healthcare organizations are unwilling to share their data that is another barrier in data collection. Patient unwilling to share their health data, Health Insurance Organization and Health Maintenance Organization don’tshare data for preserving the privacy of patients. This causes obstacle to detect fraud in the healthcare and the health insurance organizations.

ISSN 2278-5612
Volume 11 Number 2
International Journal of Management and Information Technology

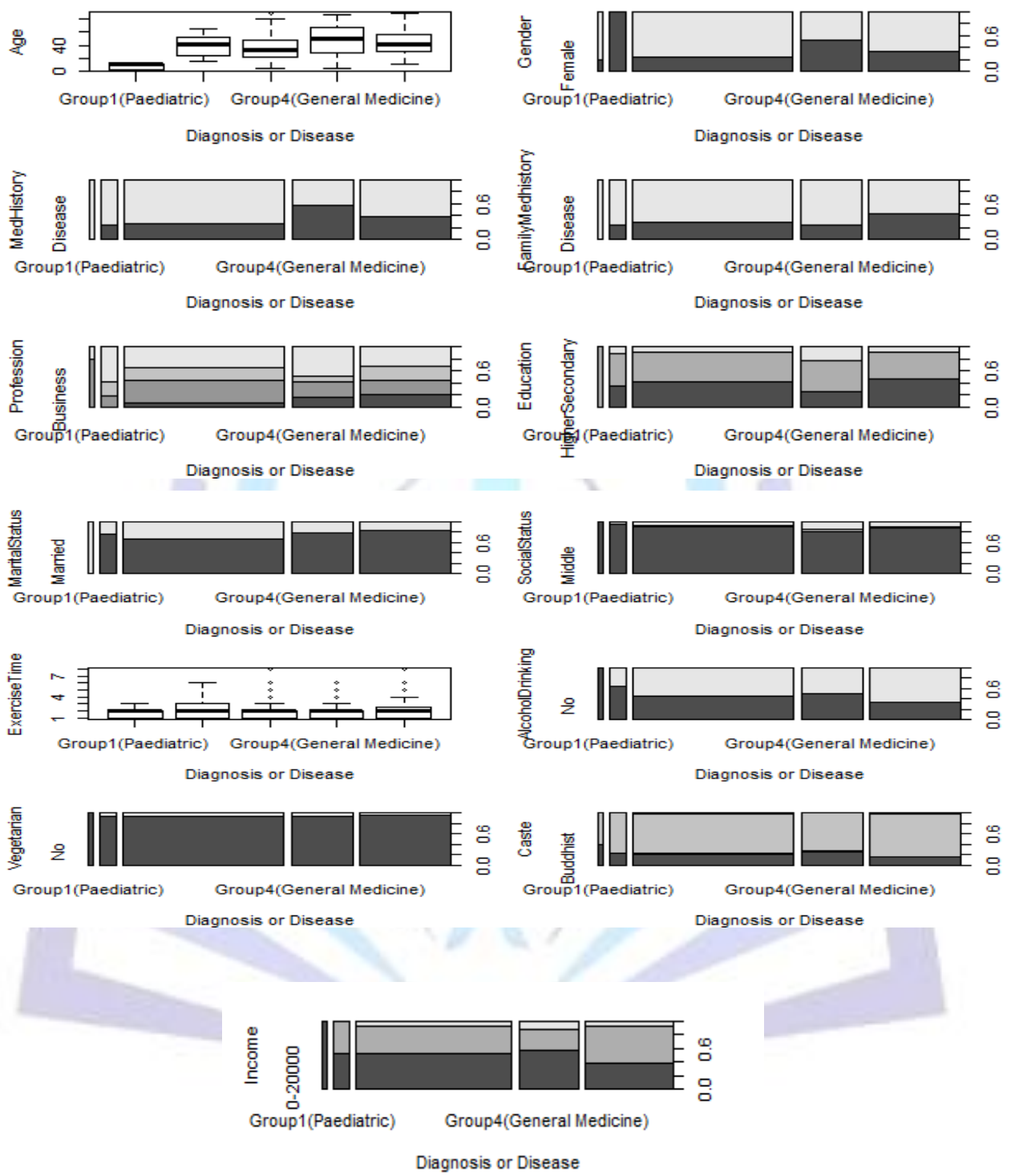


Figure 1: Statistical Descriptions of B&B Hospital Patient Data

According to Tomar (2013), data mining technologies provide benefits to healthcare organization by forming patients' group with similar types of disease or health issues so that healthcare organization can provide them effective treatments and other factors that are responsible for diseases. Similarly, in this research paper, data was collected from 351 patients (new patient, not follow-up patient) until 4 month and data mining methodology was used for data arrangement and coding. There were 475 new patients data collected. Due to error and missing data, we got 351 accurate data collected from various source such as patient interviews (education, life style, vegetarian or non-veg, caste, social status, income), hospital recorded data (age, gender, profession, marital status) and doctor analysis (diagnosis, medical history, family medical history, alcohol drinking). Before scrutiny of data, cluster analysis, data arranging, data cleaning and data processing was done. Various types of patients' diseases were diagnosed, however to make results precise diagnosed

disease had been classified into five different groups according to disease category. The groups include: group 1 (pediatric), group 2 (gynecology & obstetric), group 3 (orthopedics), group 4 (general medicine) and group 5 (surgery). The statistical summary of independent variables with dependent variable has shown in figure 1.

MLR procedures can utilize standard regression technique to select variables (Hosmer 2000, Tabatchnick 2007). Stepwise selection statistical procedures are used to select variables that make the largest contribution to prediction of outcomes variable, and the researcher determines which variables are included based on theory or practice. It is important in MLR that no two independent variables are intemperately correlated but there is no definitive test for using categorical variable. Thus, conceptually similar variables were identified and eliminated based on which one had the smallest amount of missing data (Hosmer 2000, Tabatchnick 2007).

## RESULT FINDINGS

Multinomial Logistic Regression assigns a reference group or benchmark to which all other levels of the dependent variable are compared. As steps mentioned in data analysis section, we used R for result findings. We have coded the categorical variables and used R code as steps mentioned in methodologysection of this research paper. Scrutinizing with R, the code “exp(confint(mlr))” give the confidence interval of each variable, if 1 is included in a coefficient’s confidence interval, then coefficient of this independent variable is not significant and removed it. If 1 is not included then those independent variables are significant. Following those steps, we got the final result which is shown in table 1.

**Table 1: Multinomial Logistic Regression (MLR) significant result using R**

| S.No | Benchmark | Group No. | Group Category Name | Significant Result |
|---|---|---|---|---|
| 1 | Group 5 | | Surgery | |
| | | Group 1 | Pediatric | MH, MS, ET, Veg, Caste |
| | | Group 2 | Gynecology & Obstetric | MH, Edu, MS, ET, Veg, Caste |
| | | Group 3 | Orthopedics | MH, Edu, ET, Veg, Caste |
| | | Group 4 | General Medicine | MH, ET, Caste |
| 2 | Group 4 | | General Medicine | |
| | | Group 1 | Pediatric | MH, MS |
| | | Group 2 | Gynecology & Obstetric | Age, MH, Edu, SS, Veg |
| | | Group 3 | Orthopedics | Age, MH, Edu, ET |
| | | Group 5 | Surgery | MH, ET, Caste |
| 3 | Group 3 | | Orthopedics | |
| | | Group 1 | Pediatric | MH, Prof, Edu, MS, Income |
| | | Group 2 | Gynecology & Obstetric | MH, MS, ET |
| | | Group 4 | General Medicine | Age, MH, Edu, ET |
| | | Group 5 | Surgery | MH, Edu, ET, Veg, Caste |
| 4 | Group 2 | | Gynecology & Obstetric | |
| | | Group 1 | Pediatric | MH, Prof, Edu |
| | | Group 3 | Orthopedics | MH, MS, ET |
| | | Group 4 | General Medicine | Age, MH, Edu, SS, Veg |
| | | Group 5 | Surgery | MH, Edu, MS, ET, Veg, Caste |
| 5 | Group 1 | | Pediatric | |
| | | Group 2 | Gynecology & Obstetric | MH, Prof, Edu |
| | | Group 3 | Orthopedics | MH, Prof, Edu, MS, Income |
| | | Group 4 | General Medicine | MH, MS |
| | | Group 5 | Surgery | MH, MS, ET, Veg, Caste |

Where, Edu= Education, MH= Medical History, Veg= Vegetarian (Veg or non-veg), MS= Marital Status, ET= Exercise Time, Prof= Professional, SS= Social Status.

Group 1 (Pediatric) = Pneumonia, AGN, Age, Febrile Convulsion, Seizure Disorder.

Group 2 (Gynecology & Obstetric) =PID, Fibroid Uterus, Dysmenorrhea, IDA,Ovarian Cyst, UV, Prolapse, VVF, Menorrhagia, Hyperemesis Gravid Arum, Labour, CA Cervix.

Group 3 (Orthopedics) = Fracture, Dislocation, RTA, PIVD, Chest Injury, Cut and Crush Injury, TBI, Osteomyelitis, Osteoarthritis, Osteoporosis, Spondylitis, Capsulitis, Giant Cell Tumor, Spinal Deformity.

Group 4 (General Medicine) = DM, HTN, COPD, Bronchial Asthma, Enteric Fever, Age, CVD, CVA, CKD, ALD, DVT, Pneumonia, Bronchitis, Anemia, Infraction, Disorder, Urticarial, Hypernatremia,CA Lungs, Vomiting.

Group 5 (Surgery) = Appendicitis, Cholelithiasis, Diverticulitis, Nephrolithiasis, Pancreatitis, Ureteric Calculus, Bladder Stone, Inguinal And Umbilical Hernia, Mucocele Of GB, CA Gall Bladder, Fistula in ANO, Koch Abdomen, Infective Colitis, Seminoma, CA Bladder, Peritoneal CA, Lung CA, CA Stomach, GB Polyp, BHP, Ampullary CA, Haemorrhoid.

AGN- Acute Glomeronephiritis, HTN- Hypertension, AGE- Acute Gastro Enteritis, TBI- Traumatic Brain Injury, DM- Diabetes Mellitus, COPD-Chronic Obstructive Pulmonary Disease, PID- Pelvic Inflammatory Disease, CVD-Cardio Vascular Disease, CVA-Cerebro Vascular Disease, IDA- Iron Deficiency Anemia, CKD-Chronic Kidney Disease, ALD-Acute Liver Disease, VVF- Vesiculo-Vaginal Fistula, DVT- Deep Vein Thrombosis, GB- Gall Bladder, UV- Utero Vaginal, CA- Carcinoma, RTA- Road Traffic Accident, PIVD- Prolapsed Inter Vertebral Disc.

In table 1, the noteworthy resultof each group under each reference category is shown. From analysis, we found that there are some independent variables which are not significant with any of the group i.e. gender, Family Medical History (FMH) and Alcohol Drinking. These significant results are the variables (independent variables) which effect in causalof corresponding disease (dependent variable). Author have cross checked the result and found that the analysis result is correct under the above mentioned circumstances. This result explain that if we take benchmark which is also known as reference category i.e. group 5 (surgery) then for group 1 (pediatric) patient they need to know detail account of their medical history, marital status, exercise time, vegetarian or non-vegetarian and also caste (religious factor), then according to this doctor need to do further assessment. This means preferring pediatric patient for surgery then doctor need to take record of MH, MS, ET, Veg, and Caste. Similarly, if benchmark (reference category) is surgery then for gynecology & obstetric patient specialist need to know patient medical history, education, marital status, exercise time, veg or non-veg, caste. This entails that preferring gynecology & obstetric patient for surgery then doctor need to take record (MH, Edu, MS, ET, Veg, and Caste) which shows significant relationship with surgery. If surgery is the reference category or preferring orthopedics patient for surgery then doctor need to take record of MH, Edu, ET, Veg, and Caste. Preferring general medicine patient for surgery then specialist need to take record of MH, ET, Caste.

Similarly these steps are followed for other dependent variables such as general medicine, orthopedics, pediatric and gynecology as benchmark. According to that significant result varies. Those significant result specialist need to take consideration for detailed record and according to that follow the treatment process for various disease. This analysis result will save time in treatment process, helpful for all parties associated with healthcare, reduce cost and also detect fraud.

## CONCLUSION

Healthcare data is always been toughest part for researcher to collect due to its privacy concern. Even though researchers keep trying for their best to give finest result and innovative findings. In this research according to specialist suggestion systematic data collection of B&B hospital is done. Although had problem in data collection due to patient hesitation, privacy, afraid of fraud caught etc. According Tomar and Agarwal (2013), researchers have conducted studies on healthcare sector and used various data mining techniques for data analysis but MLR methodology is still absence for data analysis.MLR analysis hasscrutinizednumerous areas of attention to healthcare practitioners and it is noticed that each of these studieswere premeditated to generate information that could immediately be integrated into appraisal, meddling or evaluation of outcomes in the clinical practice. Before applying regression analysis, data needs to be preprocessed so that the error, noise or missing data can be removed to achieve preciseresults. Statistical methods are used for distinguishing of such attributes. While doing analysis of hospital data, we tried to use some application referred from previous study and found that classification rules is used to discover the class of attributes but it does not show the relationships of attributes (Ahmad et al.2015, Tomar and Agarwal 2013). Multinomial logistic regression is the best method used for categorical data and also to show the relation between attributes. Analysis can be acquitted with sample sizes ranging from several hundred to thousandsof cases using retrospective or prospective data.

It is suggested that for categorical data analysis, finding creative and effective ways to mine the data available within existing program services. So that research can contribute to improved practice on the behalf of the medical staffs for the healthcare data. The success of healthcare data mining hinges on the availability of clean healthcare data. Rules from the various experts' knowledge could be more precise but they are hardly updated and specialized for different hospitals. From this research, it is believed that result generated using MLR methodology using R shows significant relation of attributes. This will help B&B hospital healthcare staffs (doctors, nurses) to diagnose disease theoretically and for further practical treatment process, accordingly. This result will giverelief to B&B hospital doctors tosome extent by savingtheir time and energy, saving patients 'money, and by detecting healthcare brokers' betrayal and fraud.

This verdicts will help B&B hospital in various ways as well as other healthcare organizationswill also be benefited in Nepal. It also gives idea to other researchers who are steering research on public health. There is limitation of MLRmethod: large sample sizes required in all levels of the dependent and independent variable for preciseassessment of parameters (Hossain 2002), appraising model fit is not as well advanced as in linear regression methods and rendering of the model can be challenging with several groups in the dependent variable (Hosmer 2000), the robustfact of MLR is that it hascapability to determine differential characteristics of groups through estimation of coefficients for each level of the comparison of dependent/independent variable relationships (Hosmer 2000).

## REFERENCE

1. ThanN, LofgrenC, LindholmL, ChucNT. Choice of healthcare facility following reform in Vietnam. BMC Health Ser Res 2008; 8:162.
2. YipWC, WangH, LiuY. Determinants of patient choice of medical provider: Acase study in rural China. Health Policy Plan 1998; 13:311-22.
3. Iglesias, A., Martnez, P., and Fernndez, O., 2003. Err is human: building a safer health system. Vol. 5. Athena. National Academy Press; 2000, 223–240.
4. Bellika, J. and Hartvigsen, G., 2005. The oncological nurse assistant: a web-based intelligent oncological nurse advisor. International Journal of Medical Informatics, 74, 587–595.
5. H. C. Koh and G. Tan, “Data Mining Application in Healthcare”, Journal of Healthcare Information Management, vol. 19, no. 2, (2005).
6. Tomar. D & Agarwal. S, “A survey on Data Mining approaches for Healthcare”, International Journal of Bio-Science and Bio-Technology, Vol.5, No.5, (2013), pp. 241-266. (http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25)
7. M. Silver, T. Sakara, H. C. Su, C. Herman, S. B. Dolins and M. J. O’shea, “Case study: how to apply data mining techniques in a healthcare data warehouse”, Healthc. Inf. Manage, vol. 15, no. 2, (2001), pp. 155-164.
8. R. Bellazzi and B. Zupan, “Predictive data mining in clinical medicine: current issues and guidelines”, Int. J. Med. Inform., vol. 77, (2008), pp. 81 -97
9. M. Kumari and S. Godara, “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction”, IJCST ISSN: 2229- 4333, vol. 2, no. 2, (2011) June.
10. S. Gupta, D. Kumar and A. Sharma, “Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis”, (2011).
11. RTI International. Overview of Public-Private Mix in Health Care Service Delivery in Nepal. North Carolina: Ministry of Health and Population, Government of Nepal and Research Triangle Park, 2010.
12. Li, L.X. and Xu, L.D., 1991. An integrated information system for the intervention and prevention of AIDS. International Journal of Bio-Medical Computing, 29 (3–4), 191–206.
13. Li, L., et al., 2008. Creation of environmental health information system for public health service: a pilot study. Information Systems Frontiers, 10, 531–542.
14. Yoo, S.K., Choe, J., and Kim, D.Y., 2008. Agent based architecture for secure access from multiple hospitals. In: Proceedings of the Seventh IEEE/ACIS international conference on computer and information science (ICIS 2008). Washington, DC, USA: IEEE Computer Society, 255–258.
15. Puustjarvi, J. and Puustjarvi, L., 2010. Developing interoperable semantic e-health tools for social networks. In: Proceedings 6th international conference on web information systems and technologies, Valencia, Spain. Setubal, Portugal: INSTICC Press, 305–312.
16. Dodd, M. J., Miaskowski, C., & Paul, S. M. (2001). Symptom clusters and their effect on the functional status of patients with cancer. Oncology Nursing Forum, 28(3), 465-470.
17. Gift, A. G., Jablonski, A., Stommel, M., & Given, C. W. (2004). Symptom clusters in elderly patients with lung cancer. Oncology Nursing Forum, 31 (2), 202-212.
18. Kim, H. J., McGuire, D. B., Tulman, L., &Barsevick, A. M. (2005). Symptom clusters: Concept analysis and clinical implications for cancer nursing. Cancer Nursing, 28(4), 270-282.
19. H. Hu, J. Li, A. Plank, H. Wang and G. Daggard, “A Comparative Study of Classification Methods for Microarray Data Analysis”, Proc. Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, ACS, vol. 61, (2006), pp. 33-37.
20. G. Beller, “The rising cost of health care in the United States: is it making the United States globally noncompetitive?”, J. Nucl. Cardiol., vol. 15, no. 4, (2008), pp. 481-482.
21. M. Shouman, T. Turner and R. Stocker, “Applying K-Nearest Neighbour in Diagnosing Heart Disease Patients”, International Conference on Knowledge Discovery (ICKD-2012), (2012).

22. Goharian& Grossman, Data Mining Classification, Illinois Institute of Technology, http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM-Classification.pdf, (2003).
23. M. U. Khan, J. P. Choi, H. Shin and M. Kim, "Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare", 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, (2008) August 20-24.
24. C. Chien and G. J. Pottie, "A Universal Hybrid Decision Tree Classifier Design for Human Activity Classification", 34th Annual International Conference of the IEEE EMBS San Diego, California USA, (2012) August 28-September 1.
25. D. Tomar and S. Agarwal (2013), "A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology. Vol.5. No.5. pp. 241-266. http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25
26. J. Fox, "Applied Regression Analysis, Linear Models, and Related Methods", (1997).
27. P. A. Gutiérrez, C. Hervás-Martínez and F. J. Martínez-Estudillo, "Logistic Regression by Means of Evolutionary Radial Basis Function Neural Networks", IEEE Transactions on Neural Networks, vol. 22, no. 2, (2011), pp. 246-263.
28. C. Gennings, R. Ellis and J. K. Ritter, "Linking empirical estimates of body burden of environmental chemicals and wellness using NHANES data", http://dx.doi.org/10.1016/j.envint.2011.09.002,2011.
29. Tabatchnick, B. G., &Fidell, L. S. (2007). Using multivariate statistics (5th ed., pp. 481–498). Boston: Pearson Education, Inc.
30. Hossain, M., Wright, S., & Petersen, L. A. (2002). Comparing performance of multinomial logistic regression and discriminant analysis for monitoring access to care for acute myocardial infarction. Journal of Clinical Epidemiology, 55, 400–406.
31. Allison, P. D. (1995). Survival analysis using the SAS system: A practical guide. Cary, NC: SAS Institute, Inc.
32. J. J. Tapia, E. Morett and E. E. Vallejo, "A Clustering Genetic Algorithm for Genomic Data Mining", Foundations of Computational Intelligence, vol. 4 Studies in Computational Intelligence, vol. 204, (2009), pp. 249-275.
33. D. R. Dakins, "Center takes data tracking to heart", Health Data Management, vol. 9, no. 1, (2001), pp. 32-36.
34. Y.H.Chan, "Biostatistics 305. Multinomial Logistic Regression", Singapore Med J 2005, 46(6): 259.
35. Petrucci, Carrie J. (2009),"A Primer for Social Worker Researchers on How to Conduct a Multinomial Logistic Regression", Journal of Social Service Research, 35:2,193-205. http://dx.doi.org/10.1080/01488370802678983.
36. Schwab, J. A. (2002). Multinomial logistic regression: Basic relationships and complete problems.http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems
37. Wright, R.E. 1995. Logistic regression. Pages 217-244 in L. G. Grimm, and P. R. Yarnold, Eds. Reading and understanding multivariate statistics. American Psychological Association, Washington, DC.
38. Aldrich, J.H.; Nelson, F.D. 1984. Linear probability, logit, and probit models. Newbury Park, CA: Sage Publications.
39. Hosmer, D.W.; Lemeshow, S. 2000. Applied logistic regression. New York: Wiley.
40. Long, J.S. 1997. Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage.
41. Garson, G. D. (2011). "Logistic Regression", from Stat notes: Topics in Multivariate Analysis. http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.
42. Mertler, C. &Vannatta, R. (2002). Advanced and multivariate statistical methods (2nd ed.). Los Angeles, CA: Pyrczak Publishing.
43. Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd ed.). New York: Harcourt Brace.
44. Croissant,Y.(2011).Package'mlogit'.http://cran.rproject.org/web/packages/mlogit/index.html.
45. P. Ahmad, S. Qamar, S. Q. A. Rizvi (2015). Technique of Data Mining in Healthcare: A Review. International Journal of Computer Applications. Vol. 120, no.15, June 2015, pp. 0975-8887.
46. H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005).

## Biography

Amit Yadav (amitaryan2u@yahoo.com) is a PhD scholar at University of Electronic Science and Technology of China (UESTC). He holds a B.E. in Civil & Environmental Engineering from SUST Bangladesh, Master Degree in Management Science & Engineering and is currently in the Data Mining Ph.D. program at USTC. His research interests include

transportation and traffic congestion, taxation, hazardous waste management, risk management, disease & diagnosis, supplier selection and use of data mining application in numerous areas.

Li Hui (callmelihui@gmail.com) is a Master degree student at University of Electronic Science and Technology of China (UESTC), School of Management and Economics. His area of expertise is in Data Mining and he will be graduating in 2017. Where he is working on his thesis using various MCDM method for data analysis.

Mohsin Ali (mohsinali757@gmail.com) is a PhD Scholar at University of Electronic Science and Technology of China (UESTC). He did Master of Business Administration in 2009 from International Islamic University, Islamabad, Pakistan. His research interests are risk management, decision making, data mining management and its applications.

Ashis Yadav (ashish2y@gmail.com) is a MD (Doctor of Medicine) student, specialization in radiology at Kathmandu University (KU), Kathmandu, Nepal. His research interest are in radiology, improvement in patient treatment process, decision making.

Maira Anis (maira7pk@hotmail.com) is a PhD Scholar at University of Electronic Science and Technology of China (UESTC). She did Master of Science in Mathematics from Quaid-e-Azam University, Islamabad, Pakistan. Her research interests are data mining management, class imbalance learning and its applications.