Abstract/Summary

Hepatitis C disease is chronic in nature and has reported prevalence of 3.3% worldwide and 4.9% in Pakistan. The prevalence of HCV is high (6.7%) in the province Punjab, the province under study where about 60% of the country's total population lives. In this thesis, the study of the risk factors for infection with the virus was described. As, little attention has been given to this issue in developing countries like Pakistan. Moreover, earlier studies demonstrated that 20-40% of infected HCV patients do not report a history consistent with established risk factors. Also modes of transmission vary by region. No vaccine is currently available and treatment for complications of infection is costly. This hospital based, un-matched case-control was conducted in the largest province of Pakistan i.e. Punjab. A consecutive sample of 1,400 patients with 700 cases and 700 controls was collected. Case to control ratio was taken as 1:1.

The Punjab province is divided in 9 administrative divisions and 36 districts where Divisional Headquarter Hospitals are serving the humanity in each division. Patients of both-genders, all ages and social strata, urban-rural settings visit these hospitals for their medical treatment as the Government of the Pakistan is offering free of charge treatment. The researcher visited these 9 Divisional Headquarters Hospitals and with prior permission of hospital administration, interviewed patients through a questionnaire comprised of 56 variables. Important variables related to the socio-demographic factors, behavioral characteristics, patient's medical and family history related factors were studied. The cases and controls were identified from the inpatients, outpatients, and hepatitis clinics out of each hospital. The cases were HCV positive patients determined by routine ELISA method, while controls were negative. The data were analyzed descriptively and analytically using IBM SPSS version 19.0 and OpenEpi version 2.3.1 software at the University of Auckland, New Zealand.

The specialty of this study was the model building of the risk factors of hepatitis C infection. Although, multivariate logistic regression was a conventional as well as widely known statistical technique being implied for this purpose, however this technique exhibits some practical limitations which are handling of interaction effects and missing values etc. Alternatively, other techniques were used such as artificial neural networks (ANN) and classification trees models that have now been emerged in recent years, not so much to identify risk factors, but as prediction tools. These techniques would enable us for acquiring

insight towards the recognition of key risk factors of hepatitis C infection. The present analysis constitutes the first attempt to discover most pertinent risk factors of hepatitis C infection in such a comprehensive mode. This kind of analysis would certainly encourage the medical-researchers as well as statisticians who are supposed to have little or no experience with these statistical techniques.

Initially, the descriptive of each variable was given followed by univariate and multivariate analysis. It was hypothesized that the risk factors of hepatitis C differ by gender, residential area and geographic location. On these bases, therefore, different logistic regression (Gold standard) models were run on different segments of collected data to explore pertinent risk factors at their exact place. Later on, the analysis was performed with Artificial Neural Networks and Classification Trees models only on overall data and the results were compared and contrasted with Gold standard. In a multivariate logistic regression model on the overall data, 11 potential risk factors were identified. These were patient's education (OR=0.196, 95% CI:0.150-0.254); history of local migration/frequent travelling (OR=2.777 95% CI:2.005-3.846); family history of liver disease (OR=2.646, 95% CI:1.663-4.212); endoscopy (OR=2.357, 95% CI:1.340-4.147); family history of hepatitis C (OR=2.176, 95% CI:1.584-2.989); tattooing (OR=2.175, 95% CI:1.439-3.286); blood transfusion (OR=2.043, 95% CI:1.555-2.685); minor surgery by barber (OR=1.739, 95% CI:1.094-2.762); dental surgery (OR=1.661, 95% CI:1.252-2.205); major/minor surgery (OR=1.577, 95% CI:1.182-2.105) and history of injections/intravenous drips (OR=1.519, 95% CI:1.133-2.037). Of these risk factors, only patient's education had negative association with the disease status, whilst other risk factors were positively associated with hepatitis C infection. On comparing, Logistic regression model and Artificial Neural Networks Model, it was concluded that a concordant set of risk factors, from both models were identified which implies that these Models are a useful adjunctive method to identify risk factors for hepatitis C. However, despite the fact that Artificial Neural Networks model does include interaction effects in the model, it was challenging to convey the meaning of the model characteristics. Only networks can recognize these interactions and their effects are evaluated in the model intrinsically. If someone is curious to watch interaction effects, Classification Tree models are helpful which allow the discovery of pertinent factors with their multilevel interactions. In this study, Classification Trees model has also been applied which played very informative role to explore potential risk factors. This model can have explored 6 risk factors in overall data i.e. Patient's education, Local migration/travelling, Family history of Hepatitis C, No. of

persons sharing the room, Major/Minor surgery, and receipt of a blood transfusion which were strongly associated with hepatitis C infection. Importantly, these six risk factors were also identified in logistic regression as well as Artificial Neural Networks models. And even these were the most repeated risk factors in different models obtained for different settings of data. Therefore, these were referred as the most common risk factors in the region. In addition to overall data, logistic regression models were run on different segments of data like male/female settings, urban/rural settings and North/South regions of Punjab, keeping in view the differences in biological, lifestyle, behavioral and medical health-care related factors. In all these cases, it was observed that hepatitis C risk factors differ.

This study identified several risk factors for hepatitis C infection in the province Punjab by applying an in-depth analysis which have never been witnessed before, however major sources of infection were identified as unhygienic, unsafe healthcare practices and personal behaviors & living conditions. Among these, Endoscopy, dentistry, transfusion services, surgical /gynecological procedures, minor surgery by the barbers, injected drug use and use of un-safe injections demand utmost care for not allowing hepatitis C transmission this way. Additionally, history of migration/travelling and under-shave from barber shops are also the newly identified risk factors in the region. Nosocomial infection should also be the focus of the medical professionals. This study also reveals the fact that patients living in poverty and low education were at higher risk. It is, therefore, pertinent to educate the people and disseminate awareness among them. In short, a general understanding of risk factors may guide preventive campaigns to reduce the burden of disease in the region.

Key Terms: Hepatitis C, Risk factors, Case-control, Odds Ratio, LR, ANN, CART, Punjab.