Missing Data in Pathology Databases

Master of Information Sciences (Research)

Sheikh Faisal (U3034931)
April, 2011.

Supervisors: Dr Alice Richardson, University of Canberra.

Brett A. Lidbury Ph.D., Associate Professor, Australian National University.

Abstract

Hepatitis virus is a major threat to Australia and a major health burden to the world. There are several types of this disease; the focus of this study is on Hepatitis C virus (HCV). The objective of this study is to enhance the predictive power of routinely performed diagnostic pathology laboratory results by identifying patterns in bio-markers so that the HCV infection is identified earlier rather than later, and to investigate the effects of missing values on the selection of assays. To overcome the problem of missing data, Multiple Imputation, a principled statistical imputing technique, was used to fill in the missing values. The imputed dataset was analyzed to construct predictive models using decision tree and logistics regression algorithms in R and PASW 18. ALT has been identified as the key predictor of HCV infection by all the Logistic Regression Models and all the Decision Tree Models. A higher level of ALT in the blood is indicative of HCV infection, that is, increased level of HepC (Hepatitis C antibody) in the blood. Pooled logistic regression model suggests that increased level of ALT (i.e. greater than 35 U/L) almost doubles the Odds of HCV infection. That is further affirmed by the decision tree models – all the rules in the tree models suggest increased ALT levels indicate presence of HCV infection. The study has not produced a powerful predictive model that could be used on general patients to detect the presence of HCV infection, but has provided useful information on the type of blood tests (the variables that need to be considered) to be conducted on patients who show any symptoms of HCV infection.

Certificate of Authorship of Thesis

Except where clearly acknowledged in footnotes, quotations and the bibliography, I certify that I am the sole author of the thesis submitted today entitled –

Missing Data in Pathology Databases

I further certify that to the best of my knowledge the thesis contains no material previously published or written by another person except where due reference is made in the text of the thesis.

The material in the thesis has not been the basis of an award of any other degree or diploma except where due reference is made in the text of the thesis.

The thesis complies with University requirements for a thesis as set out in *Gold Book*

Part 7: Examination of Higher Degree by Research Theses Policy, Schedule Two (S2).

Refer to http://www.canberra.edu.au/research-students/goldbook

……

Signature of Candidate

.....

Signature of chair of the supervisory panel

Date: ……..

Contents

Table of Figures

Tables

Acknowledgement

First and foremost, I would to thank the Almighty for his blessings on everyone involved in this research.

I gratefully acknowledge Dr. Alice Richardson and A. Prof. Brett Lidbury for their advice, support, encouragement and supervision from the preliminary to the concluding stages of this thesis. It has been an honour to work with my supervisors and I hope to keep up our collaboration in future.

I would like to offer my regards and blessings to the Principal and fellow staff (Faculty of Mathematics) for their support during the completion of this research.

Words fail me to express my appreciation to my wife Sainaz whose dedication, love and continued confidence in me, has taken the load off my shoulder. Farhat, thanks for being a caring and supportive sibling.

Finally, I offer my regards and blessings to all those who supported me in any respect during the completion of this research.

Sheikh Faisal.