

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

- Abstract 2
- Certificate of Authorship of Thesis 3
- Tables 7
- Acknowledgement..... 8
- Chapter 1 - Introduction..... 9
 - 1.1 Hepatitis 9
 - 1.2 Thesis Objectives 11
 - 1.3 ACT Pathology Dataset 12
 - 1.3 Thesis Outline 15
- Chapter 2 Literature Review 17
 - 2.1 Epidemiology of HCV 17
 - 2.2 Diagnosis of HCV 20
 - 2.3 Missing Data 24
 - 2.4 Data Mining and Knowledge Discovery..... 31
- Chapter 3 Methodology 34
 - 3.1 Exploratory Data Analysis..... 34
 - 3.2 Multiple Imputation (MI)..... 35
 - 3.3 Logistic Regression 40
 - 3.4 Decision Trees 42
- Chapter 4 Results 45
 - 4.1 Data Pre-processing and Cleansing 45
 - 4.2 Complete Case Analysis..... 49
 - 4.2.1 Principal Component Analysis (PCA) 49
 - 4.2.3 Multiple Logistic Regression Model 52
 - 4.2.4 Decision Tree Model..... 55
 - 4.3 Missing Value Analysis..... 58
 - 4.4 MI Results 62
 - 4.4 Analysis of Imputed Datasets 66
 - 4.4.1 First Imputed Dataset 69
 - 4.4.2 Second Imputed Dataset 75
 - 4.4.3 Third Imputed Dataset..... 79

4.4.4	Fourth Imputed Dataset	83
4.4.5	Fifth Imputed Dataset	87
Chapter 5	Discussion	91
Chapter 6	Conclusion	98
Bibliography.....		102

Table of Figures

Figure 1:(Modified from Perz JF, Farrington LA, Pecoraro C, et al. Estimated global prevalence of Hepatitis C virus infection. 42nd Annual Meeting of the Infectious Diseases Society of America; Boston, MA, USA; Sept 30–Oct 3, 2004. Data source: WHO.....	10
Figure 2: Variable Classification.....	13
Figure 3: Data Processing Flowchart	16
Figure 4: Number of notification of newly acquired HCV in Australia.....	19
Figure 5: Number of publications per year in Journals and Books	27
Figure 6: Common Patterns of Missingness.....	37
Figure 7: Imputation and Analysis Flowchart for ACT Pathology Dataset	39
Figure 8: Logistic Regression Model Building Process	42
Figure 9: Data Analysis Model	44
Figure 10 Scree Plot of the first 10 components of PCA	51
Figure 11 ROC Curve for Logistic Regression Model of Complete Cases (n = 258).....	55
Figure 12 Decision Tree for Complete Cases (n=602).....	57
Figure 13 ROC Curve of the Decision Tree Model for Complete Cases (n = 258)	58
Figure 14 Missing Patterns.....	59
Figure 15 Missing Value Pattern by Variable	63
Figure 16 Convergence Chart for HepC & ALT.....	64
Figure 17 Convergence Charts for RDW & Plt.....	65
Figure 18 ROC Curve - Logistic Regression Model of the first Imputed Dataset (n = 575)	71
Figure 19 Decision Tree Model of the first Imputed Dataset (n = 1341).....	72
Figure 20 ROC Curve - Decision Tree Model of first Imputed Dataset (n = 575)	73
Figure 21 ROC Curve - Logistic Regression Model of the Second Imputed Dataset (n = 592)	76
Figure 22 Decision Tree Model of the Second Imputed Dataset (n = 1382).....	77
Figure 23 ROC Curve - Decision Tree Model of the Second Imputed Dataset (n = 592)	78
Figure 24 ROC Curve - Logistic Regression Model for the third Imputed Dataset (n = 559).....	80
Figure 25 Decision Tree - for the Third Imputed Dataset (n = 1305)	81
Figure 26 ROC Curve - Decision Tree Model for the Third Imputed Dataset (n = 559).....	82
Figure 27 ROC Curve - Logistic Regression Model for the Fourth Imputed Dataset (n = 476).....	84
Figure 28 Decision Tree Model for the Fourth Imputed Dataset (n = 1112)	85
Figure 29 ROC Curve - Decision Tree Model for the Fourth Imputed Dataset (n = 476)	86
Figure 30 ROC Curve - Logistic Regression Model for the Fifth Imputed Dataset (n = 576).....	88
Figure 31 Decision Tree for the Fifth Imputed Dataset (n = 1344).....	89
Figure 32 ROC Curve - Decision Tree Model for the Fifth Imputed Dataset (n = 576).....	90
Figure 33 Disease Progression of HCV infection	98

Tables

Table 1: Variables Retained for Analysis.....	14
Table 2: Newly acquired HCV by age and sex, Australia Jan – Oct 2010.....	20
Table 3: Current Tests for HCV infection.....	24
Table 4: Traditional Techniques to handle Missing Data	26
Table 5: HCV positive cases by Age in the ACT Pathology Dataset.....	46
Table 6 Summary Statistics of Available Cases.....	48
Table 7 Variables retained after the first round of PCA.....	50
Table 8 Regression coefficients and the Odds Ratio for Logistic Regression Model	53
Table 9 Missingness for Each Variable for 14329 Cases	61
Table 10 Collection of Variables common in Logistic Regression Models.....	67
Table 11 Mean and Standard Deviations of Variables after MI.....	68
Table 12 Coefficients and Odds Ratios - Logistic Regression Model for the first Imputed Dataset ...	69
Table 13 Coefficients and Odds Ratios - Logistic Regression Model of the Second Imputed Dataset	75
Table 14 Coefficients and Odds Ratios of Logistic Regression Model for the Third Imputed Dataset	79
Table 15 Coefficients and Odds Ratios - Logistic Regression Model for the Fourth Imputed Dataset	83
Table 16 Coefficients and Odds Ratios - Logistic Regression Model for the Fifth Imputed Dataset ...	87
Table 17 Summary of the Models	92
Table 18: Summary of Decision Tree Models	92
Table 19 Summary of Odds Ratios	93
Table 20 Pooled Regression Coefficients of Imputed Data	94

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.