



Faculty of Commerce
Ain shams University

**Statistical Model Proposed to Predict Survival Rate
among Patients Performed Liver Transplantation
Operation in Egypt**

By

Sally Hossam EIDin Ahmed Zakria

Demonstrator - Statistics, Mathematics & Insurance Department
Faculty of Commerce – Ain Shams University

Under supervision of

Prof. Dr. Medhat Mohamed Ahmed Abdel Aal

Professor – Statistics, Mathematics & Insurance Department
Faculty of Commerce – Ain Shams University

Prof. Dr. Mohamed Amin Sakr

Professor – Tropical Medicine Department
Faculty of Medicine – Ain Shams University

A thesis submitted in partial fulfillment of the requirements for the
Master Degree in Applied Statistics

2016



Faculty of Commerce
Ain shams University

Approval sheet

Title of thesis: Statistical Model Proposed to Predict Survival Rate
among Patients Performed Liver Transplantation
Operation in Egypt

Academic Degree: M.Sc. in Applied Statistics

Name of Student: Sally Hossam ElDin Ahmed Zakria

The thesis submitted in partial fulfillment of the requirements for the
Master Degree in Applied Statistics has been approved by:

Examination Committee

- 1- Professor Dr. Mostafa Galal Mostafa**
Professor – Statistics, Mathematics & Insurance Department
Faculty of Commerce – Ain Shams University
- 2- Professor Dr. Medhat Mohamed Ahmed Abdel Aal**
Professor – Statistics, Mathematics & Insurance Department
Faculty of Commerce – Ain Shams University
- 3- Professor Dr. Mohamed Amin Sakr**
Professor – Faculty of Medicine – Ain Shams University
- 4- Professor Dr. Abdulla Ahmed Abd El - Ghaly**
Professor– Faculty of Economics & Political Science - Cairo
University

Date of Dissertation Defense / / 2016

Approval date / / 2016

Dedicated To
My Parents,
My Husband
And My Son

Acknowledgements

First and foremost, praise and thanks must be to ALLAH who guides me throughout my life and gave me the power and determination to accomplish this research work.

I would like to thank and express my deepest appreciation to **Prof. Dr. Mostafa Galal Mostafa** for all what I learnt from him throughout my academic years. Also I would like to thank him for agreeing to be a member of this committee and his valuable time and advices which added a great value to the study.

I sincerely would like to express my deepest gratitude and thanks to my supervisor Prof. **Dr. Medhat Mohamed Abdel Aal**, for his careful supervision, continuous encouragement, and great support throughout the work. He was always willing to give his time without any delay. Thanks for his generous support and sincere guidance, supervision and advices.

I am really deeply grateful to **Prof. Dr. Mohamed Amin Sakr** for his great help in the medical aspects for this work, and his assistance and efforts for getting the medical data.

I strongly acknowledge the contribution of **Prof. Dr. Abdulla Abd El - Ghaly** for his valuable time, patience and professionalism which added a great value to the study.

Finally, I am most grateful to my parents, sister and brothers they have gave me all the help and support I could possibly ask for. And very special thanks to my beloved husband, and my little baby Omar, who have always loved me and supported my dreams. I am really thankful to everyone who took part in exhibiting this work to light.

Abstract

Sally Hossam El din Ahmed Zakria Zaki

**Statistical Model Proposed to Predict Survival Rate among
Patients Performed Liver Transplantation Operation in Egypt**

Master Degree in Applied Statistics

Ain Shams University, Faculty of Commerce

Statistics, Mathematics & Insurance Department

2016.

Survival analysis refers to the general set of statistical methods developed specifically to model the timing of events. This thesis concerns a subset of those methods that deals with non informative right censored data.

The main objective of the current study is to construct statistical model that estimate the survival function of Egyptian patients performed liver transplantation operation and to determine the risk factors affecting the outcome of liver transplantation operation by using different statistical methods represented in non parametric, semi parametric and parametric methods. Also the study aimed to construct the feed-forward neural network and use it as a classifier to distinguish between censored and uncensored patients.

The study showed that the probability of 1 year survival after living donor liver transplantation was 85.76% with mean survival time 10.504 months however the probability of 2 year survival after living donor liver transplantation was 81.45% with mean survival time 20.584 months.

Also the Cox proportional hazard regression model showed that: the variables Recipient age, $MELD_3$, Ln_Creatinine, and GRWR are

statistically significant and selected as significant factors for risk of death after liver transplantation operation .

The Cox proportional hazard model displayed lack of fit in this study, since the variable Recipient age violates the proportional-hazards assumption. The stratified Cox model with interaction and with no interaction was applied and showed that the no-interaction model is acceptable at 0.05 level of significance and the variables $MELD_3$, $Ln_Creatinine$ are statistically significant and selected as significant factors for risk of death after liver transplantation operation.

In contrast of the proportional hazard model, the Accelerated Failure Time model (AFT) provides an adequate description of this data. The family of the AFT models including the exponential AFT model, Weibull AFT model, log-logistic AFT model, log-normal AFT model, was applied to this dataset. And it was concluded that the lognormal AFT model is the best model fitting and showed that the variables: Recipient age, $MELD_3$, and GRWR are statistically significant and selected as significant factors for risk of death after liver transplantation operation .

Also the Piecewise Constant Exponential model showed that the hazard is not constant over the time, and the pattern of the coefficient estimates is not monotonic.

Moreover the feed forward neural network were developed and trained using MS Excel it showed that the best network is achieved when the minimum mean square error for training data was 0.0378, and for testing data is 0.2233. It was concluded for the *training data set* the percent of correct classification was 95.0%. Also for *testing data set* the percent of correct classification was 75.80%.

Keywords: Survival analysis, Censoring, Kaplan–Meier estimate, Cox PH regression model, Stratified Cox regression model, Accelerated failure time model, Piecewise Constant Exponential model , Feed forward neural network, Multilayer perceptron, Back propagation algorithm.

Table Of Contents

CHAPTER 1: Introduction.....
1.1	INTRODUCTION..... 2
1.2	NATURE OF THE PROBLEM..... 4
1.3	THESIS IMPORTANCE 5
1.4	THESIS OBJECTIVES..... 6
1.5	VARIABLES OF THE MODEL 7
1.6	THESIS POPULATION..... 8
1.7	SOURCE OF THE DATA..... 8
1.8	THESIS OUTLINE 9
1.9	THE PREVIOUS STUDIES..... 11
1.10	MEDICAL DEFINITIONS..... 16
1.10.1	<i>Child- Pugh score:</i> 16
1.10.2	<i>MELD (Model for End-Stage Liver Disease)</i> 18
1.10.3	<i>GRWR (Graft to Recipient Body Weight Ratio)</i> 20
CHAPTER 2: Non Parametric & Semi Parametric Methods For Survival Data.....ERROR!
BOOKMARK NOT DEFINED.	
2.1	INTRODUCTION..... 23
2.2	BASIC SURVIVAL ANALYSIS DEFINITIONS 25
2.2.1	<i>The Survival Function</i> 25
2.2.2	<i>The Hazard Function</i> 27
2.3	INCOMPLETE DATA..... 30
2.4	CENSORING: 30
2.4.1	<i>Right censoring:</i> 31
2.4.2	<i>Left censoring</i> 31
2.4.3	<i>Interval censoring</i> 32
2.5	NON-INFORMATIVE CENSORING 33
2.6	LIKELIHOOD CONSTRUCTION FOR CENSORED DATA 33
2.7	NON-PARAMETRIC METHODS 34
2.8	ESTIMATING THE SURVIVAL FUNCTION 34
2.8.1	<i>Life-Table estimate of survival function</i> 34
2.8.2	<i>The Kaplan-Meier estimate of the survival function</i> 35
2.8.3	<i>Nelson - Aalen estimator of the survival function</i> 36
2.9	STANDARD ERROR FOR THE ESTIMATED SURVIVAL FUNCTION..... 37
2.10	NONPARAMETRIC COMPARISON OF SURVIVAL DISTRIBUTIONS 38
2.10.1	<i>Mantel-Haenszel log rank test</i> 38
2.11	COX REGRESSION MODEL 40

2.12	LIKELIHOOD ESTIMATION FOR THE COX PH MODEL.....	42
2.13	THE SCORE FUNCTION AND INFORMATION MATRIX:	43
2.14	HANDLING TIED FAILURE TIMES.....	45
2.15	MODEL CHECKING IN COX REGRESSION MODEL	46
2.15.1	<i>Cox-Snell residuals</i>	46
2.15.2	<i>Martingale residuals</i>	47
2.15.3	<i>Deviance residuals</i>	49
2.15.4	<i>Schoenfeld residuals</i>	49
2.15.5	<i>Proportional hazard assumption checking</i>	51
2.15.6	<i>Identification of Influential observation</i>	53
2.16	NON PROPORTIONAL HAZARDS MODEL:	54
2.16.1	<i>The Stratified Cox Regression Model:</i>	54
2.17	FRAILTY MODELS	57
2.17.1	<i>Univariate frailty models</i>	58
2.17.2	<i>Shared frailty models</i>	60
 CHAPTER 3: Parametric Models For Survival Data		
3.1	INTRODUCTION.....	63
3.2	PARAMETRIC PROPORTIONAL HAZARDS MODEL	63
3.2.1	<i>Weibull PH model</i>	65
3.2.2	<i>Exponential PH model</i>	67
3.2.3	<i>The Piece-Wise Constant Exponential Model</i>	68
3.2.4	<i>Gompertz PH model</i>	70
3.3	ACCELERATED FAILURE TIME MODEL	71
3.3.1	<i>Weibull AFT model</i>	74
3.3.2	<i>Log-Logestic AFT model</i>	75
3.3.3	<i>Lognormal AFT model</i>	77
3.3.4	<i>Gamma AFT model</i>	79
3.4	MODEL CHECKING IN PARAMETRIC MODELS.....	80
3.4.1	<i>Evaluation Criteria</i>	80
3.5	DISCRETE TIME SURVIVAL MODELS.....	81
3.5.1	<i>Discrete Logistic Model (Proportional Odds Model)</i>	83
3.5.2	<i>Discrete Survival and the C-Log-Log model</i>	84
3.5.3	<i>Discrete Time versus Continuous Time Models</i>	86
 CHAPTER 4: Artificial Neural Network For Survival Data		
4.1	INTRODUCTION.....	91
4.2	ANN ARCHITECTURE.....	92
4.3	TYPES OF ACTIVATION FUNCTIONS	92

4.3.1	<i>Identity activation function</i>	93
4.3.2	<i>Threshold (unit step) activation function</i>	93
4.3.3	<i>Piecewise-linear activation function</i>	94
4.3.4	<i>Sigmoid (logistic) activation function</i>	94
4.3.5	<i>Hyperbolic tangent function</i>	95
4.4	ANN LEARNING PROCESS	96
4.4.1	<i>Supervised learning</i>	97
4.4.2	<i>Unsupervised learning</i>	97
4.4.3	<i>Reinforcement learning</i>	97
4.5	PERCEPTRON	98
4.6	MULTILAYER PERCEPTRON.....	100
4.7	BACK-PROPAGATION ALGORITHM	101
4.8	LEARNING RATE AND MOMENTUM	103
4.9	ANN GENERALIZATION	103
4.10	WEIGHT DECAY	105
4.11	BUILDING NEURAL NETWORK BY USING MS EXCEL.....	105
CHAPTER 5: The Application Of The Statistical Techniques		
5.1	INTRODUCTION.....	121
5.2	CLINICAL DATA.....	121
5.3	DESCRIPTION OF THE VARIABLES.....	122
5.4	DESCRIPTIVE ANALYSIS:	124
5.5	KAPLAN –MEIER ESTIMATE OF THE SURVIVAL FUNCTION.....	126
5.6	COX PROPORTIONAL HAZARDS REGRESSION MODEL:	135
5.6.1	<i>Univariate Cox PH regression analysis</i>	135
5.6.2	<i>Multivariate Cox PH regression analysis</i>	139
5.7	MODEL CHECKING	144
5.7.1	<i>The PH assumption checking</i>	144
5.7.2	<i>Cox-Snell residuals</i>	147
5.8	STRATIFIED COX REGRESSION MODEL:	148
5.9	PARAMETRIC MODELS.....	150
	<i>Estimation of Accelerated Failure Time Models</i>	150
5.9.2	<i>Multivariate AFT analysis</i>	156
5.9.3	<i>Evaluation Criteria</i>	159
5.10	PIECEWISE CONSTANT EXPONENTIAL MODEL	162
5.11	DISCRETE TIME SURVIVAL MODELS:	164
5.12	FEED FORWARD NEURAL NETWORK	164

CHAPTER 6: Conclusions &Future Work	
6.1	<i>Results of the Kaplan-Meier estimate and the log rank test..... 170</i>
6.2	<i>Results of the Cox proportional hazard regression model..... 171</i>
6.3	<i>Results of the Stratified Cox regression model..... 173</i>
6.4	<i>Results of the parametric AFT models..... 174</i>
6.5	<i>Results of the Piecewise Constant Exponential model..... 175</i>
6.6	<i>Results of the Feed Forward Neural network..... 176</i>
6.7	<i>Conclusions:..... 177</i>
6.8	<i>Recommendations for Future work..... 180</i>
REFERENCES	183
APPENDIX.....	193

List of Tables

Table (1.1): The Population of the study.....	8
Table (1.2): Child-Pough Score.....	17
Table (1.3): Percentage of survival of cirrhotic liver disease	17
Table (1.4): MELD Score to predict patients survival rate after LT.....	18
Table (2.1) : Relationship between $F(t)$, $f(t)$, $h(t)$, $S(t)$	28
Table (3.1): Different error term distributions and the respective AFT models	72
Table (3.2): Comparing parametric models.....	80
Table (5.1): Descriptive statistics.....	122
Table (5.2): The overall survivor function	125
Table (5.3): The Nelson Aalen cumulative hazard estimate.....	126
Table (5.4): Summary of censored cases	127
Table (5.5): The Kaplan Meier survival function for MELD variable...	128
Table (5.6): The Nelson Aalen cumulative hazard estimate for MELD variable.....	129
Table (5.7): Mean for survival time.....	130
Table (5.8): Log-rank test for equality of survivor functions	132
Table (5.9): Log-rank test for equality of survivor functions (pairwise comparison)	132
Table (5.10): Univariate Cox PH regression analysis.....	134
Table (5.11): Multivariate Cox PH regression analysis.....	138
Table (5.12): Elimination of variable with high p- value by Stepwise..	139
Table (5.13): The final model (Cox PH model).....	140
Table (5.14): Test PH assumption by using Scaled Schoenfeld residuals.....	143
Table (5.15): Results for the No-interaction and Interaction Models...	146
Table (5.16):Univariate AFT analysis.....	149
Table (5.17):Multivariate AFT analysis:.....	155
Table (5.18): The final model (AFT model).....	156
Table (5.19): Comparing AIC scores for AFT models.....	157
Table (5.20): The piecewise constant exponential model.....	161
Table (5.21): The Training set and the Test set of the FFNN.....	162
Table (5.22): The MSE for NNs with different numbers of neurons in the hidden layer.....	163
Table (5.23-a): The classification matrix for the Training dataset.....	164
Table (5.23-b): The classification matrix for the Testing dataset.....	165
Table (5.24): The percent of incorrect predictions of theFFNN.....	165

List of Figures

Figure (1.1): Three months mortality based on listing MELD in patients on OPTN Waiting List	19
Figure (2.1) : Exponential, lognormal and Weibull survival functions.....	25
Figure (2.2): Right censored data.....	30
Figure (2.3): Left censoring data.....	31
Figure (2.4): Data with interval censoring.....	31
Figure (3.1): The hazard function of Weibull distribution	64
Figure (3.2) Exponential distribution: (a) survivorship function; (b) probability density function; (c) hazard function.....	66
Figure (3.3): The Survival curves of group 1 & group 2 under AFT model.....	71
Figure (3.4) : Summary of parametric models.....	72
Figure (3.5): The hazard function of log-logistic distribution	75
Figure (3.6): The hazard of log normal distribution.....	77
Figure (3.7) : Functional forms for the hazard rate.....	85
Figure (4.1) : ANN architecture	90
Figure (4.2) : Identity activation function.....	91
Figure (4.3) : Threshold activation function.....	91
Figure (4.4) : Piecewise-linear activation function	92
Figure (4.5) : Sigmoid activation function.....	93
Figure (4.6) : Hyperbolic activation function.....	93
Figure (4.7) : Training process of the NN.....	94
Figure (4.8) : Linearly separable patterns (a) and non linearly separable patterns (b)	96
Figure (4.9) : principle of generalization and overfitting (a) good generalization (b) Over fitted data.	102
Figure (4.10) : Scaling data by using MS Excel.	104
Figure (4.11) : Specifying the data range in scaling data.....	105
Figure (4.12) : Using the sigmoid activation function in H1.....	109
Figure (4.13) : Using the sigmoid activation function in O1.....	110
Figure (4.14) : Calculating the error function.....	111
Figure (4.15) : Calculating the Mean Squared Error.....	111
Figure (4.16) : Randomize weights.....	112
Figure (4.17) : Filling weight vector.....	112
Figure (4.18) : Selecting Excel's Solver	113
Figure (4.19): Solver parameters dialog box.....	113
Figure (4.20): Determining the objective function.....	114
Figure (4.21): Determining the adjustable cells.....	114
Figure (4.22): Adding constrains.....	115
Figure (4.23): Solver options dialoge box.....	115
Figure (4.24): Solver results dialoge box.....	116
Figure (5.1): The overall survivor function	125
Figure (5.2): The Nelson Aalen cumulative hazard function.....	126
Figure (5.3): Comparison of Kaplan–Meier survival curve & Nelson	131

Aalen cumulative hazard curve for MELD score groups.....	
Figure (5.4): Cox PH model: (a) survivorship function; (b) cumulative hazard function; (c) estimated hazard function.....	141
Figur (5.5): Log-Log Survival plot for Rec.Age and $MELD_3$	143
Figure (5.6): Scaled Schoenfeld residuals to test PH assumption.....	144
Figure (5.7): Cumulative hazard plot of the Cox-Snell residual for Cox PH model.....	145
Figure (5.8): The Cox-Snell residual for Log-normal AFT model.....	158
Figure (5.9): Log normal AFT model: (a) survivorship function; (b) cumulative hazard function; (c) estimated hazard function.....	159
Figure (5.10) : The proposed feed forward neural network.....	164

List of Abbreviations

AFT	Accelerated Failure Time Model.
AIC	Akaike Information Criterion.
ANN	Artificial Neural Network.
BMI	Body Mass Index.
CTP	Child Turcotte Pough.
ESLD	End Stage Liver Disease.
FFNN	Feed Forward Neural Network.
GLM	Generalized Linear Regression Model.
GRWR	Graft-Recipient Body weight Ratio.
HBV	Hepatitis B Virus.
HCC	Hepato-cellular Carcinoma.
HCV	Hepatitis C Virus.
INR	International Normalized Ratio.
K-M	Kaplan-Meier.
LDLT	Living Donor Liver Transplantation
LT	Liver Transplantation
MELD	Model for End Stage Liver Disease.
NN	Neural Network.
OPTN	Organ Procurement Transplantation Network.
PAT	Parenteral Antischistosomiasis Therapy .
PCE	The Piecewise-Constant Exponential Model.
PH	Proportional Hazard.
PO	Proportional Odds.
UNOS	United Nation for Organ Sharing.
WHO	World Health Organization