

PREDICTION OF RECURRENCE IN THIN MELANOMA
USING TREES AND RANDOM FORESTS

Richard M. Reiter

A Thesis Submitted to the
University of North Carolina Wilmington in Partial Fulfillment
Of the Requirements for the Degree of
Master of Science

Department of Mathematics and Statistics

University of North Carolina Wilmington

2005

Approved by

Advisory Committee

Chair

Accepted by

Dean, Graduate School

TABLE OF CONTENTS

ABSTRACT	iv
ACKNOWLEDGMENTS	v
LIST OF FIGURE.....	vi
LIST OF TABLES	vii
LIST OF ITEMS	ix
INTRODUCTION	1
TREES AND PRUNING	4
Growing a Tree	4
Tree Pruning Process	10
Testing the Tree	13
Re-substitution	13
Set Aside	13
Cross Validation.....	14
RANDOM FORESTS.....	15
PATIENT DATA.....	20
TREE AND FOREST DATA.....	23
Tree Output	24
Cross Validation Deviance and Misclassification Output	27
Random Forest Formation	28
Variable Importance Plot	31
METHODS	31
FINDINGS	40

All Patients, All Variables, Any Recurrence	40
Determining the Number of Trees	43
Determining the Number of Variables at Each Split	44
All Patients All Variables Recurrence More Than Local	46
All Patients, Leave Out Variables, Any Recurrence.....	48
All Patients, Leave Out Variables, Recurrence More Than Local	49
Limited Patients, Leave Out ‘typerec’, Any Recurrence	50
Limited Patients, Leave Out ‘typerec’, Recurrence More Than Local.....	51
Direct Comparison Random Forest and Single Trees Same Patients	51
Evaluation of Correct and Misclassified Limited Patients	52
RESULTS	54
CONCLUSIONS.....	56
BIBLIOGRAPHY.....	60
APPENDIX.....	62

ABSTRACT

In this paper, we will try to predict the recurrence of melanoma in the subset of patients with a “thin melanoma,” defined by having a Breslow Thickness less than 1.00mm. The study used the 1610 patient data base from the Duke Melanoma Clinic. The methods used to try classifying these patients as to whether or not they will suffer a recurrence are based on the statistical processes of trees and random forests introduced by Dr. L. Breiman. Successful results were achieved only when all patients were studied including those with an advanced stage of disease, but who still qualified as patients with a thin melanoma by histological measurement. The variables collected on each patient and used in this study related to the Breslow Thickness, Clark level, initial stage, age, primary site of lesion, local pathological events, histological type, sex, race and previous immunotherapy. The factors other than the initial stage and extent of disease proved too weak to give any meaningful results when only patients with local disease were studied. The results using a single tree and a random forest were compared. A detailed discussion of trees and random forest with use in the R-2.01 CRAN software packages is included. Although the results of this study did not provide the answers for this specific set of melanoma patients, it is felt that the techniques and programs written would be applicable in a differently defined set of patients. The results could be an important step in determining which patients should receive therapy and then evaluating the results of the adjuvant therapy on a more properly defined group of patients.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. Dargan Frierson of the Mathematics and Statistics Department of the University of North Carolina Wilmington for his excellent guidance helping me to complete my thesis. I appreciate the dedication and hard work of Dr. Susan Simmons and Dr. Gabriel Lugo serving as advisors for this thesis. I would also like to thank Dr. H. Seigler , Professor of Surgery and head of the Duke Melanoma Clinic, for sponsoring my application to use the D.U.M.C. patient data base and providing clinical guidance for the study. Finally I would acknowledge the expertise of Dr. L. Breiman and his staff for creating and sharing the mathematics and methods of trees and random forest.

LIST OF FIGURES

Figure	Page
1. Cross Validation Deviance Plot	63
2. Cross Validation Misclassification Plot.....	64
3. Plot of Full Tree without Text, All Patients, All Variables and Any Recurrence	64
4. Plot of Tree All Patients, All Variables and Any Recurrence. Pruned Tree k=30	65
5. Plot of Tree All Patients All Variables and Any Recurrence Pruned Tree k=8	65
6. The Variable Importance Plot of All Patients, All Variables and Any Recurrence (L) The Variable Importance Plot of All Patients, All Variables and Recurrence More Than Local (R).....	66
7. The Variable Importance Plot of All Patients, Leave Out STGGRP, DEXTEXT and TYPCAS. Any Recurrence (L). The Variable Importance Plot of All Patients, Leave Out STGGRP, DEXTEXT and TYPCAS, Recurrence More Than Local(R). ..	67
8. The Variable Importance Plot of Limited Patients, All Variables and Any Recurrence (L). The Variable Importance Plot of Limited Patients, All Variables and Recurrence More Than Local (R)	68
9. Text Full Tree All Patients All Variables Any Recurrence	69

LIST OF TABLES

Table	Page
1. 10 Consecutive Random Forests 10 Trees.....	77
2. 10 Consecutive Random Forests 25 Trees.....	77
3. 10 Consecutive Random Forests 50 Trees.....	78
4. 10 Consecutive Random Forests 100 Trees.....	78
5. 10 Consecutive Random Forests 200 Trees.....	79
6. 10 Consecutive Random Forests 300 Trees.....	79
7. 10 Consecutive Random Forests 500 Trees.....	80
8. 10 Consecutive Random Forests 1000 Trees.....	80
9. Consecutive Larger Trees 2000 3000 5000	81
10. Single Tree Results All Patients All Variables Any Recurrence.....	82
11. Single Tree Results All Patients All Variables Recurrence More Than Local.....	83
12. Single Tree Results All Patients All Variables Recurrence More Than Local 2 nd Run	84
13. Single Tree Results All Patients Leave Out Variables Any Recurrence	85
14. Single Tree Results All Patients Leave Out Variables Any Recurrence 2 nd Run.....	86
15. Single Tree Results All Patients Leave Out Variables Recurrence More Than Local	87
16. Single Tree Results All Patients Leave Out Variables Recurrence More Than Local 2 nd Run	88
17. Single Tree Results Limited Patients All Variables Any Recurrence	89
18. Single Tree Results Limited Patients All Variables Recurrence More Than Local	90

19. RF All Patients All Variables Any Recurrence	91
20. RF All Patients All Variables Recurrence More Than Local	98
21. RF All Patients Leave Out Variables Any Recurrence.....	105
22. RF All Patients Leave Out Variables More Than Local.....	109
23. RF Limited Patients Any Recurrence	113
24. RF Limited Patients Recurrence More Than Local	116
25. Tables of Direct Comparisons of Random Forests and Single Tree for the Six Different Groups	119

LIST OF ITEMS

Item	Page
1. The qqPlot and Table of YES and NONE of All Limited Patients	121
2. Error Rates of Sample Group; qqPlot and Results of YES Sample Patients	122
3. The qqPlot and Results of NONE Sample Patients	123
4. Portion of Random forest Program to Create Tables for Each Study Group.....	124
5. Programs to Create Sets of Single Trees and Random Forests for Direct Comparison on the Same Random Sample of 126 Patients. Includes Collection and Summaries from Misclassified Patients	133
6. D.U.M.C. Protocol for Thin Melanoma.....	138
7. 10 Patients with Variables from Data Base	147