

Classification and Regression Trees (CART) in Psychiatric Screening, Diagnosis and Subgroup Analysis

Mr Dean Philip McKenzie

BA (Hons)

Submitted in total fulfilment of the requirement of the degree of
Doctor of Philosophy

Monash University Faculty of Medicine, Nursing and Health Sciences

Department of Epidemiology and Preventive Medicine

December 2008

CONTENTS

Contents	2
Abstract.....	4
Candidate's General Declaration	7
Acknowledgments.....	9
Publications.....	12
Abbreviations	14
Structure of Thesis.....	17
1. Introduction	19
1.1. Conventional statistical methods	21
1.1.1. Linear decision rules	21
1.1.2. Parametric techniques	21
1.2. Machine Learning	23
1.3. Tree-building / recursive partitioning and subgroup analysis	24
1.3.1. Illustrative example of recursive partitioning	28
1.3.2. Recursive partitioning: advantages	31
1.3.3. Recursive partitioning: historical background.....	33
1.3.4. The need for tree-pruning and stopping rules	37
1.3.5. Statistical significance testing	40
1.3.6. Cross-validation: the Classification and Regression Tree (CART) algorithm	46
1.3.7. Disadvantages of recursive partitioning in general.....	55
1.3.8. Classification performance of recursive partitioning algorithms	63
1.4. Summary	65
1.5. Research aims.....	66
Background to Chapters Two and Three: The Australian Gulf War Veterans' Health Study	67
2. Introduction to Chapter Two: Hazardous or harmful alcohol use in Royal Australian Navy veterans of the 1991 Gulf War: identification of high risk subgroups	69
2.1. Hazardous alcohol consumption in the Military.....	69
2.2. Identifying potential alcohol misuse	69
2.3. Classification and Regression Tree (CART) analysis of subgroups at risk of hazardous alcohol consumption	70
3. Introduction to Chapter Three: Temporal relationships between Gulf War deployment and subsequent psychological disorders.....	87
3.1. CART and Logistic Regression Analysis of Temporal Relationships	88
4. Introduction to Chapter Four: Pessimism, worthlessness, anhedonia and thoughts of death identify DSM-IV major depression in hospitalised medically ill	122
4.1. Subtypes of depression	122
4.2. CART analysis of key symptoms of demoralization and anhedonia ...	123

5. Introduction to Chapter Five: Somatic and psychological dimensions of screening for psychiatric morbidity: a community validation of the SPHERE Questionnaire.....	149
5.1. The Somatic and Psychological Health Report (SPHERE).....	149
5.2. Screening for psychiatric disorders in young adults.....	150
5.3. Further analysis using CART.....	151
5.4. Addendum to Chapter Five: Further analysis of simple SPHERE screening rules using CART	164
5.4.1. Introduction	164
5.4.2. Method.....	165
5.4.3. Results.....	167
5.4.4. Discussion.....	168
6. Introduction to Chapter Six: Models developed by three techniques did not achieve acceptable prediction of binary trauma outcomes	171
6.1. Trauma datasets.....	171
6.2. Comparison of techniques	172
7. Discussion.....	186
7.1. Subgroups of Australian Gulf War veterans at high risk of hazardous or harmful alcohol consumption	186
7.2. Temporal progression of psychiatric disorders in Australian Gulf War veterans	187
7.3. Subtypes of DSM-IV major depression in the hospitalised medically ill....	187
7.4. Psychological and somatic symptoms of depression in young adults.	188
7.5. Prediction of binary trauma outcomes	189
7.6. Ways in which CART could be improved.....	190
7.6.1. Permutation testing of splits and final subgroups.....	190
7.6.2. Probabilistic assignment to subgroups, and more control over variables used in the growing of decision trees.....	192
7.6.3. Boolean rules within each split.....	194
7.6.4. Specification of performance criteria	194
7.7. Conclusions	195
References	199
Appendix A - AUDIT (Alcohol Use Disorders Identification Test) questionnaire	223
Appendix B - SPHERE (Somatic and Psychological Health Report) questionnaire	225

ABSTRACT

Psychiatric illnesses such as anxiety and depression are becoming increasingly more prevalent, and are associated with poor quality of life, as well as an increased risk of comorbid physical illness. Although it is important that psychiatric illness and its risk factors be clearly identified and understood, accurate screening and diagnosis can be difficult to achieve in practice.

Several statistical procedures, including simple linear decision rules based upon the number of symptoms present, and more sophisticated parametric methods such as logistic regression, have been employed in the development of screening and diagnostic tests. Such procedures make various assumptions of the data, that might not be met in practice, and give results that might be difficult to interpret by a clinician.

Alternative methods, such as tree-building or recursive partitioning techniques, make fewer assumptions of the data, can identify subgroups of individuals sharing particular combinations of potential risk factors and likelihood of outcome, and provide decision trees that are readily interpretable if they are not too large.

One of the best known and most frequently used contemporary recursive partitioning techniques is the Classification and Regression Trees (CART) algorithm. Although CART has been applied to psychiatric data since the early 1990's, it is arguably not as widely employed as it could be, particularly as an adjunct to conventional techniques. This may be due to a lack of familiarity with

CART, or it may be due to apprehension regarding the generality of results, based upon the limitations of early such techniques.

This thesis firstly looks at how recursive partitioning methods have been improved over the years, in order to better generalise their results to new datasets. CART is then applied in the screening, diagnosis and subgroup analysis of Australian Gulf War veteran, hospitalised medically ill, young adult, and physical trauma patient datasets, in conjunction with conventional techniques such as logistic regression. In all cases the performance of CART was validated on a separate subsample, or sample.

CART identified subgroups of Australian Gulf War veterans at high risk of hazardous or harmful alcohol consumption, and at high risk of developing psychiatric disorders such as major depression. Combinations of variables associated with the former included being married, and having current major depression.

In hospitalised medically ill patients, CART found two combinations of symptoms – pessimism and worthlessness, and pessimism, loss of interest in others, and thoughts of death - to be highly associated with major depression, suggesting that there may be at least two subtypes of this disorder.

CART developed a combination of simple screening rules for major depression in young adults that performed significantly better than currently employed rules. Although CART did not perform as well as logistic regression at predicting binary outcomes in trauma patients, the models that it generated were

arguably more interpretable than the logistic regression, and artificial neural network, models.

By detecting specific combinations of variables, CART allows treatment and prevention strategies to be targeted for particular subgroups. Ways in which CART could be improved are discussed.

Candidate's General Declaration

In accordance with Monash University Doctorate Regulation 17 / Doctor of Philosophy and Master of Philosophy (MPhil) regulations the following declarations are made:

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes three original papers published in peer reviewed journals and two unpublished publications (one in press with a peer reviewed journal and one submitted to a peer reviewed journal). The core theme of this thesis is the application of a computational technique known as Classification and Regression Trees (CART) to screening, diagnosis and subgroup analysis of Australian Gulf War veteran, hospitalised medically ill, young adult, and physical trauma patient datasets. I took principal responsibility for the ideas, development and writing up of all the papers in the thesis for which I was first author. I also took major responsibility for the ideas, development and writing up of those papers in the thesis with me as second author; in all instances working within the Monash University Department of Epidemiology and Preventive Medicine under the primary supervision of Professor Malcolm Sim, and co-supervisors Professor David Clarke and Professor Andrew Forbes.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of the following chapters my contribution to the work involved the following:

Thesis chapter	Publication title	Publication status*	Nature and extent of candidate's contribution
2	McKenzie DP , McFarlane AC, Creamer M, Ikin JF, Forbes AB, Kelsall HL, Clarke DM, Glass DC, Ittak P, Sim MR. Hazardous or harmful alcohol use in Royal Australian Navy veterans of the 1991 Gulf War: Identification of high risk subgroups. <i>Addictive Behaviors</i> 2006; 31: 1683-1694.	Published	I was responsible for the research question, literature review, data management and programming, statistical analyses, interpreting the results, writing the paper and undertaking any required revisions.
3	McKenzie DP , Creamer M, Kelsall HL, Forbes AB, Ikin JF, Sim MR, McFarlane AC. Temporal Relationships between Gulf War Deployment and Subsequent Psychological Disorders.	Submitted	I made a major contribution to the research question. I was responsible for the literature review, data management and programming, statistical analyses, interpreting the results, writing the paper and undertaking any required revisions.
4	McKenzie DP , Clarke DM, Forbes AB, Sim MR. Pessimism, worthlessness, anhedonia and thoughts of death identify DSM-IV major depression in the medically ill. <i>Psychosomatics</i> .	In press	I was responsible for the research question, literature review, data management and programming, interpreting the results, writing the paper and undertaking any required revisions.
5	McFarlane AC, McKenzie DP , Van Hooff M, Browne DG. Somatic and psychological dimensions of screening for psychiatric morbidity: a community validation of the SPHERE questionnaire. <i>Journal of Psychiatric Research</i> 2008; 65: 337-345.	Published	Major contribution to article concept (which follows on from a 2003 paper of which I was a co-author), literature review, introduction and discussion and revision of paper. Performed extra statistical programming and all statistical analyses. Interpreted and

			wrote results section. In addition, I was responsible for the additional literature review, CART analysis, and additional interpretation of the results in the addendum to the published article.
6	Wolfe R, McKenzie DP , Black J, Simpson P, Gabbe BJ, Cameron PA. Models developed by three techniques did not achieve acceptable prediction of binary trauma outcomes. <i>Journal of Clinical Epidemiology</i> 2006, 59, 26-35.	Published	I contributed to the article concept, as well as performed specific literature review for, and wrote the sections in, Introduction pertaining to Classification and Regression Trees (CART), and artificial neural networks, and the applications of both these methods in trauma research. Wrote section in Methods pertaining to CART, and contributed to Methods section on neural networks. Applied CART to data and interpreted the results. Wrote section in Discussion pertaining to CART, and also majorly contributed to general Discussion as to how the statistical methodology employed by CART, neural networks and logistic regression could be expanded and improved. Made a major contribution to revision of the paper.

[* For example, 'published' / 'in press' / 'accepted' / 'returned for revision']

Signed:

.....

Date:

.....

ACKNOWLEDGMENTS

The Australian Gulf War Veterans' Health Study was commissioned and funded by the Australian Government Department of Veterans' Affairs.

The research concerned with depression in the medically ill was supported by a project grant from the Australian National Health and Medical Research Council (NHMRC). The research concerned with psychiatric morbidity in young adults was supported by project and program grants from the NHMRC.

The research investigating binary trauma outcomes was funded by the Victorian Trauma Foundation.

I am grateful to the NHMRC for postgraduate Public Health scholarship funding to be able to further my research.

With regard to the Australian Gulf War Veteran's Health Study, I would like to thank the Scientific Advisory Committee and Veterans' Consultative Committee, project staff at the Department of Veterans' Affairs, in particular Drs Keith Horsley, Warren Harrex and Eileen Wilson, Health Services Australia Pty Ltd, and the following Monash University Department of Epidemiology and Preventive Medicine staff: Professors Malcolm Sim, Andrew Forbes, Michael Abramson and John McNeil; Drs Helen Kelsall, Jillian Ikin, Deborah Glass and Karin Leder; Mr Peter Ittak, Mr Ewan McFarlane, Ms Luci McFarlane, Ms Koraly Giuliano, Ms Christina Dimitriadis, Mr Anthony del Monaco, Mr Geoff Aldred, Ms Emma Conyers, Ms Jane Ball and Ms Julie Attard. I would also like to thank Professors Alexander McFarlane and Mark Creamer; and the Australian Gulf War veterans and Military Comparison Group members themselves.

With regard to the study examining depression in the medically ill, I would like to thank Professors David Clarke, Graeme Smith and Helen Herrman; Mr Kevan Pitcher and Ms Ann Silbereisen, Monash Medical Centre medical staff, and the patients themselves.

With respect to the study concerned with psychiatric morbidity in young adults, I would like to thank Professor Alexander McFarlane, Ms Miranda Van Hooff, Mr Derek Brown, the Births, Deaths and Marriages Registration Office of South Australia, the Australian Institute of Health and Welfare, the Australian Electoral Commission, teachers, principals and staff from the eight schools participating in the study, and the participants themselves.

Regarding the study concerned with prediction of binary trauma outcomes, I would like to thank Associate Professors Rory Wolfe and Peter Cameron, Drs Belinda Gabbe and James Black, Ms Pam Simpson, staff involved with the Royal Melbourne Hospital adult trauma database, and Victorian State Trauma Registry; and the patients themselves.

I would like to thank fellow decision tree researchers, in particular those 'present at the birth' of tree-building methods - Dr James Morgan and Dr Laurence Press - , as well as Professor Douglas Hawkins, Associate Professor David Dowe, Dr Dan Steinberg and Dr Robert Hogenraad for some very helpful email and in vivo discussions.

I would like to thank Professor John McNeil and the Department of Epidemiology and Preventive Medicine staff for providing an environment that is so supportive, and conducive to research; Associate Professor Rory Wolfe and Ms Kaylene Hanlon for ready help with all PhD matters; my Departmental colleagues, especially Dr Helen Kelsall, Dr Geza Benke, Dr Baki Billah, Dr Fahad Hanna, Dr Jill Ikin, Mr Peter Ittak, and fellow PhD students, particularly Mr Steve Haas, Mr Imo Inyang, Ms Margaret Stebbing, Ms Helen Walls, Ms Shelly Rodrigo and Dr Linton Harris, for all their advice and camaraderie.

My thanks also go to Professors Andrew Mackinnon, Patrick McGorry and Richard Bell, formerly at the Mental Health Research Institute; and Professor Ken Ross, Professor Anthony Jorm and Dr David Share, formerly at Deakin University – Waurin Ponds, for originally fostering my interest in computational techniques.

I would like to Ms Vanessa Murray for all her very skilful and cheerful help with the formatting of this thesis.

I would especially like to thank my supervisors Professor Malcolm Sim, Professor David Clarke and Professor Andrew Forbes, who have been such a tremendous source of support, advice and encouragement, and from whom I have learnt so very much, for which I will always be grateful.

Finally, I would especially like to thank my wonderful wife, companion and inspiration Maria; my family, particularly Diana, Stephen, Wyn, Graham, Rita, Neil, Lisa and Rob; mia seconda famiglia Rosetta, Michele, Nancy, John, Lauren and Patrick; and my friends Susan, John, Mike, Brigid, Steve and Nadia. Thank you all for your considerable love, patience, support and understanding always.

Velut arbor aevo

(Horace, Odes, Book 1, Ode 12: 'May the tree thrive')

PUBLICATIONS

In addition to the five papers included as chapters of this thesis (asterisked below), during my doctoral candidature I have played a major role in the following published, in press, or submitted papers, and published encyclopedia entry.

1. Barton CA, **McKenzie DP**, Walters EH, Abramson MJ, Victorian Asthma Mortality Study Group. Interactions between psychosocial problems and management of asthma: who is at risk of dying? *Journal of Asthma* 2005;42:249-256.
2. Bell R, Davison S, Papalia H, **McKenzie DP**, Davis S. Endogenous androgen levels and cardiovascular risk profile in women across the adult life span. *Menopause* 2007;14:630-638.
3. Creamer M, Carboon I, Forbes AB, **McKenzie DP**, McFarlane AC, Kelsall HL, Sim M. Psychiatric disorder and separation from military service: A 10 year retrospective study. *American Journal of Psychiatry* 2006;163:733-734.
4. Davis SR, Shah SM, **McKenzie DP**, Kulkarni J, Davison SL, Bell RJ. Relationship between dehydroepiandrosterone sulphate levels and cognitive function in women. *Journal of Clinical Endocrinology and Metabolism* 2008;93:801-808.
5. Glass DC, Sim MR, Kelsall HL, Ikin JF, **McKenzie DP**, Forbes AB, Ittak P. What was different about exposures reported by Australian Gulf War veterans during the 1991 Gulf War compared with exposures reported for other deployments? *Military Medicine* 2006;171:632-638.
6. Ikin J, **McKenzie D**, Creamer M, McFarlane A, Sim M, Kelsall H, Glass D, Forbes A, Horsley K, Harrex W. War zone stress without direct combat: the Australian naval experience of the Gulf War. *Journal of Traumatic Stress* 2005;18:193-204.
7. Ikin JF, Sim MR, **McKenzie DP**, Henderson S, Horsley KWA, Wilson EJ, Moore MR, Harrex WK. Korean War service impacting on quality of life fifty years on. *Journal of Epidemiology and Community Health* 2008 [In press, accepted for publication June 2007]
8. Ikin JF, Sim MR, **McKenzie DP**, Horsley KWA, Wilson EJ, Moore MR, Jelfs P, Harrex WK, Henderson AS. Anxiety, PTSD and depression in Korean War veterans fifty years after the war. *British Journal of Psychiatry* 2007;190:475-483.
9. Kelsall H, Macdonell R, Sim M, Forbes A, **McKenzie D**, Glass D, Ikin J, Ittak P. Neurological status of Australian veterans of the 1991 Gulf War and the effect of medical and chemical exposures. *International Journal of Epidemiology* 2005;34:810-819.
10. Kelsall HL, Sim MR, Ikin JF, Forbes AB, **McKenzie DP**, Glass DC, Ittak P. Reproductive health of male Australian veterans of the 1991 Gulf War. *BMC Public Health* 2007;7:79.

11. Kelsall HL, Sim MR, **McKenzie DP**, Forbes AB, Leder K, Glass DC, Ikin JF, McFarlane AC. Medically evaluated psychological and physical health of Australian Gulf War veterans with chronic fatigue. *Journal of Psychosomatic Research* 2006;60:575-584.
12. Kelsall H, **McKenzie D**, Sim M, Leder K, Ross J, Forbes A, Ikin J. Comparison of self-reported and recorded vaccinations and health effects in Australian Gulf War veterans. *Vaccine* 2008;26:4290-4297.
13. Kissane D, McKenzie M, Bloch S, Moskowitz C, **McKenzie DP**, O'Neill I. Family focused grief therapy: a randomized controlled trial in palliative care and bereavement. *American Journal of Psychiatry* 2006;163:1208-1218.
14. *McFarlane AC, **McKenzie DP**, Van Hooff M, Browne D. Somatic and psychological dimensions of screening for psychiatric morbidity: a community validation of the SPHERE Questionnaire. *Journal of Psychosomatic Research* 2008;65:337-343.
15. **McKenzie DP**. Retrospective studies. In: Everitt BS, Howell DC, eds. *Encyclopedia of Statistics in Behavioral Science*. Chichester, England: Wiley, 2005;1758-1759.
16. **McKenzie DP**. Commentary on "Coping with life-threatening events was associated with better self-perceived health in a Naval cross-sectional study" by Nils Mageroy, Trond Riise, Bjorn H. Johnsen and Bente E. Moen. *Journal of Psychosomatic Research* 2008;65:619-621.
17. ***McKenzie DP**, McFarlane AC, Creamer M, Ikin J, Forbes AB, Kelsall HL, Clarke DM, Glass DC, Ittak P, Sim MR. Hazardous or harmful alcohol use in Royal Australian Navy veterans of the 1991 Gulf War: identification of high risk subgroups. *Addictive Behaviors* 2006;31:1683-1694.
18. ***McKenzie DP**, Clarke DM, Forbes AB, Sim MR. Pessimism, worthlessness, anhedonia and thoughts of death identify major depression in hospitalized medically ill. *Psychosomatics*, in press, accepted for publication October 2008.
19. ***McKenzie DP**, Creamer M, Kelsall HL, Forbes AB, Ikin JF, Sim MR, McFarlane AC. Temporal relationships between Gulf War deployment and subsequent psychological disorders. Submitted to *Journal of Affective Disorders*, September 2008.
20. McKenzie M, Clarke D, **McKenzie D**, Smith G. Persistence of DSM-IV depressive, anxiety and somatoform disorders: which factors predict the continuing presence of psychiatric morbidity in the medically ill 3 months post hospital discharge? submitted to *Journal of Psychosomatic Research*, July 2008.
21. *Wolfe R, **McKenzie DP**, Black J, Simpson P, Gabbe BJ, Cameron PA. Models developed by three techniques did not achieve acceptable prediction of binary trauma outcomes. *Journal of Clinical Epidemiology* 2006;59:26-35.

ABBREVIATIONS

ADF	Australian Defence Force
AIC	Akaike Information Criterion
AID	Automatic Interaction Detection
AUC	Area under the Curve (often referring to the area under the Receiver Operating Characteristic curve)
AUDIT	Alcohol Use Disorders Identification Test
BIC	Bayesian Information Criterion
BDI	Beck Depression Inventory
CART	Classification and Regression Trees (registered trademark of California Statistical Software, Incorporated; used by Salford Systems, Incorporated)
CHAID	Chi-squared Automatic Interaction Detection
CI	Confidence Interval
CIDI	Composite International Diagnostic Interview
CLS	Concept Learning System
CRUISE	Classification Rule with Unbiased Interaction Selection and Estimation
C-Spine	Cervical Spine
DM	Dean McKenzie
DRG	Diagnosis-Related-Groups
DSM-III-R	Diagnostic and Statistical Manual of Mental Disorders, third edition
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, fourth edition
DVA	Department of Veterans' Affairs
et al	<i>et alii</i> or <i>et alia</i> (Latin for 'and others')
FORTTRAN,	
Fortran	FORmula TRANslation
GCS	Glasgow Coma Scale
GHQ	General Health Questionnaire
GHQ-30	General Health Questionnaire, 30 item version
GHQ-36	General Health Questionnaire, 36 item version
H-L	Hosmer-Lemeshow

ICU	Intensive Care Unit
ID3	Interactive Dichotomizer 3
ISS	Injury Severity Score
JB	James Black
LOTUS	Logistic Regression Trees with Unbiased Selection
MAP	Mood Assessment Program
MCA	Multiple Classification Analysis
MCS	Mental Component Summary (of SF-12 and SF-36)
MILP	Monash Interview for Liaison Psychiatry
MMC	Monash Medical Centre
MNA	Multivariate Nominal scale Analysis
MSEQ	Military Service Experience Questionnaire
NCO	Non-Commissioned Officer
NHMRC	National Health and Medical Research Council
NPV	Negative Predictive Value
OR	Odds Ratio
P value	Probability value
PPV	Positive Predictive Value
PS	Pam Simpson
PSYCH-6	6 item Psychological subscale of the SPHERE
PSYCH	caseness threshold for PSYCH-6
PTSD	Post-Traumatic Stress Disorder
QUEST	Quick, Unbiased, Efficient Statistical Tree
RAN	Royal Australian Navy
RMH	Royal Melbourne Hospital
ROC	Receiver Operating Characteristic
RW	Rory Wolfe
SBP	Systolic Blood Pressure
SCID	Structured Clinical Interview for DSM-III-R
SD	Standard Deviation
SF-12	Short Form Health Survey, 12 item version
SF-36	Short Form Health Survey, 36 item version
SOMA-6	6 item Somatic subscale of the SPHERE

SOMA	caseness threshold for SOMA-6
SPHERE	Somatic and Psychological Health Report
SPSS	Statistical Package for the Social Sciences
TAC	Transport Accident Commission
TBI	Traumatic Brain Injury
THAID	Theta Automatic Interaction Detection
TRISS	Trauma and Injury Severity Score
UK	United Kingdom
UN	United Nations
US	United States
VSTR	Victorian State Trauma Registry
VTF	Victorian Trauma Foundation
WHO	World Health Organization

STRUCTURE OF THESIS

Chapter One briefly examines conventional statistical techniques for developing screening and diagnostic tests, and their limitations. The historical development of tree-building or recursive partitioning methods, which make less assumptions of the data than do the above techniques, is then examined. The emphasis is on how the generality of recursive partitioning methods has been improved. Finally, the advantages and disadvantages of the two main contemporary recursive partitioning methods, as well as of recursive partitioning in general, are described.

The following chapters consist of the CART applications listed below:

- the identification of subgroups of Australian Gulf War veterans at high risk of hazardous alcohol consumption (Chapter Two),
- the examination of patterns in the development of psychiatric illness, such as affective, alcohol and anxiety disorders, in Australian Gulf War veterans, so as to ascertain which veterans are likely to develop these disorders, and when (Chapter Three),
- the identification of possible subtypes of major depression, as defined by specific combinations of depressive symptoms, in medically ill patients at an Australian general hospital, and how such combinations can aid in screening for depression (Chapter Four),

- screening for anxiety disorders and depressive disorders in a community sample of young Australian adults (Chapter Five).
- the prediction of binary outcomes according to pre-defined performance criteria, using logistic regression, artificial neural networks, and CART (Chapter Six).

The application chapters are followed by the integrative Discussion (Chapter Seven) which summarises the results of each application; examines the strengths of CART and logistic regression, and examines how CART could be improved, particularly with regard to how the two techniques above could be employed together to provide valid tests of statistical inference for subgroup comparisons.

References for all in-text citations (with the exception of those separately referenced in the articles that comprise Chapters Two through Six) are provided at the end of the thesis.

1. INTRODUCTION

Psychiatric illnesses such as anxiety and depression are becoming increasingly more prevalent, and are associated with poor quality of life, as well as an increased risk of comorbid physical illness such as cardiovascular disease ^{1,2}. It is important therefore that psychiatric illness and its risk factors be clearly identified and understood, so that effective treatment and prevention strategies can take place. Nevertheless, accurate screening and diagnosis can be difficult to achieve in practice ³, particularly in psychiatry, where there is a dearth of objective and directly measurable physical representations of mental illness such as depression ^{4,5}.

A variety of statistical and computational procedures have been employed in the development of screening and diagnostic tests. Such procedures include simple linear decision rules, based upon the number of symptoms present, as well as more sophisticated parametric methods such as linear discriminant analysis and logistic regression ⁶. The above procedures make various assumptions of the data, that might not be met in practice, and give results that might be difficult to interpret by a clinician.

Alternative methods, such as tree-building or recursive partitioning techniques ⁷, make fewer assumptions of the data, can uncover nonlinear relationships and types of statistical interdependence or interactions amongst potential risk factors, identify subgroups of individuals sharing particular combinations of potential risk factors and likelihood of outcome, and provide

graphic decision trees that are readily interpretable if they are not too large.

Recursive partitioning may be defined as the repeated subdivision of a dataset into *'nonoverlapping subregions such that cases within a subregion are more similar to each other for a specified outcome variable, than with cases in other subregions. For a classification tree, the cases within one subregion would belong predominantly to one class'*⁸ (p. 115). Thus, the goal of recursive partitioning is to identify subgroups of observations, defined by different combinations of predictor variables or risk factors, such as age, gender and medical history, which are homogenous in regard to a dependent or outcome variable, such as the presence or absence of a diagnosis of major depression.

Recursive partitioning techniques first appeared in the late 1950's⁹ and have been applied to psychological and psychiatric data since the mid 1960's¹⁰.

The next sections present a brief overview of the two conventional statistical methods that are most widely used in the construction of screening and diagnostic tests¹¹ - linear decision rules, and parametric techniques such as logistic regression - as well as their limitations. The historical development of recursive partitioning techniques is then examined, focusing on how they have been improved over the years, in order to better generalise their results to new datasets. Lack of generality was a particular bugbear of early recursive partitioning methods¹².

The two main contemporary approaches to recursive partitioning, consisting of multiple significance testing, and cross-validation, are examined, as are the

advantages and disadvantages of each approach, and of recursive partitioning in general.

1.1. Conventional statistical methods

1.1.1. Linear decision rules

For many years, psychiatric screening and diagnostic decision rules based upon instruments such as the Alcohol Use Disorders Identification Test (AUDIT)¹³, Beck Depression Inventory (BDI)¹⁴ and the General Health Questionnaire (GHQ)¹⁵ have been developed with the help of simple linear decision rules¹¹. The latter are based upon determination of whether or not the number of observed or reported symptoms exceeds a cutoff, or threshold score, representing probable caseness. Such methods are easy to perform and easy to interpret, but make the implicit assumption that the number of symptoms present is more important than the type of symptoms that are present. Such an assumption is not generally met in practice^{3,16}.

1.1.2. Parametric techniques

Parametric statistical procedures such as linear discriminant function analysis¹⁷⁻²⁰ and logistic regression^{18,21-24}, may be regarded as being more sophisticated versions of simple linear decision rules³ and have also had a long history of application to psychiatric screening and diagnosis^{11,17}. Such techniques allow the relationship between particular symptoms and/or potential risk factors and diagnoses to be readily ascertained, in the form of point estimates such as odds ratios, and also inferential statistics such as confidence

intervals and significance tests. As with all techniques, parametric methods have particular limitations, many of which are shared by simple linear decision rules.

Parametric procedures are generally used to find overall or global models. For example, logistic regression is concerned with modelling the relationship between an outcome variable, such as a diagnosis of major depression, and one or more predictor variables or risk factors. The primary objective is to identify an average set of conditions, such as values on the above variables, associated with a particular outcome ²⁵. Relationships are assumed to be linear, and additive (lacking statistical interaction), unless terms representing nonlinear effects, and/or statistical interaction between risk factors, have been specifically included in the regression model ²⁶. As defined in a standard text on logistic regression ²⁷, the presence of statistical interaction implies that ‘the association between the risk factor and the outcome differs, or depends in some way on the level of the covariate. That is, the covariate modifies the effect of the risk factor’ (p. 70). The existence of prior theory and/or the skill of the data analyst are relied upon in handling departures from such assumptions; which are often encountered in behavioural data ²⁵.

For example, there is evidence that the relationship between consumption of alcohol and diagnosis of major depression is not linear but ‘U’ or ‘J’ shaped. Depression appears to be less likely to be present in those persons who drink alcohol moderately, rather than those who drink it excessively, or not at all ²⁸. Similarly, the relationship between experiencing negative life events, such as job loss, and developing major depression, has been found to be greater for those

with an optimistic outlook on life than it is for those with a pessimistic outlook. Such a finding suggests a statistical interaction or interdependence between outlook and negative events in regard to risk for major depression²⁹. In other words, the relationship between one particular variable - negative life events - and the outcome variable - major depression - is a function of values of another variable - optimistic outlook.

Parametric techniques are also generally adversely affected by outlying observations, such as extremely low or extremely high scores. Although regression techniques that are robust against outliers are available³⁰, these methods may still make some assumptions, such as a lack of statistical interaction between predictor variables. Finally, parametric procedures generate models in the form of mathematical equations such as discriminant functions or regression equations, which may lack clear meaning to clinicians^{25,31}.

1.2. Machine Learning

The advent of electronic computers in the last five decades or so has seen the development of a field known as machine learning³²⁻³⁵. Machine learning techniques generally make fewer assumptions of the data than do conventional parametric statistical methods, and yet may provide more insight into the relationships contained within a dataset. Machine learning can initially and briefly be defined as '*The capability of a [computer] program to acquire or develop new knowledge or skills. The study of machine learning focuses on developing computation methods for discovering new knowledge from data*'³⁶ (p. 175). Machine learning is a subfield of Artificial Intelligence^{11,37}, which in turn may be

defined as 'the field concerned with developing techniques to allow computers to act in a manner that seems like an intelligent organism, such as a human would'³⁶ (p. 13).

Machine learning methods have been developed by applied statisticians, as well as cognitive psychologists, computer scientists and engineers. The goal of machine learning is to develop models, generally in the form of computer implementations such as decision trees or artificial neural networks. With a minimum of supervision and theoretical assumptions, such models learn relationships from data in order to aid in interpretation, and predict or classify new instances³².

There are many different varieties of machine learning techniques^{32,35}, including artificial neural networks^{38,39}, which attempt to model the workings of the brain. What is generally regarded as the most established, most successful and most widely used machine learning technique^{32,35,40}, is variously known as tree-building, recursive partitioning, classification and regression trees, or decision trees⁶.

1.3. Tree-building / recursive partitioning and subgroup analysis

As did other machine learning methods such as artificial neural networks⁴¹⁻⁴³, recursive partitioning methods originated in the late 1950's and early 1960's. Since the development of what is arguably the first such procedure in the 1950's by William Belson, an Australian psychologist based in London^{9,44,45}, recursive partitioning techniques have had a long history of aiding in psychiatric screening, diagnosis and subgroup analysis^{10,31}. By repeatedly or recursively dividing

datasets into homogenous subgroups, recursive partitioning methods combine aspects of regression and cluster analysis ^{46,47}.

Cluster analysis ⁴⁸⁻⁵⁰, also known as numerical taxonomy ⁵¹, may be defined as the placing of persons or objects into classes, clusters, or groups on the basis of their similarity. The number of groups as well as the fundamental attributes of the groups, including their size, are unknown. In its broadest sense, cluster analysis includes simple methods that assign each observation exclusively to one cluster, as well as more sophisticated techniques. The latter are able to assign observations probabilistically to more than one cluster. This may aid in the decision of how many clusters to retain ⁵².

Examples of simple cluster analytic techniques include K-means cluster analysis ⁵³, which assigns objects to clusters based upon their distance from each of K cluster means, with the value of K determined by the researcher. Examples of more sophisticated cluster analytic techniques are often referred to as latent class analysis or latent class cluster analysis ^{52,54-56}, grade of membership analysis ⁵⁷⁻⁵⁹ or finite mixture modelling ^{48,60-62}.

Originally, latent class procedures wholly assigned each observation to only one class ^{55,56}. Grade of membership methods, however, along with more recent latent class methods ⁵², derive the probability that an observation is a member of each class, so that each observation may be regarded as being a member of more than one class.

Classically, cluster analytic techniques usually do not involve observed outcome variables, such as the presence or absence of a diagnosis of major

depression. Rather, these approaches are employed to find groups of unknown structure. For example, latent class cluster and similar methods assume the existence of hidden, unobservable or latent classes such as varieties of mental illness, membership of which is defined by values of observed variables such as symptoms. Latent class and similar techniques have been employed in a variety of psychiatric applications, including the detection of subgroups defined by particular patterns, such as subtypes of depression that are homogenous in regard to various depressive symptoms such as loss of interest or pleasure, and feelings of worthlessness⁶³⁻⁶⁹.

Within the context of cluster analysis, finite mixture modelling methods assume that the observations come from a number of unknown populations, having different probability distributions. Such distributions might all be of the same type (e.g., all normally distributed), but have different parameters (e.g., differing mean symptom levels). The unknown populations or classes are 'mixed' in unknown proportions, and so the objective is to 'unmix' the observations into their true latent classes (assuming there is more than one) by estimating the mixing proportions, as well as the population parameters⁶¹. Latent class analysis may be regarded as a form of mixture modelling involving binary or categorical data, although the technique can also be used with dimensional observed variables, such as age. If all of the observed variables are dimensional, latent class analysis is often referred to as latent profile analysis^{56,65,70}.

The latent class model has been extended to a regression context. Also referred to as mixture regression analysis^{71,72}, latent class regression can be

employed for the analysis of observed outcome variables ^{73,74}, such as diagnosis of depression, in a similar manner to linear regression and logistic regression. The goal of latent class regression is to identify the underlying latent classes and to estimate the regression function for each class ⁷³. The clustering of observations into classes, as well as the estimation of the regression models for the prediction of the observed outcome variable within each class, is performed simultaneously ^{72,75}. Within a longitudinal context, this technique is known as latent class growth modelling or latent growth mixture modelling ^{73,76}. The existence of latent classes representing temporal patterns, such as the lifetime development of observed depression as measured by levels of observed symptoms ⁷⁷, is assumed. Unless specified to the contrary, the term latent class regression shall, in this thesis, refer to the use of latent class modeling techniques with observed outcome variables ⁷³.

Although it could also be described as combining aspects of regression and cluster analysis, latent class regression is concerned with the homogeneity of subgroups in regard to regression coefficients, rather than the observed outcome variables. Subgroup members may therefore not necessarily be homogenous on values of the latter ⁷². In contrast, recursive partitioning methods directly seek to find subgroups, defined by different combinations of predictors or risk factors, which are homogenous on observed outcome variables, although not necessarily homogenous on latent variables ⁸. Recursive partitioning methods can therefore be used to identify subgroups of observations sharing similar

outcomes, such as a low risk, or a high risk, of developing depression ⁷⁸, or who are most likely, or least likely, to seek treatment for mental illness ⁷⁹.

In addition, by graphically providing results in the form of a decision tree, recursive partitioning methods clearly show which specific predictor variables or risk factors are associated with each subgroup ⁴⁷.

1.3.1. Illustrative example of recursive partitioning

An example of a simple decision tree, constructed using the Classification and Regression Tree (CART) ^{80,81} technique, described in more detail below, is given in Figure 1. The example dataset uses information collected in the 1991 US General Social Survey, supplied with the Statistical Package for the Social Sciences (SPSS) computer program ⁸² and was chosen merely in order to illustrate recursive partitioning techniques. The outcome variable consists of self-rated happiness, coded as 'very happy', 'pretty happy' and 'not too happy'. The first two categories were together recoded for convenience into 'happy'.

The predictor variables in this example consist of sex, and occupational category - coded 'managerial and professional', 'technical', 'service', 'farming, forest and fishing', 'precision production, craft and repair' and 'operation, fabrication and general labor'. The sample consists of 1504 observations.

As can be seen in Figure 1, CART firstly split the dataset by occupational category, the predictor (out of the two employed) with the largest decrease in the impurity of the subgroups, as represented by the tree branches or nodes formed at each split. By default CART chooses splits based on minimising the Gini index

^{80,83,84}, a measure of node impurity. The Gini index is equal to zero when each node of a binary split contains only members of one class or outcome.

In the present example, occupational category has six categories, the frequencies of which are given in Table 1. CART combined these categories into two; this algorithm being restricted to binary splits throughout the tree-building process. CART can repeatedly split the same variable if necessary, using a sequence of binary splits.

For each binary split involving categorical predictors, CART performs a search for the best way of merging categories into two, with the goal of merging those categories with similar values on the outcome variable. In the example, service and manufacturing categories have been merged into one subgroup; and managerial, technical, farming and craft categories have been merged into another. The former subgroup exhibits a higher frequency of self-rated unhappiness (14.8%) than that observed for the total sample (10.6%), while the other subgroup, consisting mainly of white collar occupations, exhibits a slightly lower frequency (9.5%) of unhappiness. The service / manufacturing group is not able to be further split, whereas the other subgroup is able to be further split by sex, so that white collar females tended to report more unhappiness in this dataset than did white collar males.

Figure 1. Recursive partitioning / decision tree analysis of self-rated happiness, 1991 US General Social Survey

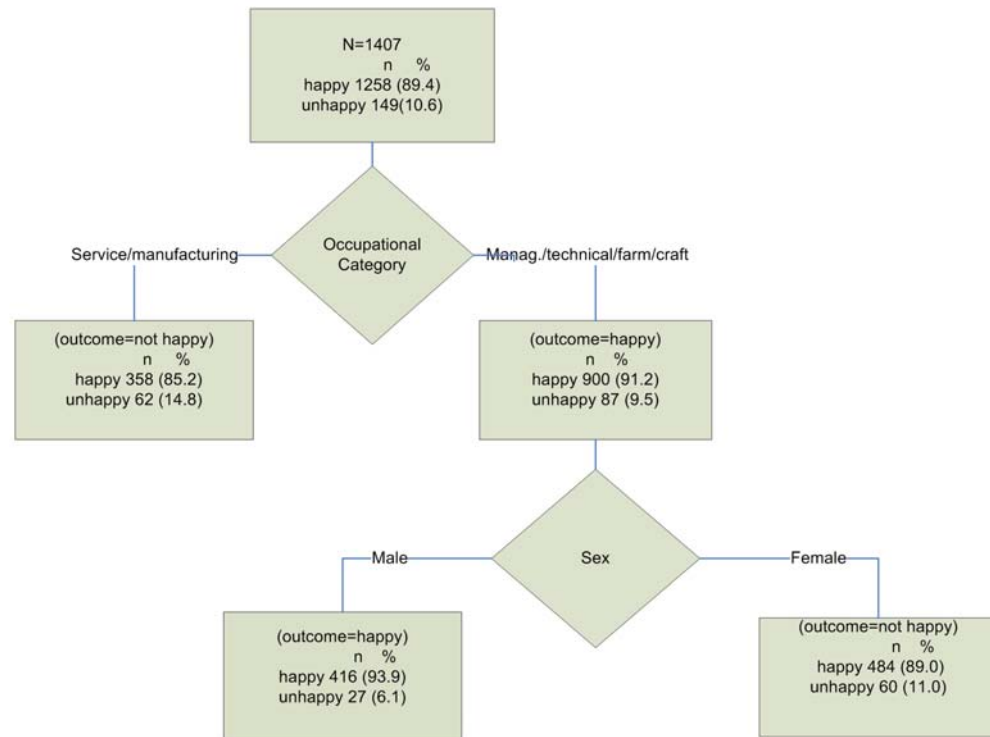


Table 1: Frequencies of Happy and Not Happy self-ratings by Occupational Category

Occupational Category	N	n Happy	%	n Not too Happy	%
Managerial/professional	337	307	91.1	30	8.9
Technical	452	408	90.3	44	9.7
Service	203	177	87.2	26	12.8
Farming/Forest/Fishing	36	35	97.2	1	2.8
Precision production, craft	162	150	92.6	12	7.4
Operation, fabrication and general labour	217	181	83.4	36	16.6

1.3.2. Recursive partitioning: advantages

As shall be seen, various early recursive partitioning algorithms were designed to be used in conjunction with parametric statistical techniques. Although they also have their disadvantages, which will be returned to in far more detail below, recursive partitioning generally makes fewer assumptions of both the dataset, and the person analysing it, than do conventional statistical techniques. Recursive partitioning may therefore be regarded as a nonparametric statistical technique ⁶.

Recursive partitioning only examines relationships between the outcome variable, and one or more predictor variables, within that subgroup of observations that is defined at each stage of the tree growing process. The method can therefore be thought of as a local, rather than a global, one. As described earlier, recursive partitioning attempts to detect subgroups of individuals typified by various combinations of symptoms and or risk factors, such subgroups being homogenous on the observed outcome variable. The risk factors defining each subgroup are clearly shown in the resulting tree structure.

Recursive partitioning is able to uncover, and help researchers and clinicians to interpret, various types of statistical interaction between predictor variables and the outcome. Indeed, one of the first recursive partitioning computer programs was originally known as the Automatic Interaction Detector (AID) ⁸⁵.

In terms of the decision tree example shown in Figure 1, there appears to be a statistical interaction between occupational category and sex. Those respondents in the service and manufacturing occupational categories tended to report not being happy, regardless of what sex they were. In the case of the predominantly white collar categories however, females reported unhappiness more often than did males. Thus, intervention strategies to decrease unhappiness could usefully be directed at members of the above two subgroups.

Recursive partitioning does not assume linear relationships ²⁵, and can be used to identify statistical outliers ³⁰ through the generation of subgroups having small numbers of observations with atypical scores ^{85,86}. Provided that the final tree is not too large (has too many levels or contains too many end-groups, 'leaves' or terminal nodes to be readily interpreted), the graphic output of a recursive partitioning analysis is readily understood by researchers and clinicians ^{25,87} and conveniently shows how the dependent variable is related to the predictor variables. The resulting tree can also be used as a 'pen and paper' method of making predictions about future cases, or form the basis of a computerised classification system ⁸⁸.

1.3.3. Recursive partitioning: historical background

The next section provides a comprehensive outline of tree-building or recursive partitioning techniques. These techniques can be applied to the interpretation and prediction of ordinal (e.g., 'never', 'sometimes', 'always') or dimensional outcomes (e.g., scores on a personality test), as well as nominal or categorical (e.g., type of psychiatric diagnosis), including binomial or binary (e.g., presence or absence of major depression) outcomes^{89,90}. Only categorical outcomes will be examined here however, as they are the most relevant in terms of psychiatric screening and diagnosis.

A brief historical review of recursive partitioning techniques will be presented in order to put the underlying methodologies of contemporary procedures such as CART into context. Existing literature reviews in this area⁸⁹⁻⁹⁵ tend to concentrate on methods developed in the United States. As discussed earlier, the first publication of a recursive partitioning algorithm involving a separate outcome variable appeared in the UK⁹. Other techniques have been developed in several countries including Australia, Canada, France, Korea, Singapore, South Africa, Taiwan, the UK and the US.

What appears to be the very first recursive partitioning technique, known as the Stable Correlate method^{9,44}, was developed in order to statistically match groups on various criteria such as demographics, with the resulting weighted samples then analysed by conventional parametric procedures such as linear regression. At around the same time, broadly similar techniques for the clustering of binary attributes, and which did not involve any outcome variables, had been developed in Australia⁹⁶ and in the UK^{97,98}. The Stable Correlate technique

employed binary splits, based upon a simple distance measure related to the conventional chi-square statistic ^{9,99}, and was mainly applied in marketing and social survey applications ^{44,100-103}.

1.3.3.1. The Concept Learning System (CLS)

In what was largely a separate thread of research to that being conducted in the area of survey statistics ^{88,104}, cognitive psychologists and computer scientists in the US (and later in Australia) developed automated decision trees in the late 1950's in an attempt to understand human information processing ^{10,105}. Working in the early 1960's at the University of California, and later at the University of Sydney, Earl Hunt, Janet Marin and Philip Stone developed varieties of recursive partitioning techniques, together known as the Concept Learning System (CLS) in order to model how humans learn concepts. The latter were defined as the 'acquisition, or utilization, or both, of a common identifying response to dissimilar stimuli' ¹⁰⁵, (p. 2). An early application of the CLS to psychiatric data ¹⁰ was distinguishing genuine from simulated suicide notes ¹⁰⁶.

1.3.3.2. Automatic Interaction Detection (AID)

The early development of recursive partitioning procedures is most closely associated with survey researchers James Morgan's and John Sonquist's Automatic Interaction Detection (AID) computer program. AID was developed in the US in the early 1960's, inspired by the Stable Correlate technique ^{85,107}. AID was originally intended to be used in conjunction with a statistical method known as multiple classification analysis (MCA) ^{108,109}, an early multivariate procedure for analysing categorical predictor variables. MCA makes the assumption that

there are no statistical interactions between the predictor variables. It was originally intended that derived variables representing possible interactions uncovered by AID could then be analysed using MCA ¹¹⁰.

Although the earlier Stable Correlate technique was primarily developed for use with binary outcomes, AID was intended for dimensional outcomes, in what can be likened to an extension of one way analysis of variance ^{107,111}. Predictors were required to be either categorical or ordinal. Dimensional predictors were first automatically divided into equally sized categories or n-tiles by the AID computer program. As did the Stable Correlate technique and many of the procedures that followed it, AID generated only binary trees. For each predictor with more than two categories, the latter were merged into the best binary split at each stage in the tree. This merging was done to increase understanding of the resulting models, as well as to maximise the number of persons in each resulting end-group or terminal node of the tree ¹¹².

AID could also be applied to binary outcomes, although the procedure had a tendency to split the sample into outlying subgroups of small size ⁸⁵. Nevertheless, the original AID technique was employed by various psychiatric researchers to assist in the development of screening instruments for outcomes such as alcohol misuse ¹¹³⁻¹¹⁵ and stress reactions to trauma ^{116,117}. Perhaps the most famous and enduring legacy of the original AID, however, is that a version of this procedure was used in the US to create diagnosis-related-groups (DRG's), representing a taxonomy based upon lengths of hospital stay for various

illnesses¹¹⁸⁻¹²⁰. Early Australian research into ascertaining the average length of stay for various psychiatric illnesses also employed AID¹²¹.

1.3.3.3. THAID: AID for categorical outcomes

In order to apply the underlying precepts of AID to categorical outcomes, the Theta Automatic Interaction Detection (THAID) procedure was later developed in the US by the originators of AID^{122,123}. THAID was intended to be used in conjunction with a categorical analogue of multiple classification analysis; multivariate nominal scale analysis¹²⁴ (MNA). Theta referred to a measure based on prediction to the modal (most commonly occurring) outcome category, related to Goodman and Kruskal's lambda statistic¹²⁵. Theta was of limited utility, however, when dealing with largely frequent or infrequent events. In such situations the modal outcome values for a given table will all tend to be the same. In the case of the example above, the most frequent, or modal, outcome category for each occupational group is 'happy'.

THAID could also choose splits based on the delta statistic or absolute 'city-block' distance measure, related to the chi-square statistic¹²⁶. Use of the delta statistic was intended to generate end-groups or terminal nodes that had maximally different values on the outcome variable, even if all such groups had the same modal outcome category.

Unlike AID, THAID does not appear to have been widely applied to psychiatric, or indeed medical, data¹²⁷. The two techniques were eventually incorporated into the Search procedure, which has recently been described by one of the original AID developers⁸⁹. Search has dropped the categorical theta

and delta criteria in favour of the far better known chi-square statistic, which had been employed in other early computer tree-building programs ^{10,128}.

Other pioneering recursive partitioning methods, similar in rationale to AID/THAID, were developed in France ^{128,129} and the UK ^{130,131}, as well as in the US ¹³²⁻¹³⁶.

1.3.3.4. Inductive Data Exploration and Analysis (IDEA)

Although it was never made generally available, one of the most powerful early recursive partitioning procedures was the Inductive Data Exploration and Analysis (IDEA) method, developed in the US in the mid-1960's by Laurence Press ^{137,138}. IDEA was unique in that it was able to merge categories of predictor variables without being limited to binary splits.

Interactive tree-building procedures, in which the analyst is able to directly select the variable on which to split the sample at a particular point in the tree, had been described earlier ^{10,139}. The IDEA algorithm was, however, unique in the context of the 1960's in that it allowed use of display screens and a hand-held computer mouse ¹⁴⁰. Further details of many of the pioneering recursive procedures are provided in an early review ⁹⁹.

1.3.4. The need for tree-pruning and stopping rules

'The object of pruning Fruit Trees is for the proper regulation of the branches, so as to encourage the production of blossom, and the maturing of heavy crops of good fruit. Many trees will crop heavily enough without much attention, but the quality is often very poor. If the pruning be too severe the tree will grow to wood instead of fruit. On the other hand, if the branches are

left too thick they overshadow those beneath them, excluding the light and air and encouraging a great growth of leaves, but very little fruit'. ¹⁴¹ p. 217

1.3.4.1. Overfitting by AID

Early tree-building methods such as AID relied on forward pruning or 'stopping rules' to terminate the tree-growing process. This is a similar strategy to that used by forward entry stepwise regression ^{142,143}. In the latter case, the computer program adds a variable to a model at each step, until there are no more variables that meet one or more criteria, such as statistical significance. The major stopping rule or criterion used by AID to terminate the tree growing process was based upon the reduction of error variance by a minimum value. Guidelines for such minimum values were suggested based on the results of simulation studies ¹¹⁰. In contrast to the contemporary recursive partitioning techniques discussed below, AID model-building was an extremely ad hoc process with few built-in safeguards.

The original AID software documentation ⁸⁵ clearly stated that AID trees needed to be validated on datasets not used to create the models in the first place, a practice also suggested for use with other early recursive partitioning procedures ^{131,135}. As a further safeguard against overfitting models to data, AID was intended to be applied to large datasets, containing at least 1,000 observations ¹¹². Unfortunately, however, such advice was not always followed by users of the technique.

Particularly when they were employed by less experienced analysts ¹⁴⁴, AID and THAID could, and frequently did, severely overfit models to data. In other words, the techniques would generate models that described one set of

data very well, but did not generalise well to new data. This understandably led to a great deal of criticism ^{12,144}. For example, in a study concerned with the application of AID to data involving a dimensional outcome, AID accounted for over 30% of the variance in the data used to create the tree. However, less than 3% of the variance in a fresh set of data was accounted for. Similarly, an AID tree accounted for over 30% of the variance in random data ¹².

In practice therefore, AID and THAID offered little protection against overfitting models to data, in that it was difficult to know when to halt the tree-growing process. Simple models often generalise to new data better than do more complex models, as the former are less likely to be merely fitting noise in a particular sample of data used to train the model in the first place. In the words of the original AID developers, 'One purchases completeness with the coin of instability' ⁸⁵ (p. 135).

1.3.4.2. Scientific parsimony

Simple models are also in keeping with the scientific concept of parsimony, as typified in the application of 'Occam's Razor'. The fourteenth century philosopher William of Occam (or Ockham) proposed something to the effect of 'pluralitas non est ponenda sine necessitate' ('plurality should not be posited without necessity'). In other words, the number of entities needed to explain something should not be unnecessarily increased ^{145,146}.

Occam's Razor does not necessarily mean that 'less is better'. It merely proposes that complexity should not be increased unless this is necessary, for example, to increase the generality of a tree by increasing its size. As is

commonly, but perhaps mistakenly ¹⁴⁷ believed to have been first suggested by Albert Einstein, 'everything should be made as simple as possible, but not simpler'. Although very simple decision trees have been found to generalise very well to unseen data ¹⁴⁸, this is not always the case ¹⁴⁹.

It is therefore prudent to aim for a trade-off between model complexity and model performance in the hope of developing models that will exhibit satisfactory performance on unseen data, as well as on the original data. Unfortunately however, AID and THAID and the other early recursive partitioning programs offered little help in achieving such a trade-off. Indeed it was to be almost two decades from the publication of the original AID computer program manual in 1964 before such methods were to become comparatively readily available ^{80,150}.

1.3.5. Statistical significance testing

It is obviously crucial to employ a recursive partitioning procedure that protects against the overfitting of data, particularly when attempting to identify possible subgroups. Even the analysis of subgroups that have been specified in advance by the researcher can be problematic ¹⁵¹, let alone those subgroups that are identified ad hoc by a computer program. Two main strands of research into tree-pruning methods that are not prone to overfitting can be identified ¹⁵² – those based upon statistical significance testing, and those based upon cross-validation.

1.3.5.1. Early recursive partitioning techniques using statistical significance testing

Several pioneering tree-building techniques ^{96,103,130} incorporated statistical significance testing of the chi-square statistic used to determine tree-

splitting. Goodall's procedure ⁹⁶ also included a Bonferroni multiple comparison technique ^{111,153}, which involved testing at a level of statistical significance equal to the nominal level (e.g., 0.05) divided by the total number of binary variables included in the analysis. This was done in an attempt to reduce the possibility of false alarms or type one errors - results appearing to be statistically significant due simply to the large numbers of comparisons performed, and which may not be replicated on other data.

The Bonferroni procedure can however be highly conservative, especially when a large number of statistical comparisons is performed ^{154,155}. The inclusion of significance testing had originally been considered by the developers of AID ⁸⁵ but was rejected due to the large numbers of possible binary splits for predictors with many categories. The current version of Search includes statistical significance testing, but does not adjust for the number of splits performed ⁸⁹.

The IDEA method ^{137,138}, described earlier, attempted to reduce the number of categories for a given predictor by combining those categories that had similar values on the outcome variable. Such reduction would reduce the associated degrees of freedom, and so increase the statistical significance of a given chi-square statistic, for the merged predictor. However, IDEA did not take into account the fact that chi-square was being calculated using empirically determined categories that had been merged on the basis of an intensive search procedure, rather than using preexisting categories. The application of standard tests of significance, involving statistics such as chi-square or the odds ratio, to

empirically determined comparisons can give highly misleading (generally too liberal, or too many false positives) results ^{156,157}.

1.3.5.2. CHi-square Automatic Interaction Detection (CHAID)

Early attempts at preventing overfitting, by adjusting for the number of statistical significance tests performed during the tree-splitting process, were described by various researchers in the 1970's ^{158,159}. The most comprehensive body of research in this area was that undertaken in South Africa by Gordon Kass ¹⁶⁰. Maintaining the original naming conventions, Kass developed the CHi-square Automatic Interaction Detection (CHAID) algorithm for the analysis of categorical data ¹⁵⁰.

As with the earlier IDEA procedure ^{137,138}, CHAID is able to split k categories of a predictor into a smaller number of categories without being limited to binary splits. Although binary splits can be performed repeatedly for a particular predictor, they are not guaranteed to find the optimal splitting of categories, if the optimal split involves more than two categories ¹⁶¹⁻¹⁶³. Unlike IDEA, CHAID adjusts the statistical significance of the chi-square statistic representing the association between the outcome and the merged categories of the predictor variable. Merging involves combining those categories that are not statistically significant from one another in regard to values on the outcome variable, at a given (unadjusted) level of significance, generally 0.05.

CHAID adjusts for the number of splits using the Bonferroni multiple comparison technique outlined earlier. The adjustment factor is based upon the maximum number of ways in which the k original categories of the predictor can

be partitioned into c merged categories ¹⁶⁴. Tree-growing stops when the number of observations in a subgroup becomes smaller than a user-specified minimum, or if the adjusted statistical significance exceeds a user-specified maximum. It was later found that the Bonferroni statistical adjustment used by CHAID is highly conservative in the case of predictor variables with ten or more categories ¹⁶⁵. A less conservative modified Bonferroni method is included in an extended version of CHAID known as KnowledgeSEEKER ¹⁶⁵, developed in Canada by Barry de Ville and others.

As an alternative to Bonferroni adjustment, it has recently been proposed ¹⁶³ that the statistical significance of CHAID splits be assessed by the use of Monte Carlo permutation tests ¹⁶⁶⁻¹⁶⁸. Every possible ordering or permutation ^{169,170}, or several thousand random samples thereof ^{171,172}, for assigning c (after merging) category labels to individuals is generated. A chi-square statistic for each contingency table is obtained. The probability of obtaining a chi-square as large as the one obtained for the original data, simply by chance, is then given by the number of chi-square statistics for the permuted data that are equal to, or larger than, the chi-square obtained for the actual data. The permutation procedure does not require Bonferroni adjustment, but nevertheless reduces the chance of incorrectly concluding that the merged categories are different from each other ^{156,173}. The permutation testing of decision trees is an active and fast-growing area of research ¹⁷⁴⁻¹⁷⁶.

1.3.5.3. Interactive variable selection in CHAID

Allowing the researcher to interactively select which variables to split the dataset on, perhaps from a list of those meeting certain criteria, can in theory be used with any tree-building technique, with varying degrees of success. However, the forward pruning or stopping rules used by CHAID readily lend themselves to such a process. Inspired by the earlier IDEA procedure^{137,138} KnowledgeSEEKER^{165,177} was the first interactive tree-building program for microcomputers that, at each stage of the tree-growing procedure, enabled the analyst to interactively choose the more theoretically interesting variables out of, in this case, those found to be statistically significant (e.g.,^{88,178}). Other interactive tree-building procedures are also now available²⁵.

1.3.5.4. Applications of CHAID to psychiatric data

The first application of the KnowledgeSEEKER¹⁶⁵ extension of CHAID to psychiatric screening and diagnosis or subgroup identification was described in the early 1990's by McKenzie et al., in the area of schizophrenia diagnosis^{152,179}. Other applications of CHAID and its extensions have been described in the specific areas of alcohol and other substance abuse¹⁸⁰⁻¹⁸³, bereavement¹⁸⁴, life course of positive and negative affect²⁵; pathological gambling¹⁸⁵, psychiatric service utilisation¹⁸⁶⁻¹⁸⁸, psychogeriatrics¹⁸⁹⁻¹⁹¹, psychological distress¹⁹², and vulnerability to mental health problems¹⁹³.

1.3.5.5. Controversies of statistical significance testing

CHAID is one of the few contemporary tree-building algorithms that is able to merge categories of categorical predictors without being limited to binary splits. The technique is conceptually simple to understand, in that it only involves

merging categories that are not significantly different from one another, with regard to values on the outcome variable. A Bonferroni adjustment to the statistical significance of the resulting chi-square statistic is then performed. Chi-square and statistical significance testing, including Bonferroni adjustments, are well known to most psychiatric researchers, although CHAID's reliance on statistical significance may not be desirable.

Aware of the criticisms of the earlier ad hoc AID procedures, the developers of CHAID attempted to provide the new algorithm with a sound theoretical footing, based upon statistical significance testing. The latter has, however, long been criticised ¹⁹⁴⁻¹⁹⁶ for encouraging a simplistic view of conducting research, with an emphasis on the statistical significance (or otherwise) of the results, rather than on the magnitude of the results. There was a resurgence of criticisms of significance testing in the mid 1990's ^{197,198}, with the American Psychological Association (APA) even considering banning the use of significance testing in its journals. The APA eventually developed guidelines encouraging the use of effect sizes and confidence intervals, in conjunction with, or instead of, tests of statistical significance ¹⁹⁹⁻²⁰¹.

Statistical significance testing, if used at all, is arguably best used in conjunction with measures of the practical significance or magnitude of the results ¹⁹⁹. However, CHAID relies solely on statistical significance in choosing variables for tree-growing. Furthermore, the implications of group size on the merging process are rarely discussed in the CHAID literature. If sample sizes were very large, groups that had only trivial differences on the outcome variable

might not be merged, because these differences could still be deemed to be statistically significant.

The use of adjusted significance levels to try and take into account the many statistical comparisons performed by CHAID, is itself controversial. As mentioned earlier, the Bonferroni adjustment used by CHAID is highly conservative when employed with large numbers of categories ¹⁶⁵. Some researchers, particularly those within epidemiology, have suggested that statistical significance testing only be used to test clearly defined and biologically plausible hypotheses, without adjusting for multiple comparisons ^{155,202}. This is not to suggest that procedures such as CHAID abandon adjusted significance testing in favour of unadjusted significance testing. What might be of more relevance to the user of a recursive partitioning procedure, however, would be some actual indication of how well the resulting tree will generalise to other sets of data.

1.3.6. Cross-validation: the Classification and Regression Tree (CART) algorithm

As mentioned earlier, the developers of early recursive partitioning algorithms such as AID and THAID advised dividing datasets into two, consisting of a learning or training subsample and a 'hold-out' or validation subsample. Trees were to be generated on the first subsample and then applied to the second in order to obtain a more accurate estimate of classification performance. The latter would tend to be biased upwards, or optimistically, if estimated from the subsample used to create the tree in the first place ^{85,131,135}. Even if this advice was actually followed by researchers, it involved using datasets large

enough to be split into two, while still ensuring that the learning or training sample was large enough to be able to construct a representative tree model. What was obviously required was a method of using as much of the original data as possible to create the original tree model, and yet still be able to test the generality of the model on fresh data. A statistical technique known as cross-validation^{203,204} seemed to provide a means of doing this.

An early form of cross-validation, known as 'leave-out-one', and originally applied to parametric procedures such as linear discriminant analysis²⁰⁵ and linear regression^{206,207} consisted of leaving out each of the n observations of a dataset in turn, and then estimating a parameter such as classification accuracy for the remaining $n-1$ observations. Thus, the technique involved using all of the available data.

Cross-validation was eventually employed with a recursive partitioning procedure developed in the UK by a team of researchers²⁰⁸ including Mervyn Stone, one of the originators of cross-validation^{209,210}. This procedure used cross-validation to assess the performance of the variables chosen at each point in the tree however, rather than to determine the optimal size of the tree.

A comprehensive approach to recursive partitioning, using 'built-in' cross-validation to ascertain how large a particular tree should be, based upon an estimate of how well it would generalise to similar datasets, was undertaken in the US in the late 1970's and early 1980's by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone^{80,211}. Such an approach led to the development of the Classification and Regression Trees (CART) algorithm.

CART is one of the best known recursive partitioning techniques and appears to have played a major part in legitimising such methods amongst the wider statistical community^{6,212}.

1.3.6.1. Backward pruning

CART retains the binary splits used by AID/THAID, but unlike those early techniques and the contemporary CHAID, CART appears to be the first tree-building technique to employ backward pruning. The latter is similar in concept to the backward elimination strategy often used with logistic and linear regression^{27,143}. CART deliberately develops an oversized tree which is then pruned, attempting a trade-off between cost - the performance of a given tree or sub-tree, and complexity - the number of nodes in a given tree or sub-tree. Such pruning helps to protect against the possibility that it may be necessary to make an apparently uninformative split, in order to uncover a highly informative split further down the resulting tree.

CART firstly grows a tree, using the entire dataset, until the subgroups become too small for further splitting. This final, 'fully expanded' tree, which is generally very large, is then pruned back, using 'cost-complexity' or 'weakest-link' pruning^{80,213}. The weakest link is defined as that subtree whose deletion leads to the smallest change in the apparent classification error rate at that point in the tree. Such change takes into account the error rate (the cost) and the number of non-terminal nodes in the subtree (the complexity). In this fashion, different sized trees or subtrees are constructed, ranging from the unpruned, fully expanded tree, to a tree consisting of only two nodes (the first split chosen by CART), and

one node (the null tree, equivalent to assigning all unknown observations to the modal, or most frequent, outcome). Each successively smaller tree is a subtree of a previous tree, resulting from successive pruning.

CART next assesses the generality of each sized tree through cross-validation. For a large dataset, by default one that is greater than 3000 observations, CART simply randomly divides it into a learning subsample and a validation subsample, similar to what was suggested for early tree-building algorithms^{85,131}. For smaller datasets, by default those less than 3000 observations, CART uses v-fold cross-validation, in which datasets are randomly divided into v equally-sized subsets. By default v is equal to 10, this value being empirically determined by the developers of CART to give good overall performance⁸⁰.

For each of v cross-validations, a fully expanded classification tree is generated on v-1 subsets combined, and then tested on the vth or 'hold-out' subset. For larger samples a full expanded tree is grown using the learning subsample and then tested on the validation subsample. For each subsample, pruning of each fully expanded tree is continued until there is a sequence of subtrees of approximately equal size to the sequence of subtrees grown on the complete sample. The error rate for each sized tree is then averaged across the application of each sized tree to each validation dataset, and a standard error obtained.

Each different sized tree will therefore end up with an associated average error and standard error, based upon cross-validation. Recent implementations

of CART choose by default that sized tree with the lowest error, as averaged across the validation subsamples. The original CART monograph⁸⁰, as well as subsequent discussions²¹³, suggest however that the smallest sized tree with an error rate within one standard error of the tree with the smallest error rate be chosen. Such a rule is in keeping with the principle of Occam's Razor, defined above, that simpler explanations are often to be preferred over more complex ones. The one standard error rule also takes into account the variability that is inherent even in cross-validated estimates of classification error.

The developers of CART found that pruning based upon cross-validation gave better results on validation datasets than pruning based upon the related bootstrap technique²¹⁴⁻²¹⁶ which involves the repeated resampling, with replacement, from a particular dataset. Within a specific bootstrap sample therefore, a particular observation may appear more than once, or not at all. Variants of CART that employ new developments in bootstrap techniques^{217,218} have been developed^{219,220}, but have not yet been widely tested or used.

1.3.6.2. Splitting criteria

CART generally chooses splits based on minimising the Gini index or coefficient^{80,83,84}, which the algorithm employs as a measure of node purity. In tree-building applications the Gini index will be equal to zero when each node is 'pure', or only contains members of one class. Ideally, at least the modal or most numerous outcome category would be classified correctly (be represented by pure or homogenous nodes), but otherwise splits leading to at least one pure node may be chosen. Such a strategy may lead to what the developers of CART

termed an 'end cut preference', the choosing of splits leading to subgroups that are small but consist mainly of observations with a particular outcome. Other indices such as the information or entropy statistic²²¹⁻²²³ and the twoing criterion⁸⁰, are also available.

The information statistic is conceptually similar, and mathematically identical, to the likelihood ratio statistic²²⁴⁻²²⁶, except that the former generally uses logarithms of base two. Splitting based on minimising the information statistic will attempt to achieve homogeneity of outcome in as many tree nodes as possible⁸⁰. Twoing, on the other hand, attempts to group both the outcome and predictor categories into two new 'super classes'. At each stage of the tree-growing process, the tree will be split as if the two current super classes are the only classes of interest. The twoing criteria is equal to the Gini index in the case of binary outcomes, and is related to the delta statistic employed by THAID^{122,123}. The information statistic had earlier been employed in several tree-building procedures^{98,131,227}.

Use of the Gini index in recursive partitioning gives very good performance overall, while the developers of CART⁸⁰ point out that tree-pruning criteria should be considered as being more important than splitting criteria, which has also been shown empirically^{228,229}. The latter study found that various splitting criteria performed about equally well on over 50 real datasets.

CART performs an exhaustive search for the optimum binary split for categorical predictors with up to 15 categories. Recent versions of CART use proprietary techniques for merging larger numbers of categories, when outcomes

are also categorical. In the case of binary outcomes, a heuristic method based upon sorting predictor categories, similar to that used by AID^{230,231} is employed. For dimensional predictors, CART chooses the optimum cut-point of the form 'assign to the left node if score $\leq X$ ', with nonlinear relationships of ten being represented by multiple binary splits of a particular risk factor.

1.3.6.3. Interactive variable selection in CART

Recent versions of CART allow the researcher to specify which predictor variables are to be used to split the dataset at the root node (the top or start of the tree) and the first two daughter nodes or subgroups⁸¹, but this must be done before running the program. There is currently no facility in CART for interactively selecting variables at each step of the tree-building.

1.3.6.4. Choice of prior probabilities and misclassification costs

By default, CART assumes that the outcome categories have equal probability, and that both false negatives (incorrectly classifying a case as a non-case) and false positives (incorrectly classifying a non-case as a case) are equally to be avoided. Probabilities can optionally be based upon the frequency of each outcome however, while the costs of different types of error can be specified. Thus, when developing a screening test, the researcher could direct the CART algorithm to give higher weight to false negatives than to false positives.

1.3.6.5. Applications of CART to psychiatric data

One of the first applications of CART (and CHAID) to psychiatric data, in the area of schizophrenia diagnosis, was described by McKenzie et al. in 1993

¹⁵². Other psychiatric applications of CART and its variants ²⁵ (CART is a registered trademark, and strictly, refers only to the computer programs that are direct descendants of the early CART software outlined in the original monograph ^{80,81}) have been described in the specific areas of alcohol and other substance abuse ²³²⁻²³⁴, fatigue ²³⁵, major depression ^{78,236,237}, panic disorder ^{238,239}, neuropsychiatry ²⁴⁰, psychiatric service utilisation ²⁴¹⁻²⁴³, psychogeriatrics ^{78,244,245} schizophrenia ^{246,247}, suicide ²⁴⁸ and treatment seeking by PTSD sufferers ⁷⁹.

1.3.6.6. Other backward pruning recursive partitioning procedures

There are a variety of other recursive partitioning procedures based upon backward pruning and cross-validation ^{92,95} which seek to extend CART. An alternative method of backward pruning, which does not involve cross-validation, is employed by the C4.5 algorithm (and its commercial extension C5.0 (www.rulequest.com, accessed 30 November 2008)). C4.5 was developed in Australia by Ross Quinlan ^{249,250}, as an extension of his earlier Interactive Dichotomizer 3 (ID3) ^{227,251}. The latter in turn grew out of the Concept Learning System (CLS) ¹⁰ described earlier.

C4.5 chooses variables based on information gain ^{249,252}, as measured by the information statistic described above. Trees are grown until they run out of cases and then pruned back if this leads to a decrease in a 'pessimistic' estimate of the true error rate. This estimate is based upon the upper limit of a 50% confidence interval around the classification error rate, estimated from the learning or training data. Although this heuristic appears to work well in practice

^{249,253}, it does not have the formal statistical underpinnings of the approach taken by CART, or indeed CHAID.

Other backward pruning procedures, based upon information theory ²²², which attempt to balance model complexity against model performance and so quantify Occam's Razor ^{254,255}, have also been described. Such techniques include those developed by Quinlan ²⁵⁶, and Christopher Wallace and his colleagues, in Australia ²⁵⁷ and Anthony Ciampi and his colleagues in Canada ^{258,259}. The latter technique employs the Akaike Information Criterion (AIC) ²⁶⁰, which is often employed in applications such as regression model selection ²⁶¹ analysis. The Quinlan and Wallace recursive partitioning techniques employ somewhat more complex measures that take into the precision of the model estimates as well as sample size ²⁶²⁻²⁶⁵.

All of the recursive partitioning methods described in this section have so far been only rarely applied to psychiatric data ^{152,266-268}.

1.3.6.7. Advantages of cross-validation

CART protects against overfitting without relying on the statistical significance testing employed by CHAID, or the somewhat statistically unorthodox backward pruning heuristics employed by C4.5. In addition, CART provides an estimate, based upon cross-validation, of how well a tree will perform on future data.

1.3.6.8. Disadvantages of cross-validation

Cross-validation is more computer resource intensive than other pruning procedures such as the significance testing approach used by CHAID, or the

pessimistic pruning approach used by C4.5. Cross-validation can also be more complex to understand than the latter approaches⁴⁶. Although not directly concerned with CART, recent evidence suggests that the repeated subsampling from a given dataset, as used by cross-validation and related subsampling or resampling techniques such as the bootstrap²¹⁵ outlined earlier, may give misleading results on small samples. Testing the resulting model on a separate or external sample is therefore strongly suggested^{269,270}.

1.3.7. Disadvantages of recursive partitioning in general

As described above, recursive partitioning techniques can help identify subgroups, defined by particular combinations of predictor variables or risk factors, that are homogenous on outcome variables, as well as help uncover nonlinear relationships, and statistical interactions. Recursive partitioning provides decision trees that are readily interpretable if not too large. If these techniques are to be widely used however, their disadvantages also need to be understood. Apart from overfitting data if tree models are not pruned correctly, recursive partitioning procedures suffer from several other major problems²⁶¹, such as bias in variable selection, sharp decision boundaries and the lack of adjustment for potential confounders, and the use of only one predictor at each stage of the tree-growing process. Recent developments in recursive partitioning methodology that attempt to overcome these problems shall now be examined, and related to the approaches described above.

1.3.7.1. Variable selection bias

Variable selection bias refers to the tendency of tree-building algorithms to select predictor variables with large numbers of categories. It has been empirically observed^{85,126,175,252,271,272} that binary splitting procedures such as AID and CART, as well as procedures such as CLS¹⁰ that do not merge categories, tend to exhibit a bias in terms of choosing predictors with large numbers of categories over those with small numbers of categories. This is because there are more ways in which variables with many categories can be split^{85,272}, and also because impurity indices are lower for large numbers of categories²⁵². For example, a variable representing unique case identification or ID, in which the number of categories is equal to N, the number of observations, would have zero impurity used to split the dataset, as each branch would only have one category.

A possible solution to the above bias, proposed by various authors^{10,85,252}, is to replace categories with binary variables, one variable per predictor category. However, such a solution increases the number of predictors to be analysed. Alternatively, algorithms such as CHAID, that employ adjusted multiple significance testing, incorporate a more severe adjustment for predictor variables with many categories. The original CHAID algorithm may exhibit a bias in the other direction however, by being less likely to choose predictors with many categories. This is due to the overly conservative nature of the standard Bonferroni adjustment¹⁶⁵. The form of Bonferroni adjustment employed in the KnowledgeSEEKER procedure¹⁶⁵ does not appear to exhibit such a bias.

There are several recursive partitioning algorithms that are based upon cross-validation, but which attempt to reduce variable selection bias. Such algorithms include Quick, Unbiased, Efficient Statistical Tree (QUEST), developed by Wei-Yin Loh in the US and Yu-Shan Shih in Taiwan ²⁷², Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE), co-developed by Loh, and Hjunjoong Kim in Korea ^{273,274}, and Logistic Regression Trees with Unbiased Selection (LOTUS), co-developed by Loh, and Kim-Yee Chan in Singapore ^{275,276}.

The above algorithms perform variable selection separately from split selection. The best predictor overall is firstly chosen, based on chi-square for categorical predictors, and one-way analysis of variance for dimensional predictors. Cluster analytic techniques are then used to identify the best binary (QUEST, LOTUS) or multi-way (one branch for each category of the outcome variable, CRUISE) split for the variable chosen in the above step.

In regard to other contemporary recursive partitioning techniques, C4.5 attempts to avoid selection bias by selecting variables for splitting based upon the increase or gain in information ^{249,252}, as described above, while other techniques take into account the number of categories of a particular predictor when attempting to balance complexity against performance ^{257,265}.

Although the potential for variable selection bias must be kept in mind, in practice most of the predictor variables used in psychiatric research do not have large numbers of categories.

1.3.7.2. Splitting on only one predictor

Most recursive partitioning algorithms are limited to splitting datasets on one variable at a time. It may not be possible to find the best split while controlling for the possible effects of other variables, as described below, while some datasets may be better described by splits involving linear combinations of predictor variables such as

‘branch left if $-0.148 * \text{variable_A} - 0.989 * \text{variable_B} < 0$, else branch right’

Linear combination splits can be difficult to understand, and take longer to generate, compared with univariate splits and also conventional regression equations. Although linear combination splits may aid in the uncovering of linear structure, this type of split may also have a higher complexity that is not accompanied by a commensurate increase in performance. Indeed, linear combinations found for a particular dataset may not generalise well to other datasets⁸⁰.

An alternative to linear combination splits that is applicable to binary predictors, is the use of Boolean expressions involving lists of items linked by logical OR’s or AND’s. Such expressions are common in psychiatry^{16,277}. For example, in order to meet the diagnostic criteria for major depression listed in the fourth edition of the Diagnostic and Statistic Manual of Mental Disorders (DSM-IV)²⁷⁸, either ‘markedly diminished interest or pleasure’, or ‘depressed mood’, must be present, along with another four out of a further seven symptoms such as ‘feelings of worthlessness or excessive or inappropriate guilt’. Decision trees can be viewed as a type of Boolean expression, in that various combinations of

items linked by logical 'AND' may be associated with a similar outcome, and so be linked by logical 'OR' ²¹³.

In terms of the running example, increased risk of unhappiness is associated with occupational categories of service or production (regardless of gender); OR occupational categories of managerial, technical, farming or craft AND being female. It can be seen that each rule must contain a value for variables used to split the dataset further up a particular branch, i.e. all rules in the running example must include values for occupational category, which was used to split the dataset in the first place.

Decision trees generally find it difficult to represent logical relations that may not be mutually exclusive, and which do not share items, such as 'outcome is present if A and B present, OR if C and D present' ^{213,279,280}. Instead, procedures such as CART would generate rules such as 'outcome is present if A is absent AND C is present AND D is present, OR if A is present AND B is present' ²¹³.

The original CART monograph ⁸⁰ described a technique for combining binary predictors using logical AND's and OR's, but which is not available in the actual computer implementations of CART. Other machine learning methods for generating Boolean decision rules, rather than trees, have been described ^{88,279,281-287}, as have recursive partitioning approaches that directly employ Boolean rules within particular nodes of a tree ²⁸⁸. The above techniques, however, are not yet as widely tested, and developed, as is CART.

1.3.7.3. Sharp decision boundaries, and lack of adjustment for confounders

Recursive partitioning algorithms generally assign every observation within a particular subgroup to the same class or outcome, or assign the same class probability⁸⁰. For example, a split might consist of

‘branch left if (age less than or equal to 49.5), else branch right’

Persons aged 49 years would therefore have the same class probability as those aged 39 years, or 29 years, or 19 years, unless this subgroup was further split by age. Such an approach is useful within a tree-building context, in that it may otherwise be difficult to split datasets by dimensional variables such as age. The use of indicative, but often essentially arbitrary, cut-points is common in psychiatry, with scores above a particular threshold on a test such as the GHQ denoting possible caseness, for example³. The use of cut-points may, however, lead to over-simplification of the results. In contrast, a logistic regression approach would allow the predicted odds of a particular outcome to be estimated for any given age or GHQ score, rather than assuming constant odds for all those values equal to or below, or above, a given cut-point.

If a dimensional variable represents scores on a particular test of a particular characteristic or trait, such as level of psychological distress, which is assumed to be latent or not readily observable, then latent trait, or item response theory, techniques^{289,290} can also be employed. One of the best known latent trait approaches is the Rasch model, originating in the field of educational testing²⁹¹. The basic Rasch model assumes that the probability of a ‘correct’ response

to a given binary test item depends on both the individual's 'ability' and on the 'difficulty' of the item.

Within a psychiatric setting, the probability of a symptom, such as a symptom of anxiety, being reported as being present would be assumed to be a function of the individual's severity of illness and on the level of illness expressed by that symptom²⁹². The greater the severity of illness and the lower the level of illness expressed by a particular symptom, the higher the probability that a given individual will report that item as being present. The original Rasch model has been extended to include ordinally scaled items such as 'never', 'sometimes', 'always'^{293,294} and employed in various psychiatric applications^{292,295-297}.

. As a possible means of avoiding sharp decision boundaries within recursive partitioning, methods that fit a linear or logistic regression equation, in the leaves of a tree structure, have been developed. For example, LOTUS outlined above^{275,276} allows predictors to be designated as splitting and/or regression variables. If age was designated as a regression variable, for instance, the predicted odds of an outcome could be estimated for any age. In contrast to the simple linear combination splits outlined above, the use of regression equations within trees allows each observation to be assigned probabilities of branching to the left or to the right at a given node of a tree^{298,299}. Although the use of a decision tree structure means that the branching probabilities are contingent upon the probability of splits made earlier in the tree-building sequence, the inclusion of regression equations allows probabilistic assignment to each subgroup. Indeed, it has been proposed that such

probabilistic recursive partitioning methods are related to latent class and similar techniques described earlier ²⁶¹.

Whether employed within a cluster analysis, or a regression context, latent class analysis assigns each observation a probability of belonging to each and every class, whether the observed variables are dimensional or binary. Latent class Rasch models, which assume the existence of latent classes, within which each observation is assumed to have the same ability, or severity of illness, have also been developed ^{300,301}. Recursive partitioning algorithms incorporating Rasch modelling, and which detect subgroups that are homogenous in regard to ability, are currently under development ^{299,302}.

The use of regression equations within trees may also allow the effects of other variables to be controlled for. For example, the best split on a single variable could be sought, while controlling or adjusting for the effects of one or more other variables on the outcome variable. CART does not incorporate regression equations, apart from the facility for simple linear combinations described above. It has been suggested ^{7,303}, however, that trees developed by CART be further analysed using logistic regression. Although this would not allow probabilities of outcome to be computed for each specific age in the above example, it would still allow the overall assignment to each outcome category to be probabilistic for each subgroup. Most importantly, the subsequent logistic regression analysis of trees developed by CART, or other recursive partitioning procedure, allows the subgroups to be compared while adjusting for those predictor variables available, but not chosen, in recursive partitioning analyses

^{237,303}. For example, the subgroups could be compared using logistic regression, while adjusting for the effects of possible confounders such as age and education. These variables may not have been selected in a particular tree, but may still have a bearing on the outcome.

1.3.8. Classification performance of recursive partitioning algorithms

There is some evidence that contemporary recursive partitioning techniques such as CART clearly exhibit worse performance at classifying hold-out or validation data than do conventional parametric techniques, such as logistic regression, or other machine learning techniques, such as artificial neural networks ^{270,304-307}. On the other hand, there is also evidence that recursive partitioning techniques do about as well ^{87,308-310}, or better ^{31,47,235,268,311-313}, than the above procedures. In addition, recursive partitioning methods provide graphic output in the form of a decision tree that is generally more easily interpretable by researchers and clinicians than regression equations or interconnected artificial neural network weights ³².

One of the most comprehensive performance comparisons of various machine learning and statistical procedures, applied to over 30 datasets, found that logistic regression, linear discriminant analysis, QUEST, C4.5, and a public domain variant of CART, all performed very well at classifying unseen or validation data ³¹⁴. The highest performing algorithm was POLYCLASS ³¹⁵, a form of polytomous logistic regression (for more than two outcomes) ²⁷, with automated selection of variables and nonlinear functions thereof. POLYCLASS

could take hours to run however, while there were no statistically significant differences in performance between the top twenty algorithms.

CHAID was not included in the above comparison. Studies utilising a much smaller number of datasets have found CART and CHAID to exhibit similar performance^{152,316}, or for the former to be more conservative with datasets of known structure than the latter³¹⁷. CART has also been found to generate smaller trees, but exhibit lower classification performance on validation data, than C4.5²⁵³. Such a result suggests that CART may ‘overprune’.

The results of several studies indicate that backward pruning, as used by most contemporary recursive partitioning algorithms such as CART and C4.5, leads to better performance on validation samples than does forward stopping^{174,249,253}. More recent results suggest, however, that backward pruning by cross-validation performs about as well as forward stopping using permutation testing, which was defined earlier¹⁷⁵.

Few studies have compared the performance of recursive partitioning techniques with that of latent class regression methods. There is some evidence that the latter are better at predicting known outcomes than are the former³¹⁸. There is also some evidence that newly developed recursive partitioning techniques, that fit regression equations within tree nodes, and allow probabilistic assignment to subgroups, perform better than existing recursive partitioning methods such as CART^{298,299}.

A general interpretation of what is known in machine learning research as the ‘No Free Lunch Theorem’³¹⁹, proposes that there is no such thing as a ‘best’

algorithm or procedure. For any given procedure, there are datasets on which it will do well, and other datasets where it will do less well ³².

1.4. Summary

Conventional parametric techniques continue to perform very well on most occasions. Nevertheless, there are situations where recursive partitioning techniques may suggest nonlinearities, statistical interactions, or subgroups within data, and so may be a useful alternative, or adjunct, to the former methods.

For example, recursive partitioning may uncover relationships that can be further examined using logistic regression. The latter readily provides odds ratios and inferential statistics such as confidence intervals, and can also adjust for possible confounders, that may not have been selected by the recursive partitioning technique, but which may still be related to the outcome.

CART is a complex procedure that may be overly conservative when compared with other recursive partitioning techniques, and may exhibit a bias in selecting variables with large numbers of categories, if such variables are employed in a particular application. On the other hand, CART is also one of the best known, and one of the most frequently used, recursive partitioning techniques. It does not rely on statistical significance testing, as does CHAID, or theoretically uncertain (albeit successful in practice) 'pessimistic' pruning heuristics, as does C4.5.

Although CART has been applied to psychiatric data since the early 1990's, it is arguably not as widely employed as it could be, particularly as an

adjunct to conventional techniques. This may be due to a lack of familiarity with CART, or it may be due to apprehension regarding the generality of results, based upon the limitations of early recursive partitioning techniques such as AID. The following chapters are concerned with employing CART in a variety of practical and clinically relevant applications concerned with screening, diagnosis and subgroup identification.

1.5. Research aims

The aims of this thesis are twofold,

- to investigate the advantages of employing CART, along with conventional techniques, in a variety of applications.
- to examine specific ways in which CART can be improved in order to increase its utility.

BACKGROUND TO CHAPTERS TWO AND THREE: THE AUSTRALIAN GULF WAR VETERANS' HEALTH STUDY

The following two chapters are concerned with the Classification and Regression Tree (CART) analysis of data that were collected as part of the Australian Gulf War Veterans' Health Study. Chapter Two is specifically concerned with the identification of Gulf War veteran subgroups at high risk of hazardous alcohol consumption. Chapter Three is specifically concerned with the overall timing and ordering of psychiatric disorders in Gulf War veterans, as well as whether there are any subgroups with particular developmental patterns of these disorders. The Australian Gulf War Veterans' Health Study itself, and Australia's military contribution to the 1990-1991 Gulf War, will firstly be briefly described.

As a result of the invasion of Kuwait by Iraq on the 2nd of August, 1990, Australia participated in a large multinational response in support of United Nations (UN) Security Resolutions, consisting of 41 countries, and almost one million military personnel, forming the Coalition forces ³²⁰. The Australian military contingent eventually totalled 1871 military personnel, most of whom (84%) were in the Royal Australian Navy. After the so-called 'Air War,' commencing on the 16th of January 1991, and a massive Coalition ground attack, beginning on the 24th of February 1991, the Iraqi forces surrendered a few days later.

After the War, military personnel from several countries that had participated began to report a wide range of physical and psychological health complaints. Initial research studies on these problems were predominantly

conducted on veterans from the UK and US. These countries had mainly deployed Army, rather than Navy, forces. After an open tendering process, the Commonwealth Department of Veterans' Affairs (DVA), the sole funding body, selected the Department of Epidemiology and Preventive Medicine at Monash University in conjunction with Health Services Australia, to carry out a study of the physical and psychological health of Australian Gulf War veterans. This study was conducted for DVA, as well as the Commonwealth Department of Defence.

The Australian Gulf War Veterans' Health Study was carried out at various testing sites throughout Australia during the period 2000-2002. Assessment was performed using a detailed postal questionnaire, as well as a comprehensive medical examination and a structured psychological interview. The latter employed Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV)²⁷⁸ criteria, and was conducted by psychologists administering the computer-assisted version of the Composite International Diagnostic Interview (CIDI)^{321,322}. To facilitate comparison of veterans with an age and service matched military comparison group, the above information was also collected on persons who had been serving in the military at the time of the Gulf War but had not been deployed there.

2. INTRODUCTION TO CHAPTER TWO: HAZARDOUS OR HARMFUL ALCOHOL USE IN ROYAL AUSTRALIAN NAVY VETERANS OF THE 1991 GULF WAR: IDENTIFICATION OF HIGH RISK SUBGROUPS

2.1. Hazardous alcohol consumption in the Military

As a result of several factors, including exposure to stressful situations, deployment to isolated areas, and the presence of large numbers of single young people ³²³, excessive alcohol consumption or alcohol use disorder has been a particular problem in the serving military ³²⁴, as well as veteran ³²⁵ populations ever since warriors have toasted their victories or attempted to drown their sorrows.

Elevated levels of alcohol use disorders have recently been observed in veterans of Korea ³²⁶, Vietnam ³²⁷, Bosnia ³²⁸, as well as the 1991 Gulf War ^{329,330}.

2.2. Identifying potential alcohol misuse

Earlier analyses found that Australian Gulf War veterans had higher rates of DSM-IV alcohol abuse and dependence compared with the military comparison group ³²⁹. DSM-IV diagnoses require structured and validated clinical interviews to be administered by trained personnel, which can be time-consuming, especially when employed with large groups. On the other hand, screening for possible alcohol use disorders can be conducted using brief self-report instruments such as the World Health Organization's Alcohol Use Disorders Identification Test (AUDIT) ¹³. It was previously established that

Australian Gulf War veterans had higher levels of 'caseness' on the AUDIT, indicating higher risks of hazardous or harmful alcohol use, than the military comparison group³³¹.

2.3. Classification and Regression Tree (CART) analysis of subgroups at risk of hazardous alcohol consumption

In order to help identify those veterans at highest risk of alcohol misuse, and so facilitate screening and treatment, the study presented in Chapter Two employed Classification and Regression Tree (CART)⁸⁰ analysis to look for possible high risk subgroups within 1201 male Royal Australian Navy (RAN) Gulf War veterans who had completed both the AUDIT questionnaire and the CIDI. Analyses were restricted to male RAN veterans because the vast majority (98%) of the participating 1456 veterans (80.5% of those veterans who were eligible) were male, and in the above service (85.5%) at the time of the conflict. Potential risk factors, identified in prior research, consisted of military rank at time of conflict, age, education, marital status and smoking status at time of interview, whether or not criteria for current (past 12 months) DSM-IV PTSD or major depression were met, and whether or not criteria for pre-War (2nd August, 1990) DSM-IV alcohol abuse or dependence were met.

Although CART had previously been used in the study of alcohol misuse in civilians²³², the study presented in Chapter Two represents the first application of CART to Gulf War data. This study is also one of the first to perform subgroup analysis of alcohol misuse in veterans as previous studies were concerned only with subgroups defined by a limited number of risk factors, such as physical disablement³³². Logistic regression was employed to examine all risk factors

simultaneously, as well as to compare the subgroups generated by CART. The latter analysis controlled for the possible effects of any risk factors that were available to, but not chosen by, CART. As outlined in the Introduction (Chapter One), the formal statistical comparison of empirically derived subgroups can be highly problematic¹⁵⁶. Therefore, the results of the logistic regression analysis of the CART subgroups should be seen as being primarily descriptive, allowing the magnitude of the risk of alcohol misuse to be assessed for each subgroup, while controlling for the potential effects of the other risk factors.

Declaration for Thesis Chapter 2

McKenzie DP, McFarlane AC, Creamer M, Ikin JF, Forbes AB, Kelsall HL, Clarke DM, Glass DC, Ittak P, Sim MR. Hazardous or harmful alcohol use in Royal Australian Navy veterans of the 1991 Gulf War: Identification of high risk subgroups. *Addictive Behaviors* 2006; 31: 1683-1694.

Declaration by candidate

In the case of Chapter 2, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
I was responsible for the research question, literature review, data management and programming, statistical analyses, interpreting the results, writing the paper and undertaking any required revisions	75%

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
A. McFarlane	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A
M. Creamer	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A
J. Ikin	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	4%
A. Forbes	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper. Statistical consultation.	N/A
H. Kelsall	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A
D. Clarke	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A
D. Glass	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the	N/A

	written paper.	
P. Ittak	Interpretation of results and critical review of the written paper.	N/A
M. Sim	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A

Candidate's
Signature

Neer Mehta

Date
16/12/2008

Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s)

Monash University Department of Epidemiology and Preventive Medicine, Alfred Hospital

[Please note that the location(s) must be institutional in nature, and should be indicated here as a department, centre or institute, with specific campus identification where relevant.]

Signature 1

AEMCFarlane

16/12/2008

Signature 2

Signature 3

J. Ittak

17/12/2008

Signature 4

Signature 5

Signature 6

David Clarke

08/12/2008

Signature 7

Signature 8

D. O'Neil

16/12/2008

Signature 9

[Signature]

15/12/2008

Signature 10

[Signature]

	written paper.	
P. Ittak	Interpretation of results and critical review of the written paper.	N/A
M. Sim	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A

**Candidate's
Signature**

Shen M. Sim

Date

16-12-2008

Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s)

Monash University Department of Epidemiology and Preventive Medicine, Alfred Hospital

[Please note that the location(s) must be institutional in nature, and should be indicated here as a department, centre or institute, with specific campus identification where relevant.]

Signature 1

	Date
--	-------------

Signature 2

<i>Mark Geane</i>	09/12/2008
-------------------	------------

Signature 3

--	--

Signature 4

<i>Alfred</i>	10/12/08
---------------	----------

Signature 5

<i>SS Silsall</i>	15/12/2008
-------------------	------------

Signature 6

<i>David Clarke</i>	08/12/2008
---------------------	------------

Signature 7

--	--

Signature 8

--	--

Hazardous or harmful alcohol use in Royal Australian Navy veterans of the 1991 Gulf War: Identification of high risk subgroups

Dean P. McKenzie^{a,*}, Alexander C. McFarlane^b, Mark Creamer^c, Jillian F. Ikin^a,
Andrew B. Forbes^a, Helen L. Kelsall^a, David M. Clarke^d, Deborah C. Glass^a,
Peter Ittak^a, Malcolm R. Sim^a

^a *Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia*

^b *Department of Psychiatry, University of Adelaide, Adelaide, Australia*

^c *Australian Centre for Posttraumatic Mental Health, University of Melbourne, Melbourne, Australia*

^d *Department of Psychological Medicine, Monash University, Australia*

Abstract

Elevated alcohol use disorders have been observed in 1991 Gulf War veterans from a variety of countries. This study used a self-report instrument, the Alcohol Use Disorders Identification Test (AUDIT), to ascertain whether any subgroups of 1232 male Royal Australian Navy (RAN) Gulf War veterans were at higher risk of hazardous or harmful alcohol use. Recursive partitioning/classification and regression tree (CART) analysis, followed by logistic regression, found five subgroups among the veterans, with differing risks of AUDIT caseness. The highest risk subgroup comprised current smokers. The other two high risk groups both consisted of former or never smokers of lower rank who were (1) not married, or (2) married, with a current diagnosis of major depression. The above subgroups were over three times as likely to exhibit AUDIT caseness than those who were former or never smokers of higher rank. The findings have important implications for effective development of public health initiatives designed to encourage safe alcohol use among veterans.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Alcohol; Hazardous drinking; Major depression; Posttraumatic stress disorder; Veterans; Classification and regression trees

* Corresponding author. Monash University, Department of Epidemiology and Preventive Medicine, Alfred Hospital, Commercial Road, Melbourne, Victoria, 3004 Australia. Tel.: +61 3 9903 0555; fax: +61 3 9903 0556.

E-mail address: Dean.Mckenzie@med.monash.edu.au (D.P. McKenzie).

1. Introduction

As discussed in a recent review (Hotopf & Wessely, 2005), 1991 Gulf War veterans from a variety of countries have demonstrably lower levels of psychological and physical health than those in military control or comparison groups that did not deploy to the Gulf. Several studies have observed higher levels of alcohol abuse or dependence in Gulf War veterans, measured using both diagnostic (Black et al., 2004; Ikin et al., 2004) and self-report (The Iowa Persian Gulf Study Group, 1997) instruments, compared with military comparison groups. Alcohol abuse or dependence is also commonly found among Gulf War veterans seeking medical treatment (Engel et al., 1999).

In order to ascertain which veterans may most benefit from additional care and treatment facilities, those at highest risk of exhibiting hazardous or harmful alcohol use must be identified. Previous studies have performed only limited subgroup analyses. For example, disabled Gulf War veterans have been shown to be more likely to exhibit alcohol-related disorders than non-disabled veterans (Ismail et al., 2002).

Our group recently conducted the Australian Gulf War Veterans' Health Study, comparing the health of Australian Gulf War veterans with that of an age and service matched military comparison group that had not deployed to the Gulf (Forbes et al., 2004; Glass et al., *in press*; Ikin et al., 2004; Kelsall et al., 2004; McKenzie et al., 2004). Using a structured clinical interview, we found significantly higher levels of alcohol abuse and dependence among the Gulf veterans, in the period since their deployment (Ikin et al., 2004). While structured and validated clinical interviews provide a high degree of clinical accuracy in making a diagnosis, they require administration by trained personnel and are time-consuming when employed with large groups. Brief self-report instruments such as the Alcohol Use Disorders Identification Test (AUDIT) (Babor, Fuente, Saunders, & Grant, 1989) can, therefore, be useful in screening for elevated alcohol use disorders.

Having identified high rates of substance use problems in deployed veterans 10 years after the Gulf War, we sought in the present study to explore those factors that increased the risk of being identified as a "case" on the AUDIT. In addition to employing logistic regression we applied an exploratory statistical technique known as classification and regression trees (CART) (Breiman, Friedman, Olshen, & Stone, 1984). The former procedure is generally used to find global or overall relationships between potential risk factors and outcomes, whereas CART and other tree-building or recursive partitioning methods are used to examine local (subgroup) relationships (Zhang & Singer, 1999). Although not previously employed in Gulf War analyses, CART and similar techniques have been applied in studies of alcohol (Bailey, 1999; Barnes, Welte, & Dintcheff, 1991; Schwan et al., 2004) and other addictive behaviours (Welte, Barnes, Wiczorek, & Tidwell, 2004), as well as in general applications (Barton, McKenzie, Walters, Abramson, & Victorian Asthma Mortality Study Group, 2005; Craig, Siegel, Hopper, Lin, & Sartorius, 1997; McKenzie et al., 1993; Schmitz, Kugler, & Rollnik, 2003). For example, Barnes et al. (1991) used CART to find subgroups typified by heavier (such as young male students) or lighter (such as married or widowed females on lower incomes) alcohol consumption.

Although few studies have actually done so, it is important to include diagnoses such as posttraumatic stress disorder (PTSD) and depression when analysing post-combat alcohol disorders (Rundell & Ursano, 1996). Such diagnoses, as well as other anxiety disorders, have been associated with alcohol disorders in veteran (McLeod et al., 2001; Tomlinson, Tate, Anderson, McCarthy, & Brown, *in press*) as well as general (Libby, Orton, Stover, & Riggs, 2005; McFarlane, 1998) populations. Rundell and Ursano (1996) have also proposed that pre-deployment alcohol use be examined, because alcohol problems may be present before conflict is encountered.

The aim of this paper, therefore, was to extend the overall findings of Ikin et al. (2004) regarding elevated levels of alcohol abuse or dependence in Australian Gulf War veterans by ascertaining which patterns of psychiatric co-morbidity and demographic variables are associated with higher risk of hazardous or harmful alcohol use. The latter was assessed using the AUDIT, intended for the early identification of such behaviour (Babor et al., 1989).

2. Method

2.1. Participants

The participating Gulf War veteran study group consisted of 1456 veterans (80.5% of those eligible) of whom 1249 (85.8%) were Royal Australian Navy (RAN), 95 (6.5%) were Australian Army and 112 (7.7%) were Royal Australian Air Force personnel. As the numbers of females (only 2%), and participants from the Army and Air Force were comparatively small, the current analysis was restricted to the 1232 male RAN Gulf War veterans. The mean age at time of study (2000 to 2002) for the above group was 37.36 years (S.D. = 6.08).

2.2. Measures and procedure

The AUDIT core questionnaire was developed by the World Health Organization (WHO) to identify persons whose drinking of alcohol has become hazardous or harmful to their health (Babor et al., 1989). The instrument consists of ten questions pertaining to alcohol consumption, alcohol dependence, and alcohol-related problems during the past 12 months. The AUDIT performs similarly to, or better than, other self-report alcohol screening tests (Rumpf, Hapke, Meyer, & John, 2002), and has previously been applied in Gulf War research (e.g., Haley et al., 1997).

The AUDIT is generally employed with the recommended WHO cut-off score of eight (Babor et al., 1989). Several optimal cut-off scores, as large as 16 (Pal, Jena, & Yadav, 2004), have since been reported. We empirically determined the optimal cut-off for the AUDIT.

In our study, the 'standard drink' referred to in the AUDIT was defined as being one containing 10 g or 12.5 ml (0.44 fluid ounces) of pure alcohol (National Health & Medical Research Council, 2001).

A trained psychologist conducted an interview at which the presence of DSM-IV (American Psychiatric Association, 1994) psychiatric diagnoses was evaluated using the computer-assisted version of the Composite International Diagnostic Interview (CIDI-Auto) (Robins et al., 1988; World Health Organization Collaborating Centre for Mental Health & Substance Abuse, 1997). Data collection also involved a comprehensive postal questionnaire comprising a range of demographic and service history information, as well as the AUDIT. Details of the data collection process can be found in Ikin et al. (2004).

2.3. Statistical analyses

2.3.1. Receiver Operating Characteristic analysis

Receiver Operating Characteristic (ROC) (Kraemer, 1992) analysis was used to establish the optimal AUDIT cut-off score for the RAN Gulf War veterans. The criterion diagnosis was defined

as the presence of current (past 12 months prior to interview) DSM-IV alcohol abuse or dependence. ROC analysis was undertaken using SPSS 12 (SPSS Inc., 2003). Confidence intervals (CI) for classification accuracy or diagnostic efficiency, sensitivity and specificity (Kraemer, 1992) were obtained using a Fortran 90 programme (McKenzie, Vida, Mackinnon, Onghena, & Clarke, 1997).

2.3.2. Risk factors for AUDIT caseness

Logistic regression was employed to perform an overall analysis of risk factors of AUDIT caseness, defined in Section 3.1 below. Current DSM-IV (past 12 month) diagnoses of PTSD, any other anxiety disorder, major depression, and pre-1991 Gulf War (prior to invasion of Kuwait on 2 August 1990) diagnosis of alcohol abuse or dependence were included as possible risk factors.

Other possible risk factors included in the analyses were age at 2 August 1990, highest education level (≤ 10 years of schooling, 11–12 years, certificate or diploma, university or college degree), marital status (married/de facto, separated/divorced/widowed, single/never married), smoking status (never, former, current smoker), and military rank as at 2 August, 1990. Rank was categorised as officer, other rank—supervisory (at or above the rank of Leading Seaman) and other rank—non-supervisory. These categories are comparable to those of officer, non-commissioned officer and enlisted employed in other Gulf War research.

Unless otherwise specified, all statistical analyses were performed using Stata 8 (StataCorp, 2004).

2.3.3. CART subgroup analysis

We used the CART 4 binary tree-building procedure (Breiman et al., 1984; Salford Systems, 2001) to identify possible subgroups of Gulf War veterans at high risk of hazardous or harmful alcohol use, as determined using AUDIT caseness. All of the variables employed in the logistic regression analysis were available for selection by CART. Unlike early recursive partitioning algorithms (Morgan, 2005), CART explicitly validates the generality of its tree structures, using by default 10-fold cross-validation (Breiman et al., 1984). The dataset is randomly divided into 10 subsets, each subset in turn being used to test the performance of the tree created with the other nine subsets.

Contemporary recursive partitioning techniques have been found to give satisfactory performance when applied to different sub-samples of a given dataset (James, White, & Kraemer, 2005). The performance of CART is comparable to that of other recursive partitioning techniques (Lim, Loh, & Shih, 2000; McKenzie et al., 1993). We further tested the generality of CART using a separate ‘hold-out’ or validation subset (Bleeker et al., 2003). SPSS randomly divided those Gulf War veterans with non-missing AUDIT information into a learning subset of approximately 75% and a validation subset of approximately 25% of the observations, with similar levels of AUDIT caseness. The classification tree obtained for the learning subset was then applied to the validation subset (James et al., 2005). The classification accuracy obtained for each subset was compared using a two (learning and validation subsets) by four (true positives, true negatives, false positives, false negatives) Chi-squared test. If the results of this test did not approach statistical significance, the subsets were combined. Finally, the final subgroups or ‘terminal nodes’ of the CART tree structure were themselves entered into a logistic regression analysis (Zhang & Singer, 1999), adjusting for the effects of possible risk factors available to, but not chosen by, CART.

3. Results

3.1. AUDIT cut-off

A total of 1201 (97.5%) of the Gulf War veterans in the study group completed the AUDIT questionnaire and the CIDI. In determining the optimal cut-off score for the AUDIT, the prevalence of the criterion diagnosis of current DSM-IV alcohol use or dependence was 4.5%. ROC analysis found the optimal cut-off to be 10 or greater, therefore this score was chosen to determine AUDIT caseness. The area under the ROC curve was 0.88 (chance performance=0.50), with a 95% confidence interval (CI) of 0.84 to 0.92. Sensitivity using the above cut-off score was 0.85 (95% CI=0.73 to 0.93), specificity was 0.77 (95% CI=0.75 to 0.79), and overall classification accuracy was 0.77 (95% CI=0.75 to 0.80).

3.2. Overall risk factors for AUDIT caseness

Of the 1201 RAN Gulf War veterans analysed, 25.7% (309) exhibited AUDIT caseness as defined in Section 3.1 above. Table 1 shows the results of the logistic regression analysis of possible risk factors for AUDIT caseness. The results for age ($p=0.31$), number of active deployments ($p=0.21$) and current (past year) DSM-IV diagnosis of other anxiety disorder ($p=0.73$) are clearly not statistically significant

Table 1

Relationships of DSM-IV diagnoses and other possible risk factors to hazardous or harmful alcohol use (AUDIT caseness) in male Royal Australian Navy Gulf War veterans

	AUDIT caseness (%)	OR	Adjusted OR ^a	<i>p</i> value	95% CI
Education				0.15 ^b	
≤ 10 years	34.6	1.0	1.0		
11–12 years	25.7	0.7	0.8	0.32	0.5, 1.2
Certificate/diploma	25.4	0.6	0.8	0.34	0.6, 1.2
University/college	11.0	0.2	0.5	0.02	0.2, 0.9
Rank		1.6	1.5 ^c	0.01	1.1, 2.1
Officer	12.8	1.0	1.0		
Other rank—supervisory	25.3	2.3	1.4	0.25	0.8, 2.5
Other rank—non-supervisory	31.3	3.1	2.2	0.03	1.1, 4.3
Marital status				<0.001 ^b	
Married/de facto	21.9	1.0	1.0		
Separated/divorced/widowed	39.0	2.3	1.9	0.002	1.3, 2.8
Single/never married	36.1	2.0	1.9	0.001	1.3, 2.9
Smoking status				<0.001 ^b	
Never	17.4	1.0	1.0		
Former	24.7	1.6	1.4	0.042	1.01, 2.0
Current	39.6	3.1	2.4	<0.001	1.7, 3.3
PTSD	47.5	2.8	1.8	0.052	0.99, 3.3
Major depression	44.9	2.6	1.7	0.03	1.1, 2.7
Pre-August 2, 1990, alcohol abuse or dependence	34.0	1.7	1.6	0.004	1.2, 2.2

^a Odds ratios were obtained using logistic regression, adjusting for the other variables in the model.

^b Omnibus test of statistical significance for overall difference between variable categories, adjusted for the other variables in the model.

^c Dose-response slope is the expected proportionate increase in the odds ratio per decrease in rank category.

and so are not shown. It can be seen in Table 1 that those who were separated/divorced/widowed, single/never married, former smokers, current smokers, and those with current DSM-IV diagnoses of major depression, and pre-deployment alcohol use or dependence, were significantly more likely to be AUDIT cases. The association between current DSM-IV diagnosis of PTSD and AUDIT caseness narrowly failed to achieve statistical significance ($p=0.052$).

There was a statistically significant ($p<0.05$) negative dose–response relationship between rank and AUDIT caseness. The odds of the latter would be expected to increase by 50% with each decrease in rank category. In other words, the odds of AUDIT caseness for the other rank—non-supervisory category would be expected to be 50% higher than the odds of AUDIT caseness for the other rank—supervisory category. Finally, although the overall association between education level and AUDIT caseness was not statistically significant ($p=0.15$), those veterans with a university or college degree were significantly less likely ($p<0.05$) to be AUDIT cases.

3.3. CART subgroup analysis

The results of the CART recursive partitioning subgroup analysis of AUDIT caseness are given in Fig. 1. There was no statistically significant difference (Chi-squared=2.03, $df=3$, $p=0.61$) between the performance of the classification tree obtained for the learning subset, and the performance of that tree applied to the validation subset. Therefore, the two subsets were combined.

By default CART chooses splits by minimising the Gini impurity criterion (Breiman et al., 1984), a measure of variability. To aid interpretation we compared pairs of CART subgroups using odds ratios obtained using logistic regression, adjusted for those variables entered into the CART analyses but not selected. This procedure is similar to that employed by Schmitz et al. (2003).

The variable first chosen by CART to split the dataset was smoking status—with the best binary merging of categories being former or never smokers versus current smokers. Of the 870 former or never smokers, 178 (20.5%) were AUDIT cases. Of the 331 current smokers, 131 (39.6%) were AUDIT cases. The current smoker subgroup was more than twice as likely to exhibit AUDIT caseness than the former or never smoker subgroup (39.6% versus 20.5%, OR=2.5, adjusted OR=2.1, 95% CI=1.6, 2.9). CART then split the latter group by military rank—with the optimal merging of categories being officers versus other rank—supervisory and other rank—non-supervisory. The latter were over twice as likely as the higher ranking subgroup to be AUDIT cases (23.0% versus 8.4%, OR=3.2, adj. OR=2.1, 95% CI=1.1, 4.3).

The former or never smoker, other rank—supervisory and non-supervisory subgroup was then split by marital status. Veterans who were not married/de facto were over twice as likely to be AUDIT cases than those who were (35.8% versus 19.2%, OR=2.3, adj. OR=2.3, 95% CI=1.5, 3.4). The latter subgroup was split by current diagnosis of DSM-IV major depression. Subgroup members who had this diagnosis were over three times more likely to be AUDIT cases than were those who did not (38.7% versus 18.1%, OR=2.9, adj. OR=3.4, 95% CI=1.4, 8.0). The odds ratios given above facilitate the comparison of subgroups at each stage of the tree, but do not readily allow identification of the highest risk subgroups. Three subgroups had over three times the risk of AUDIT caseness than the reference group or subgroup with the lowest risk of AUDIT caseness, which comprised the former or never smokers who were officers. The highest risk was observed for current smokers (39.6% versus 8.4% for the reference group, OR=7.1, adj. OR=4.6, 95% CI=2.4, 9.1). Of the former or never smokers of other rank—supervisory or non-supervisory, statistically significant increased risks of AUDIT caseness were observed for not

married/de facto (35.8% versus 8.4%, OR=6.0, adj. OR=4.2, 95% CI=2.1, 8.7) or married/de facto with current diagnosis of major depression (38.7% versus 8.4%, OR=6.8, adj. OR=3.8, 95% CI=1.4, 10.5). There was no statistically significant increased risk for those veterans who were married/de facto

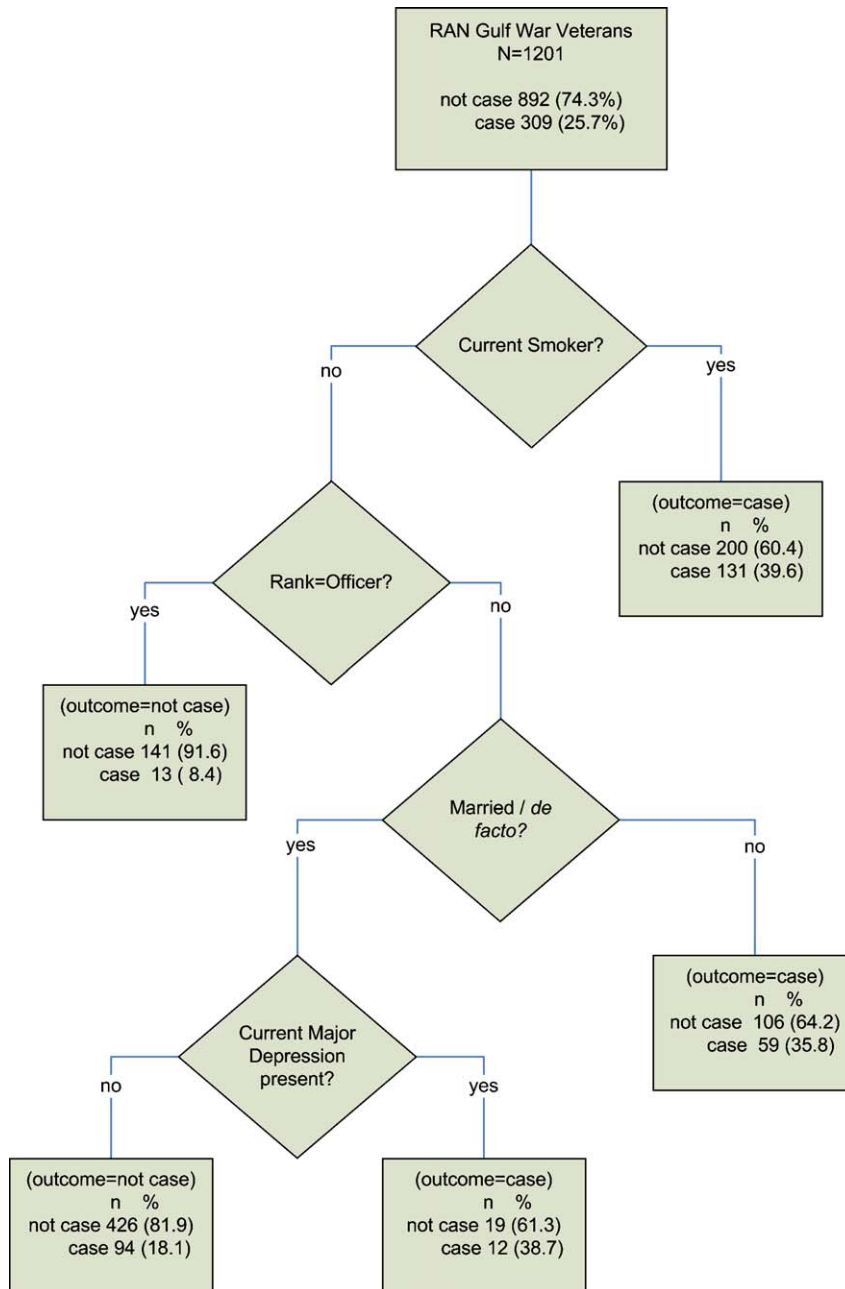


Fig. 1. Recursive partitioning/classification and regression tree (CART) analysis of hazardous or harmful alcohol use (AUDIT caseness) in male Royal Australian Navy (RAN) Gulf War veterans.

but who did not have a current diagnosis of major depression (18.1% versus 8.4%, OR=2.4, adj. OR=1.7, 95% CI=0.9, 3.4). PTSD was not selected by CART, even if major depression was temporarily excluded from the analysis.

4. Discussion

An overall logistic regression analysis indicated that smoking status, military rank, marital status, current DSM-IV diagnosis of major depression and pre-deployment diagnosis of alcohol use or dependence were significantly associated with AUDIT caseness. The association with current PTSD narrowly missed statistical significance ($p=0.052$). CART recursive partitioning analysis found five subgroups of Gulf War veterans with differing risks of AUDIT caseness, defined by specific combinations of smoking status, military rank, marital status and current diagnosis of major depression.

Current smoking was shown to be a major risk factor by itself, regardless of the other variables selected in the CART analysis. It has long been established that nicotine and alcohol dependence often co-occur, while current smoking is also associated with hazardous alcohol consumption (Kranzler et al., 2002). However, little is known about the aetiology of dual-dependency, and concurrent, as well as individual, treatment can be difficult (Stotts, Schmitz, & Grabowski, 2003). Nevertheless, the current data support the case for public health initiatives aimed at those personnel who are both smokers and heavy drinkers.

For the former or never smoking group, those who had lower rank during the 1991 Gulf War were more likely to exhibit current AUDIT caseness when compared with those who were officers. This is consistent with a study of referrals for alcohol misuse in UK Royal Naval personnel which found that officers were under-represented (Micklewright, 1996). Several Gulf War studies have found an inverse relationship between rank and psychological ill-health (Hotopf et al., 2004; Ismail et al., 2000; McKenzie et al., 2004). Ismail et al. (2000) proposed that rank is a proxy for socio-economic status, inversely related with psychological ill-health in the general population. Although education level was not selected by CART, the overall logistic regression indicated that having a university or college degree, another proxy for socio-economic status, was associated with lower levels of AUDIT caseness.

Among those veterans who had lower rank, those that were currently single, separated or widowed were more likely to be AUDIT cases than were those who were married or in a de facto relationship. The transition to marriage is generally accompanied by a reduction in alcohol use, perhaps as a result of greater social responsibility, and the potentially destructive effects of heavy drinking on marital quality and stability (Leonard & Rothbard, 1999). Single persons, and those who become separated from their partners, are more likely to be problem drinkers, although other factors apart from marital status such as individual and social predisposition may be more important longitudinally (Matzger, Delucchi, Weisner, & Ammon, 2004). Micklewright (1996) found that Royal Navy personnel who were not married were more likely to be referred for alcohol problems than those who were married. Again, the data provide support for aiming messages around safe alcohol use towards single personnel, with particular emphasis on those recently separated from a stable relationship.

Married/de facto, not being a current smoker, and higher military rank would appear to be protective factors for AUDIT caseness. However, those veterans within the former or never smoking, lower rank, married/de facto subgroup with a current (past 12 months) diagnosis of major depression were at higher risk of AUDIT caseness. Cause and effect between alcohol use and depression is of course difficult to

establish (Libby et al., 2005). The association between depression and AUDIT caseness may be explained by so-called self-medication (Tomlinson et al., *in press*), with sufferers trying to reduce the symptoms of depression through increased use of alcohol. Alternatively, substance use disorders and depression may be independent, with the possibility of both the above developing in response to traumatic exposure such as combat or military service. Finally, of course, depression may develop secondarily to, and as a direct result of, the substance use disorder (Libby et al., 2005; Tomlinson et al., *in press*).

Self-report can be misleading and it is possible that married veterans, and those who had been officers during the 1991 Gulf War understated their actual alcohol use. However, this would not explain why married veterans with a current diagnosis of major depression exhibited slightly higher AUDIT caseness than those who were not married.

Comparisons of three of the above subgroups with the lowest risk subgroup—former or never smoker, officers—remained statistically significant even when other diagnoses and other confounding variables listed earlier were controlled for. It should be re-emphasised however that CART utilises cross-validation, not statistical significance, to build and test its tree structures. Statistical significance was assessed using logistic regression after the classification tree was constructed.

Current DSM-IV diagnosis of PTSD was not selected by CART, and narrowly failed to achieve statistical significance in the overall logistic regression ($p=0.052$). PTSD is generally correlated with alcohol use (McFarlane, 1998), although a recent study of US Gulf War veterans (Shipherd, Stafford, & Tanner, 2005) failed to find such a relationship. Future studies need to further examine the effects of comorbidity, longitudinally as well as cross-sectionally. For example, some researchers have observed that it is the changes in levels of PTSD symptoms, rather than the levels themselves, that are associated with alcohol use (Forbes, Creamer, Hawthorne, Allen, & McHugh, 2003; Read, Brown, & Kahler, 2004).

In conclusion, it was possible to identify several clear subgroups of Royal Australian Navy 1991 Gulf War veterans at increased risk of hazardous or harmful alcohol use as measured by the AUDIT. Current smoking is highly associated with AUDIT caseness, regardless of other factors. Higher military rank, higher education level, and marriage appeared to be protective factors, but not when the latter was accompanied by a current (past 12 months) diagnosis of major depression.

Our study illustrates the application of recursive partitioning techniques to subgroup analysis in Gulf War veterans. Although generalising from the current results to other branches of the military or to other populations should be done cautiously, important data about subgroups of Gulf War veterans at high risk of hazardous or harmful alcohol use have been obtained. This information could be incorporated into education and other public health initiatives, as well as assisting civilian and military medical personnel in identifying high risk individuals in routine medical settings. Our findings highlight potential risk factors, particularly current diagnosis of major depression in married veterans, which should be the focus of further research aimed at elucidating the underlying mechanisms.

Acknowledgements

The Australian Gulf War Veterans' Health Study was funded by the Australian Government Department of Veterans' Affairs. The study was overseen by a Scientific Advisory Committee, headed by Professor Terry Dwyer, and by a veterans' Consultative Forum. We are grateful to members of both groups for their contribution and support. We gratefully acknowledge the contribution of Dr. Keith

Horsley, Dr. Warren Harrex, Mr. Bob Connolly and his contact and recruitment team at the Department of Veterans' Affairs, and the staff at Health Services Australia who conducted the medical and psychological assessments. Lastly, but most importantly, we wholeheartedly thank the participants in the study.

Declaration: This research was funded by the Australian Government Department of Veterans' Affairs. Professor A.C. McFarlane is Chair of the Mental Health Consultative Group to the Director General of the Health Service Branch of the Australian Defence Forces. The Australian Centre for Posttraumatic Mental Health is partially funded by the Department of Veterans' Affairs.

References

- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (DSM-IV)* (4th ed.). Washington, DC: American Psychiatric Association.
- Babor, T., Fuente, J., Saunders, J., Grant, M. (1989). *The alcohol use disorders identification test: Guidelines for use in primary health care*. World Health Organization, Geneva: Division of Mental Health.
- Bailey, S. L. (1999). The measurement of problem drinking in young adulthood. *Journal of Studies on Alcohol*, 60, 234–244.
- Barnes, G. M., Welte, J. W., Dintcheff, B. (1991). Drinking among subgroups in the adult population of New York State: A classification analysis using CART. *Journal of Studies on Alcohol*, 52, 338–344.
- Barton, C. A., McKenzie, D. P., Walters, E. H., Abramson, M. J., & Victorian Asthma Mortality Study Group. (2005). Interactions between psychosocial problems and management of asthma: Who is at risk of dying? *Journal of Asthma*, 42, 249–256.
- Black, D. W., Carney, C. P., Fornan-Hoffman, L., Letuchy, E., Peloso, P., Woolson, R. F., et al. (2004). Depression in veterans of the first Gulf War and comparable military controls. *Annals of Clinical Psychiatry*, 16, 53–61.
- Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R. T., Derksen-Lubsen, G., Grobbee, D. E., et al. (2003). External validation is necessary in prediction research: A clinical example. *Journal of Clinical Epidemiology*, 56, 826–832.
- Breiman, L., Friedman, J., Olshen, R. A., Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.
- Craig, T. J., Siegel, C., Hopper, K., Lin, S., Sartorius, N. (1997). Outcome in schizophrenia and related disorders compared between developing and developed countries: A recursive partitioning reanalysis of the WHO DOSMD data. *British Journal of Psychiatry*, 170, 229–233.
- Engel, C. C. J., Ursano, R., Magruder, C., Tartaglione, R., Jing, Z., Labbate, L. A., et al. (1999). Psychological conditions diagnosed among veterans seeking Department of Defense care for Gulf War-related health concerns. *Journal of Occupational and Environmental Medicine*, 41, 384–392.
- Forbes, D., Creamer, M., Hawthorne, G., Allen, N., McHugh, T. (2003). Comorbidity as a predictor of symptom change after treatment in combat-related posttraumatic stress disorder. *Journal of Nervous and Mental Disease*, 191, 93–99.
- Forbes, A. B., McKenzie, D. P., Mackinnon, A. J., Kelsall, H. L., McFarlane, A. C., Ikin, J. F., et al. (2004). The health of Australian veterans of the 1991 Gulf War: Factor analysis of self-reported symptoms. *Occupational and Environmental Medicine*, 61, 1014–1020.
- Glass, D. C., Sim, M. R., Kelsall, H. L., Ikin, J. F., McKenzie, D. P., Forbes, A. B., et al. (in press). What was different about exposures reported by Australian Gulf War veterans during the 1991 Gulf War compared with exposures reported for other deployments? *Military Medicine*.
- Haley, R. W., Hom, J., Roland, P. S., Bryan, W. W., Van Ness, P. C., Bonte, F. J., et al. (1997). Evaluation of neurologic function in Gulf War veterans. A blinded case-control study. *Journal of the American Medical Association*, 277, 223–230.
- Hotopf, M., David, A., Hull, L., Nikalaou, V., Unwin, C., Wessely, S. (2004). Risk factors for continued illness among Gulf War veterans: A cohort study. *Psychological Medicine*, 34, 1–8.
- Hotopf, M., Wessely, S. (2005). Can epidemiology clear the fog of war? Lessons from the 1990–1991 Gulf War. *International Journal of Epidemiology*, 34, 791–800.
- Ikin, J. F., Sim, M. R., Creamer, M. C., Forbes, A. B., McKenzie, D. P., Kelsall, H. L., et al. (2004). War-related psychological stressors and risk of psychological disorders in Australian veterans of the 1991 Gulf War. *British Journal of Psychiatry*, 185, 116–126.

- Ismail, K., Blatchley, N., Hotopf, M., Hull, L., Palmer, I., Unwin, C., et al. (2000). Occupational risk factors for ill health in Gulf veterans of the United Kingdom. *Journal of Epidemiology and Community Health*, 54, 834–838.
- Ismail, K., Kent, K., Brugha, T., Hotopf, M., Hull, L., Seed, P., et al. (2002). The mental health of UK Gulf War veterans: Phase 2 of a two phase cohort study. *British Medical Journal*, 325, 525–576.
- James, K. F., White, R. F., Kraemer, H. C. (2005). Repeated split sample validation to assess logistic regression and recursive partitioning: An application to the prediction of cognitive impairment. *Statistics in Medicine*, 24, 3019–3035.
- Kelsall, H. L., Sim, M. R., Forbes, A. B., McKenzie, D. P., Glass, D. C., Ikin, J. F., et al. (2004). Respiratory health status of Australian veterans of the 1991 Gulf War and the effects of exposure to oil fire smoke and dust storms. *Thorax*, 59, 897–903.
- Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage.
- Kranzler, H. R., Amin, H., Conney, N. L., Conney, J. L., Burleson, J. A., Petry, N., et al. (2002). Screening for health behaviors in ambulatory clinical settings. Does smoking status predict hazardous drinking? *Addictive Behaviors*, 27, 739–749.
- Leonard, K. F., Rothbard, J. C. (1999). Alcohol and the marriage effect. *Journal of Studies on Alcohol. Supplement*, 13, 139–146.
- Libby, A. M., Orton, H. D., Stover, S. K., Riggs, P. D. (2005). What came first, major depression or substance use disorder? Clinical characteristics and substance use comparing teens in a treatment cohort. *Addictive Behaviors*, 30, 1649–1662.
- Lim, T.-S., Loh, W.-Y., Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40, 203–228.
- Matzger, H., Delucchi, K., Weisner, C., Ammon, L. (2004). Does marital status predict long-term drinking? Five-year observations of dependent and problem drinkers. *Journal of Studies on Alcohol*, 65, 255–265.
- McFarlane, A. C. (1998). Epidemiological evidence about the relationship between PTSD and alcohol use: The nature of the association. *Addictive Behaviors*, 23, 813–825.
- McKenzie, D. P., Ikin, J. F., McFarlane, A. C., Creamer, M., Forbes, A. B., Kelsall, H. L., et al. (2004). Psychological health of Australian veterans of the 1991 Gulf War: An assessment using the SF-12, GHQ-12 and PCL-S. *Psychological Medicine*, 34, 1419–1430.
- McKenzie, D. P., McGorry, P. D., Wallace, C. S., Low, L. H., Copolov, D. L., Singh, B. S. (1993). Constructing a minimal diagnostic decision tree. *Methods of Information in Medicine*, 32, 161–166.
- McKenzie, D. P., Vida, S., Mackinnon, A. J., Onghena, P., Clarke, D. M. (1997). Accurate confidence intervals for measures of test performance. *Psychiatry Research*, 69, 207–209.
- McLeod, D. S., Koenen, K. C., Meyer, J. M., Lyons, M. J., Eisen, S., True, W., et al. (2001). Genetic and environmental influences on the relationship among combat exposure, posttraumatic stress disorder symptoms, and alcohol use. *Journal of Traumatic Stress*, 14, 259–275.
- Micklewright, S. (1996). Problem drinking in the Naval Service: A study of personnel identified as alcohol abusers. *Journal of the Royal Naval Medical Service*, 82, 34–40.
- Morgan, J. N. (2005). History and potential of binary segmentation for exploratory data analysis. *Journal of Data Science*, 3, 123–136.
- National Health and Medical Research Council. (2001). *Australian alcohol guidelines: Health risks and benefits*. Canberra: Commonwealth of Australia.
- Pal, H. R., Jena, R., Yadav, D. (2004). Validation of the Alcohol Use Disorders Identification Test (AUDIT) in urban community outreach and de-addiction center samples in north India. *Journal of Studies on Alcohol*, 65, 794–800.
- Read, J. P., Brown, P. J., Kahler, C. W. (2004). Substance use and posttraumatic stress disorders: Symptom interplay and effects on outcome. *Addictive Behaviors*, 29, 1665–1672.
- Robins, L. N., Wing, J., Wittchen, H. U., Helzer, J. E., Babor, T. F., Burke, J., et al. (1988). The Composite International Diagnostic Interview: An epidemiological instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry*, 45, 1069–1077.
- Rumpf, H., Hapke, U., Meyer, C., John, U. (2002). Screening for alcohol use disorders and at-risk drinking in the general population: Psychometric performance of three questionnaires. *Alcohol and Alcoholism*, 37(3), 261–268.
- Rundell, J. R., Ursano, R. J. (1996). Psychiatric responses to war trauma. In R. J. Ursano, & A. E. Norwood (Eds.), *Emotional aftermath of the Persian Gulf War: Veterans, families, communities and nations* (pp. 43–81). Washington, DC: American Psychiatric Press.
- Salford Systems. (2001). *CART for Windows, version 4 [computer software]*. San Diego, CA: Salford Systems.

- Schmitz, N., Kugler, J., Rollnik, J. (2003). On the relationship between neuroticism, self-esteem, and depression: Results from the National Comorbidity survey. *Comprehensive Psychiatry*, 44, 169–176.
- Schwan, R., Albuissou, E., Malet, L., Loiseaux, M. N., Reynard, M., Schellenberg, F., et al. (2004). The use of biological laboratory markers in the diagnosis of alcohol misuse: An evidence-based approach. *Drug and Alcohol Dependence*, 74, 273–279.
- Shipherd, J. C., Stafford, J., Tanner, J. (2005). Predicting alcohol and drug abuse in Persian Gulf War veterans: What role do PTSD symptoms play? *Addictive Behaviors*, 30, 595–599.
- SPSS Inc. (2003). *SPSS for Windows, version 12.0 [computer software]*. Chicago, IL: SPSS Inc.
- StataCorp. (2004). *Stata Statistical Software, version 8.2 [computer software]*. College Station, TX: Stata Corporation.
- Stotts, A. L., Schmitz, J. M., Grabowski, J. (2003). Concurrent treatment for alcohol and tobacco dependence: Are patients ready to quit both? *Drug and Alcohol Dependence*, 69, 1–7.
- The Iowa Persian Gulf Study Group. (1997). Self-reported illness and health status among Gulf War veterans: A population-based study. *Journal of the American Medical Association*, 277(3), 238–245.
- Tomlinson, K. L., Tate, S. R., Anderson, K. G., McCarthy, D. M., Brown, S. A. (in press). An examination of self-medication and rebound effects: Psychiatric symptomatology before and after alcohol or drug relapse. *Addictive Behaviors* (Available online 21 June 2005).
- Welte, J. W., Barnes, G. M., Wiczorek, W. F., Tidwell, M. C. (2004). Gambling participation and pathology in the United States—a sociodemographic analysis using classification trees. *Addictive Behaviors*, 29, 983–989.
- World Health Organization Collaborating Centre for Mental Health and Substance Abuse. (1997). *Composite International Diagnostic Interview: CIDI Auto version 2.1 Administrator's guide and reference*. Sydney: World Health Organization Collaborating Centre for Mental Health and Substance Abuse.
- Zhang, H., Singer, B. (1999). *Recursive partitioning in the health sciences*. New York: Springer-Verlag.

3. INTRODUCTION TO CHAPTER THREE: TEMPORAL RELATIONSHIPS BETWEEN GULF WAR DEPLOYMENT AND SUBSEQUENT PSYCHOLOGICAL DISORDERS

Over a decade after the ceasefire of the 1991 Gulf War, Australian Gulf War veterans reported higher levels of physical ³³³⁻³³⁸ and psychological ill-health, especially in regard to DSM-IV diagnoses of alcohol disorders, anxiety (including posttraumatic stress disorder (PTSD)) disorders, affective (including major depression) disorders ³²⁹, as well as self-reported PTSD, alcohol misuse ³³¹, psychological distress, and poor health-related quality of life ³³⁹, compared with the military comparison group. These results are in close agreement with the results of studies conducted overseas, including those in Canada ³⁴⁰, Denmark ³⁴¹, France ³⁴², the UK ³⁴³ and the US ³⁴⁴.

There is little doubt that going to war is a highly stressful experience, even if direct combat is not actually encountered ³⁴⁵. After an initial 'honeymoon phase' of welcome and rejoicing, the veteran's eventual return to his or her family, friends and colleagues involves a process of readjustment, which can be a stressful experience for all concerned ³⁴⁶. Although the formal study of psychological illness in serving military and veteran populations has been occurring since at least the early 1900's ³⁴⁷, there has been little research into when psychiatric disorders such as PTSD, major depression and alcohol use disorders develop, and in what order. Clearly, knowledge of the timing and co-

occurrence of such disorders is vital in order to facilitate their identification, treatment, and prevention.

3.1. CART and Logistic Regression Analysis of Temporal Relationships

The study presented in Chapter Three examined the temporal progression and ordering of psychological disorders in 1197 male RAN Gulf War veterans who had completed the CIDI, and provided information on the age of onset of symptomatology. Temporal progression was assessed using discrete time survival analysis, a statistical technique which allows for the examination of specific events, such as the onset of psychiatric symptomatology, occurring in discrete time intervals such as the number of years post-Gulf War³⁴⁸. This procedure can be carried out using standard logistic regression techniques³⁴⁸.

CART was used to examine whether there were any subgroups defined by combinations of several potential risk factors, identified in prior research, including age, military rank, education, serving status and whether or not veterans had deployed to the Gulf during the Air War, with regard to patterns of development of psychiatric disorders. Discrete time survival analysis was then employed for the further descriptive analysis of CART subgroups.

To summarise, the study presented in Chapter Two employs CART and logistic regression to identify subgroups of Australian Gulf War veterans at risk of hazardous alcohol consumption. The study presented in Chapter Three employs discrete time survival analysis and CART to examine the overall timing and

ordering of psychiatric disorders, as well as to ascertain whether there are any subgroups exhibiting particular developmental patterns of these disorders.

Declaration for Thesis Chapter 3

McKenzie DP, Creamer M, Kelsall HL, Forbes AB, Ikin JF, Sim MR, McFarlane AC. Temporal Relationships between Gulf War Deployment and Subsequent Psychological Disorders. Submitted to Journal of Affective Disorders, 11 September, 2008.

Declaration by candidate

In the case of Chapter 3, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
I made a major contribution to the research question. I was responsible for the literature review, data management and programming, statistical analyses, interpreting the results, writing the paper and undertaking any required revisions.	75%

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
M. Creamer	Major contribution to the research question. Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A
H. Kelsall	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A
A. Forbes	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper. Statistical consulting.	N/A
J. Ikin	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A (not a student at time of paper submission)
M. Sim	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A
A. McFarlane	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A

Candidate's

	Date 16-12-2008
---	-----------------

Signature

--	--

Declaration by co-authors

The undersigned hereby certify that:


- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

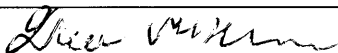
Location(s)

Monash University Department of Epidemiology and Preventive Medicine, Alfred Hospital

[Please note that the location(s) must be institutional in nature, and should be indicated here as a department, centre or institute, with specific campus identification where relevant.]

Signature 1

	Date
	09/12/2008
Signature 2	
Signature 3	15/12/08
Signature 4	10/12/08
Signature 5	17/12/2008
Signature 6	
Signature 7	

Signature		16-12-2008
-----------	--	------------

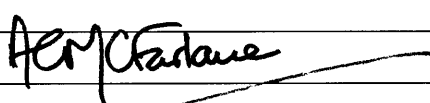
Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s)	Monash University Department of Epidemiology and Preventive Medicine, Alfred Hospital
-------------	---

[Please note that the location(s) must be institutional in nature, and should be indicated here as a department, centre or institute, with specific campus identification where relevant.]

Signature 1		Date
Signature 2		
Signature 3		
Signature 4		
Signature 5		
Signature 6		
Signature 7		10/12/2008

Temporal Relationships between Gulf War Deployment and Subsequent Psychological Disorders

Dean P. McKenzie ^{a,*}, Mark Creamer ^b, Helen L. Kelsall ^a, Andrew B. Forbes ^a, Jillian F. Ikin ^a, Malcolm R. Sim ^a, Alexander C. McFarlane ^c

^a *Department of Epidemiology & Preventive Medicine, Monash University, Melbourne, Australia*

^b *Australian Centre for Posttraumatic Mental Health, University of Melbourne*

^c *Department of Psychiatry, University of Adelaide, Adelaide, South Australia*

*Corresponding author. Monash University, Department of Epidemiology & Preventive Medicine, Alfred Hospital, Commercial Road, Melbourne, Victoria 3004, Australia.

E-mail address: dean.mckenzie@med.monash.edu.au (D.P. McKenzie)

1 table, 4 figures, approximately 5350 words

Abstract

Background: Although much has been published on the effects of the 1990/1991 Gulf War on the psychological health of veterans, few studies have addressed the pattern and timing of post-war development of psychological disorders. Our study examines the most common psychological disorder to first appear post-Gulf War, the period of peak prevalence, and the sequence of multiple psychological disorders.

Methods: The temporal progression of psychological disorders in male Australian naval Gulf War veterans with no prior psychological disorders was calculated across each year of the post-Gulf War period. DSM-IV diagnoses were obtained using the Composite International Diagnostic Interview (CIDI).

Results: Psychological disorders peaked in the first two years (1991-1992) following the Gulf War. Alcohol use disorders were the most likely to appear first. Recursive partitioning / classification and regression tree (CART) analysis found that risk of disorder was exacerbated if veterans had been exposed to a high number of potential psychological stressors during their military service. Lower military rank was associated with increased risk of alcohol disorders, particularly during the first two years post-Gulf War. In veterans with two or more disorders, anxiety disorders and alcohol disorders tended to appear before affective disorders.

Limitations: A retrospective design was used, rendering age of onset vulnerable to recall bias. The CIDI records age of onset in whole years, therefore the minimum post-Gulf War window that can be studied is one year. Only the first appearance of symptomatology is recorded, and so it was not possible to examine recurrent episodes of disorders.

Conclusions: Psychological disorders often occur in sequence following Gulf War deployment. Our findings may help clinicians to anticipate, and better manage, multiple symptomatology. The findings may also assist veteran and defense organisations in planning effective mental health screening, management, and prevention policy.

1. Introduction

Many studies have found increased levels of psychological disorders in veterans of the 1990/1991 Gulf War compared with military controls (e.g., (Fiedler et al., 2006; Ikin et al., 2004; Kelsall et al., 2006; McKenzie et al., 2004; Salamon et al., 2006)). Reviews of the Australian (Sim and Kelsall, 2006), UK (Wessely, 2004) and US (Friedman, 2005) contexts have shown considerable consistency in the frequency of these disorders post-Gulf War. However, little is known about the pattern of onset and temporal progression of single and comorbid psychological disorders in Gulf War, and other veteran, populations. Such understanding is important in order to plan for the longer term mental health outcomes of military deployments, with a view to aiding in the development of screening, treatment and intervention programs.

A small number of studies has investigated the course of psychological disorders, including comorbid ones, in veterans of the 1990/1991 Gulf War (Orcutt et al., 2004), the Vietnam war (Mellman et al., 1992), and the ongoing conflict in Iraq (Milliken et al., 2007). There has also been limited research into temporal patterns of psychological disorders in other specific occupational groups exposed to high levels of stress, such as fire fighters (McFarlane, 1988a), and in victims of traumatic injuries (Carty et al., 2006).

The findings of a majority of these studies suggest that levels of posttraumatic stress disorder (PTSD) symptoms peak approximately two (McFarlane, 1988a; Wolfe et al., 1999a) to four (Bremner et al., 1996) years after having experienced traumatic events. The findings also suggest that when PTSD and major depression are comorbid, specific symptoms of PTSD precede specific symptoms of depression.

Limitations of the above studies include small samples and / or the use of self-report measures of psychological disorders. A recent study involving large samples and formal diagnostic criteria employed the computer-assisted version of the Composite International Diagnostic Interview (CIDI) (World Health Organization Collaborating Centre for Mental Health and Substance Abuse,

1997) age of onset data to examine the prevalence of psychological disorders at three time-points; prior to the Gulf War, around the time of the Gulf War (1991-1993) and 10 years post-Gulf War (Toomey et al., 2007). Prevalences of depression, PTSD and any anxiety disorders were lower 10 years post-Gulf War than they were at the time of the war, however the course of disorders across each year of the post-Gulf War period was not examined.

Large population-based epidemiological studies can provide an important context in which to understand the results of studies of military and veteran groups. The recent US National Comorbidity Survey Replication (NCS-R) (Kessler et al., 2005) for example, found that the age of onset of anxiety disorders is usually around 11 years. Age of onset for substance disorder, however, tends to be approximately 20 years, while for affective disorders it is approximately 30 years. Importantly for the question of comorbidity, the NCS-R found that later onset disorders generally occur as secondary comorbid conditions. If anxiety disorders and affective disorders occur comorbidly for example, symptoms of anxiety will usually occur first, at least in the general population.

Our study examines the temporal progression of psychological disorders in Australian Gulf War veterans. We investigate the most common psychological disorder to first appear post-Gulf War, the period of peak prevalence, and the ordering or sequence of any multiple (comorbid) psychological disorders. As there is evidence that the development of disorders over time may differ across individuals (Orcutt et al., 2004), we also examine whether there are subgroups of veterans, defined by particular combinations of variables such as military rank, combat exposure, and education, with different patterns of development. We performed subgroup analysis using a recursive partitioning / classification and regression trees (CART) procedure (Breiman et al., 1984). CART has been employed in a variety of psychiatric applications (Craig et al., 1997; Mann et al., 2008; McKenzie et al., 1993; Schmitz et al., 2003), including those involving military personnel (Fikretoglu et al., 2006), and Gulf War veterans (McKenzie et al., 2006).

2. Methods

2.1. Recruitment and data collection

The participating Gulf War veteran study group consisted of 1456 veterans, 80.5% of those who were eligible. 1249 (85.8%) of the study group served in the Royal Australian Navy (RAN) at the time of the Gulf War, 95 (6.5%) served in the Australian Army, and 112 (7.7%) served in the Royal Australian Air Force. Comprehensive details of recruitment and demographics have been reported previously (Ikin et al., 2004) and are only briefly summarised here. Our analyses were restricted to 1,204 male Royal Australian Navy (RAN) 1990/1991 Gulf War veterans for whom CIDI DSM-IV (American Psychiatric Association, 1994) diagnostic information was available, due to there being only small numbers of females, male Air Force and male Army personnel. Seven veterans who met criteria for a post-Gulf War psychiatric diagnosis, but reported CIDI-supplied age ranges such as “between 16 and 25” due to uncertainty about age of onset, were excluded from statistical analyses.

Participants were recruited via mailed invitation in 2000-2002. At the time of the study, the study group averaged 37.4 years of age ($SD = 6.1$; range = 27 - 61), was predominantly married (895 or 74.8%), and with a slight majority not currently serving in the military (711 or 59.4%). At the time of the Gulf War, the study group was primarily of other ranks - supervisory (i.e., non-commissioned officers) (585 or 48.9%) and other ranks - non-supervisory (e.g., enlisted personnel) (433 or 36.2%).

The primary purpose of our study is to examine the effects of the Gulf War on hitherto psychologically healthy veterans. Almost two thirds (782 or 65.0%) of the study group ($n=1197$) did not meet criteria for any pre-Gulf CIDI DSM-IV diagnoses whatsoever. With regard to specific diagnostic categories 874 (73.0%) veterans did not meet DSM-IV criteria for pre-Gulf alcohol abuse or dependence, 1084 (90.6%) did not meet criteria for any pre-Gulf anxiety disorder, and 1152 (96.2%) did not meet criteria for any pre-Gulf affective disorder.

The analyses of comorbid diagnoses utilised only those 782 veterans who did not meet criteria any pre-Gulf War CIDI DSM-IV diagnoses. The analyses of broad DSM-IV diagnostic categories such as any anxiety disorder only excluded veterans for whom the onset of symptomatology for that particular category occurred prior to the Gulf War. Any veteran with an age of onset of symptomatology for a disorder that was earlier than their age (rounded up to the next highest integer) at first deployment to the Gulf War was excluded from analysis of that disorder. The maximum number of veterans in any statistical analysis was 1152, comprising those who did not meet criteria for pre-Gulf affective disorder.

2.2. Assessment

Participants were evaluated for any history of affective, alcohol, and anxiety disorders, according to CIDI DSM-IV criteria, as assessed using the computer-assisted version of the CIDI, administered by psychologists. The CIDI includes questions regarding the age, in whole years, of onset of psychological symptomatology.

2.3. Statistical analyses

The peak prevalence and temporal course of each psychological disorder were assessed using discrete time survival analysis, which can be carried out using standard logistic regression techniques (Singer and Willett, 2003). Discrete time survival analysis allows examination of the probability of events, such as the onset of psychological symptomatology, occurring in discrete time intervals such as the number of post-Gulf War years (e.g., 1991, 1992). This technique has previously been applied to the present data as part of a wider analysis of separation from military service (Creamer et al., 2006), and in other recent psychological applications (Armeli et al., 2008; Breslau et al., 2007; Degenhardt et al., 2008).

Person-years were accrued for each calendar year, and for the elapsed time since the onset of first post-Gulf War symptomatology, for each diagnosis. The number of years from 1990, the start of the Gulf War, to the onset of

psychological symptomatology, or to the end of the study (2002), whichever came first, was analysed. The maximum number of years actually observed before onset of symptomatology was 12 years (to the year 2001) in the case of any affective disorders.

In order to aid interpretation of the results of discrete time survival analysis, elapsed time is often grouped into a smaller number of categories (Singer and Willett, 2003). As there is evidence from both Gulf War (Wolfe et al., 1999b) and civilian (McFarlane, 1988a) studies that the likelihood of psychological disorders such as PTSD may be high in the first two years following traumatic events, we divided the number of years elapsed since 1990 into four phases, first 2 years (1990-1991), second 2 years (1992-1993), medium phase (1994-1997), and late phase (1998 to the end of the study).

As well as elapsed time phase, we examined several other potential risk factors for differential temporal patterns of psychological disorders, based upon our prior research (Creamer et al., 2006; Ikin et al., 2004; Kelsall et al., 2006; McKenzie et al., 2004). The variables consisted of age at the start of the Gulf War on 2nd August 1990 (divided into three categories, < 25, >25 to < 30, 30+ years), military rank at that time (other ranks - non-supervisory; other ranks - supervisory; officer; corresponding to enlisted, non-commissioned officer, officer), marital status at time of interview (married/de facto, separated/divorced/widowed, single/never married), serving status at time of interview (whether still serving in the Australian Defence Force), schooling at time of interview (10 years, 11 or 12 years, certificate / diploma, undergraduate / postgraduate degree), whether deployed in the Gulf during the Air War, and total number of potential psychological stressors experienced during military deployment, divided into three approximately equal-sized categories or tertiles (< 6, >6 to 11, > 11). In order to have been deemed present during the Air War, the deployment commencement date for a particular veteran must have been on or before the formal ceasefire on 28th February 1991, while the deployment end date had to have been on or after the Air War began on 17th January 1991.

Number of potential psychological stressors was assessed using the Military Service Experience Questionnaire (Ikin et al., 2005; Ikin et al., 2004) (MSEQ), which consists of 44 items, each describing a potentially stressful experience for Australian Gulf War veterans, such as having to board hostile ships at sea. The MSEQ total score was calculated by summing the 44 binary coded items.

Rather than attempting to statistically adjust for the effects of the potential risk factors outlined above, we examined whether there were any subgroups defined by combinations of these factors in regard to patterns of development of psychological disorders. We performed subgroup analysis using Classification and Regression Tree analysis (CART) (Breiman et al., 1984, Salford Systems, 2006), within a discrete time survival analysis framework, in a similar fashion to that used in the logistic regression analyses.

Unlike early such procedures (Morgan and Sonquist, 1963), CART explicitly validates the generality of its tree models, using by default 10-fold cross-validation (Breiman et al., 1984), and has been empirically shown to be comparable to, or more conservative than, similar techniques (Hawkins and McKenzie, 1995; Lim et al., 2000).

The subgroups of Gulf War veterans identified by CART were further analysed using logistic regression, once again within a discrete time survival analysis framework. The statistical comparison of empirically defined groups can be misleading (Austin and Goldwasser, 2008). The subsequent analysis of CART subgroups should therefore be viewed as being primarily descriptive, allowing the magnitude of the risk of developing psychological disorders to be assessed for different patterns of variables, controlling for variables available to, but not selected by, CART (McKenzie et al., 2006 ; Schmitz et al., 2003).

Individual DSM-IV diagnoses were collapsed into the broad categories of “any alcohol disorder”, “any affective disorder” and “any anxiety disorder” for purposes of statistical analyses, in view of the relatively small number of participants with individual psychological disorders commencing after the Gulf deployment.

To allow for possible imprecision in recall of age of onset (Farrer et al., 1989), statistical analyses of the sequence of onset of comorbid affective, alcohol and anxiety disorders were restricted to those veterans having symptoms of two or more disorders with onset at least one year apart. Frequencies of the ordering of each pair of diagnoses (e.g., any affective disorder followed by any anxiety disorder, versus the reverse pattern) were compared using exact binomial tests (Armitage et al., 2002), implemented in the StatXact 4 (CYTEL Software Corporation, 2000) computer program. Unless stated otherwise, all statistical analyses were conducted using SPSS, version 15 (SPSS Inc., 2006).

3. Results

3.1. Frequencies of psychological disorders to first appear

Of the 1197 eligible veterans, 415 (34.7%) developed one or more psychological disorders following the Gulf War, with 326 (27.2%) developing one or more diagnoses from the broad categories of alcohol abuse or dependence, any affective disorder, and any anxiety disorder. Alcohol abuse or dependence was the most frequent (14.6%; N=114) post-Gulf War CIDI DSM-IV diagnosis to first appear, followed in frequency by any affective disorder (9.8%, N=77), consisting mostly of major depression (9.0%, N=70), and any anxiety disorder (5.9%, N=46), consisting mostly of PTSD (4.0%, N=31).

3.2. Development of psychological disorders: overall

As shown in Table 1, the prevalence of onset of symptomatology of each broad DSM-IV diagnostic category peaks in the first two years following the Gulf War, and then subsides. This pattern is particularly noticeable in the case of alcohol disorders, with an initial rate of 4.1 per 100 person years in the first two years post-Gulf War, sharply dropping to just over half this value in the next two

years, whereas the fall in rates of development of affective disorders across the post-Gulf War period was more gradual. The overall effect of time phase was statistically significant for alcohol disorders ($p < 0.001$) and any anxiety disorder ($p < 0.001$), and approached statistical significance for any affective disorder ($p = 0.06$).

[insert Table 1 about here]

3.3. Development of psychological disorders: subgroup analysis

As we were specifically interested in the interrelationships between time phase and other variables such as military rank and military experience, phase was manually selected to be the first splitting variable in each tree.

In regard to alcohol disorders, as shown in Figure 1 CART split phase into the first category, representing the first two years post-Gulf War, versus the rest. Veterans who developed disorders in the first two years were then split by categorised MSEQ score. Those veterans with MSEQ scores in the first or second tertiles were then split by military rank, into other ranks - non-supervisory (enlisted) versus higher ranks. Neither subgroup was split any further. For post-Gulf War periods of longer than two years, CART also split the sample by military rank, using the same split-point as above.

CART readily allows the identification of those subgroups at the greatest risk of developing post-Gulf War psychological disorders. Discrete time survival analysis, performed using logistic regression, was used to compare each CART subgroup with a reference subgroup with the lowest rate of disorders, adjusting for potential demographic and military service confounders. The reference group exhibited a rate of 0.8 alcohol disorders per 100 person years. The highest rate of DSM-IV alcohol disorders, a 10-fold increase relative to that for the reference group, was observed in the first two years post-Gulf War for those veterans with a MSEQ score in the highest tertile (11 or greater) (rate per 100 person years = 8.4, OR = 12.10, adjusted OR = 10.02, 95% CI = 6.15 to 16.31). The next highest rates, relative to the reference group, were in the first two years post-Gulf War for those veterans with an MSEQ score in the first and second tertiles and of

other, non-supervisory (enlisted) rank (rate = 4.5, OR = 6.13, adjusted OR = 4.73, 95% CI = 2.49 to 9.01); and three years post-Gulf War or later of other - non-supervisory rank, regardless of MSEQ score (rate = 2.8, OR = 3.73, adjusted OR = 2.89, 95% CI = 1.72 to 4.85). The rate of alcohol use disorders for the higher ranks, in the first two years, was similar to that for the reference group (rate = 1.1, OR = 1.52, adjusted OR = 1.55, 95% CI = 0.74 to 3.25).

[insert Figure 1 about here]

The CART tree for any affective disorders is shown in Figure 2. CART split phase into eight years or less post-Gulf War, versus nine or more years. The earlier phase was then split by categorised MSEQ into the highest tertile versus the rest. The later phase was split by marital status into separated/divorced/widowed versus the rest. No subgroups were further split.

Two subgroups - nine or more years post-Gulf War, and married/de facto/single, and eight years or less, with low or medium MSEQ scores, had the lowest rates of any affective disorders (1.0 per 100 person years). The former subgroup was arbitrarily designated as the reference group, and did not significantly differ in rate from the latter (rate = 1.0, OR = 1.05, adjusted OR = 1.04, 95% CI = 0.69 – 1.60). The highest rate of any affective disorders, relative to the reference group, was observed in the first eight years post-Gulf, accompanied by MSEQ scores in the highest tertile (rate = 3.7, OR = 3.93, adjusted OR = 3.81, 95% CI = 2.55 to 5.70), followed by the nine or more years period, accompanied by separated/divorced/widowed marital status (rate = 3.3, OR = 3.47, adjusted OR = 3.81, 95% CI = 2.55 to 5.70).

[insert Figure 2 about here]

As shown in Figure 3, the CART tree for any anxiety disorders firstly split phase into two years or less post-Gulf War versus more than two years. Both the early and later phases were then split by categorised MSEQ, into the highest tertile versus the rest. Three years and later, post-Gulf War phases, accompanied by low or medium MSEQ scores, exhibited the lowest rate of any anxiety disorder (0.2 per 100 person years). The highest rate of any anxiety

disorder, relative to the above reference group was observed in the first two years post-Gulf, accompanied by MSEQ scores in the highest tertile (rate = 4.5, OR = 19.77, adjusted OR = 20.21, 95% CI = 10.74 to 38.04), followed by the later post-Gulf War period, accompanied by MSEQ scores in the same tertile (rate = 1.0, OR = 4.21, adjusted OR = 4.41, 95% CI = 2.31 to 8.37). The first two years post-Gulf, accompanied by MSEQ scores in the first or second tertiles, exhibited a higher rate of any anxiety disorder than the reference group, but this difference was not statistically significant (rate = 1.0, OR = 1.94, adjusted OR = 1.93, 95% CI = 0.79 to 4.70).

[insert Figure 3 about here]

3.4. Comorbidity of psychological disorders

66 veterans met criteria for more than one DSM-IV diagnosis with onset after the Gulf War. Of these veterans, 40 had comorbid diagnoses with onset more than one year apart, consisting of 26 veterans meeting diagnostic criteria for two disorders, and 14 veterans meeting criteria for all three broad diagnostic categories. As shown in Figure 4, 16 veterans developed an anxiety disorder and an affective disorder concurrently (i.e., within one year of each other). Of the 15 veterans who met criteria for both of the above disorders with onset at least one year apart, a higher number developed the anxiety disorder, followed by the affective disorder (86.7%, N = 13), than the reverse ordering (13.3%, N = 2). This difference was statistically significant ($p = 0.007$).

16 veterans developed an alcohol use disorder and an affective disorder concurrently. Of the 20 veterans who developed these disorders with onset more than one year apart, more veterans developed the alcohol use disorder before the affective disorder (80%, N = 16) than the reverse ordering (20%, N = 4). This difference was statistically significant ($p = 0.012$). Finally, 8 veterans developed an alcohol use disorder and an anxiety disorder concurrently, while 10 veterans developed the two disorders more than one year apart. Within the latter group, developing the anxiety disorder before the alcohol use disorder (60%, N = 6) was

slightly more frequent than the reverse ordering (40%, $N = 4$). This difference was not statistically significant however ($p = 0.75$).

[insert Figure 4 about here]

4. Discussion

4.1 Onset of psychological disorders

Our finding that onset of psychological disorders is greatest in the first two years post-Gulf War and then subsides is consistent with previous research demonstrating that psychological symptoms peak in the first few years after traumatic exposure such as combat (Mellman et al., 1992; Milliken et al., 2007; Orcutt et al., 2004; Toomey et al., 2007), or natural disasters (McFarlane, 1988a). This pattern is also consistent with the suggestion that deployment to the 1990/1991 Gulf War represented a major psychosocial stressor, contributing to the subsequent onset of psychological disorders in a significant minority of deployed personnel (Friedman, 2005).

CART analyses found that the risk of developing psychological disorders during the first two years is greatest for those veterans who have experienced a high (> 11) number of potential psychological stressors during military service, as measured by the MSEQ. This result is in keeping with the results of other studies of Gulf War (Ikin et al., 2005; King et al., 2008; Orcutt et al., 2004; Stein et al., 2005) and other (Fontana and Rosenheck, 1994) veterans showing increased likelihood of PTSD and other psychological disorders following increased combat exposure or perceived threat or fear of harm. However, there is also evidence that the relationship may proceed in the other direction, with the severity of PTSD symptoms influencing recall of combat exposure (Koenen et al., 2007; McFarlane, 1988b; Southwick et al., 1997).

CART also found that veterans with lower numbers of potential psychological stressors, but of lower ranks (other ranks - non-supervisory) also exhibited increased risk of alcohol use disorders, compared with higher ranks. Previous analysis of these data showed an increased risk of hazardous drinking

behavior amongst lower ranks (McKenzie et al., 2006), and so it is possible that excessive drinking is more common amongst this group, or more likely to be reported, than amongst the higher ranks. It has been suggested that military rank is a proxy for education (Ismail et al., 2000), however analysis of the above subgroups statistically controlled for differences in level of schooling. Those of higher rank may have greater access to support services for alcohol problems, while high levels of drinking may be regarded as being more acceptable among those of lower rank.

The CART results also suggested that the rate of affective disorders increased mainly after eight years, particularly for veterans who are divorced or separated. Marital breakdown is an established risk factor for affective and other disorders in the general (Rotermann, 2007), as well as Gulf War veteran (Fiedler et al., 2006) populations, while there is also evidence that combat exposure can itself lead to marital breakdown (Prigerson et al., 2004).

The above results indicate that there are subgroups of veterans in regard to the development of disorders, providing support for earlier research (Orcutt et al., 2004), with important implications for treatment and identification, rather than adopting a 'one size fits all' approach.

Although earlier studies of Gulf War and other veterans have found PTSD to be the first disorder to develop following combat (Mellman et al., 1992), we found alcohol disorders to be the first to develop followed in terms of frequency by any affective disorders, and then any anxiety disorders, including PTSD. It has recently been pointed out that PTSD is often not the most prevalent or severe psychological disorder to be found in military personnel, and that more attention should be paid to depression and alcohol disorders (Wessely et al., 2005). Our finding must be seen in the context of the study population – predominantly young, male naval personnel – and a culture in which heavy alcohol use was not unusual, particularly amongst the lower ranks.

It is very difficult to prove causal relationships between psychological disorders, especially those relationships involving alcohol abuse or dependence

(Brady et al., 2007), as the presence of the latter may mask other psychological symptoms (McFarlane, 1998). Nevertheless, the above finding raises the important possibility that alcohol was used in the hope of reducing unpleasant psychological reactions (Armeli et al., 2008) and that it may have masked or delayed the onset of other DSM-IV Axis I conditions.

4.2. Comorbidity of psychological disorders

A slightly different picture emerges when only those individuals with more than one condition are examined. In those cases, the pattern is consistent with the broader community data. In general, anxiety disorders appear first, followed by substance use disorders, followed by affective disorders. In Gulf War veterans with both anxiety and affective disorders, for example, anxiety disorders tended to precede affective disorders. Although we did not include PTSD as a separate diagnosis, in order to maximise numbers, over two thirds of those with an anxiety disorder had PTSD. Our results are thus consistent with those found in civilian (Kessler et al., 2005; McFarlane, 1988a), and other veteran studies (Erickson et al., 2001; Mellman et al., 1992) suggesting that the onset of PTSD precedes the onset of depression.

The numbing symptoms of PTSD – which tend to overlap most strongly with the symptoms of depression – tend to be characteristic of more chronic forms of the condition (Breslau et al., 2005; Gamez et al., 2007). Research with civilian populations has also shown that, as time progresses, the two conditions become almost indistinguishable (O'Donnell et al., 2004). Our finding may be a reflection of increased levels of depressive symptoms in more chronic forms of PTSD and other anxiety disorders. Finally, PTSD and other anxiety disorders are associated with avoidance, including withdrawal from social situations. Such withdrawal, leading to isolation and loneliness, may, in turn, increase depressive symptomatology, as recently examined in other studies of Gulf War veteran psychological health (Erickson et al., 2001; Toomey et al., 2007).

The sequence of diagnoses is less clear in regard to anxiety and alcohol use. Although onset of anxiety disorders tended to precede onset of alcohol disorders

in veterans with both disorders, this finding was not statistically significant. More consistent was the finding that onset of alcohol disorders precedes the onset of affective disorders which once again is consistent with the broader epidemiological data. One possible interpretation could be that the social, physiological and behavioural problems encountered as a result of alcohol use disorders lead to, or modify, the subsequent onset of depression (Vaillant, 1995). Conversely, the use of alcohol could delay reaching the diagnostic threshold of a depressive disorder by, for example, assisting with sleep.

4.4. Strengths and limitations

A limitation of the present study is the reliance on retrospective information on onset of psychological symptoms, a methodology that is known to be problematic (Jorm, 2006), although by no means uncommon in epidemiological research (Degenhardt et al., 2008). Reference to specific life events such as getting married demonstrably aids recall (Kessler et al., 2005), which may partly explain why incidences of psychological disorders are elevated close to the time of the Gulf War. Veterans may simply be placing onset of symptoms in terms of such a major event, although this does not explain why studies using sophisticated recall prompting techniques (Bremner et al., 1996), or prospective designs (Solomon and Mikulincer, 2006), also found increased psychological symptomatology in the first few years following combat.

Both retrospective and prospective studies of psychological disorders also face the potential problem that the severity and development of such disorders may themselves be influenced by major events such as those of the 11th September, 2001 (9/11), leading to a rise or 'spike' in incidence rates (Weissman et al., 2003). It is possible that events such as 9/11, and their media coverage, may be associated with the severity and development of psychological disorders in the present study, although only symptomatology associated with any affective disorder was reported after 2000.

Another limitation of our study is that the CIDI measures age at onset in whole years. It was therefore not possible to examine whether psychological

disorders developed in the first few weeks or months after the Gulf War, although such a precision is difficult to obtain retrospectively (Farrer et al., 1989). A further limitation was that it was not possible to study multiple episodes of a specific disorder using the CIDI-derived age of onset, as information is only obtained for the onset of the first symptoms for each psychological disorder. The above limitations also apply to other studies that employ CIDI age of onset information (Toomey et al., 2007).

Finally, since our aim was to investigate the sequencing of psychological disorders developing since the Gulf War in hitherto psychologically healthy veterans, we did not investigate the sequence of comorbid symptomatology in those veterans who met DSM-IV diagnostic criteria prior to the Gulf War. Psychopathology following combat and other traumatic events may be amplified when there is a pre-existing psychological condition (Toomey et al., 2007), and so this may be something to be addressed in other studies. This would have been difficult to address here in that only a little over a third of our study group met criteria for any DSM-IV diagnosis before the Gulf War, and only 11.5% met criteria for any anxiety disorder or any affective disorder before the Gulf War (the remainder being substance use disorders).

Two major strengths of our study are the use of a large and comparatively homogenous population, and a comprehensive instrument with formal diagnostic criteria, in the form of the CIDI. In addition, our CART analysis which unlike other computer intensive model search and clustering techniques, includes in-built cross-validation, suggests the presence of at risk subgroups, particularly amongst those veterans with high numbers of potential psychological stressors, within the first two years. In the later years following the Gulf War, low military rank at time of Gulf War appears to be a specific risk factor for alcohol disorders, while being divorced or separated is a specific risk factor for affective disorders, although it is difficult to establish the direction of this association, as information on time of marital breakdown was not available.

4.5. Conclusions

In conclusion, we found clear patterns in the development of psychological disorders in Gulf War veterans following their deployment. For a proportion of individuals, especially those experiencing a large number of potential military stressors, or of lower military rank, or who have experienced a marital breakdown, these problems are likely to continue during the development of other, possibly more treatment resistant, Axis I conditions. Alcohol abuse problems following deployment should not be ignored or explained away as a military discipline problem. Indeed, it has recently been shown that US combat veterans frequently report alcohol problems and yet are very rarely referred for treatment (Milliken et al., 2007).

Although anxiety disorders develop less frequently in the first few post-Gulf War years than alcohol use disorders, our results suggest that the former are likely to progress to, or accompany additional and more complex psychiatric problems such as major depression, which itself tends to become more severe over time (Kessing, 2008).

In the majority of cases with multiple conditions, anxiety disorders were the first to appear. Although the utility of psychological screening being applied after, and particularly before, traumatic events such as combat is a contentious issue (Rona et al., 2006; Wessely, 2003), it is important that routine health checks among military personnel pay due attention to anxiety, as well as alcohol misuse, symptoms. Assuming that such symptoms would be reported, which might be problematic within a military culture (Rona et al., 2004), early identification and referral for treatment might be facilitated. Persons with PTSD subsequently followed by alcohol dependence, for example, have been shown to respond better to cognitive-behavioural therapy than those with alcohol dependence subsequently followed by PTSD (Back et al., 2005).

The findings of our study could be used to inform planning for the level and type of health services required in future for the 1990/1991 Gulf War veteran community, as well as other veteran populations. The results also have broader

implications for our general understanding of the mechanisms underlying the development of psychological disorders over time, especially those following stressful events.

References

- American Psychiatric Association, 1994. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*. American Psychiatric Association, Washington, DC.
- Armeli, S., Todd, M., Conner, T.S., Tennen, H., 2008. Drinking to cope with negative moods and the immediacy of drinking within the weekly cycle among college students. *Journal of Studies on Alcohol and Drugs* 69, 313-322.
- Armitage, P., Berry, G., Matthews, J.N.S., 2002. *Statistical methods in medical research*. Blackwell, Oxford, UK.
- Austin, P.C., Goldwasser, M.A., 2008. Pisces did not have increased heart failure: data driven comparison of binary proportions between levels of a categorical variable can result in increased significance levels. *J. Clin. Epidemiol.* 61, 295-300.
- Back, S.E., Jackson, J.L., Sonne, S., Brady, K.T., 2005. Alcohol dependence and posttraumatic stress: differences in clinical presentation and response to cognitive-behavioral therapy by order of onset. *J. Subst. Abuse Treat.* 29, 29-37.
- Brady, K.T., Tolliver, B.K., Verduin, M.L., 2007. Alcohol use and anxiety: diagnostic and management issues. *Am. J. Psychiatry* 164, 217-221.
- Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 1984. *Classification and regression trees*. Wadsworth, Belmont, CA.
- Bremner, J.D., Southwick, S.M., Darnell, A., Charney, D.S., 1996. Chronic PTSD in Vietnam combat veterans: course of illness and substance abuse. *Am. J. Psychiatry* 153, 369-375.
- Breslau, J., Aguilar-Gaxiola, S., Borges, G., Kendler, K.S., Su, M., Kessler, R.C., 2007. Risk for psychiatric disorder among immigrants and their US-born descendants: evidence from the National Comorbidity Survey Replication. *J. Nerv. Ment. Dis.* 195, 189-195.
- Breslau, N., Reboussin, B.A., Anthony, J.C., Storr, C.L., 2005. The structure of posttraumatic stress disorder: latent class analysis in 2 community samples. *Arch. Gen. Psychiatry* 62, 1343-1351.
- Carty, J., O'Donnell, M.L., Creamer, M., 2006. Delayed-onset PTSD: a prospective study of injury survivors. *J. Affect. Disord* 90, 257-261.
- Craig, T.J., Siegel, C., Hopper, K., Lin, S., Sartorius, N., 1997. Outcome in schizophrenia and related disorders compared between developing and developed countries: a recursive partitioning reanalysis of the WHO. *Br. J. Psychiatry* 170, 229-233.

- Creamer, M., Carboon, I., Forbes, A.B., McKenzie, D.P., McFarlane, A.C., Kelsall, H.L., Sim, M., 2006. Psychiatric disorder and separation from military service: A 10 year retrospective study. *Am. J. Psychiatry* 163, 733-734.
- CYTEL Software Corporation, 2000. StatXact 4 for Windows [computer software]. CYTEL Software Corporation, Cambridge, Massachusetts.
- Degenhardt, L., Chiu, W.T., Conway, K., Dierker, L., Glantz, M., Kalaydjian, A., Merikangas, K., Sampson, N., Swendsen, J., Kessler, R.C., 2008. Does the 'gateway' matter? Associations between the order of drug use initiation and the development of drug dependence in the National Comorbidity Study Replication. *Psychol. Med.*, 1-11.
- Erickson, D.J., Wolfe, J., King, D.W., King, L.A., Sharkansky, E.J., 2001. Posttraumatic stress disorder and depression symptomatology in a sample of Gulf War veterans: A prospective analysis. *J. Consult. Clin. Psychol.* 69, 41-49.
- Farrer, L.A., Florio, L.P., Bruce, M.L., Leaf, P.J., Weissman, M.M., 1989. Reliability of self-reported age at onset of major depression. *J. Psychiatr. Res.* 23, 35-47.
- Fiedler, N., Ozakinci, G., Hallman, W., Wartenberg, D., Brewer, N.T., Barrett, D.H., Kipen, H.M., 2006. Military deployment to the Gulf War as a risk factor for psychiatric illness among US troops. *Br. J. Psychiatry*, 453-459.
- Fikretoglu, D., Brunet, A., Schmitz, N., Guay, S., Pedlar, D., 2006. Posttraumatic stress disorder and treatment seeking in a nationally representative Canadian military sample. *J. Trauma. Stress* 19, 847-858.
- Fontana, A., Rosenheck, R., 1994. Traumatic war stressors and psychiatric symptoms among World War II, Korean and Vietnam War veterans. *Psychol. Aging* 9, 27-33.
- Friedman, M.J., 2005. Veterans' mental health in the wake of war. *N. Engl. J. Med.* 352, 1287-1290.
- Gamez, W., Watson, D., Doebbeling, B.N., 2007. Abnormal personality and the mood and anxiety disorders: implications for structural models of anxiety and depression. *J. Anxiety Disord.* 21, 526-539.
- Hawkins, D.M., McKenzie, D.P., 1995. A data-based comparison of some recursive partitioning procedures. Statistical Computing Section, American Statistical Association. American Statistical Association, Raleigh, North Carolina, pp. 245-252.
- Ikin, J.F., McKenzie, D.P., Creamer, M.C., McFarlane, A.C., Kelsall, H.L., Glass, D.C., Forbes, A.B., Horsley, K.W.A., Harrex, W.K., Sim, M.R., 2005. War zone stress without direct combat: the Australian naval experience of the Gulf War. *J. Trauma. Stress* 18, 193-204.
- Ikin, J.F., Sim, M.R., Creamer, M.C., Forbes, A.B., McKenzie, D.P., Kelsall, H.L., Glass, D.C., McFarlane, A.C., Abramson, M.J., Ittak, P., Dwyer, T., Blizzard, L., Delaney, K.R., Horsley, K.W.A., Harrex, W.K., Schwarz, H., 2004. War-related psychological stressors and risk of psychological disorders in Australian veterans of the 1991 Gulf War. *Br. J. Psychiatry* 185, 116-126.

- Ismail, K., Blatchley, N., Hotopf, M., Hull, L., Palmer, I., Unwin, C., David, A., Wessely, S., 2000. Occupational risk factors for ill health in Gulf veterans of the United Kingdom. *J. Epidemiol. Community Health* 54, 834-838.
- Jorm, A.F., 2006. National surveys of mental disorders: are they researching scientific facts or constructing useful myths? *Aust. N. Z. J. Psychiatry* 40, 830-834.
- Kelsall, H., Sim, M., McKenzie, D., Forbes, A., Leder, K., Glass, D., Ikin, J., McFarlane, A., 2006. Medically evaluated psychological and physical health of Australian Gulf War veterans with chronic fatigue. *J. Psychosom. Res.* 60, 575-584.
- Kessing, L.V., 2008. Severity of depressive episodes during the course of depressive disorder. *Br. J. Psychiatry* 192, 290-293.
- Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Arch. *Gen. Psychiatry* 62, 593-602.
- King, L.A., King, D.W., Bolton, E.E., Knight, J.A., Vogt, D.S., 2008. Risk factors for mental, physical, and functional health in Gulf War veterans. *J. Rehabil. Res. Dev.* 45, 395-408.
- Koenen, K.C., Stellman, S.D., Dohrenwend, B.P., Sommer, J.F., Stellman, J.M., 2007. The consistency of combat exposure reporting and course of PTSD in Vietnam War veterans. *J. Trauma. Stress* 20, 3-13.
- Lim, T.-S., Loh, W.-Y., Shih, Y.-S., 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning* 40, 203-228.
- Mann, J.J., Ellis, S.P., Waternaux, C.M., Liu, X., Oquendo, M.A., Malone, K.M., Brodsky, B.S., Haas, G.L., Currier, D., 2008. Classification trees distinguish suicide attempters in major psychiatric disorders: a model of clinical decision making. *J. Clin. Psychiatry*, e1-e9.
- McFarlane, A.C., 1988a. The aetiology of posttraumatic stress disorders following a natural disaster. *Br. J. Psychiatry* 152, 116-121.
- McFarlane, A.C., 1988b. The longitudinal course of post-traumatic morbidity: the range of outcomes and predictors. *J. Nerv. Ment. Dis.* 176, 30-39.
- McFarlane, A.C., 1998. Epidemiological evidence about the relationship between PTSD and alcohol use: the nature of the association. *Addict. Behav.* 23, 813-825.
- McKenzie, D.P., Ikin, J.F., McFarlane, A.C., Creamer, M., Forbes, A.B., Kelsall, H.L., Glass, D.C., Ittak, P., Sim, M.R., 2004. Psychological health of Australian veterans of the 1991 Gulf War : An assessment using the SF-12, GHQ-12 and PCL-S. *Psychol. Med.* 34, 1419-1430.
- McKenzie, D.P., McFarlane, A.C., Creamer, M., Ikin, J., Forbes, A.B., Kelsall, H.L., Clarke, D.M., Glass, D.C., Ittak, P., Sim, M.R., 2006. Hazardous or harmful alcohol use in Royal Australian Navy veterans of the 1991 Gulf War: identification of high risk subgroups. *Addict. Behav.* 31, 1683-1694.
- McKenzie, D.P., McGorry, P.D., Wallace, C.S., Low, L.H., Copolov, D.L., Singh, B.S., 1993. Constructing a minimal diagnostic decision tree. *Methods Inf. Med.* 32, 161-166.

- Mellman, T.A., Randolph, C.A., Brawman-Mintzer, O., Flores, L.P., Milanese, F.J., 1992. Phenomenology and course of psychiatric disorders associated with combat-related posttraumatic stress disorder. *Am. J. Psychiatry* 149, 1568-1574.
- Milliken, C.S., Auchterlonie, J.L., Hoge, C.W., 2007. Longitudinal assessment of mental health problems among active and reserve component soldiers returning from the Iraq war. *Journal of the American Medical Association* 298, 2141-2148.
- Morgan, J.A., Sonquist, J.N., 1963. Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association* 58, 415-434.
- O'Donnell, M.L., Creamer, M., Pattison, P., 2004. Posttraumatic stress disorder and depression following trauma: understanding comorbidity. *American Journal of Psychiatry* 161, 1390-1396.
- Orcutt, H.K., Erickson, D.J., Wolfe, J., 2004. The course of PTSD symptoms among Gulf War veterans: A growth mixture modeling approach. *J. Trauma. Stress* 17, 195-202.
- Prigerson, H.G., Maciejewski, P.K., Rosenheck, R.A., 2004. Population attributable fractions of psychiatric disorders and behavioral outcomes associated with combat exposure among US men. *Am. J. Public Health* 92, 59-63.
- Rona, R.J., Hooper, R., Jones, M., Hull, L., Browne, T., Horn, O., Murphy, D., Hotopf, M., Wessely, S., 2006. Mental health screening in armed forces before the Iraq war and prevention of subsequent morbidity: follow-up study. *Br. Med. J.* 333, 991-995.
- Rona, R.J., Jones, M., French, C., Hooper, R., Wessely, S., 2004. Screening for physical and psychological illness in the British Armed Forces: I: The acceptability of the programme. *J. Med. Screen.* 11, 148-152.
- Rotermann, M., 2007. Marital breakdown and subsequent depression. *Health Rep.* 18, 33-44.
- Salamon, R., Verret, C., Jutand, M.A., Begassat, M., Laoudj, F., Conso, F., Brochard, P., 2006. Health consequences of the first Persian Gulf War on French troops. *Int. J. Epidemiol.* 35, 479-487.
- Salford Systems, 2006. CART for Windows, version 6 [computer software]. Salford Systems, San Diego, California.
- Schmitz, N., Kugler, J., Rollnik, J., 2003. On the relationship between neuroticism, self-esteem, and depression: results from the National Comorbidity survey. *Compr. Psychiatry* 44, 169-176.
- Sim, M., Kelsall, H., 2006. Gulf War illness: a view from Australia. *Philosophical Transactions of the Royal Society of London. B* 361, 619-626.
- Singer, J.D., Willett, J.B., 2003. *Applied longitudinal data analysis : modeling change and event occurrence*. Oxford University Press, New York.
- Solomon, Z., Mikulincer, M., 2006. Trajectories of PTSD: a 20-year longitudinal study. *Am. J. Psychiatry* 163, 659-666.
- Southwick, S.M., Morgan, C.A., III., Nicolaou, A.L., Charney, D.S., 1997. Consistency of memory for combat-related traumatic events in veterans of Operation Desert Storm. *Am. J. Psychiatry* 154, 173-177.

- SPSS Inc., 2006. SPSS for Windows, version 15.0 [computer software]. SPSS Inc., Chicago, Illinois.
- Stein, A.L., Tran, G.Q., Lund, L.M., Haji, U., Dashevsky, B.A., Baker, D.G., 2005. Correlates for posttraumatic stress disorder in Gulf War veterans: a retrospective study of main and moderating effects. *J. Anxiety Disord.* 19, 861-876.
- Toomey, R., Kang, H.K., Karlinsky, J., 2007. Mental health of US Gulf War veterans 10 years after the war. *Br. J. Psychiatry* 190, 385-393.
- Vaillant, G.E., 1995. The natural history of alcoholism revisited. Harvard University Press, Cambridge, Massachusetts.
- Weissman, E.M., Kushner, M., Marcus, S.M., Davis, D.F., 2003. Volume of VA patients with posttraumatic stress disorder in the New York metropolitan area after September 11. *Psychiatr. Serv.* 54, 1641-1643.
- Wessely, S., 2003. The role of screening in the prevention of psychological disorders arising after major trauma : pros and cons. In: Ursano, R.J., Fullerton, C.S. and Norwood, A.E. (Eds.), *Terrorism and disaster : Individual and community mental health interventions*. Cambridge University Press, Cambridge, pp. 121-145.
- Wessely, S., 2004. The long aftermath of the 1991 Gulf War. *Ann. Intern. Med.* 141, 155-156.
- Wessely, S., Hyams, K.C., Rona, R.J., 2005. Screening for psychological illness in the military [reply]. *Journal of the American Medical Association* 294, 43.
- Wolfe, J., Erickson, D.J., Sharkansky, E.J., King, D.W., King, L.A., 1999a. Course and predictors of posttraumatic stress disorder among Gulf War veterans: A prospective analysis. *J. Consult. Clin. Psychol.* 67, 520-528.
- Wolfe, J., Proctor, S.P., Erickson, D.J., Heeren, T., Friedman, M.J., Huang, M.T., Sutker, P.B., Vasterling, J.J., White, R.F., 1999b. Relationship of psychiatric status to Gulf War veterans' health problems. *Psychosom. Med.* 61, 532-540.
- World Health Organization Collaborating Centre for Mental Health and Substance Abuse, 1997. Composite International Diagnostic Interview: CIDI-Auto 2.1 - Administrator's guide and reference. World Health Organization Collaborating Centre for Mental Health and Substance Abuse, Sydney.

Table 1. Discrete time survival analysis of four time phases since Gulf War and onset of psychological disorder in male Royal Australian Navy (RAN) Gulf War veterans (N=1197)

DSM-IV diagnosis	Phase since Gulf War	n^a	Rate (per 100 person years)	OR	95% CI	p value
Any Affective Disorder (including Major Depression)	1-2 years	47	2.1	1.0		
	3-4 years	36	1.6	0.8	0.5-1.2	
	5-8 years	70	1.7	0.8	0.6-1.2	
	9-12 years	46	1.2	0.6	0.4-0.9	
	diagnosis absent	953				0.057 ^b
Any Anxiety Disorder (including PTSD)	1-2 years	35	1.6	1.0		
	3-4 years	6	0.3	0.2	0.1-0.4	
	5-8 years	25	0.6	0.4	0.2-0.6	
	9-12 years	9	0.3	0.2	0.1-0.4	
	diagnosis absent	1009				< 0.001 ^b
Alcohol Abuse or Dependence	1-2 years	71	4.1	1.0		
	3-4 years	35	2.2	0.5	0.3-0.8	
	5-8 years	38	1.3	0.3	0.2-0.4	
	9-12 years	23	1.1	0.3	0.2-0.4	
	diagnosis absent	707				< 0.001 ^b

^a Numbers of Gulf War veterans who first met diagnostic criteria within the specified time period.

^b Omnibus test of time phase since Gulf War, i.e. a test of whether odds ratios for each time phase are equal to one.

Figure 1. Recursive partitioning /classification and regression tree (CART) analysis of onset of DSM-IV Alcohol Abuse or Dependence in male Royal Australian Navy (RAN) Gulf War veterans

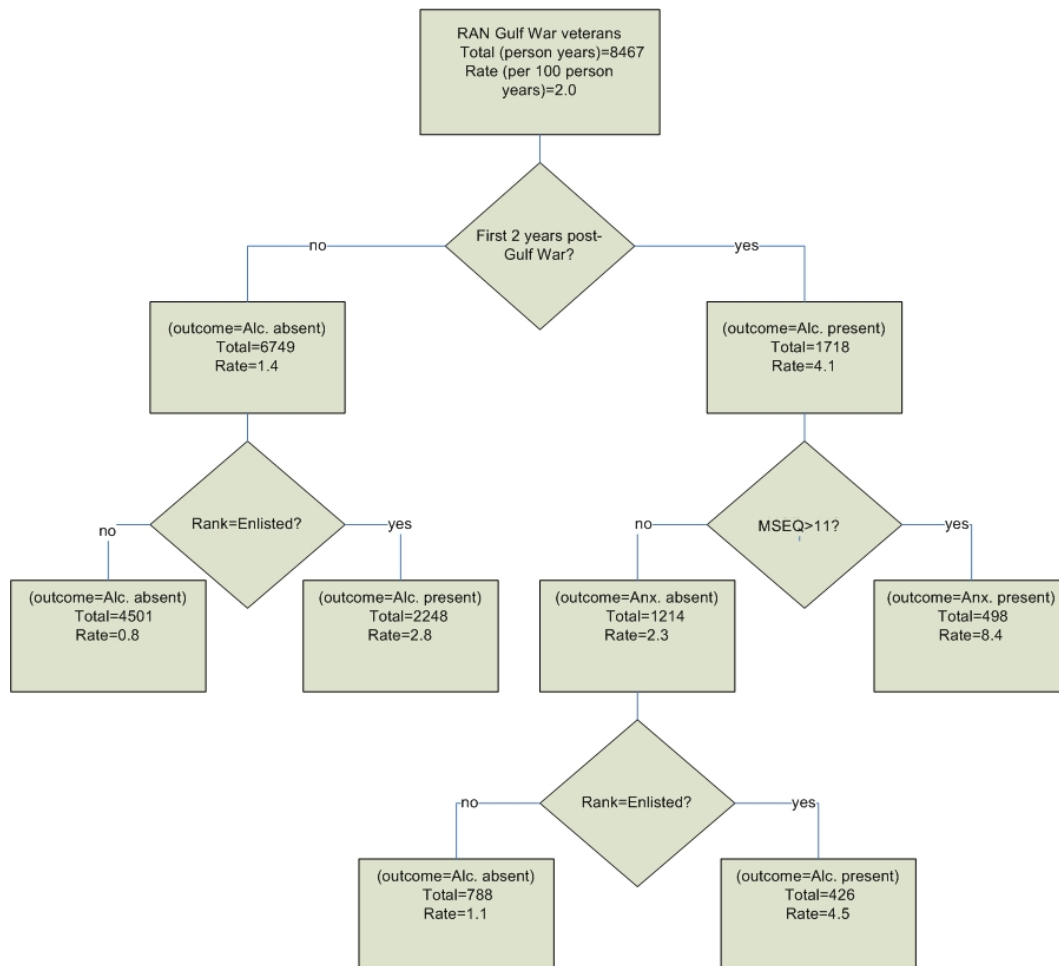


Figure 2. *Recursive partitioning /classification and regression tree (CART) analysis of onset of DSM-IV Any Affective Disorders in male Royal Australian Navy (RAN) Gulf War veterans*

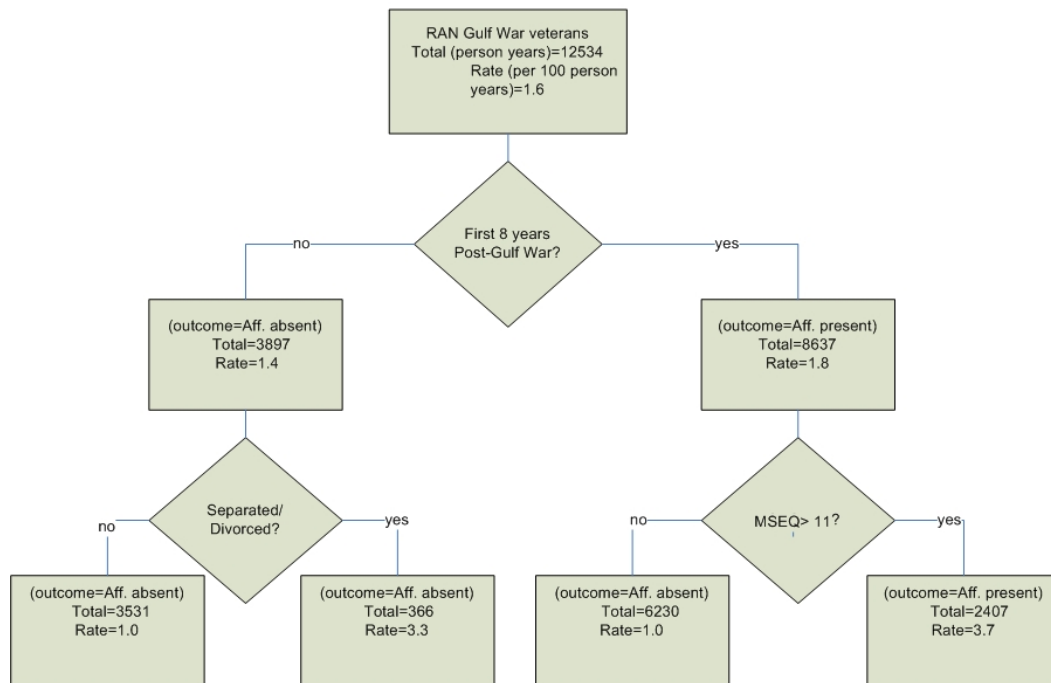


Figure 3. *Recursive partitioning /classification and regression tree (CART) analysis of onset of DSM-IV Any Anxiety Disorder in male Royal Australian Navy (RAN) Gulf War veterans*

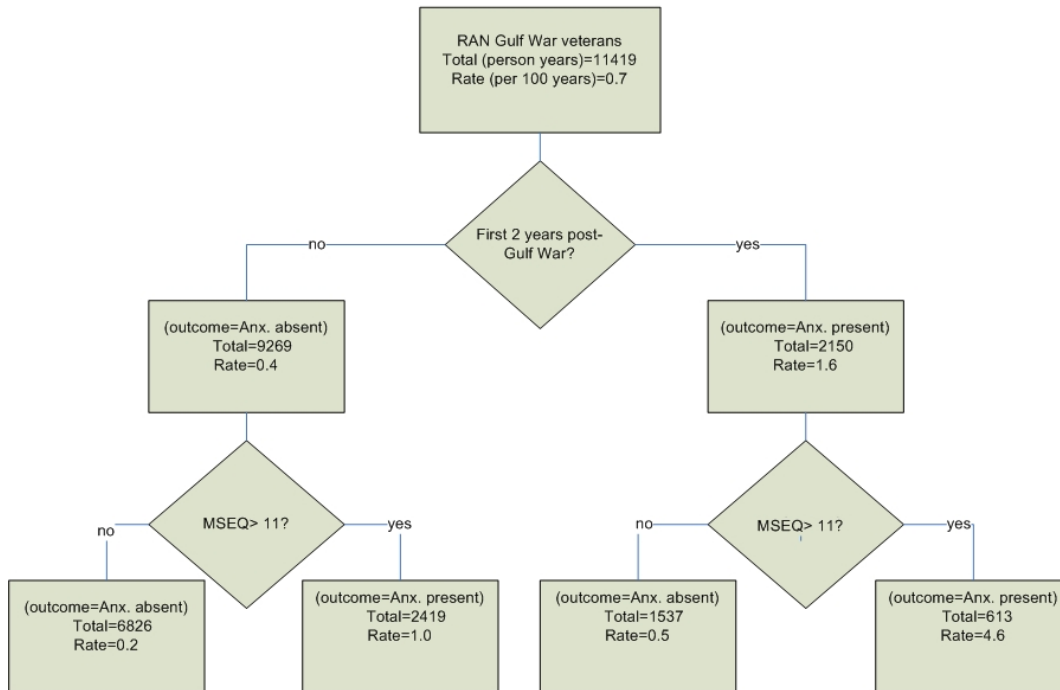


Figure 3. *Recursive partitioning /classification and regression tree (CART) analysis of onset of DSM-IV Any Anxiety Disorder in male Royal Australian Navy (RAN) Gulf War veterans*

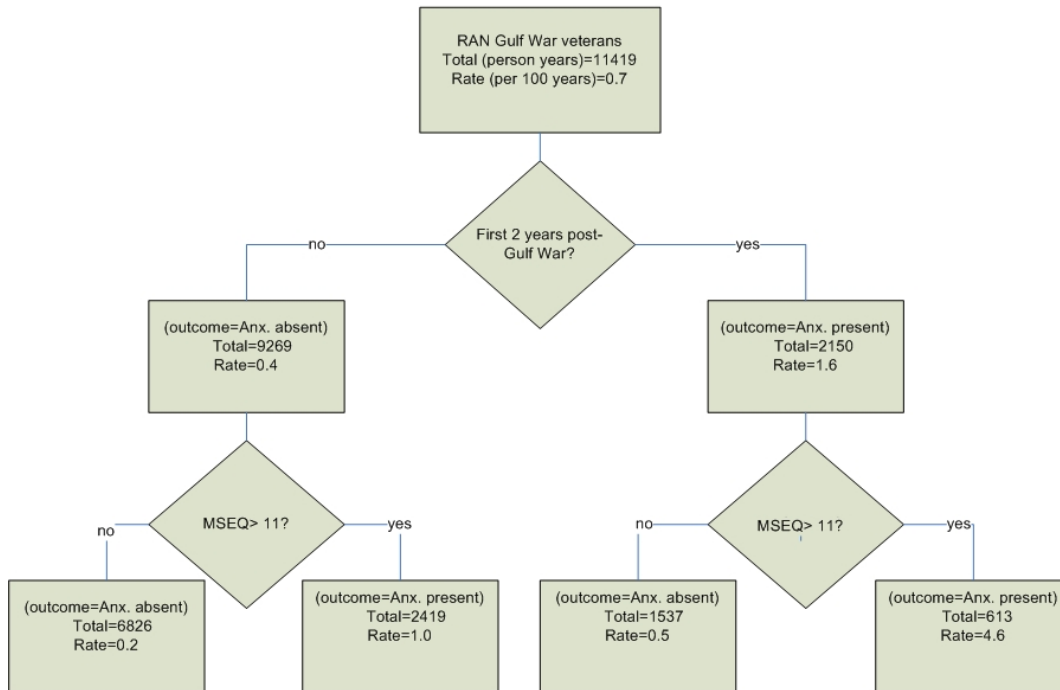
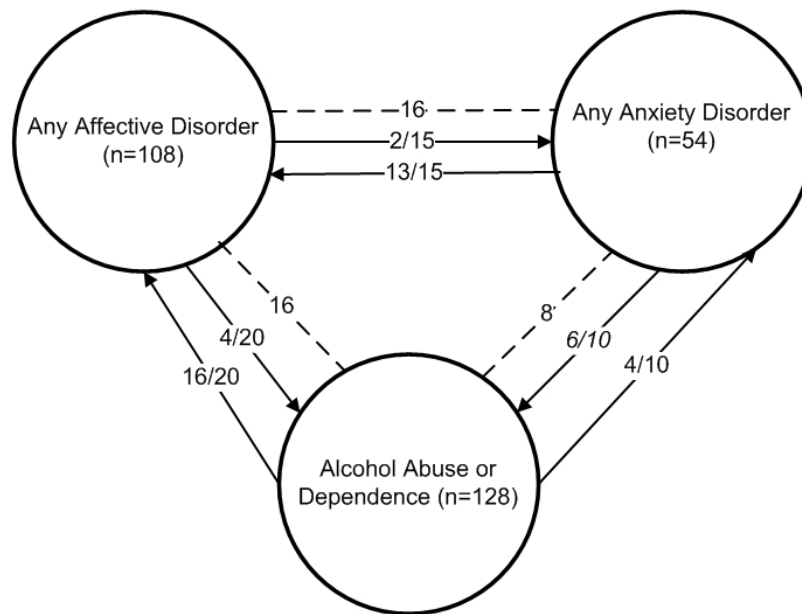


Figure 4. Frequency of sequence of post-Gulf War DSM-IV Any Affective Disorder, Any Anxiety Disorder and Alcohol Abuse Dependence in 66 veterans with comorbid psychological disorders. Forty veterans had onset of comorbid disorders at least one year apart. Note: Dashed lines represent number of comorbid disorders with onset less than or equal to one year apart



4. INTRODUCTION TO CHAPTER FOUR: PESSIMISM, WORTHLESSNESS, ANHEDONIA AND THOUGHTS OF DEATH IDENTIFY DSM-IV MAJOR DEPRESSION IN HOSPITALISED MEDICALLY ILL

The following chapter is concerned with the recognition of depression in the medically ill, and whether there are subgroups with regard to the type of depressive symptoms exhibited. As a result of physical illness, such as stroke and heart disease, and/or the sense of isolation that comes from being in hospital³⁴⁹, depression is common in the hospitalised medically ill. The prevalence of this major depression in this population has been reported to be approximately 20% to 30%³⁵⁰. Depression can be difficult to recognise in the hospitalised medically ill³⁵¹, due to a variety of reasons. For example, depression may be mistakenly regarded as being an expected or 'natural' part of illness or hospital stay. Moreover, diagnostic systems such as the DSM-IV²⁷⁸ regard depression as being a single condition, and so clinicians may find it difficult to differentiate between the various mood states that are commonly encountered in the medically ill.

4.1. Subtypes of depression

Prior research looked for subtypes of depression in the medically ill, and found several classes or clusters of patients⁶³, typified by scores on five traits or factors²⁹⁵, including demoralization (feelings of helplessness and hopelessness)³⁵² and anhedonia (lack of interest or pleasure)³⁵³. DSM-IV major depression was found to be highly prominent in clusters with high scores on the former

factor, and to a slightly lesser extent, in clusters with high scores on the latter factor ⁶³.

4.2. CART analysis of key symptoms of demoralization and anhedonia

The study presented in Chapter Four represents further analysis of the above sample, comprising 312 medically ill patients admitted to Monash Medical Centre ³⁵⁴, a university-affiliated major metropolitan general hospital in Melbourne, Australia. These patients met criteria for probable psychiatric caseness, according to the 36 item version of the General Health Questionnaire (GHQ-36), a commonly employed measure of psychological distress ³⁵⁵. Logistic regression and CART analysis were used to examine the relationship between key symptoms of demoralization and anhedonia and DSM-IV major depression. Some of the key symptoms of demoralization (e.g., ‘feelings of worthlessness’, and ‘thoughts of death’) and anhedonia (‘markedly diminished interest or pleasure’) are part of the DSM-IV criteria for major depression ²⁷⁸.

CART and other tree-building techniques have previously been employed to look at the relationship between symptoms and diagnosis of other types of psychiatric illness such as schizophrenia ¹⁵², and autistic disorders ³⁵⁶. The study presented in Chapter Four represents the first time that CART has been used to examine the relationship between DSM-IV depression symptoms and DSM-IV depression diagnosis, as well as the relationship between the symptoms of demoralization and anhedonia and DSM-IV depression diagnosis. The use of CART allows local or subgroup relationships between depression symptoms and diagnoses to be ascertained, recent logistic regression ³⁵⁷ and artificial neural

network³⁵⁸ applications in this area having focussed on overall or global relationships. The identification of subgroups exhibiting particular combinations of symptoms may aid in the better understanding, identification, diagnosis and treatment of major depression in the hospitalised medically ill.

Declaration for Thesis Chapter 4

McKenzie DP, Clarke DM, Forbes AB, Sim MR. Pessimism, worthlessness, anhedonia and thoughts of death identify DSM-IV major depression in the medically ill. *Psychosomatics* (in press, accepted 29 October 2008).

Declaration by candidate

In the case of Chapter 4, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
I was responsible for the research question, literature review, data management and programming, interpreting the results, writing the paper and undertaking any required revisions.	80%

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
D. Clarke	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A
A. Forbes	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper. Statistical consulting.	N/A
M. Sim	Responsible for reviewing and approving all aspects of the research methodology and design, interpretation of results and critical review of the written paper.	N/A

Candidate's
Signature

 Date 16-12-2008

Declaration by co-authors

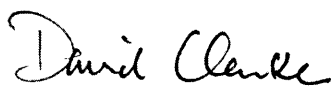
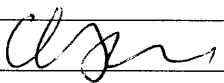
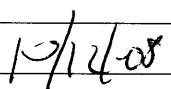
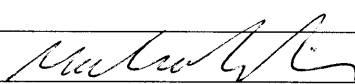
The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;

- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s) **Monash University Department of Epidemiology and Preventive Medicine, Alfred Hospital**

[Please note that the location(s) must be institutional in nature, and should be indicated here as a department, centre or institute, with specific campus identification where relevant.]

Signature 1		Date 08/12/2008
Signature 2		
Signature 3		
Signature 4		

Pessimism, worthlessness, anhedonia and thoughts of death identify DSM-IV major depression in hospitalized medically ill

**DEAN P. MCKENZIE, B.A. (Hons.)^{a*}, DAVID M. CLARKE Ph.D. FRANZCP^b,
ANDREW B. FORBES Ph.D.^a, MALCOLM R. SIM Ph.D. FAFOM FFOM^a**

^a **Monash Centre for Occupational and Environmental Health, Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia**

^b **Psychological and Behavioural Medicine Unit, School of Psychiatry, Psychology and Psychological Medicine, Monash University, Monash Medical Centre, Clayton, Australia**

*Corresponding author. Monash Centre for Occupational and Environmental Health, Department of Epidemiology and Preventive Medicine, Monash University, Alfred Hospital, Commercial Road, Melbourne, Victoria, 3004 Australia

Tel: +61 3 9903 0555; fax: +61 3 9903 0556

Email address: dean.mckenzie@med.monash.edu.au

(Approximately 4060 words)

ACCEPTED BY PSYCHOSOMATICS 29 October 2008

Published by American Psychiatric Publishing, Virginia USA

2007 impact factor 2.199

Background: Major depression can be difficult to diagnose in the medically ill, as distinct mood states may not be adequately differentiated. Our prior research found several dimensions of mood states, including demoralization (hopelessness-helplessness) and anhedonia (inability to experience pleasure), upon which we constructed a taxonomy of depression. DSM-IV major depression was highly prevalent in the clusters of participants typified by high levels of demoralization, and to a slightly lesser extent, anhedonia. **Objective:** The present study provides a further analysis of 312 medically ill patients, examining how key individual symptoms of demoralization and anhedonia relate to DSM-IV major depression. **Method:** Logistic regression and Classification and Regression Tree (CART) analysis were employed. **Results:** Two combinations of symptoms – pessimism and worthlessness; and pessimism, loss of interest in others, and thoughts of death – were highly associated with major depression. **Conclusion:** The identification of key symptoms, particularly those involving pessimism, may aid clinical understanding and treatment of depression.

Depression is common in the hospitalized medically ill, with prevalence reported at approximately 20 to 30%. ¹ This disorder is associated with increased health care costs, reduced compliance, and increased morbidity and mortality ²⁻³

In order for depression to be treated, it must first be identified, which poses a particular problem in the medically ill. Depression may be regarded by some clinicians, as well as by patients and relatives, as an expected or ‘natural’ consequence of illness or hospital stay ⁴ – therefore the boundary between clinical depression and normal sadness is unclear. There is also growing opinion that, within clinical depression, traditional diagnostic systems do not allow differentiation between different mood states commonly experienced in the medically ill. ⁵ For example, the criteria for major depression employed by the fourth edition of the Diagnostic and Statistical Manual (DSM-IV), ⁶ essentially views depression as a single condition varying only in severity, yet it has been found to be heterogeneous with regard to clinical presentation, course, treatment, genetics and neurobiology. ⁷

In our previous research, we found evidence for subtypes of depression in the medically ill. We did this by firstly exploring the dimensional structure of both psychological and somatic symptoms, and identified five traits or factors, which

we termed demoralization, anhedonia, autonomic anxiety, somatoform symptoms and grief.⁸ Anhedonia was typified by diminished interest and the inability to experience pleasure,⁹ while demoralization was characterized by feelings of hopelessness, helplessness and despair.¹⁰

In subsequent research, we found several classes or clusters of patients with similar scores on the above five factors.¹¹ DSM-IV major depression was highly prominent in the clusters typified by high scores on the demoralization factor, and to a slightly lesser extent, high scores on the anhedonia factor. Further evidence for the distinction and importance of demoralization and anhedonia was provided in our later study of 271 severely medically ill patients, where we used self-report measures of depression.¹² Together these results suggest that there are different dimensions or types of depression, primarily distinguished by levels of demoralization and anhedonia.

Although both anhedonia^{13,14} and demoralization^{15,16} have been operationalized, and observed in other populations such as adolescents,¹⁷ as well as the medically ill,¹⁸⁻²¹ there has been little formal research into the relationship between them and the commonly employed category of major depression. It is evident that there is overlap between these constructs however, which should be further investigated. For example, ‘feelings of worthlessness’, a key feature of demoralization,¹⁰ is a component of the DSM-IV major depression symptom of ‘feelings of worthlessness or excessive or inappropriate guilt’. Similarly, ‘markedly diminished interest or pleasure’, fundamental to the concept of anhedonia⁹ is, along with depressed mood, a core symptom of DSM-IV major depression. In addition, loss of interest or pleasure, along with lack of reactivity to usually pleasurable stimuli, is a core feature of DSM-IV major depression with melancholic features.

There has recently been renewed interest in major depression at the symptom level, including examination of which specific symptoms are the most important in making the diagnosis.^{22,23} The present study seeks to extend our previous findings regarding constructs of demoralization and anhedonia, being

those mood states most highly associated with major depression in the medically ill. We further analyze the sample of 312 medically ill patients in order to identify which specific key symptoms of our previously identified demoralization and anhedonia factors⁸ are most highly associated with a diagnosis of DSM-IV major depression. Both logistic regression and a computational method known as classification and regression trees (CART),²⁴ are employed. We firstly used logistic regression to examine the relationship between DSM-IV major depression and the key symptoms of demoralization and anhedonia taken individually and together. CART is then employed to look for subgroups of patients defined by particular combinations of key symptoms, which are then further analyzed using logistic regression. The identification of combinations of specific symptoms within the major depression construct will clarify whether or not the latter is truly heterogeneous, and may clarify different qualitative subtypes of depression. This could have implications for clinical treatment. Finally, core symptoms or groups of symptoms that identify major depression especially well in the medically ill, might be used as markers or screening tests for depression in this group.

METHOD

Recruitment and Screening

The study was conducted at Monash Medical Centre, a university-affiliated general hospital in Melbourne, Australia. A sample of 312 patients who provided written informed consent was recruited from consecutive admissions to the medical and surgical wards, after screening for probable psychiatric caseness. The latter was defined as exceeding a threshold or cut-off of 21 or greater, on the 36 item version of the General Health Questionnaire (GHQ-36),²⁵ scored using the chronic binary method.²⁶ This cut-off was used on the basis that it has demonstrated good sensitivity for a broad range of disturbances in a medical population.⁸ The sample was restricted to probable cases in order to approximate the type of clinical sample that would generally be referred to consultation-liaison psychiatry within a general hospital.

Patients were administered a structured psychiatric interview described below, and completed a self-report questionnaire. Exclusion criteria comprised cognitive impairment and insufficient fluency in the English language. The 312 patients had a mean age of 47.5 years (range 18 to 85 years) and were predominantly female (61%), these characteristics being similar to those of the general hospital population.²⁷ Patients had been admitted to hospital with a range of medical conditions; predominantly cardiovascular (22%), gastrointestinal (17%), respiratory (15%), rheumatological (13%) and neurological (11%). The mean severity of illness, assessed in consultation with hospital medical staff using a four point Likert scale with a range of 1-4, was 2.4. Further details on the patient sample and study design are available in our earlier analyses of these data.^{8,11}

Evaluation

Structured interviews were performed using the Monash Instrument for Liaison Psychiatry (MILP),²⁸ which includes information necessary to make DSM-IV diagnoses, as well as a broader enquiry of symptom data relevant to the medically ill. The MILP has been shown to have high interrater reliability, with a kappa coefficient^{29,30} of 0.87 for DSM-III-R³¹ major depression and 0.79 for DSM-IV major depression.³² In regard to procedural validity, the MILP shows close agreement on DSM-III-R major depression with that diagnosed by DTREE,³³ a computer assisted decision tree, but lower agreement with that diagnosed by the Structured Clinical Interview for DSM-III-R (SCID).³⁴ Obtained kappa values were 0.74 and 0.54 respectively. It is thought that the lower agreement between the MILP and SCID is due to the former being stricter in its judgement concerning the attribution of cause of symptoms, and the exclusion of symptoms thought to be of organic origin.³² Administration of the MILP was performed by two experienced psychologists. Judgement of attribution of cause of symptoms was guided by consultation with medical staff if required. For a symptom to be judged as being present, its cause could not be attributed to physical illness or injury, or medication, drugs, or alcohol.

The demoralization factor found in the earlier analyses consisted of 28 items.⁸ In the present analyses eight of these symptoms were chosen to be most representative of demoralization, on the basis of high loadings on this factor, and because they are key features of demoralization as clinically described.³⁵ The key symptoms of demoralization were discouragement / despondency, pessimism, hopelessness, feeling unable to cope, helplessness, worthlessness, loss of confidence, and thoughts of death.

The anhedonia factor found in the earlier analysis⁸ consisted of four items, all of which were individually employed in the present analyses. The key symptoms of anhedonia were loss of interest in activities with others, being unable to enjoy activities with others, loss of interest in solitary activities, and being unable to enjoy solitary activities.

Possible demographic and medical confounders for major depression³⁶ were included in the statistical analyses. These consisted of age, gender, years of schooling, marital status (formerly or currently married versus never married), presence of past psychiatric history and severity of illness.

Data Analysis

The logistic regression procedure of SPSS³⁷ was employed to perform an overall analysis of the key symptoms, with current (in the past month) DSM-IV major depression as the outcome variable. Initially, the 12 key symptoms listed above were examined separately, adjusting for the potential confounders outlined previously. Subsequently, a single logistic regression was performed with all key symptoms entered into the regression model.

In order to look for particular combinations of symptoms highly associated with DSM-IV major depression, binary recursive partitioning or classification and regression tree (CART) analysis was used^{24,38} This procedure can be likened to a cross between cluster analysis and regression, in that it identifies subgroups of persons defined by particular combinations of variables that are associated with a particular outcome, in this case major depression. CART has been applied previously to a variety of psychiatric data,³⁹⁻⁴² including depression.⁴³

CART explicitly validates the generality of its tree models, using by default 10-fold cross-validation.²⁴ Datasets are randomly divided into 10 subsets, with each subset in turn being used to test the performance of the tree created with the other nine subsets. We further tested the generality of the CART tree by using a totally separate validation subset. SPSS was employed to randomly divide the patients into a learning subset of approximately 75% and a validation subset of approximately 25% of the observations, with similar levels of DSM-IV major depression in each subset. The classification accuracy obtained for each subset was then compared using a two (learning and validation subsets) by four (true positives, false positives, true negatives, false negatives) chi-square test. If the results of the latter did not approach statistical significance the subsets were combined.⁴²

The subgroups of patients identified by CART were further analyzed using logistic regression. The statistical comparison of empirically defined groups can be problematic.⁴⁴ The logistic regression analysis of CART subgroups should therefore be seen primarily as being descriptive, allowing the magnitude of the risk of major depression to be assessed for each pattern of symptoms, while controlling for effects of the potential demographic and medical confounders.^{42,45}

The screening performance of the key symptoms and their combinations selected by CART was evaluated using positive predictive value (the probability of having a positive diagnosis of major depression amongst those patients with a positive test result), negative predictive value (probability of negative diagnosis amongst those with a negative test result), sensitivity (probability of positive test result amongst those with a positive diagnosis) and specificity (probability of having a negative test result amongst those with a negative diagnosis).³⁰ The above statistics and their associated 95% confidence intervals were obtained using a custom computer program.⁴⁶

RESULTS

Of the original sample of 312 hospital patients, 12 (3.8%) had missing diagnostic information for DSM-IV major depression and so were excluded from

the statistical analyses. Of the remaining 300 patients, 57 (19.0%) met current (past month) diagnostic for DSM-IV major depression.

Analysis of key symptoms of demoralization and anhedonia

Table 1 shows the results of the logistic regression analyses of the relationship between the key symptoms of demoralization and anhedonia, and current (past month) DSM-IV major depression. As the crude and adjusted odds ratios were similar in magnitude only the latter are shown throughout. The most prevalent symptoms were discouragement / despondency (75.3%), unable to cope (58.3%), pessimism (56.7%), and loss of interest in activities with others (55.7%). When examined individually, using separate logistic regressions, the symptoms most strongly associated with major depression were pessimism, worthlessness, discouragement / despondency, thoughts of death, feeling unable to cope, being unable to enjoy the company of others and experiencing less enjoyment in solitary activities, although all associations with DSM-IV major depression were statistically significant.

When all key symptoms were analyzed using a single logistic regression model, taking any inter-correlation between the symptoms into account, only the key demoralization symptoms of pessimism, worthlessness and thoughts of death were significantly associated with DSM-IV major depression. Of the key symptoms of anhedonia, only one symptom, showing less interest in activities with others, was significantly associated with DSM-IV major depression.

Classification tree of key symptoms of demoralization and anhedonia

The results of the CART recursive partitioning analysis are given in Figure 1. All twelve key symptoms of demoralization and anhedonia were made available to CART. There was no statistically significant difference (chi-square = 3.6, df = 3, p = 0.31) between the performance of the classification tree obtained for the learning subset, and the performance of that tree applied to the validation subset. The two subsets were therefore combined.

CART first split the dataset by feelings of pessimism. The group with this symptom absent was not able to be split any further. The pessimism present subgroup was then split by feelings of worthlessness. The pessimism present, worthlessness present subgroup was not further split.

The pessimism present, worthlessness absent subgroup was split by loss of interest in activities with others. The loss of interest in activities with others subgroup was then split by thoughts of death.

Logistic regression analysis of classification tree subgroups

CART readily allows the identification of those subgroups, defined by key symptoms of demoralization and anhedonia, at the greatest risk of current DSM-IV major depression. Logistic regression was used to compare each subgroup with a reference subgroup with no key symptoms of demoralization or anhedonia, adjusting for potential demographic and medical confounders. The reference group exhibited a prevalence of DSM-IV major depression of 3.1%. The highest risk of DSM-IV major depression was observed for the CART subgroup with symptoms of pessimism and worthlessness both present (50.6%, adjusted OR = 28.66, 95% CI = 9.19 – 89.36).

The combination of symptoms with the next highest risk of DSM-IV major depression was pessimism present, worthlessness absent, loss of interest in others present and thoughts of death present (38.9%, adjusted OR = 22.80, 95% CI = 4.97 – 104.54). There was no statistically significant difference in the risk of DSM-IV major depression between the above two subgroups (50.6% versus 38.9%, adjusted OR = 1.21, 95% CI = 0.34 – 4.26).

There was a moderate increase in risk of DSM-IV major depression for the combination of pessimism present, worthlessness absent, loss of interest in others present and thoughts of death absent (10.0%, adjusted OR = 3.71, 95% CI = 0.72 – 19.15), although this result was not statistically significant. There was a very slight decrease in risk for those patients with pessimism present, but worthlessness and loss of interest both absent (2.6%, adjusted OR = 1.08, 95% CI = 0.11 – 10.43), however this result was likewise not statistically significant.

Screening test performance

The screening test performance of the key symptoms of demoralization and anhedonia, and their combinations, selected by CART, is given in Table 2. Of the individual symptoms, worthlessness and thoughts of death exhibited the highest positive predictive power, while pessimism and worthlessness exhibited the highest negative predictive power. The combination of pessimism present and worthlessness present exhibited a similar performance to the total CART tree based upon all four items, except that the latter had higher sensitivity (86.0 versus 73.7) but lower specificity (78.6 versus 83.1) than the former. The combination of pessimism present, worthlessness absent, loss of interest in others present and thoughts of death present exhibited the lowest sensitivity, but the highest specificity, of any single symptom or combination of symptoms. The latter subgroup was generated at the bottom of the CART tree, based upon 18 observations. At each split CART attempts to maximize the difference between the two subgroups created, in terms of frequency on the outcome variable, in this case the presence of major depression. As the tree-growing proceeds, the two subgroups being compared at each split become smaller. This reduction in size lowers the sensitivity for a screening rule based upon a specific combination of items, but increases the specificity as the small subgroups become increasingly homogenous.

[insert Table 1 about here]

[insert Figure 1 about here]

[insert Table 2 about here]

DISCUSSION

This study examined the relationship between the key symptoms of demoralization and anhedonia and a diagnosis of DSM-IV major depressive disorder in the hospitalized medically ill. All of the key symptoms of demoralization and anhedonia were significantly associated with a diagnosis of DSM-IV major depressive disorder when analyzed individually.

Although demoralization had been previously discussed by Donald Klein in the context of depression and helplessness,⁴⁷ a more comprehensive discussion was provided by Jerome Frank.¹⁰ Frank viewed demoralization as being typified by feelings of hopelessness and helplessness, brought about by not being able to meet demands made by the environment, from which persons are not able to disengage.

Coined in the late nineteenth century by Theodule Ribot, anhedonia is typified by the inability to experience pleasure and diminished interest in things, in an analogy with analgesia, a diminished ability to experience pain.⁹ In the words of William James, who suggested that it was a type of “pathological depression”, anhedonia “is mere passive joylessness and dreariness, discouragement, dejection, lack of taste and zest and spring” (p. 127).⁴⁸ Although, for many years anhedonia has been recognized as an accompaniment, if not a key symptom of, schizophrenia⁴⁹ it has also been proposed as a signifier of a subtype of depression, termed endogenomorphic depression.⁵⁰

Currently, the place of anhedonia in major depression as defined by DSM-IV remains somewhat ambiguous, it being an important core symptom, but not necessary to the diagnosis. Demoralization, on the other hand, is often considered as being separate to major depression,^{51, 52} and described as non-specific distress,¹⁵ or a form of minor depression.²¹ Although very few studies have examined both demoralization and anhedonia together, there is empirical evidence from our previous studies of the medically ill,^{8,11,12} as well as other studies,¹⁷ that the two are separate constructs, and represent subtypes of major depression as defined by DSM-IV.¹¹ Further evidence for these subtypes has been presented here.

On examining the results of the key symptoms analyzed simultaneously, pessimism is identified as a very important symptom, accounting for an eight-fold increased risk of diagnosis of major depression. This result is quite surprising, in that pessimism is not one of the DSM-IV criteria for major depression. However, pessimism has been described by some researchers as a core feature of the

experience of depression.⁵³ As well as being associated with a higher number of depression symptoms,⁵⁴ pessimism is a specific risk factor for heart disease,⁵⁵ with a pessimistic view having greater negative consequences than an optimistic view.⁵⁶ Pessimism is included as a clinical symptom of depressive episode in ICD-10,⁵⁷ and, given our findings, could be considered a useful inclusion into future DSM criteria.

When the key symptoms of demoralization and anhedonia were analyzed simultaneously, feelings of worthlessness, loss of interest in activities with others, and thoughts of death are important, each being associated with a statistically significant three-fold rise in the likelihood of having major depression. The presence of worthlessness/guilt in children and adolescents has been found to be predictive of adult depression in longitudinal studies.⁵⁸ Furthermore it is associated with persistence of depression⁵⁹ and a greater risk of in-hospital mortality.⁶⁰ These associations with persistence of depression and mortality also apply to the symptom of thoughts of death.⁵⁸

Loss of interest in activities with others is a key component of anhedonia, specifically social anhedonia, defined as a deficit in the ability to experience pleasure deriving from social relationships, in contrast with physical anhedonia.¹³ DSM-IV does not distinguish between these two types of anhedonia. The other anhedonia symptoms were clearly not statistically significant when analyzed together, although they were significant when examined individually. This suggests that social anhedonia is the more important feature of the type of depression commonly seen in the physically ill, or defined by DSM major depression.

The results of the CART analysis suggest two subtypes of depression within DSM-IV major depressive disorder. Both subtypes have pessimism as a foundation. One is then characterized by worthlessness; the other by loss of interest and thoughts of death. Patients with 'worthless depression' were almost thirty times as likely to have current DSM-IV major depression as those with no key symptoms. Patients with loss of interest and thoughts of death were over

twenty times as likely to have current DSM-IV major depression as were those in the symptom-free reference group.

We did not find helplessness and hopelessness to be statistically significant when all symptoms were analyzed together, despite these two symptoms being important markers of demoralization.¹⁰ Another recent study conducted in psychiatric outpatients has found that although an item representing helplessness and hopelessness was significantly associated with DSM-IV major depression, the association was not as strong as that for symptoms of worthlessness and loss of interest or pleasure.⁶¹ Although symptoms were analyzed individually in the latter study, a multivariate analysis of general population data found hopelessness to be only weakly associated with DSM-III major depression.⁶²

Hopelessness has nevertheless been proposed as an important subtype of depression,⁶³ and to be an independent risk factor for heart disease,⁵⁵ other types of chronic illness⁶⁴ and in-hospital mortality,⁶⁰ in medical patients. Hopelessness depression has been defined as involving an expectation that highly desired outcomes, in many areas of life, will not occur or that undesirable outcomes will occur, regardless of any action that the individual may take.⁶³ On the other hand, these researchers have defined pessimism, or at least 'circumscribed pessimism' as applying to more limited areas of life, and being associated with less severe symptoms of depression, although pessimism regarding extremely important outcomes may be associated with more severe symptoms. Hopelessness is seen as being associated with global attributions for negative life events, pessimism with more specific attributions. Although severity of illness was controlled for in the logistic regression analyses it is possible that patients were pessimistic regarding their chances of recovery from illness, leading to depression, although of course the relationship may be circular.⁵³

The mixed empirical results regarding hopelessness suggest that whilst helplessness-hopelessness is an important construct, it may not be one that is captured fully by DSM defined depression. Perhaps, in any revision of categories

of depression, a form of 'hopelessness depression' ⁶³ needs to be described that is distinct from other forms of depression characterized by anhedonia, or worthlessness.

The validity and utility of categories representing different forms of depression of course is yet to be proven. This is a major deficiency in our current state of knowledge. The use of pharmacotherapy and psychotherapy for medically ill patients with significant depression is advised, ⁶⁵ but clinicians should be wary of a 'one size fits all' approach to treatment. ⁶⁶ Although demoralization has been observed to respond to psychotherapy ^{10 67 47} and anhedonia has been considered the hallmark of a biogenic depression responsive to drug treatment ^{47,50}, there is a dearth of research to inform us as to which forms of depression will respond to which treatments. If we are to understand depression we need to understand more about the symptoms of depression, including those such as pessimism that are not included in DSM-IV, and how they co-occur.

Employing a recursive partitioning technique, in addition to logistic regression, our study extends the methodology of previous studies that have looked at how individual DSM-IV depression symptoms, or their combinations, apply to the total sample. ^{23,61,62,68} Although the findings need to be replicated on other samples of medically ill patients, the cross-validation employed by CART, as well as the testing of the CART results on a validation subset, suggests generality of these results. The resulting four item CART tree exhibits similar screening performance to other, longer, depression measures intended for use with the medically ill. ⁶⁹ Using an approach such as ours, clinicians and researchers could look for subgroups typified by particular combinations of specific symptoms. Tailored treatments could, if necessary, be offered to particular subgroups of patients, defined by particular collections of symptoms. Our results suggest that pessimism, worthlessness, loss of interest in activities with others, and thoughts of death, are strongly associated with major depression in the hospitalized medically ill, although pessimism is not part of the DSM-IV

diagnostic criteria. Pessimism should be strongly considered for inclusion in screening or diagnostic tests intended for use with this population.

This research, and the first author, were supported by the National Health and Medical Research Council of Australia. The authors thank Kevan Pitcher and Anne Silbereisen for conducting the patient interviews. We thank Monash Medical Centre medical staff for rating illness severity, and most of all, we are grateful to the patients for their participation in the study.

References

1. Clarke DM, Minas IH, Stuart GW: The prevalence of psychiatric morbidity in general hospital patients. *Aust NZ J Psychiatry* 1991;25:322-329
2. Gehi A, Haas D, Pipkin S, et al: Depression and medication adherence in outpatients with coronary heart disease: findings from the Heart and Soul Study. *Arch Intern Med* 2005;165:2508-13
3. Strain JJ, Blumenfield M: Challenges for consultation-liaison psychiatry in the 21st century. *Psychosomatics* 2008;49:93-96
4. Simon GE, Von Korff M, Lin E: Clinical and functional outcomes of depression treatment in patients with and without chronic medical illness. *Psychol Med* 2005;35:271-9
5. Grassi L, Mangelli L, Fava GA, et al: Psychosomatic characterization of adjustment disorders in the medical setting: Some suggestions for DSM-V. *J Aff Disorders* 2007;101:251-254
6. American Psychiatric Association: *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*. 4th Edition. Washington, DC: American Psychiatric Association, 1994
7. Rush AJ: The varied clinical presentations of major depressive disorder. *J Clin Psychiatry* 2007;68, Supplement 8:4-10
8. Clarke DM, Mackinnon AJ, Smith GC, et al: Dimensions of psychopathology in the medically ill: a latent trait approach. *Psychosomatics* 2000;41:418-425
9. Willner P: Anhedonia, in *Symptoms of Depression*. Edited by Costello CG. New York, NY, Wiley, 1993
10. Frank JD, Frank JB: *Persuasion and Healing: A Comparative Study of Psychotherapy*, 2nd Edition. Baltimore, MD, John Hopkins University Press, 1973

11. Clarke DM, Smith GC, Dowe DL, et al: An empirically derived taxonomy of common distress syndromes in the medically ill. *J Psychosom Res* 2003;54:323-330
12. Clarke DM, Kissane DW, Trauer T, et al: Demoralization, anhedonia and grief in patients with severe physical illness. *World Psychiatry* 2005;4:96-105
13. Chapman LJ, Chapman JP, Raulin ML: Scales for physical and social anhedonia. *J Abnorm Psychol* 1976;85:374-382
14. Fawcett J, Clark D, Scheftner W, et al: Assessing anhedonia in psychiatric patients: the pleasure scale. *Arch Gen Psychiatry* 1983;40:79-84
15. Dohrenwend BP. Nonspecific psychological distress and other dimensions of psychopathology. *Arch Gen Psychiatry* 1980;37:1229-1236
16. Kissane DW, Wein S, Love A, et al: The demoralization scale: a report of its development and preliminary validation. *J Palliat Care* 2004;20:269-276
17. Weber S: Factor structure of the Reynolds Adolescent Depression Scale in a sample of school-based adolescents. *J Nurs Meas* 2000;8:23-40
18. Boscaglia N, Clarke DM: Sense of coherence as a protective factor for demoralisation in women with a recent diagnosis of gynaecological cancer. *Psycho-Oncol* 2007;16:189-195
19. Clark DA, Cook A, Snow D: Depressive symptom differences in hospitalized, medically ill, depressed psychiatric inpatients and nonmedical controls. *J Abnorm Psychol* 1998;107:38-48
20. Marchesi C, Maggini C: Socio-demographic and clinical features associated with demoralization in medically ill in-patients. *Soc Psychiatry Psychiatr Epidemiol* 2007;42:824-9
21. Rowe SK, Rapaport MH: Classification and treatment of sub-threshold depression. *Curr Opin Psychiatr* 2006;19:9-13
22. Parker GB: Commentary on diagnosing major depressive disorder: ask less that we embrace major depression and ask more what the concept does for us. *J Nerv Ment Dis* 2006;194:155-157.
23. Zimmerman M, Chelminski I, McGlinchey JB: Diagnosing major depressive disorder X: can the utility of the DSM-IV symptom criteria be improved? *J Nerv Ment Dis* 2006;194:893-7
24. Breiman L, Friedman J, Olshen RA, et al: Classification and Regression Trees. Belmont, CA, Wadsworth, 1984
25. Goldberg DP, Williams P: A User's Guide to the General Health Questionnaire. Windsor, United Kingdom, NFER-Nelson, 1988.
26. Goodchild M, Duncan-Jones P: Chronicity and the General Health Questionnaire. *Brit J Psychiatry* 1985;146:55-61
27. Clarke DM, Smith GC: Consultation-liaison psychiatry in general medical units. *Aust NZ J Psychiatry* 1995;29:424-432

28. Clarke DM, Smith GC, Herrman DP, et al: The Monash Interview for Liaison Psychiatry (MILP); development, reliability and procedural validity. *Psychosomatics* 1998;39:318-328.
29. Cohen J: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;20:37-46
30. Kraemer HC: *Evaluating Medical Tests : Objective and Quantitative Guidelines*. Newbury Park, CA, Sage, 1992
31. American Psychiatric Association: *DSM-III-R: Diagnostic and Statistical Manual of Mental Disorders*. 3rd, Revised Edition. Washington, DC:, American Psychiatric Association, 1987
32. Clarke D, Smith G, Herrman H, et al: The Monash Interview for Liaison Psychiatry (MILP): development, reliability and procedural validity. *Psychosomatics*. 1998;39:318-328
33. First MB, Opler LA, Hamilton RM, et al: Evaluation in an inpatient setting of DTREE: a computer-assisted diagnostic assessment procedure,. *Comp Psychiatry* 1993;34:171-175
34. Spitzer RL, Williams JBW, Gibbon M, et al: The Structured Clinical Interview for DSM-III-R (SCID) I: history, rationale and description. *Arch Gen Psychiatry* 1992;49:624-629
35. Clarke DM, Kissane DW: Demoralization: its phenomenology and importance. *Aust NZ J Psychiatry* 2002;36:733-742
36. Stansfield S, Rasul F: Psychosocial factors, depression and illness, In *Depression and Physical Illness*. Edited by Steptoe A. Cambridge, United Kingdom, Cambridge University Press, 2007
37. SPSS for Windows, version 15.0 [computer software]: Chicago, IL, SPSS Incorporated, 2006
38. Salford Systems. *CART for Windows*, version 6 [computer software]: San Diego, CA, Salford Systems, 2006
39. Craig TJ, Siegel C, Hopper K, et al: Outcome in schizophrenia and related disorders compared between developing and developed countries: a recursive partitioning reanalysis of the WHO DOSMD data. *Br J Psychiatry* 1997;170:229-233
40. Drozd EM, Gage B, Maier J: Patient casemix classification for Medicare Psychiatric prospective payment. *A J Psychiatry* 2006;163:724-732
41. McKenzie DP, McGorry PD, Wallace CS, et al: Constructing a minimal diagnostic decision tree. *Methods Inf Med* 1993;32:161-166
42. McKenzie DP, McFarlane AC, Creamer M, et al: Hazardous or harmful alcohol use in Royal Australian Navy veterans of the 1991 Gulf War: identification of high risk subgroups. *Addict Behav* 2006;31:1683-1694

43. Schmitz N, Kugler J, Rollnik J: On the relationship between neuroticism, self-esteem, and depression: results from the National Comorbidity survey. *Comp Psychiatry* 2003;44:169-176
44. Austin PC, Goldwasser MA: Pisces did not have increased heart failure: data driven comparison of binary proportions between levels of a categorical variable can result in increased significance levels. *J Clin Epidemiol* 2008;61:295-300
45. Zhang H, Singer B: *Recursive Partitioning in the Health Sciences*. New York, NY: Springer-Verlag, 1999
46. McKenzie DP, Vida S, Mackinnon AJ, et al: Accurate confidence intervals for measures of test performance. *Psychiatry Res* 1997;69:207-209
47. Klein DF, Davis JM: *Diagnosis and Drug Treatment of Psychiatric Disorders*. Baltimore, MD, Williams and Wilkins, 1969
48. James W: *The Varieties of Religious Experience: A Study in Human Nature*. New York, NY, Collier, 1902/1961
49. Meehl PE: Schizotaxia, schizotypy, schizophrenia: *Am Psychol* 1962;17:827-838
50. Klein DF: Endogenomorphic depression. *Arch Gen Psychiatry* 1974;31:447-454
51. de Figueiredo JM: Depression and demoralization: phenomenologic differences and research perspectives. *Comp Psychiatry* 1993;34:308-311
52. Kissane DW, Clarke DM, Street AF: Demoralization syndrome - a relevant psychiatric diagnosis for palliative care. *J Palliat Care* 2001;17:12-21
53. Seligman MEP: *Helplessness: On Development, Depression and Death*. 2nd Edition. New York, NY, WH Freeman, 1992
54. Isaacowitz DM, Seligman ME: Is pessimism a risk factor for depressive mood among community-dwelling older adults? *Behav Res Therapy* 2001;39:255-72
55. Kubzansky LD, Davidson KW, Rozanski A: The clinical impact of negative psychological states: expanding the spectrum of risk for coronary artery disease. *Psychosom Med* 2005;67(suppl 1):S10-4
56. Brink E, Grankvist G: Associations between depression, fatigue, and life orientation in myocardial infarction patients. *J Cardiovasc Nurs* 2006;21:407-11
57. World Health Organization: *ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Categories*. Geneva, Switzerland, World Health Organization, 1992
58. Wilcox HC, Anthony JC: Child and adolescent clinical features as forerunners of adult-onset major depressive disorder: retrospective evidence from an epidemiological sample. *J Affect Disorders* 2004;82:9-20

59. Mojtabai R, Olfson M: Major depression in community-dwelling middle-aged and older adults: prevalence and 2- and 4-year follow-up symptoms. *Psychol Med* 2004;34:623-34
60. Furlanetto LM, von Ammon Cavanaugh S, Bueno JR, et al: Association between depressive symptoms and mortality in medical inpatients. *Psychosomatics* 2000;41:426-32
61. McGlinchey JB, Zimmerman M, Young D, et al: Diagnosing major depressive disorder VIII: are some symptoms better than others? *J Nerv Ment Dis* 2006;194:785-90
62. Nair J, Nair SS, Kashani JH, et al: Analysis of the symptoms of depression--a neural network approach. *Psychiatry Res* 1999;87:193-201
63. Abramson LY, Metalsky GI, Alloy LB: Hopelessness depression: a theory-based subtype of depression. *Psychol Rev* 1989;96:358-372
64. Katon W, Lin EH, Kroenke K: The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. *Gen Hosp Psychiatry* 2007;29:147-55
65. Rosenstein DL, Soleymani K, Cai J: Chronic depression in patients with medical illness, In *Handbook of Chronic Depression: Diagnosis and Therapeutic Management*. Edited by Alpert JE, Fava M. New York, NY: Marcel Dekker, 2004
66. Parker G, Fletcher K: Treating depression with the evidence-based psychotherapies: a critique of the evidence. *Acta Psychiatr Scand* 2007;115:352-9
67. Griffith JL, Gaby L: Brief psychotherapy at the bedside: countering demoralization from medical illness. *Psychosomatics* 2005;46:109-16
68. Andrews G, Slade T, Sunderland M, et al: Issues for DSM-V: simplifying DSM-IV to enhance utility: the case of major depressive disorder. *Am J Psychiatry* 2007;164:1784-5
69. Wilhelm K, Kotze B., Waterhouse M, et al: Screening for depression in the medically ill: a comparison of self-report measures, clinician judgement, and DSM-IV diagnoses. *Psychosomatics* 2004;45:461-469

Table 1: Relationship of key demoralization and anhedonia symptoms to DSM-IV major depression

Symptom	Prevalence (%)	Adjusted OR ^a	95% CI	Adjusted OR ^b	95% CI
Discouragement / despondency	75.3	7.92	2.26-27.70	2.67	0.49-14.62
Pessimism	56.7	15.10	5.02-45.23	8.26	1.95-34.96
Hopelessness	45.3	5.52	2.61-11.69	0.72	0.22-2.35
Unable to cope	58.3	7.52	2.98-19.00	2.21	0.65-7.49
Helplessness	49.7	4.03	1.94-8.34	1.09	0.38-3.17
Worthlessness	39.0	8.90	4.09-19.37	3.60	1.39-9.31
Loss of confidence	38.3	4.16	2.08-8.35	2.37	0.84-6.68
Thoughts of death	42.3	7.74	3.53-16.96	3.23	1.24-8.42
Less interest in activities with others (anhedonia)	55.7	4.71	2.09-10.60	3.64	1.10-12.04
Unable to enjoy activities with others (anhedonia)	47.0	5.01	2.41-10.42	1.01	0.31-3.31
Less interest in solitary activities (anhedonia)	41.7	3.69	1.86-7.29	1.08	0.31-3.72
Less enjoyment in solitary activities (anhedonia)	45.0	5.05	2.49-10.40	1.77	0.51-6.16

^a Adjusted odds ratios were obtained using logistic regression, on each symptom separately.

Adjusted odds ratios adjust for possible confounders - age, gender, years of schooling, marital status (ever married, never married), past psychiatric history and severity of physical illness.

^b Adjusted odds ratios were obtained using logistic regression, with all symptoms entered simultaneously into the regression model. 95% confident intervals refer to these values. Adjusted odds ratios adjust for possible confounders - age, gender, years of schooling, marital status (ever married, never married), past psychiatric history and severity of physical illness.

Figure 1. Recursive partitioning / classification and regression tree (CART) analysis of risk of current major depression in hospitalized medically ill patients

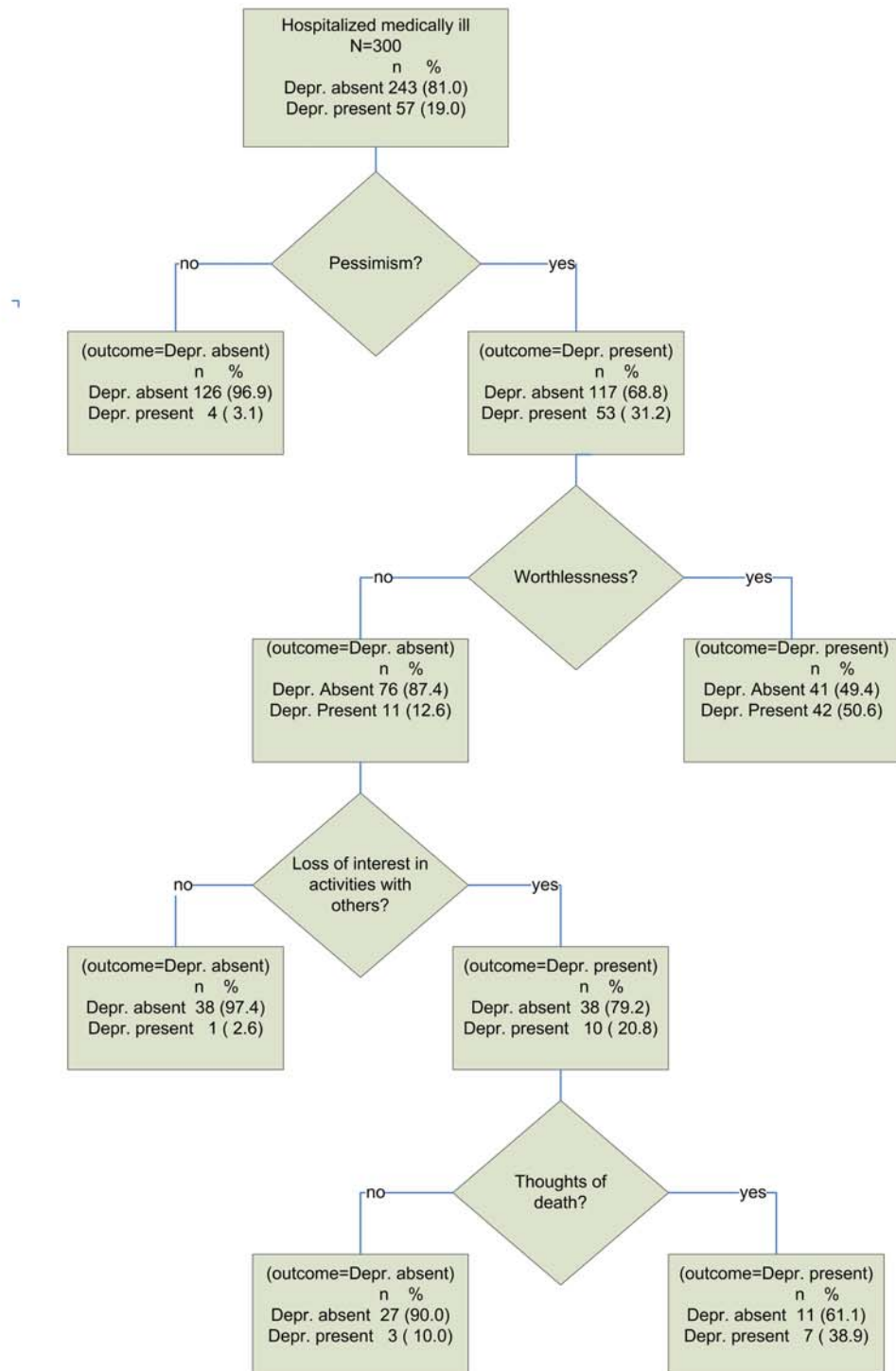


Table 2: Performance of CART-selected key demoralization and anhedonia symptoms and their combinations at screening for DSM-IV major depression

Symptom	Positive Predictive Value (%)	95% CI	Negative Predictive Value (%)	95% CI	Sensitivity (%)	95% CI	Specificity (%)	95% CI
Pessimism	31.2	24.3-38.7	96.9	92.3-99.2	93.0	83.0-98.1	51.9	45.4-58.3
Worthlessness	39.3	30.4-48.8	94.0	89.5-97.0	80.7	68.1-90.0	70.8	64.7-76.4
Less interest in activities with others	28.7	22.0-36.2	93.2	87.5-95.9	84.2	72.1-92.5	51.0	44.6-57.5
Thoughts of death	34.6	26.4-43.6	92.5	87.5-96.9	77.2	64.2-87.3	65.9	59.5-71.8
Pessimism present and Worthlessness present	50.6	39.4-61.8	93.1	88.9-96.1	73.7	60.3-84.5	83.1	77.8-87.6
Pessimism present, Worthlessness absent, Less interest in activities with others present, Thoughts of death present	38.9	17.3-64.3	82.3	77.3-86.5	12.3	5.1-23.7	95.5	92.0-97.0
Total CART tree	48.5	38.5-58.7	95.9	92.2-98.3	86.0	74.2-93.7	78.6	72.9-83.6

5. INTRODUCTION TO CHAPTER FIVE: SOMATIC AND PSYCHOLOGICAL DIMENSIONS OF SCREENING FOR PSYCHIATRIC MORBIDITY: A COMMUNITY VALIDATION OF THE SPHERE QUESTIONNAIRE

The following chapter is concerned with screening for psychiatric disorders such as anxiety and depression, using both psychological symptoms such as feeling miserable or sad, and somatic (physical, or pertaining to the body) symptoms such as sleeplessness and headaches. There is a strong association between somatic symptoms and anxiety and depression, as well as between psychological symptoms and these disorders ³⁵⁹. For example, the commonly employed DSM-IV diagnostic criteria ³⁶⁰ lists no less than four somatic symptoms – loss of weight or appetite, sleep disturbance, agitation or retardation, and fatigue, amongst the nine symptoms of major depression. Disorders such as depression can be difficult to recognise in primary care ³⁶¹, particularly when sufferers present with predominantly somatic symptoms ³⁶².

5.1. The Somatic and Psychological Health Report (SPHERE)

A self-report instrument that separately examines both psychological and somatic dimensions of anxiety and depression has recently been developed in Australia in an effort to improve the accuracy of screening for mental illness. This instrument, known as the Somatic and Psychological Health Report (SPHERE) ^{363,364}, was developed from other screening instruments such as the 30 item version of the General Health Questionnaire (GHQ-30) ³⁵⁵. A reanalysis of

published general practice data by Clarke and McKenzie ³⁶⁵; the results of which later appeared in a lead feature in the Melbourne Age newspaper ³⁶⁶, indicated that the SPHERE exhibited low specificity, and greatly overestimated the number of persons with possible psychiatric illness, leading to possible feelings of distress and stigma by those misclassified ³⁶⁷. Although the evidence for the efficacy of the SPHERE as a screening instrument in primary care is mixed, there has been a dearth of research into the performance of the SPHERE in community samples.

5.2. Screening for psychiatric disorders in young adults

Somatic symptoms of psychiatric illness are often shared with those of physical illness and this relationship increases with age ³⁶⁸. The study described in Chapter Five examines the screening and diagnostic performance of the SPHERE in 821 young Australian adults aged 22 to 34 (mean age = 28.23 years, SD = 2.29), who, due to their youth, would be presumed to have a low prevalence of comorbid physical illness. Participants had originally been recruited for a large-scale longitudinal study of the psychological effects of disaster exposure in childhood, following the major Ash Wednesday bushfires that occurred in Victoria and South Australia in February 1983, representing one of Australia's costliest natural disasters ³⁶⁹.

The study presented in Chapter Five looks at screening for lifetime and current (past month) any psychiatric diagnosis, as well as current depressive disorder and current anxiety disorder. Psychiatric diagnoses were obtained using the computer-assisted version of the CIDI ³²². Screening instruments consisted of

the PSYCH-6 and the SOMA-6 subscales of the 34 item SPHERE. The PSYCH-6 consists of six psychological symptoms of anxiety and depression, which are not listed in Chapter Five and hence are listed here - 'feeling nervous or tired', 'feeling unhappy and depressed', 'feeling constantly under strain', 'everything getting on top of you', 'losing confidence' and 'being unable to overcome difficulties'. The SOMA-6 consists of six somatic symptoms - 'muscle pain after activity', 'needing to sleep longer', 'prolonged tiredness after activity', 'poor sleep', 'poor concentration' and 'tired muscles after activity' ³⁶⁴.

Measures of test performance, including positive predictive value or power, negative predictive value, sensitivity, specificity, kappa, and efficiency or accuracy ³, were calculated based upon the published PSYCH-6 caseness threshold, (henceforth termed PSYCH), the published SOMA-6 caseness threshold, (henceforth SOMA), and their Boolean combinations (PSYCH or SOMA, PSYCH and SOMA). Thresholds consisted of scores of two or more for the psychological subscale, and three or more for the somatic subscale. Each item of the SPHERE is scored on a three point Likert scale ranging from 0, denoting 'never or some of the time', to 2, denoting 'most of the time'.

5.3. Further analysis using CART

The published paper presented in Chapter Five includes six tables of results, predominantly concerned with the specific measures of test performance mentioned above. In addition, overall measures of test performance comprising areas under the Receiver Operating Characteristic (ROC) curve ³⁷⁰ and kappa coefficients were statistically compared. The latter comparison employed the first

published method for comparing kappa coefficients obtained from a single sample, developed by McKenzie et al in 1996¹⁶⁷.

In a follow-up to the original study, described in an addendum to Chapter Five due to space limitations in the published paper, the CART technique was used to examine combinations of simple screening rules. In summary, the performance of simple screening rules involving caseness according to psychological and/or somatic criteria and their combinations, used in screening for psychiatric illness, was evaluated and compared.

Declaration for Thesis Chapter 5

McFarlane AC, **McKenzie DP**, Van Hooff M, Browne DG. Somatic and psychological dimensions of screening for psychiatric morbidity: a community validation of the SPHERE questionnaire. *Journal of Psychiatric Research* 2008; 65: 337-345.

Declaration by candidate

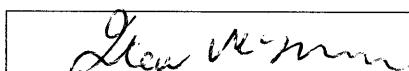
In the case of Chapter 5, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Major contribution to article concept (which follows on from a 2003 paper of which I was a co-author), literature review, introduction and discussion and revision of paper. Performed extra statistical programming and all statistical analyses. Interpreted and wrote results section. In addition, I was responsible for the additional literature review, CART analysis, and additional interpretation of the results in the addendum to the published article.	45%

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
A. McFarlane	Lead role in article concept, writing and revising of paper.	N/A
M. Van Hooff	Contributed to article concept, writing and revising of paper.	N/A
D. Browne	Contributed to article concept, writing and revising of paper.	N/A

Candidate's
Signature



Date

16-12-2008

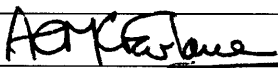
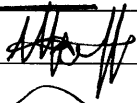
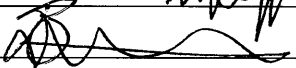
Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s) **Monash University Department of Epidemiology and Preventive Medicine, Alfred Hospital**

[Please note that the location(s) must be institutional in nature, and should be indicated here as a department, centre or institute, with specific campus identification where relevant.]

Signature 1		Date
Signature 2		11.12.08
Signature 3		10.12.08
Signature 4		10.12.08

Somatic and psychological dimensions of screening for psychiatric morbidity: A community validation of the SPHERE Questionnaire

Alexander C. McFarlane^a, Dean P. McKenzie^b, Miranda Van Hooff^{a,*}, Derek Browne^a

^aThe Centre for Military and Veterans' Health, University of Adelaide, Adelaide, South Australia

^bDepartment of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia

Received 12 July 2007; received in revised form 22 November 2007; accepted 6 February 2008

Abstract

Objective: Nonspecific somatic symptoms play an important role in the manifestation of psychiatric morbidity. This study examined the psychometric performance of an instrument developed to improve the rates of identification of psychiatric disorders, which incorporated somatic and psychological dimensions of distress (the Somatic and Psychological Health Report, or SPHERE). **Methods:** Eight hundred twenty-one adults who were participating in an epidemiological longitudinal study of the psychological impact of childhood disaster exposure (mean age of 28.23 years, S.D. of 2.29, range of 22–34) were recruited out of the original cohort of 1531. The 34-item SPHERE was administered in a self-report booklet, and the subjects were interviewed using the Composite International Diagnostic Interview to ascertain current and lifetime psychiatric diagnoses. **Results:** While the negative predictive power was high (96.1% for current disorder and 81.7% for lifetime disorder),

the positive predictive power was low (56.8% for a lifetime disorder and 27.5% for current disorder). This was despite 61.6% of lifetime sufferers and 78.6% who met current criteria for any disorder screened positive using the SPHERE. Sensitivity was highest when the broad “PSYCH or SOMA” screen was used (78.6%). Specificity of 89.5% was obtained for the “PSYCH and SOMA” scale. **Conclusions:** In this population of young adults, where age limited the prevalence of comorbid physical disease, the SPHERE was an acceptable screening measure. The psychometric performance was better for lifetime than current disorder. The psychometric characteristics of this instrument indicate that its particular use may be in defining individuals who need a more detailed assessment in a clinical setting.

Crown Copyright © 2008 Published by Elsevier Inc. All rights reserved.

Keywords: Somatic; Screen; Psychiatric disorder; Psychometric performance

Introduction

Epidemiological research has shown that there is a strong association between nonspecific somatic symptoms and psychiatric morbidity [1]. In fact, one of the common reasons for missing psychiatric diagnoses in general practice settings is the somatic presentation of anxiety and depression. This study examined the psychometric performance of an instrument developed to improve the rates of identification

of psychiatric morbidity by independently examining both psychological and somatic dimensions of distress (the Somatic and Psychological Health Report, or SPHERE). While this instrument has been administered in a variety of settings [2,3], particularly in primary care and general practice [4,5], to date there has been no published study comparing its psychometric performance against a structured diagnostic interview in a community-based sample.

It is well known that the rates of mental disorder in the general community are significantly different from those who present for treatment. Less than half of those living with mental disorders in the community, for example, receive treatment [6] in part because they are not detected in general practice settings [7]. One of the challenges, particularly because of the frequent comorbidity between physical illness

* Corresponding author. Centre for Military and Veterans' Health, University of Adelaide, Level 2/122 Frome Street, Adelaide 5000, South Australia. Tel.: +61 8 8303 5356; fax: +61 8 8303 5368.

E-mail address: miranda.vanhooff@adelaide.edu.au (M. Van Hooff).

and psychiatric/psychological disorder, is determining what role somatic symptoms could play in the more accurate identification of common psychiatric morbidities [8].

Hickie et al. [5] developed the 34-item SPHERE questionnaire “based upon the assumption that mental disorders in general practice are best characterized by some mix of psychological, and somatic distress”. The SPHERE was developed from the General Health Questionnaire, 30-item version (GHQ-30) [9], the Schedule of Fatigue and Anergia, and the Illness, Fatigue and Irritability Questionnaire [5]. Hickie et al. reported that 49% of 46,515 patients attending Australian general practitioners were classified as having a common mental disorder using the SPHERE (25% using the PSYCH and SOMA classification, 12% using PSYCH only, and 12% using the SOMA-only scale). In a substudy of 364 patients who were additionally examined using the Composite International Diagnostic Interview (CIDI), the PSYCH and SOMA scales correctly identified 40% of participants with a CIDI diagnosis in the past 12 months as having a disorder and 27% of those meeting criteria for a disorder in the past month [10,11]. The PSYCH-alone scale correctly identified 28% (12 months) and 17% (1 month), while the SOMA-alone scale performed the worst, identifying only 14% and 9% of 12-month and 1-month CIDI cases, respectively. This result was seen to demonstrate the utility of this instrument and particularly the benefit of focusing on both psychological and somatic symptoms. However, the precise psychometric performance of the instrument was not reported in this study.

Clark and McKenzie [12] reexamined Hickie et al.’s data in regard to the screening performance and the efficiency of the SPHERE compared with the GHQ-30. They found that the SPHERE had a very high false/positive rate. In one sample, 83% screened positive on the SPHERE, yet only 27% had a current psychiatric diagnosis. This raised some doubt as to the instrument’s appropriateness for use in a general practice setting.

One of the problems confronting somatic measures of psychological distress is that the physical symptoms of organic disease are shared with the nonspecific markers of psychological distress. The probability of this confounding relationship increases with the age of the sample being investigated. Equally, this confound is a particular challenge within general practice populations, given the fact that people are seeking assistance with physical disease. The use of psychometric screens in general practice populations [13] and at-risk population samples, such as defense forces [14], where nonspecific symptoms of somatic distress are common, highlights the need to establish the performance of a measure that also taps into somatic distress [15,16].

Against this background, this study examined the psychometric performance of the SPHERE when compared with the CIDI in a population sample of young adults, where age, which can impact on psychometric performance, is a protective factor against comorbid physical disorder. In this setting, the instrument was examined for its capacity to

screen for the presence of psychiatric disorder, in a young sample of adults who presumably, by virtue of their age alone, would have a low prevalence of physical illness.

Given the findings of Clarke and McKenzie [12], it was hypothesized that the SPHERE may be more suitable for population screening in conjunction with an interview rather than being used as a primary instrument for the detection of psychiatric caseness.

Methodology

Participants were 821 Australian adults comprising of 382 (46.5%) males and 439 (53.5%) females. The mean age of the sample was 28.23 years (S.D.=2.29; range, 22–34). Of the sample, 40.6% were married, 29.4% were never married, 26.4% were in de facto or common law relationships, 2.1% were separated, and 1.6% were divorced. The majority of participants (65.7%) were employed fulltime, 16.1% were working part time, or casual, and 10.2% reported home duties.

All participants were originally recruited as part of a large-scale longitudinal follow-up of children living in the Southeast of South Australia at the time of the Ash Wednesday Bushfires in 1983. All participants completed a self-report booklet (which included the SPHERE) and were interviewed over the telephone using the computerized version of the CIDI (CIDI-Auto Version 2.1) [10,11]. The total interview took approximately 1 h to complete and was conducted by experienced research psychologists at a time nominated by the participant.

The SPHERE-34 is a self-report screening tool for common mental disorders most commonly used in medical settings. Although composed of 34 items, the scoring is based on a subset of 12 items in order to create two subscales, PSYCH-6 (comprised of six items assessing psychological symptoms of depression and anxiety) and SOMA-6 (comprised of six items assessing somatic symptoms such as fatigue and pain). SPHERE questions are scored on a Likert scale with a score of 0 for “never or some of the time”, 1 for “a good part of the time”, and 2 for “most of the time”. The timeframe applied to all questions is “the last few weeks”.

Both the PSYCH-6 and SOMA-6 subscales have been reported to have high internal consistency (PSYCH-6: 0.90; SOMA-6: 0.80) and test–retest reliability (PSYCH-6: 0.81; SOMA-6: 0.80) [17]. Cutoffs for determining caseness on both subscales were derived by Hickie et al. [5] with a score of ≥ 2 identifying a patient who is a PSYCH-6 case and a score of ≥ 3 identifying a patient who is a SOMA-6 case. Participants are categorized according to their combination of psychological and somatic symptoms into “PSYCH or SOMA” (the broadest screen that identifies participants who are a case on either the PSYCH scale or the SOMA scale), “PSYCH and SOMA” (the narrowest screen, identifying participants who are a case on both), PSYCH only (which

identifies participants who are only a PSYCH case and not a SOMA case), and SOMA only (identifying participants who are only a SOMA case but not a PSYCH case) [5].

Diagnostic and Statistical Manual for Mental Disorders—Fourth Revision (DSM-IV) disorders were assessed using the computer assisted Composite International Diagnostic Interview (CIDI-Auto Version 2.1) [10,11]. The CIDI is a structured, standardized, and comprehensive interview used to assess current and lifetime prevalence of mental disorders in adults, based on the *DSM-IV* [18].

Symptoms were scored according to the standard CIDI scoring criteria of 0 (“indeterminate diagnosis”), 1 (“criteria not met”), 3 (“positive criteria met but exclusion not met”), and 5 (“all diagnostic criteria met”). The participants were diagnosed with a lifetime CIDI disorder if they scored 3 or 5 at any time during their lifetime. A current CIDI diagnosis was given if the participant exhibited the full symptom profile for that disorder in the month prior to assessment.

Studies have found that the CIDI has excellent interrater reliability, and satisfactory test–retest reliability and validity in a variety of settings worldwide [19,20]. To ensure reliability and validity in the current study, research psychologists who had extensive experience and training in the CIDI, telephone recruitment, interviewing, and psychiatric assessment conducted the interviews. A panel consisting of a psychiatrist and three research psychologists reviewed the scoring of structured interviews on a weekly basis to ensure interrater reliability.

Disorders were grouped into depressive disorders, anxiety disorders, and any disorder. Depressive disorders included 296.2×, 296.3×, 311, and 300.4. Anxiety disorders included 300.01, 300.21, 300.22, 300.29, 300.23, 300.3, 300.02, 300.00, and 309.81. Any disorder included any depressive disorder, any anxiety disorder, bipolar disorders (296.0×, 296.40, 296.4×, 296.6×, 296.5×, 296.7, 296.89, 301.13, 296.80), and eating disorders (307.1, 307.51, 307.50).

It is difficult to assess the performance of a screening test with just a single statistic and so a variety of measures are generally used [21]. In order to examine overall test performance, the area under the receiver operating characteristic (ROC) curve [22] was computed. This area can be interpreted as the probability that the test (e.g., in this case SPHERE) score of a randomly selected person with a diagnosis (in the present example) will be higher than the test score of a randomly selected person without a diagnosis [21]. Chance performance is equal to 0.50 (or 50.0%). Although generally used to measure chance-corrected agreement between raters, the κ coefficient [23] can also be employed to measure chance-corrected agreement between diagnostic instruments and screening tests [21]. The above statistics provide a general overview of how well screening tests perform, but they do not provide specific information on the type of errors (e.g., false positives and false negatives) made.

Specific measures of test performance including sensitivity (the probability of having a positive test result among

patients with a positive diagnosis), specificity (the probability of having a negative test result among patients with a negative diagnosis), positive predictive power (the probability of having a positive diagnosis among those patients having a positive test result), and negative predictive power (the probability of having a negative diagnosis among those patients with a negative test) [23] were therefore computed. The above statistics and their associated confidence intervals were calculated using the computer program developed by Mackinnon [24] and employed by Clarke and McKenzie [12]. To reduce the likelihood of obtaining statistical significance merely due to a wide range of comparisons being performed, only the overall measures of test performance such as area under the ROC curve (AUC) and the κ coefficient were statistically compared across different screening (e.g., PSYCH-alone vs. SOMA alone) and diagnostic tests (e.g., lifetime vs. current). Areas under the ROC curve were calculated and compared using version 9 of the Stata statistical package [25]. κ Coefficients were compared using a method introduced by McKenzie et al. [26]. Based upon the bootstrap resampling procedure [27], this method has recently been employed in a variety of applications [28,29] and is implemented in an updated version of an earlier computer program [30].

Results

Performance of the broadest method of screen (PSYCH or SOMA) in detecting current and lifetime “any” psychiatric disorder is reported in Table 1.

Of the participants, 34.1% screened positive on the SPHERE, 11.9% received a current psychiatric diagnosis, and 31.4% reported a lifetime history of psychiatric disorder.

Table 1 indicates that 61.6% of lifetime sufferers and 78.6% meeting criteria for “any” current disorder screened positive using the SPHERE. The likelihood of a person screening positive and being a true case was higher for lifetime disorder compared to current disorder (positive predictive power: 56.8% for a lifetime disorder and 27.5% for current disorder).

Overall negative predictive power was high for both current and lifetime disorder (current 96.1% and lifetime 81.7%).

κ Results indicate fair (κ values of 0.21–0.40) to moderate (κ values of 0.41 to 0.60) agreement (as arbitrarily but conventionally defined by Landis and Koch 1977 [31]) between the SPHERE and CIDI. Comparison of κ values for current vs. lifetime diagnoses showed a statistically significant difference (−11.2, 95% CI=−17.9 to 4.5), suggesting that the SPHERE PSYCH or SOMA screen is significantly better at detecting lifetime psychiatric disorder than current disorder.

Table 2 repeats this procedure for the narrower screen (PSYCH and SOMA), the PSYCH-alone scale and the SOMA-alone scale for “any” current diagnosis.

Table 1
Derived screening characteristics for SPHERE (PSYCH or SOMA) for “any diagnosis” (current vs. lifetime)

	Any diagnosis, current			Any diagnosis, lifetime		
	Yes	No		Yes	No	
SPHERE +	77	203	<i>n</i>	159	121	<i>n</i>
(PSYCH or SOMA) –	21	520	<i>n</i>	99	442	<i>n</i>
			95% CI			95% CI
CIDI diagnosis rate	11.9%		9.8–14.4%	31.4%		28.3–34.7%
Screen-positive rate	34.1%		30.9–37.5%	34.1%		30.9–37.5%
Sensitivity	78.6%		69.1–86.2%	61.6%		55.4–67.6%
Specificity	71.9%		68.5–75.2%	78.5%		74.9–81.8%
Positive predictive power	27.5%		22.4–33.1%	56.8%		50.8–62.7%
Negative predictive power	96.1%		94.1–97.6%	81.7%		78.2–84.9%
Overall efficiency	72.7%		69.5–75.7%	73.2%		70.0–76.2%
κ	28.0		21.8–34.2	39.2		32.6–45.9
AUC—SPHERE PSYCH	80.1		75.2–85.0	72.9		69.3–76.5
AUC—SPHERE SOMA	76.1		71.3–81.0	71.5		67.8–75.2

Sensitivity remained highest when the broad PSYCH or SOMA screen was utilized (78.6%), however positive predictive power rose to 37.7% when the PSYCH and SOMA screen was used.

Specificity rose to 90.7% for the PSYCH-alone scale, to 91.7% for the SOMA-alone scale, and to 89.5% for the PSYCH and SOMA scales. Negative predictive power remained highest for the broad PSYCH or SOMA screen but ranged between 88.2% and 92.6% for the three other screens.

Overall, the highest level of agreement between the SPHERE and current CIDI diagnosis was exhibited for the PSYCH and SOMA screen (κ , 32.9) followed closely by the PSYCH or SOMA scale (κ , 28). Agreement between the SPHERE and CIDI when the PSYCH-alone or SOMA-alone screen was only slight (PSYCH-alone κ is 13.9 and SOMA-alone κ is –0.1). In fact, the negative κ value for the soma scale alone indicates the level of agreement is less than would be expected by chance. The overall efficiencies of the four levels of screen ranged from the 72.7% (PSYCH or SOMA) to 84.4% (PSYCH and SOMA).

For current depressive disorder, the same pattern of results emerged (Table 3). As expected, the CIDI diagnosis rate for current depressive disorder (4.4%) was lower than the diagnosis rate of “any” current disorder (11.9%). Sensitivity for the PSYCH OR SOMA scale and the

narrower PSYCH and SOMA scales rose to 88.9% and 72.2%, respectively, whereas sensitivity for the PSYCH-alone scale dropped to 16.7%. Positive predictive power dropped by 11–18% across all screens. Interestingly, none of the participants with a current depressive disorder screened positive on the SOMA-alone scale; however, negative predictive power remained over 95% for all screens.

A comparison of the AUCs for SPHERE psychological and SPHERE somatic for current depressive disorder was statistically significant ($P=.013$), with SPHERE psychological having a significantly higher AUC than SPHERE somatic (90.7 vs. 83.4, $P<.05$).

Overall negative predictive power was very high for all screens, whereas positive predictive power was very low (21% for current depressive disorder). Overall efficiency was highest when the narrower screens were used.

For current anxiety disorder, the CIDI diagnosis rate was 10.4%, higher than the rate for current depression, but the performance (Table 3) closely resembled that for “any” current disorder. As with “any” depression and “any” disorder, the broad PSYCH or SOMA screen correctly identified the most cases (77.7%) of anxiety. Compared to

Table 2
Comparison of SPHERE screening for “any current diagnosis”

	PSYCH or SOMA	PSYCH alone	SOMA alone	PSYCH and SOMA
CIDI diagnosis rate	11.9%	11.8%	11.8%	11.9%
Screen-positive rate	34.1%	10.9%	8.3%	14.9%
Sensitivity	78.6%	22.7%	8.3%	46.9%
Specificity	71.9%	90.7%	91.7%	89.5%
Positive predictive power	27.5%	24.7%	11.8%	37.7%
Negative predictive power	96.1%	89.7%	88.2%	92.6%
Overall efficiency	72.7%	82.7%	81.8%	84.4%
κ	28.0	13.9	–0.1	32.9

Table 3
Comparison of SPHERE screening for “current depressive disorder”

	PSYCH or SOMA	PSYCH alone	SOMA alone	PSYCH and SOMA
CIDI diagnosis rate	4.4%	4.4%	4.4%	4.4%
Screen-positive rate	34.1%	10.9%	8.3%	14.9%
Sensitivity	88.9%	16.7%	0.0%	72.2%
Specificity	68.6%	89.4%	91.3%	87.8%
Positive predictive power	11.4%	6.7%	0.0%	21.3%
Negative predictive power	99.3%	95.9%	95.2%	98.6%
Overall efficiency	69.4%	86.2%	87.3%	87.1%
κ	13.6	3.6	–6.1	28.0
AUC—SPHERE PSYCH	90.7			
	(85.3–96.1)			
AUC-SPHERE SOMA	83.4			
	(77.3–89.5)			

Table 4

Comparison of SPHERE screening for “current anxiety disorder (including PTSD)”

	PSYCH or SOMA	PSYCH alone	SOMA alone	PSYCH and SOMA
CIDI diagnosis rate	10.4%	10.4%	10.3%	10.4%
Screen-positive rate	34.2%	10.9%	8.3%	14.9%
Sensitivity	77.7%	22.6%	9.5%	44.7%
Specificity	70.8%	90.5%	91.8%	88.5%
Positive predictive power	23.6%	21.4%	11.8%	31.2%
Negative predictive power	96.5%	91.1%	89.9%	93.3%
Overall efficiency	71.5%	83.5%	83.4%	83.4%
κ	24.1	12.7	1.5	27.9
AUC—SPHERE PSYCH	77.9 (72.5–83.3)			
AUC—SPHERE SOMA	74.3 (69.4–80.2)			

depression, sensitivity rose to 22.6% for the PSYCH-alone scale but dropped to 44.7% for the narrowest PSYCH and SOMA scales. Unlike in the current depressive disorder group, 9.5% of those with a current anxiety disorder were correctly identified using the SOMA-alone scale.

Positive predictive power was higher for current anxiety disorder compared to current depressive disorder across all screens ranging from 31.2% for the PSYCH and SOMA scales to 11.8% for the SOMA-alone screen.

As was the case in Tables 2 and 3, specificity was highest for the SOMA alone screen (91.8%) and lowest for the broadest screen PSYCH or SOMA screen (70.8%), and negative predictive power was highest for the broadest screen PSYCH or SOMA (96.5%).

There were no significant differences in κ values for PSYCH or SOMA vs. PSYCH and SOMA (difference=−3.8, 95% CI=11.5 to 3.96) or in the AUCs for SPHERE PSYCH vs. SPHERE SOMA for current anxiety disorder ($P=.25$) Table 4.

Performance of the SPHERE using empirical cutoffs derived from the AUC, for “any” current diagnosis, current depressive disorder, and current anxiety disorder, is reported in Table 5 and supported the need for unique cutoffs for different diagnostic categories. Optimum cutoffs for detecting cases of depression (≥ 3 for both the PSYCH and the SOMA scales) were higher than for anxiety (≥ 1 for the PSYCH scale and ≥ 2 for the SOMA scale). For “any”

current disorder, optimum cutoffs were ≥ 2 for the PSYCH scale and ≥ 2 for the SOMA scale.

For all three categories, any current disorder, current depressive disorder, and current anxiety disorder, the PSYCH scale performed significantly better overall at detecting current disorder than the SOMA scale, with efficiencies being highest for current depression (difference in κ for any current disorder=14.2, 95% CI=7.3–20.2; difference in κ for depression=10.4, 95% CI=4.3–16.6; difference in κ for anxiety=5.8, 95% CI=0.05–11.1).

Finally, Table 6 shows the derived screening characteristics for the SPHERE (PSYCH or SOMA) for any diagnosis (current and lifetime) based on empirical cutoffs (2+ for PSYCH and 2+ for SOMA). Using these new cutoffs, 70.2% of lifetime sufferers and 83.7% of those with a current disorder now screened positive using the SPHERE, which is slightly higher than the sensitivity derived using Hickie et al.’s [5] cutoffs; however, positive predictive power, specificity, and negative predictive power decreased with the new cutoffs.

Discussion

In general, the SPHERE performed as an acceptable screening instrument in this population of young adults. The psychometric performance was generally better when looking at a lifetime diagnosis of psychiatric disorder, in contrast to any current disorder. For example, in this population, the CIDI identified 11.9% as having a current psychiatric disorder. In contrast, 34.1% of the assessments screened positive on the SPHERE. However, of the 31.4% participants who reported a lifetime history of a psychiatric disorder, the specificity increased from 71.9% for any current disorder to 78.5% for lifetime disorder. The nature of the performance of the SPHERE as a screening instrument is possibly best demonstrated by the low positive predictive power, high negative predictive power, and overall efficiency of approximately 73%.

Using the PSYCH or SOMA method of scoring, the sensitivities and specificity of the scale were adequate. However, its low positive predictive power, especially for any current disorder, indicated that a significant number of individuals without a disorder were still being defined as a

Table 5

Comparison of SPHERE screening for “any current disorder,” “current depressive disorder,” and “current anxiety disorder (including PTSD)” using empirical cutoffs derived from AUC

	Any current disorder		Current depressive Disorder		Current anxiety disorder	
	PSYCH ≥ 2	SOMA ≥ 2	PSYCH ≥ 3	SOMA ≥ 3	PSYCH ≥ 1	SOMA ≥ 2
Sensitivity	70.4%	71.1%	86.1%	72.2%	78.8%	67.9%
Specificity	80.2%	66.8%	84.2%	79.1%	67.4%	65.8%
Positive predictive power	32.6%	22.3%	20.0%	13.7%	21.9%	18.5%
Negative predictive power	95.2%	94.5%	99.3%	98.4%	96.5%	94.7%
Overall efficiency	79.1%	67.3%	84.3%	78.8%	68.6%	66.0%
κ	33.7	19.5	27.2	16.9	21.2	15.4

Table 6

Derived screening characteristics for SPHERE (PSYCH or SOMA) for “any diagnosis” (current vs. lifetime) based on empirical cutoffs (2+ for psych, 2+ for soma)

	Any diagnosis, current			Any diagnosis, lifetime		
	Yes	No		Yes	No	
SPHERE +	82	283	N	181	184	N
(PSYCH or SOMA) –	16	440	N	77	379	N
			95% CI			95% CI
CIDI diagnosis rate	11.9%		9.8–14.4	31.4%		28.3–34.7
Screen-positive rate	44.5%		41.0–47.9	44.5%		41.0–47.9
Sensitivity	83.7%		74.8–90.4	70.2%		64.2–75.7
Specificity	60.9%		57.2–64.4	67.3%		63.3–71.2
Positive predictive power	22.5%		18.3–27.1	49.6%		44.3–54.8
Negative predictive power	96.5%		94.4–98.0	83.1%		79.4–86.4
Overall efficiency	63.6%		60.2–66.9	68.2%		64.9–71.4
κ	20.4		15.5–25.4	33.7		27.4–40.0
AUC—SPHERE PSYCH	80.1		75.2–85.0	72.9		69.3–76.5
AUC— SPHERE SOMA	76.1		71.3–81.0	71.5		67.8–75.2

case on the SPHERE. The finding that instrument had better psychometric performance for lifetime than current disorder suggests many of the screen-positive individuals who did not have a current diagnosis, the false positives, did, in fact, have a lifetime history of disorder. One benefit of the SPHERE is its ability to identify individuals with a lifetime vulnerability to psychiatric disorder as the group requiring further assessment for current disorder using a more thorough interview.

The potential benefit of an instrument such as the SPHERE might be that it tends to define those individuals who require a more detailed assessment, given its high negative predictive power and low positive predictive power. However, this raises the question as to the nature of those individuals who screen negative on the SPHERE. We examined the types of disorder experienced by this subgroup and found that the majority of these individuals had specific phobias, which are not typically associated with substantial levels of distress or disability.

In general, the sensitivities of the PSYCH-alone or SOMA-alone scales were low (see Table 2). The sensitivity of the PSYCH-alone scale, for any current disorder, any current depressive disorder, or any current anxiety, was much lower than in the study of Clarke and McKenzie [12]. The differences in the results of these two studies are indicative of the different performance of the SPHERE in a general practice as against a community sample of young adults.

As with all psychometric instruments, there is a tradeoff between specificity and sensitivity. κ Statistics suggest that the use of the PSYCH AND SOMA method of scoring, where caseness is defined if participant scores above the cutoff on both scales, was best for any current disorder, although this led to a substantial decrease in the sensitivity with only 14.9% of the population being screened positive compared with 34.1% for the PSYCH OR SOMA method of scoring. The latter method of scoring is possibly more equivalent to the rates of caseness obtained using a structured diagnostic interview such as the CIDI, suggesting that the

method of scoring should be used when the aim is to make population prevalence estimates.

In general, for both current depressive and current anxiety disorders, the overall efficiency and κ statistics suggested that the inclusion of the fatigue component may be an advantage when combined with the “AND” method of scoring in contrast to the “OR” method. However, as suggested by Clarke and McKenzie [12], the overall efficiency of the PSYCH and SOMA scales can be improved by modifying the threshold for both scales to a score of greater than or equal to 2 for any current disorder and greater than or equal to 3 for current depressive disorder. A smaller degree of improvement was found by modifying the thresholds (Table 5) for any current anxiety disorder to greater than or equal to 1 for the PSYCH scale and greater than or equal to 2 for the SOMA scale. These results suggest that the current cutoffs, as suggested by Hickie et al. [5], of 2 for the PSYCH scale and 3 for the SOMA scale may not be appropriate in all settings, especially in the detection of anxiety disorders in a younger, community population, where age serves as a protective factor against comorbid physical disorder.

The different methods of scoring the SPHERE were examined to determine the benefit of adding the somatic items to those characterizing psychological distress. The recommended scoring (PSYCH OR SOMA) had a sensitivity of 78.6% for those with a current psychiatric disorder. The PSYCH-alone scale had a sensitivity of 22.7%, the SOMA-alone scale had a sensitivity of 8.3%, and the PSYCH and SOMA scales had a sensitivity of 46.9%. Such results argue for the use of a scale that focuses on a combination of psychological and somatic symptoms, given that the PSYCH and SOMA scales performed better than either the PSYCH scale or the SOMA scale alone. In this study, the SOMA-alone scale only defined a small subgroup of participants (8.3%); however, when combined with the PSYCH-alone scale, the sensitivity increased substantially. These findings support the rationale for the development by

Hickie et al. [5] of the SPHERE to address the issue of somatic manifestations of psychiatric disorder.

These findings are similar to those of Smith et al. [32]. In this study of high-utilizing primary care patients with medically unexplained symptoms, they found that this population is better characterized by diagnoses of anxiety and depression rather than somatoform disorders. Other studies have reported a linear relationship between the number of unexplained physical symptoms and the severity of the anxiety/depression [33–35]. In general, it appears that anxiety disorders are the group where modifications of the cutoffs for the SPHERE might offer the most in terms of psychometric performance. This is important given the greater prevalence of anxiety disorders than depressive disorders among patients with prominent somatic symptoms [36]. Our data take these findings further with only those with a current anxiety disorder screening positive on the SOMA-alone scale. No participants with current depression met criteria for the SOMA-alone scale (Table 3), despite 4.4% prevalence of depressive disorders. One possible explanation for these findings is that somatic distress in young adults, in contrast to older age groups, seldom characterizes depressive disorder.

In general, Lowe et al. [37] have argued about the relative strategies for determining cutoffs. They have suggested that with a one-stage approach, such as that used in epidemiological and research studies, the best strategy is to have a good tradeoff between false positives and false negatives. On the other hand, in a setting where screening will lead to diagnosis and the instigation of treatment, a two-stage strategy is better suited. This implies identification of the optimal sensitivity for case detection. In this regard, the SPHERE generally performs well for the identification of current depression but less effectively for anxiety. However, the cases that are being missed with the anxiety disorders are likely to be specific phobia, as discussed above.

The methodological limitations of the study are that the CIDI interviews were administered by telephone in contrast to other studies that have been conducted in clinical settings with face-to-face interviews. This may have reduced the accuracy of the screening diagnoses and increased false-negative rates on the CIDI. Telephone interviews however allowed greater flexibility for participants possibly increasing response rates, as interviews could be conducted in their own homes at a time most suitable to them. The ability of the CIDI-Auto to accurately capture lifetime diagnoses where the participant is required to retrospectively report incidence, and age of onset of symptoms is also questionable. This, however, is an inherent problem in any instrument relying on recall to establish a diagnosis and is not a specific limitation of the CIDI.

Although the CIDI is one of the most common diagnostic instruments used in epidemiological research, the poor level of agreement between expert clinician's diagnoses and CIDI-Auto generated diagnoses requests acknowledgement [38–41]. In such a large-scale survey, however, it would

be impractical to use experienced clinicians to conduct interviews given the expense and time commitment required. A computerized instrument such as the CIDI-Auto, therefore, which has the additional strength of eliminating clinician bias, being cost-effective, time-efficient, and successful in eliminating data entry and scoring errors [42], is a valid and suitable choice for this type of research.

The size of the sample and the prevalence of cases, however, were sufficient to adequately test the psychometric properties of these questionnaires. In general, the findings of the present study should not be applied to more general screening populations as our study focused on young adults who, by virtue of their age, did not have the usual prevalence of somatic disease that would be present in a stratified population sample. The false positives on the SOMA scale due to somatic disease were not excluded, which may impact upon the psychometric performance of this instrument. Our study also did not utilize the entire 34-item scale, rather the scoring was limited to that suggested by Hickie et al. [5]. It is possible that an exploration of the factor structure of the questionnaire in a community population could improve its psychometric performance.

The psychometric performance of any instrument should be judged against the purpose for which it is being used. For example, if an instrument such as the SPHERE were being used in a questionnaire survey of a general population sample, which is aiming to measure prevalence, a high false-positive rate would be a significant disadvantage. In contrast, if the SPHERE instrument was being used as a primary screen leading to a secondary diagnostic interview, false-positive rate at the higher margins of tolerance may have some advantages in identifying an at-risk group that can be then subjected to a more thorough examination. This pattern of performance has been argued to justify the utility of other similar screening measures, such as the Hopkins Symptom Checklist-25, that pick up both distress and disorder [43]. However, it is beyond the scope of this article to compare the relative merits of the SPHERE with other available measures.

In conclusion, somatic symptoms were prevalent within this community. It appears that the addition of somatic items in a psychological screening questionnaire adds to the false-positive rates. This should be taken account of in the decision whether or not to use such an instrument. Goldberg and Bridges' [44] warning about the independence of somatic symptoms from anxiety and depression should remain until further evidence emerges to support the approach advocated by the use of instruments such as the SPHERE.

Such an instrument may be best suited for screening a general practice population for further diagnostic clarification, as it will miss relatively few cases. However, the high false-positive rate, using the suggested PSYCH or SOMA scoring cutoffs, will overestimate prevalence if it were used in an epidemiological study, and in this setting, the PSYCH and SOMA scoring method should be adopted.

Acknowledgments

This research was supported by a project grant from the Australian National Health and Medical Research Council (NHMRC Project Grant ID 201813 and program Grant 300403). The authors also wish to acknowledge the following organizations and personnel for their assistance and support with the project: The Births Deaths and Marriages Registration Office of South Australia; Australian Institute of Health and Welfare; the Australian Electoral Commission; and the teachers, principals and staff from the eight schools that participated in the study. A special thanks to all participants who generously gave their time, support, and cooperation with the project.

References

- [1] Katon W, Lin EH, Kroenke K. The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. *Gen Hosp Psychiatry* 2007;29:147–55.
- [2] Bennett B, Goldstein D, Lloyd A, Davenport T, Hickie I. Fatigue and psychological distress—exploring the relationship in women treated for breast cancer. *Eur J Cancer* 2004;40:1689–95.
- [3] Butow P, Meiser B, Price M, Bennett B, Tucker K, Davenport T, et al. Psychological morbidity in women at increased risk of developing breast cancer: a controlled study. *Psychooncology* 2005;14:196–203.
- [4] Wijeratne C, Hickie I, Davenport T. Is there an independent somatic symptom dimension in older people? *J Psychosom Res* 2006;61:197–204.
- [5] Hickie IB, Davenport TA, Hadzi-Pavlovic D, Koschera A, Naismith SL, Scott EM, et al. Development of a simple screening tool for common mental disorders in general practice. *Med J Aust* 2001;175 (Suppl):S10–7.
- [6] Henderson S, Andrews G, Hall W. Australia's mental health: an overview of the general population survey. *Aust N Z J Psychiatry* 2000;34:197–205.
- [7] Australian Bureau of Statistics. National Health Survey: summary of results (1989–90); 1991. Canberra: Australian Bureau of Statistics. Report No.: 4364.0.
- [8] Goldberg D. A dimensional model for common mental disorders. *Br J Psychiatry Suppl* 1996;44–9.
- [9] Goldberg DP, Williams P. A User's Guide to the General Health Questionnaire. Windsor, England: NFER-Nelson, 1988.
- [10] World Health Organization. Composite International Diagnostic Interview (CIDI-AUTO). Geneva: Author, 1997.
- [11] World Health Organization Collaborating Centre for Mental Health and Substance Abuse. Composite International Diagnostic Interview: CIDI-Auto 2.1—Administrator's guide and reference. Sydney: World Health Organization Collaborating Centre for Mental Health and Substance Abuse, 1997.
- [12] Clarke DM, McKenzie DP. An examination of the efficiency of the 12-item SPHERE questionnaire as a screening instrument for common mental disorders in primary care. *Aust N Z J Psychiatry* 2003;37:236–9.
- [13] Biddle L, Gunnell D, Sharp D, Donovan JL. Factors influencing help seeking in mentally distressed young adults: a cross-sectional survey. *Br J Gen Pract* 2004;54:248–53.
- [14] Rayner S. Prevalence of psychological trauma in operationally deployed Navy personnel: a baseline surveillance report. *ADF Health* 2005;6:81–4.
- [15] Forbes AB, McKenzie DP, Mackinnon AJ, Kelsall HL, McFarlane AC, Ikin JF, et al. The health of Australian veterans of the 1991 Gulf War: factor analysis of self-reported symptoms. *Occup Environ Med* 2004;61:1014–20.
- [16] Ristkari T, Sourander A, Ronning J, Helenius H. Self-reported psychopathology, adaptive functioning and sense of coherence, and psychiatric diagnosis among young men—a population-based study. *Soc Psychiatry Psychiatr Epidemiol* 2006;41:523–31.
- [17] Hickie IB, Koschera A, Davenport TA, Naismith SL, Scott EM. Comorbidity of common mental disorders and alcohol or other substance misuse in Australian general practice. *Med J Aust* 2001;175 (Suppl):S31–6.
- [18] American Psychiatric Association. Diagnostic and statistical manual of mental disorders: *DSM-IV*. Washington, DC: American Psychiatric Association, 1994.
- [19] Andrews G, Peters L. The psychometric properties of the Composite International Diagnostic Interview. *Soc Psychiatry Psychiatr Epidemiol* 1998;33(2):80–8.
- [20] Wittchen HU, Robins LN, Cottler LB, Sartorius N, Burke JD, Regier D. Cross-cultural feasibility, reliability and sources of variance of the Composite International Diagnostic Interview (CIDI). The Multicentre WHO/ADAMHA Field Trials. *Br J Psychiatry* 1991;159:645–653,658.
- [21] Swets J. Signal Detection theory and ROC analysis in psychology and diagnostics: collected papers. Mahwah (NJ): Erlbaum, 1995.
- [22] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
- [23] Kraemer HC. Evaluating medical tests: objective and quantitative guidelines. Newbury Park (Calif): Sage, 1992.
- [24] StataCorp. Stata statistical software version 8. College Station (Tex): Stata Corporation, 2002.
- [25] Mackinnon AJ. A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. *Comput Biol Med* 2000;30:127–34.
- [26] McKenzie DP, Mackinnon AJ, Peladeau N, Bruce PC, Onghena P, Clarke DM, et al. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? *J Psychiatr Res* 1996;30:483–92.
- [27] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall, 1993.
- [28] Elder D, Abramson M, Fish D, Johnson A, McKenzie D, Sim M. The SABRE (Surveillance of Australian workplace-Based Respiratory Events) Scheme: notifications for the first 3.5 years and results of a validation study for occupational asthma. *Occup Med* 2004;54:395–9.
- [29] Foody G. Thematic map comparisons: evaluating the statistical significance of differences in classification accuracy. *Photogramm Eng Remote Sensing* 2004;70:627–33.
- [30] McKenzie DP, Mackinnon AJ, Clarke DM. KAPCOM: a program for the comparison of kappa coefficients obtained from the same sample of observations. *Percept Mot Skills* 1997;85:899–902.
- [31] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
- [32] Smith RC, Gardiner JC, Lyles JS, Sirbu C, Dwamena FC, Hodges A, et al. Exploration of *DSM-IV* criteria in primary care patients with medically unexplained symptoms. *Psychosom Med* 2005;67:123–9.
- [33] Katon WJ, Walker EA. Medically unexplained symptoms in primary care. *J Clin Psychiatry* 1998;59(Suppl 20):15–21.
- [34] Kroenke K, Spitzer RL, Williams JB, Linzer M, Hahn SR, deGruy FV, et al. Physical symptoms in primary care. Predictors of psychiatric disorders and functional impairment. *Arch Fam Med* 1994;3:774–9.
- [35] Kisely S, Goldberg D, Simon G. A comparison between somatic symptoms with and without clear organic cause: results of an international study. *Psychol Med* 1997;27:1011–9.
- [36] Henningsen P, Jakobsen T, Schiltenswolf M, Weiss MG. Somatization revisited: diagnosis and perceived causes of common mental disorders. *J Nerv Ment Dis* 2005;193:85–92.

- [37] Lowe B, Spitzer RL, Grafe K, Kroenke K, Quenter A, Zipfel S, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord* 2004; 78:131–40.
- [38] Komiti AA, Jackson HJ, Judd FK, Cockram, Kyrios M, Yeatman R, et al. A comparison of the Composite International Diagnostic Interview (CIDI-Auto) with clinical assessment in diagnosing mood and anxiety disorders. *Aust N Z J Psychiatry* 2001;35:224–30.
- [39] Peters L, Clark D, Carroll F. Are computerized interviews equivalent to human interviewers? CIDI-Auto versus CIDI in anxiety and depressive disorders. *Psychol Med* 1998;28:893–901.
- [40] Rosenman SJ, Korten AE, Levings CT. Computerised diagnosis in acute psychiatry: validity of CIDI-Auto against routine clinical diagnosis. *J Psychiatr Res* 1997;31:581–92.
- [41] Peters L, Andrews G. Procedural validity of the computerized version of the Composite International Diagnostic Interview (CIDI-Auto) in the anxiety disorders. *Psychol Med* 1995;25:1269–80.
- [42] Erdman HP, Klein MH, Greist JH. Direct patient computer interviewing. *J Consult Clin Psychol* 1985;53:760–73.
- [43] Sandanger I, Moum T, Ingebrigtsen G, Sorensen T, Dalgard OS, Bruusgaard D. The meaning and significance of caseness: the Hopkins Symptom Checklist-25 and the Composite International Diagnostic Interview. II. *Soc Psychiatry Psychiatr Epidemiol* 1999;34: 53–9.
- [44] Goldberg D, Bridges K. Minor psychiatric disorders and neurasthenia in general practice. In: Gastpar M, Kielholz P, editors. *Problems of Psychiatry in General Practice*. Lewiston (NY): Hogrefe & Huber, 1991. pp. 79–88.

5.4. Addendum to Chapter Five: Further analysis of simple SPHERE screening rules using CART

5.4.1. Introduction

The preceding paper looked at the performance of SPHERE screening rules for any current (past month) diagnosis, current depressive disorder, and current anxiety disorder. The simple conjunctive decision rule, PSYCH and SOMA ('diagnosis present if PSYCH and SOMA caseness both present'), exhibited the highest positive predictive value and kappa coefficient for these three diagnoses. Sensitivity for this rule was low, however, in the case of any current diagnosis (46.9%) and current anxiety disorder (44.7%), limiting its utility as a possible screening instrument. On the other hand, the simple disjunctive decision rule, PSYCH or SOMA ('diagnosis present if either PSYCH or SOMA, or both, present'), exhibited higher sensitivity, but lower specificity and positive predictive value, than the conjunctive rule.

The preceding paper recommended that PSYCH or SOMA was the simple rule best suited for use in screening for possible mental illness within a general practice population. Although not found to be as low as that observed in an earlier study³⁶⁵, the low specificity of the above rule led to the suggestion that PSYCH and SOMA would be better suited for estimating prevalence of mental illness within an epidemiological setting.

In an effort to provide an additional perspective on the results obtained in the preceding paper, an examination of whether some combination of the above simple decision rules could perform better than a single rule alone was

performed. Classification and Regression Tree (CART) analysis was used as it constructs readily interpretable decision trees, which can be used as screening and diagnostic tests, and because it automatically tests the generality of its results.

5.4.2. Method

CART was employed in an identical fashion to that described in Chapters Two through Four, including dividing the original dataset into a learning subsample of 75% of the observations, and a validation subsample of 25% of the observations. The classification performance obtained with the tree grown using the learning subsample was compared with that tree applied to a validation subsample of 25% of the dataset. If the difference in performance, assessed using the chi-square statistic, did not approach statistical significance, the two subsamples were combined.

Unlike most other decision tree-building procedures⁹⁵, CART includes a built-in facility for assigning costs to the different types of classification errors made. Thus, for example, false negatives could be weighted more highly than false positives, if a test with high specificity was required. Each CART analysis was performed with three sets of costs, 1/ equal (the default), 2/ increased sensitivity, with the cost of false negatives weighted 75% higher than the cost of false positives (this value was also employed in the study presented in Chapter Six), and 3/ increased specificity, with the cost of false positives weighted 75% higher than the cost of false negatives.

CART analyses were conducted for any current psychiatric diagnosis, current depressive disorder, and current anxiety disorder separately. The four simple rules analysed by CART consisted of PSYCH and SOMA, PSYCH or SOMA, PSYCH alone (caseness only on the Psychological subscale of the SPHERE, not the Somatic subscale) and SOMA alone (caseness only on the Somatic subscale of SPHERE, not the Psychological subscale). The young adults whose SPHERE subscale scores did not meet PSYCH caseness, or SOMA caseness, would have a value denoting 'absent' for each of the above simple rules.

The preceding paper compared various kappa coefficients, including those obtained for the best screening rule, PSYCH or SOMA, versus those obtained for the best diagnostic rule, PSYCH and SOMA, using McKenzie et al's ¹⁶⁷ bootstrap resampling procedure for comparing kappa coefficients obtained from a single sample. This method has been extended for the present analyses in order to compare screen positive rate, sensitivity, specificity, positive predictive value, negative predictive value and efficiency as well as kappa, between the combination rule generated by CART and the simple rules described in the preceding paper. As with the previous analyses, an updated version of an earlier computer program for comparing correlated kappa coefficients ³⁷¹ was used.

CART analyses were performed using version 6 of the CART software ⁸¹. Chi-squared analyses, and the splitting the dataset into learning and validation subsets, were undertaken using version 15 of the SPSS statistical package ⁸².

5.4.3. Results

With regard to any current psychiatric diagnosis and current anxiety disorder, CART firstly split the sample by PSYCH or SOMA. Subsequent splits were found not to lead to an increase in cross-validated classification accuracy and so were not retained. The single split on PSYCH or SOMA was obtained for all three sets of costs defined above. With respect to any current depressive disorder, weighting for increased specificity again led CART to split the sample by PSYCH and SOMA, with subsequent splits not leading to an increase in cross-validated accuracy. The other two sets of weights led to an identical tree being generated, shown in Figure 2. There was no statistically significant difference ($\chi^2 = 0.9$, $df = 3$, $p = 0.82$) between the overall performance of this tree obtained for the learning subset, and the overall performance of the tree applied to the validation subset. The two subsets were therefore combined.

As can be seen in Figure 2, CART first split the dataset by PSYCH and SOMA. The subgroup for which this rule was true could not be split any further. Just over twenty one percent (21.3%) of those young adults for whom this rule was true had a current CIDI DSM-IV diagnosis of depressive disorder, compared with 1.4% of those for whom this rule was not true. The latter subgroup was then split according to whether or not PSYCH alone was present. Almost seven percent (6.7%) of young adults exhibiting caseness solely on the PSYCH-6 subscale had a current CIDI DSM-IV diagnosis of depressive disorder, compared with 0.7% in those for whom this rule was not true. The latter subgroup consisted of those who met caseness solely on the SOMA-6 subscale, as well as those

who did not meet caseness using either subscale. No further splits could be performed.

The performance of the overall CART decision tree for current depressive disorder described above, obtained using the default weighting of errors, as well as weighting for increased sensitivity, is given in Table 2. The screening and diagnostic performance of PSYCH or SOMA, the optimum screening rule (highest sensitivity) found in the preceding paper is repeated for ready comparison (CART found the optimum diagnostic rule to be PSYCH and SOMA). As shown in Table 2, the CART screening rule performed significantly better (the 95% confidence interval does not include zero) than PSYCH or SOMA on all measures of test performance apart from sensitivity, where it equalled the performance of the latter screening rule.

5.4.4. Discussion

CART generated a simple decision tree that first ascertained whether PSYCH and SOMA are both present, and then if not, ascertained whether PSYCH alone is present. A screening rule based upon this tree exceeded or equalled the performance of the PSYCH or SOMA screening rule on all measures of test performance. The CART tree might therefore be useful as a screening rule, although more testing in other populations is required, while the positive predictive value is still somewhat low (15.2%).

In the case of any current disorder and any anxiety disorder, CART merely selected PSYCH or SOMA, which was found to be the best screening rule in the preceding paper. Similarly, in the case of any current depressive disorder,

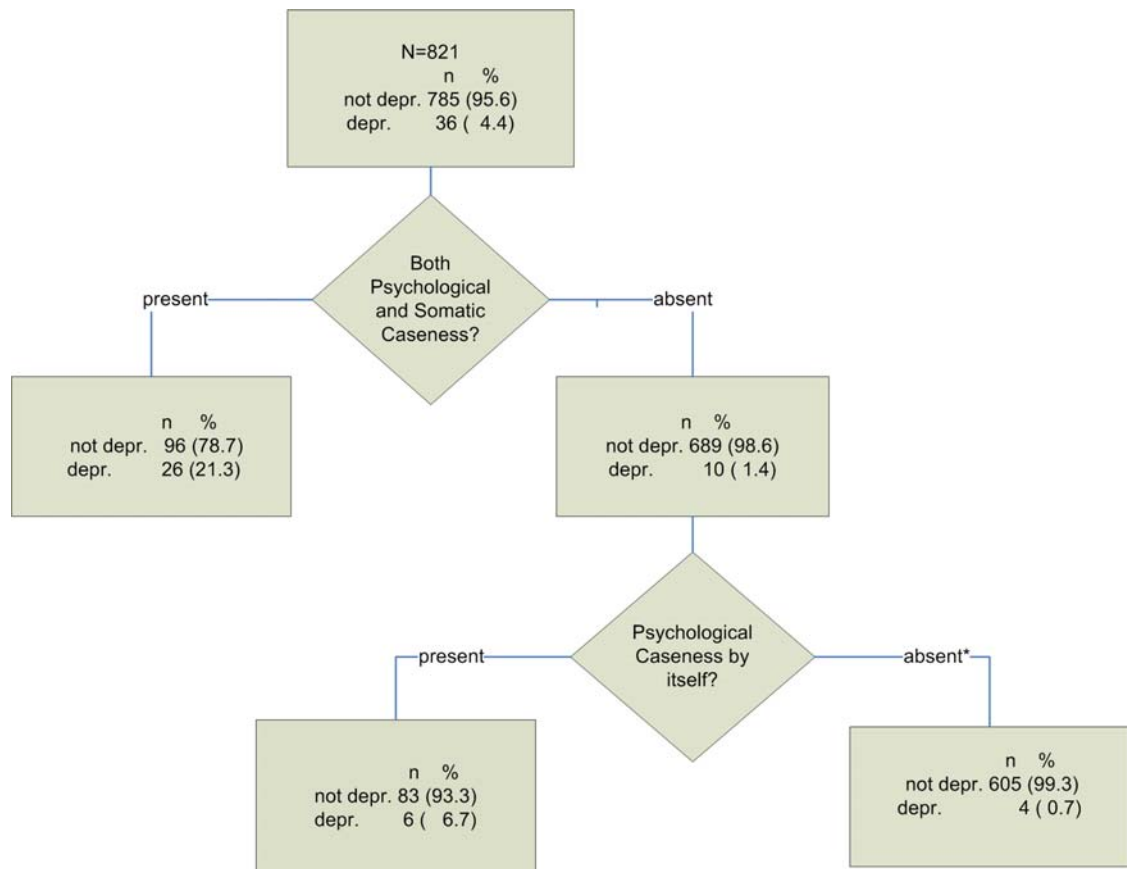
weighted so as to increase specificity, CART merely selected PSYCH and SOMA, which was found to be the best diagnostic rule in the preceding paper.

Although it appears that Boolean combinations of thresholds perform better than thresholds employed alone, pairs of thresholds linked by AND as well as OR were themselves entered into CART, rather than the procedure finding such combinations by itself. Ideally, CART could search for combinations of elements at a particular point in the tree, as described in the original CART manuscript⁸⁰ however this has not yet been implemented.

Comparison with other brief screening tests for anxiety and depression³⁷², including those based upon the GHQ³⁷³ or indeed the GHQ itself^{355,374} is required in future research, as it is not clear that the SPHERE performs significantly better than the GHQ³⁶⁵. Furthermore, caution should be exercised in primary care settings to ensure that the focus is not on symptoms and signs of psychological distress to the exclusion of obtaining information on patients' life circumstances. Both symptoms and life circumstances should be considered by clinicians when developing treatments³⁶⁷. I will return to this point in Chapter Seven.

CART was able to significantly improve upon a simple screening rule for current depressive disorder, as well as offering a ready facility for cross-validating models. The technique should arguably be incorporated as part of the general psychiatric test constructor's toolkit.

Figure 2. Classification and Regression Tree (CART) analysis of current depressive disorder in a community sample of young Australian adults.



(* this subgroup includes those with Somatic Caseness present, as well those who don't meet Psychological Caseness or Somatic Caseness)

Table 2: Comparison of CART tree and PSYCH or SOMA rule for screening for 'Current Depressive Disorder'

	CART tree	PSYCH or SOMA	Difference	95% CI
Screen Positive Rate	25.7%	34.1%	-8.4%	-10.2%--6.5%
Sensitivity	88.9%	88.9%	0.0%	
Specificity	77.2%	68.6%	8.6%	6.7%-10.7%
Positive Predictive Value	15.2%	11.4%	3.8%	2.4%-5.3%
Negative Predictive Value	99.3%	99.3%	0.1%	0.02%-0.2%
Overall Efficiency	77.7%	69.4%	8.2%	6.4%-10.2%
Kappa Coefficient	19.9%	13.6%	6.3%	4.2%-8.6%

6. INTRODUCTION TO CHAPTER SIX: MODELS DEVELOPED BY THREE TECHNIQUES DID NOT ACHIEVE ACCEPTABLE PREDICTION OF BINARY TRAUMA OUTCOMES

The previous chapters have been concerned with the identification of psychiatric illness. The study presented in Chapter Six relates to the identification of physical trauma. This study is primarily methodological, and techniques for predicting binary outcomes according to clearly defined clinical guidelines are also relevant in psychiatric applications such as the presence of depression. Moreover, there is an increasing amount of research into the relationship between physical trauma such as traumatic brain injury, and psychiatric disorders such as depression and PTSD ³⁷⁵, particularly amongst Iraq War veterans ³⁷⁶.

Predicting which persons will exhibit a particular binary outcome, such as developing an illness, or requiring hospital stay, or who will develop further complications or even die, is a fundamental issue in health research. As highlighted in the Introduction (Chapter One), classification and prediction models must not only exhibit satisfactory performance on the data used to develop them, they must be generalisable to other datasets.

6.1. Trauma datasets

The study conducted in Chapter Six employed various statistical and machine learning methods to predict two separate types of binary outcome – admission to an Intensive Care Unit, and patient survival. Models were developed using 4014 blunt (not involving intrusion through the skin) trauma

cases from the adult trauma database of the Royal Melbourne Hospital (RMH), a university-affiliated major metropolitan general hospital in Melbourne, Australia. Models were tested or validated using 3205 blunt trauma cases from the Victorian State Trauma Registry (VSTR). The use of an external validation dataset is not common in trauma research, with a systematic review published in 2008 revealing that only four of the 31 trauma studies reviewed had employed such a dataset ³⁷⁷.

6.2. Comparison of techniques

The studies comparing various statistical and machine learning methods outlined in Chapter One, have generally used one or more researchers to apply all such methods. The design of the study presented in Chapter Six is unique in that statistical analyses were carried out by researchers who were highly skilled in, and had published papers on the application of, a particular technique. Thus two of the researchers applied backward elimination stepwise logistic regression ²⁷, one researcher applied artificial neural networks consisting of multiple layers of inter-connected neurons ^{38,39}; and one researcher applied Classification and Regression Trees (CART). The VSTR validation data were only made available to the researchers after submission of their final models developed using the RMH data. As is discussed in more detail in the published paper comprising Chapter Six, all three of the above techniques have previously been used in the analysis of trauma data. In addition, another type of recursive partitioning algorithm ¹⁶⁵ was recently employed to create the Canadian C-Spine (cervical spine) rule for determining the use of radiography in trauma patients ^{378,379}. The

latter point was omitted in Chapter Six due to space restrictions on the published paper.

Predictors employed in the study presented in Chapter Six, selected on the basis of prior research, included age, respiratory rate, systolic blood pressure, cause of injury, severity of injury and a widely used measure of coma and impaired consciousness, the Glasgow Coma Scale (GCS) ³⁸⁰. In summary, logistic regression, artificial neural networks and CART are applied in the prediction of binary outcomes, according to pre-specified performance criteria.

Declaration for Thesis Chapter 6

Wolfe R, **McKenzie DP**, Black J, Simpson P, Gabbe BJ, Cameron PA. Models developed by three techniques did not achieve acceptable prediction of binary trauma outcomes. *Journal of Clinical Epidemiology* 2006, 59, 26-35.

Declaration by candidate

In the case of Chapter 6, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
I contributed to the article concept, as well as performed specific literature review for, and wrote the sections in, Introduction pertaining to Classification and Regression Trees (CART), and artificial neural networks, and the applications of both these methods in trauma research. Wrote section in Methods pertaining to CART, and contributed to Methods section on neural networks. Applied CART to data and interpreted the results. Wrote section in Discussion pertaining to CART, and also majorly contributed to general Discussion as to how the statistical methodology employed by CART, neural networks and logistic regression could be expanded and improved. Made a major contribution to revision of the paper.	45%

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
R. Wolfe	Lead role in article concept and writing and revising of the paper. Responsible for reviewing and approving all aspects of the research methodology and design, Performed logistic regression analyses and interpreted the results.	N/A
J. Black	Performed neural network analyses and interpreted the results. Contributed to drafting of the paper.	N/A
P. Simpson	Performed logistic regression analyses and interpreted the results. Contributed to drafting of the paper.	N/A
B. Gabbe	Data collection and cleaning. Responsible for interpretation of results in the context of the trauma registry data and contributed to drafting of paper.	N/A
P. Cameron	Responsible for clinical interpretation of results and contributed to drafting of paper.	N/A

Candidate's
Signature

[Handwritten Signature]

Date
16-12-2008

Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s)

Monash University Department of Epidemiology and Preventive Medicine, Alfred Hospital

[Please note that the location(s) must be institutional in nature, and should be indicated here as a department, centre or institute, with specific campus identification where relevant.]

Signature 1

[Handwritten Signature]

Date 15/12/08

Signature 2

Signature 3

[Handwritten Signature]

15/12/2008

Nb: Signed by Prof Michael Abramson, Deputy Head, Department of Epidemiology and Preventive Medicine, as Dr Black was overseas and out of email contact at the time of signing. Proof of this can be supplied if required.

Signature 4

[Handwritten Signature]

17/12/2008

Signature 5

[Handwritten Signature]

15/12/2008

Signature 6

[Handwritten Signature]

15/12/2008

Models developed by three techniques did not achieve acceptable prediction of binary trauma outcomes

Rory Wolfe^{a,*}, Dean P. McKenzie^a, James Black^b, Pam Simpson^a,
Belinda J. Gabbe^a, Peter A. Cameron^a

^a*Department of Epidemiology and Preventive Medicine, Monash University, Central and Eastern Clinical School, Melbourne, Victoria 3004, Australia*

^b*NHMRC Centre for Clinical Research Excellence in Infectious Diseases, Victorian Infectious Diseases Service, Royal Melbourne Hospital, Melbourne, Victoria, Australia*

Accepted 31 May 2005

Abstract

Background and Objectives: To develop prediction models for outcomes following trauma that met prespecified performance criteria. To compare three methods of developing prediction models: logistic regression, classification trees, and artificial neural networks.

Methods: Models were developed using a 1996–2001 dataset from a major trauma center in Victoria, Australia. Developed models were subjected to external validation using the first year of data collection, 2001–2002, from a state-wide trauma registry for Victoria. Different authors developed models for each method. All authors were blinded to the validation dataset when developing models.

Results: Prediction models were developed for an intensive care unit stay following trauma (prevalence 23%) using information collected at the scene of the injury. None of the three methods gave a model that satisfied the performance criteria of sensitivity >80%, positive predictive value >50% in the validation dataset. Prediction models were also developed for death (prevalence 2.9%) using hospital-collected information. The performance criteria of sensitivity >95%, specificity >20% in the validation dataset were not satisfied by any model.

Conclusion: No statistical method of model development was optimal. Prespecified performance criteria provide useful guides to interpreting the performance of developed models. © 2006 Elsevier Inc. All rights reserved.

Keywords: Prediction model; Logistic regression; Classification and regression trees; Neural networks; Performance criteria; External validation

1. Introduction

Interest in developing prognostic models for binary outcomes is widespread, and guidelines exist for their creation [1] and validation [2] as applied to health outcomes. A key element of model creation is the use of a large database containing variables that will be available if the new model is applied in routine practice. This creation step involves trade off between (1) complexity of model; the more complex the model, the more attuned it becomes to observed features of the database from which it is developed, and (2) transportability of the model, that is, how well it performs with different databases and when it is used in practice. Validation involves two critical elements. First, “external validation”: the performance of the newly created model must be tested on a second dataset that was not a part

of model creation but that is representative of the same population [3,4]. Second, the model must be validated in the context of prespecified utility scores for correct predictions and/or indicators that represent acceptable model performance when used in routine practice [2]. These indicators represent minimum performance thresholds for the model to be of practical use when applied.

Numerous statistical methods have been applied to create prognostic models for binary outcomes. Many comparisons between different methods exist [4–11]; however, there is no consensus as to an optimal method. It is therefore prudent to explore different methods.

We present a case study of prognostic model development for binary outcomes that involves separate model-creation and validation datasets and prespecified performance indicators. Three statistical methods were employed for the creation of a prognostic model; logistic regression, classification trees, and neural networks. Logistic regression has become a mainstay of medical research; the latter two techniques are commonly used with data involving interactions

* Corresponding author. Tel.: 3 990 30594; fax: 3 990 30556.

E-mail address: Rory.Wolfe@med.monash.edu.au (R. Wolfe).

and nonlinear relationships [12] and are long established [13,14]. Each method was applied by a different investigator, blinded to the validation dataset when creating their models.

2. Case study

Data from the adult trauma database of the Royal Melbourne Hospital (RMH; a major metropolitan trauma service of Melbourne) for the period of January 1, 1996 to April 30, 2001, was used for developing models. A total of 4,014 blunt trauma cases were in the dataset; penetrating trauma cases were excluded because of their differing clinical presentation, management, and likely outcomes. To evaluate the performance of newly developed prognostic models, the Victorian State Trauma Registry (VSTR) was accessed for all blunt trauma cases in its first year of data collection (July 1, 2001 to June 30, 2002).

Table 1 describes the RMH and VSTR datasets. The demographics were typical of blunt trauma, with mean age around 40 years and approximately two-thirds of patients being male. Almost one in five patients experienced complications related to their injuries. Approximately half the patients were covered by Transport Accident Commission (TAC; third-party insurer) funding, which is a consequence of the similar proportion injured on a road, street, or highway.

Although the two datasets summarized in Table 1 are broadly similar, VSTR patients tended to have greater severity of injury than RMH patients as indicated by: a higher proportion of patients classified as resuscitation or emergency according to triage categories; a higher proportion of patients with an injury severity score (ISS) greater than 15 indicating a severe injury; a higher proportion of deaths and Intensive Care Unit (ICU) admissions.

The outcomes for which prognostic models were developed were: ICU stay (No = 0, Yes = 1) and survival status (Survival = 0, Death = 1). Table 2 lists the variables considered for inclusion in the two models. The model for ICU stay was based only on information collected at the scene of the injury. The model for death also took into account hospital-collected data including the variables used in existing prediction models for mortality.

There was a problem with missing data in the RMH and VSTR datasets (see Table 1), particularly for information collected at the scene of the injury. For model creation and validation we included only those patients with complete data on each of the variables listed in Table 2, separately for each outcome.

3. Methods

3.1. Logistic regression

Logistic regression is common in the analysis of medical data [15]. A statistical model is specified for the probability

Table 1

Description of the model development dataset (RMH) and validation dataset (VSTR)

Variable	Subgroup	RMH		VSTR	
		<i>N</i> ^a	%	<i>N</i> ^a	%
Sex	Male	4,014	63.7	3205	69.1
Triage category	Resuscitation	3,396	21.8	2927	32.7
	Emergency		21.1		32.3
	Urgent		35.0		25.2
	Semi- or non-urgent		22.1		9.8
Cause of injury	Motor vehicle	3,674	23.0	3165	33.7
	Motorcycle		7.2		12.2
	Pedestrian		9.7		8.9
	Low fall		26.9		15.2
	Other		33.2		30.0
Intent of injury	Intentional	3,221	9.7	3115	6.5
Place of injury	Home/residential	2,917	23.4	2830	20.3
	institution				
	Road/street/highway		52.5		61.0
	Work place		4.8		7.9
	Other		19.3		10.8
Funding source	TAC funded	3,029	44.3	2994	53.8
Complications	Yes	4,014	19.8	3205	16.5
Comorbidities	Yes	4,014	28.0	2385	30.3
ISS ^b	> 15	4,014	24.9	2442	41.9
Survival status	Dead	4,014	3.3	3205	6.3
ICU ^c admission	Yes	3,989	14.1	3205	24.3
Variable		<i>N</i> ^a	Mean (SD)	<i>N</i> ^a	Mean (SD)
Age, yr		4014	45 (22)	3204	39 (23)

^a*N* = number of patients with nonmissing information on specified variable, % = proportion of patients in specified subgroup.

^b Injury severity score.

^c Intensive care unit.

of an outcome event, *Y*, based on a function of predictor variables, *X*, including covariates such as age: probability of $Y = \exp(\mathbf{X}\boldsymbol{\beta}) / [1 + \exp(\mathbf{X}\boldsymbol{\beta})]$. These models give a predicted probability of *Y* for any future individual based on their covariate values and the parameter vector, $\boldsymbol{\beta}$. By assessing whether an individual's predicted probability is greater than or less than some arbitrary reference probability, a binary prediction can be generated. We obtained maximum-likelihood estimates of $\boldsymbol{\beta}$ [16].

The first step of analysis was to combine categories where there were no events in a predictor variable category. The next step was to perform exploratory analyses of the relationship between each continuous predictor variable and the log-odds of each outcome using nonparametric smoothed plots [17] and fractional polynomial models [18]. These analyses led to consideration of quadratic relationships between log-odds of outcome and SBP, respiratory rate and pulse rate, with centering of covariates around their mean, for example, $\mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1[\text{SBP} - \text{mean}(\text{SBP})] + \beta_2[\text{SBP} - \text{mean}(\text{SBP})]^2$. We considered a linear relationship only for age. The parameters for linear and quadratic terms were assessed for inclusion in a final model together with terms for categorical covariates using a backward elimination stepwise procedure [19] based on likelihood ratio (LR) statistics (terms removed if the *P*-value is greater than or equal to .15; terms reentered

Table 2

Variables considered for inclusion in prognostic models for intensive care unit stay (prehospital variables) and death following trauma (in-hospital variables)

Variables	Prehospital	In-hospital	Scale of measurement
Age	Yes	Yes	Years
GCS; verbal response	Yes	Yes	1: Normal (score of 5) 0: Abnormal (score of 1/2/3/4)
GCS; eye opening response	Yes	Yes	1: Normal (score of 4) 0: Abnormal (score of 1/2/3)
GCS; motor response	Yes	Yes	1: Normal (score of 6) 0: Abnormal (score of 1/2/3/4/5)
Cause of injury	Yes	Yes	0: Motor vehicle 1: Motorcycle 2: Pedestrian 3: Other
Systolic blood pressure, SBP	Yes	Yes	mmHg
Respiratory rate	Yes	Yes	Breaths/minute
Pulse rate	Yes	Yes	Beats/minute
Head injury status	Yes	Yes	1: Injury to head region with an Abbreviated Injury Score, AIS > 2 0: Injury to head region with AIS ≤ 2 or no injury to head region
Place of injury	No	Yes	0: Home/residential institution 1: Sport or recreation area 2: Road/street/highway 3: Work place 4: Other
Injury severity score, ISS	No	Yes	1: > 15 0: ≤ 15
Intent of injury	No	Yes	1: Intentional 0: Nonintentional
Funding status	No	Yes	1: Transport Accident Commission (TAC) 0: Other
Comorbidity before trauma	No	Yes	1: ≥ 1 comorbidity 0: No comorbidity present
Complications during treatment	No	Yes	1: ≥ 1 complication 0: No complications
Triage category	No	Yes	0: Resuscitation 1: Emergency 2: Urgent 3: Semi- or nonurgent

Abbreviation: GCS, Glasgow coma scale.

if the *P*-value is less than 0.05). Linear terms were considered for removal from the model only after their corresponding quadratic term was removed.

3.2. Classification tree modeling

Classification tree or recursive partitioning (RP) algorithms developed from research into the analysis of large survey datasets [13,20]. RP attempts to find interactions between predictor variables through identifying subgroups, represented as end points or nodes on a decision tree. If left to “grow” unchecked, these trees become “bushy,” with nodes having few cases, and may not generalize well; several “pruning” techniques are available [21]. We employed an implementation [22] of the CART (Classification and Regression Trees) algorithm [12,23], previously applied to trauma data [24–30]. When employed with categorical outcomes, CART by default minimizes the Gini impurity criterion [23], a measure of variability within the subgroups created at each stage of the tree.

CART employed a backward pruning tree-building strategy, which has a similar aim to backward elimination stepwise regression. CART first grew a tree until it ran out of cases, subject to a minimum of five cases per node. This maximally sized tree was then pruned, trading cost (performance based on crossvalidation) against complexity (the number of nodes). CART tested each tree using 10-fold crossvalidation whereby the sample was divided into 10 subsamples; the model developed using nine of the subsamples and tested on the tenth. This process was repeated with a different subsample “held out” and used for testing each time.

In contrast with other tree-building procedures [31] CART performs only binary splits, although variables could have been split more than once (e.g., age split into ≤ 40 years, > 40 years, and then further divided into ≤ 23.5 years, > 23.5 years, and so on). CART examined all the different possible cut points (e.g., ≤ 10 years, ≤ 11 years). Following usual practice, the benefits of positive predictive and negative predictive values were equally weighted. For prehospital prediction of ICU stay, however, a second model was

explored on the basis of a high sensitivity and high PPV compromise, achieved by setting the cost of a false negative to be 75% greater than the cost of a false positive. The final trees were expressed as if-then rules, with each combination of model rules giving a predicted outcome for future individuals.

3.3. Artificial Neural Network modeling

Artificial neural nets or networks developed from psychologic [14] and engineering [32] research and attempt to mimic the action of biologic neurons in software. One of the most popular type of networks is known as a multi-layered perceptron, trained by an iterative process termed back propagation of errors [33,34]. Neural networks are made up of individual units or “neurons” arranged in interconnected layers, generally comprising an input layer, one or more hidden layers, and an output layer. Each neuron in a layer receives inputs from all the neurons in the previous layer, calculates its own activation level, and passes this on as an input to the next layer [35]. Back propagation neural networks have been applied to trauma data in a variety of studies [8–10,36–40].

Neural networks were constructed using the NeuroShell 2 [41] program. Several different network architectures and training strategies were tried in the exploratory phase. The selected models comprised feed-forward neural networks trained by back propagation using momentum and learning rate terms of 0.1 each, employing the mean squared error of the training data as the training criterion. The networks used “jump connection” (i.e., layer connected to every other layer), of three hidden layers, with the logistic function as the activation function for each [35]. An early stopping technique was used to avoid overtraining. Ten percent of the available RMH training data were randomly selected and reserved before training began using the remainder. After each 200 learning events (i.e., after 200 patterns had been presented for training) the current state of the network was applied to the reserved data and the mean squared error calculated. This was continued until 100,000 learning events had occurred without further improvement in the error on the reserved test dataset. The final “trained” network selected was the one that gave the lowest error on the latter.

After training in this way a second training run was performed without a test set, but only training until the mean squared error on the training data had reached the same level as it had for the earlier best network. For models predicting ICU stay this gave a slightly better overall result on the VSTR data, but this was not the case for the models predicting death.

3.4. Experiment design

In our study, in an attempt to avoid investigator bias, different investigators developed models using the three different statistical methods. Investigators were experienced

in the application of their preferred method, having published application papers in peer-reviewed health research journals. All investigators remained blinded to the validation dataset until after final models had been developed using the training dataset. The VSTR registry is confidential until an official release of the data and no release had occurred when the logistic regression models were presented in an interim report to the Victorian Trauma Foundation. Hence, the two investigators (R.W., P.S.) fitting logistic regression models and involved in coordination of the project were blinded to the validation data set. The investigator (D.M.) who fitted classification trees submitted final models at the same time as the above models. The investigator (J.B.) who fitted neural networks did so at a later date, but was based at a separate institution and was only provided with the validation dataset after final models were lodged with investigator R.W.

3.5. Model evaluation

The clinical usefulness of each model was evaluated in the training (RMH) and validation (VSTR) datasets using positive predictive value (PPV) and negative predictive value (NPV) [15]. Calibration of the developed models was measured with the following statistics [15]: sensitivity, specificity, classification accuracy (the proportion of patients for whom the prediction matched their outcome), and Hosmer-Lemeshow (H-L) statistic [19] for which 10 groups of patients were defined according to deciles of predicted risk probability. Discriminative ability of the models was evaluated using area under the receiver-operating characteristic curve (AUC). The AUC can be interpreted as the probability that a pair of individuals, one experiencing the outcome and one not, will be ranked correctly on the predicted probability of experiencing the outcome by the final model [42]. A value of .5 for AUC represents no discriminatory ability; a value of one indicates perfect discrimination.

3.6. Performance indicators

Performance indicators, representing lower limits of “acceptable” performance for newly developed prediction models, were specified in advance of the development of our prediction models [2]. These minimum criteria formed the primary test of performance of the newly developed models.

The aim of prehospital prediction models is to identify patients likely to require treatment at a major trauma service without overtriaging, that is, overloading the system with patients who, ultimately, do not require such specialized treatment. For prediction of an ICU stay using information available prehospital, the performance indicators were sensitivity > 85% and PPV > 50%. It is generally accepted by emergency medicine clinicians that correct identification of 85% of all patients with major trauma as

requiring attention at a Major Trauma Service with an overtriage rate of 50% (overtriage rate equals one minus PPV) is consistent with good prehospital triage [43].

A prediction model for death using hospital-collected data is useful as an audit tool to evaluate patient care and trauma center performance in the context of what is “expected.” The primary criterion for audit tools is the ability to correctly predict all patients that experience an adverse outcome, because unexpected adverse outcomes require further investigation. Second, the tool should be well-calibrated such that the predicted prevalence of adverse outcomes for a patient group is close to the observed prevalence, that is, that overall performance is as expected. Hence, for our prediction models of death we considered the performance indicators sensitivity > 95% and specificity > 20%. In consultation with an emergency medicine expert it was decided that 95% sensitivity represented the minimal acceptable “miss” rate for this serious adverse event.

For choosing among competing cut points on a predicted probability scale the most simplistic of utility scores were used. For prehospital prediction of an ICU stay the chosen cut point was that which satisfied sensitivity > 85%, had maximum PPV, and hence, minimum overtriage. For auditing death, the chosen model was that which satisfied sensitivity > 95% and had maximum specificity.

4. Results

4.1. Prehospital prediction of ICU stay

Of the 4,014 blunt trauma cases, 1,324 (33%) had complete data for the creation of prognostic models for an ICU stay. The prevalence of ICU stay was 301/1,324 (23%). The final logistic regression model was

$$\begin{aligned} X\hat{\beta} = & -2.643 \\ & + 0.972 \cdot I(\text{GCS eye opening response} = \text{abnormal}) \\ & + 1.998 \cdot I(\text{GCS motor response} = \text{abnormal}) \\ & + 0.320 \cdot I(\text{Cause of injury} = \text{pedestrian}) \\ & + 0.941 \cdot I(\text{Cause of injury} = \text{motorcycle}) \\ & + 0.448 \cdot I(\text{Cause of injury} = \text{vehicle}) \\ & - 0.00826 \cdot (\text{SBP}-126) \\ & + 0.000198 \cdot (\text{SBP}-126)^2 \\ & + 0.0488 \cdot (\text{Respiratory rate}-19) \\ & + 0.00301 \cdot (\text{Pulse rate}-89) \\ & + 0.000241 \cdot (\text{Pulse rate}-89)^2 \end{aligned}$$

where $I(\cdot)$ represent indicator functions with values 0/1 if the statement in parentheses is False/True. The choice of cutoff for prediction of a positive outcome using predicted probabilities from this logistic regression model was based on Fig. 1, and it can be seen that for the RMH data, the two performance criteria could not be satisfied simultaneously. As no utility scores had been defined to choose among models failing to satisfy the performance criteria, a cutoff

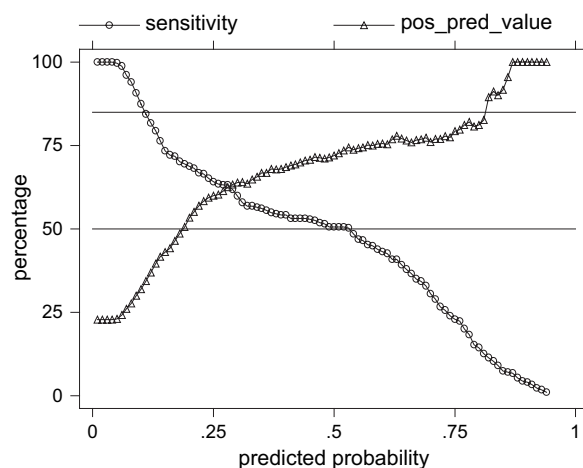


Fig. 1. Plot of sensitivity and positive predictive value (pos.pred.value) vs. predicted probability for model predicting an intensive care unit stay from prehospital data. Performance criteria were positive predictive value > 50% and sensitivity > 85% and are indicated by horizontal lines.

of > 0.10 was made to meet sensitivity criteria and maximize PPV.

The chosen classification tree model (Table 3) arose when opting for a high PPV, high sensitivity compromise within the tree growing process. For the RMH data this model had sensitivity = 80%, PPV = 47%. These figures were considered a better balance than sensitivity = 70%, PPV = 59% achieved by a 12-node model fitted when false positive and false negative predictions were considered to be of equal importance. The node with greatest discrimination and prediction of ICU stay was node 12, representing patients with abnormal GCS motor score.

The complex nature of neural networks precludes them from being represented concisely. A “jump connection” network was selected on the fourth attempt comprising three hidden layers, all layers using logistic activation function, 11, 13, 13, 13, and 1 neurons for the input, three hidden layers, and output layer respectively, momentum 0.1, learning rate 0.1, trained until the mean squared error < 0.065. (This was decided by the training characteristics of the first attempt, which involved a similar network trained using early stopping, with a random 10% test set, calibration interval of 200, and trained until 300,000 events without improvement.) From Table 4 we see that the criteria were not met in the RMH data and a cutoff of > 0.125 was chosen to meet sensitivity criteria while maximizing PPV.

Table 5 summarizes the performance of the models obtained using the three different statistical methods. One thousand fifty-five patients from the VSTR dataset had complete data and were used in validation, and the prevalence of ICU stay among these patients was 27%, a slightly higher prevalence than in the model-development dataset. No model satisfied the criteria of sensitivity > 85% and PPV > 50%. Performance of the three models was similar;

Table 3

A 12-node classification tree model for intensive care unit (ICU) stay chosen for high sensitivity, high positive predictive value compromise; predicted outcome for individuals allocated to each node and split of patients from the model development dataset at each node.

Node	Criteria for allocation of patient to node	Outcome for model development patients		
		ICU Stay	No ICU	Predicted outcome
1	Cause of injury = Other; GCS ^a motor = Normal; SBP ^b ≤ 106.5	5	59	No ICU
2	Cause of injury = Vehicle, motorcycle or pedestrian; GCS motor = Normal; SBP ≤ 106.5; Pulse rate ≤ 60.5	0	11	No ICU
3	Cause of injury = Vehicle, motorcycle or pedestrian; GCS motor = Normal; SBP ≤ 106.5; Pulse rate > 60.5	47	102	ICU stay
4	GCS motor = Normal; SBP > 106.5; GCS eye = Normal; Respiratory rate ≤ 23; Pulse rate ≤ 121	52	598	No ICU
5	SBP > 106.5; GCS eye = Normal; Respiratory rate ≤ 23; Pulse rate > 121	3	5	ICU stay
6	Cause of injury = Pedestrian; GCS motor = Normal; SBP > 106.5; GCS eye = Normal; Respiratory rate > 23	0	21	No ICU
7	Cause of injury = Vehicle, motorcycle or other; GCS motor = Normal; SBP > 106.5; GCS eye = Normal; Respiratory rate > 23; GCS verbal = Normal; Age ≤ 24.46	1	18	No ICU
8	Cause of injury = Vehicle, motorcycle or other; GCS motor = Normal; SBP > 106.5; 19 GCS eye = Normal; Respiratory rate > 23; GCS verbal = Normal; Age > 24.46	19	58	ICU stay
9	Cause of injury = Vehicle, motorcycle or other; GCS motor = Normal; SBP > 106.5; 1 GCS eye = Normal; Respiratory rate > 23; GCS verbal = Abnormal	1	21	No ICU
10	GCS motor = Normal; GCS eye = Abnormal; SBP > 106.5 and ≤ 135	17	33	ICU stay
11	GCS motor = Normal; GCS eye = Abnormal; SBP > 135	2	29	No ICU
12	GCS motor = Abnormal	154	68	ICU stay

^a Glasgow coma scale (GCS).

^b Systolic blood pressure (SBP).

gains in sensitivity were offset by reductions in PPV. The three methods had almost identical values for the average of sensitivity and specificity, that is, calculating AUC for the receiver operating characteristic (ROC) curve defined by the single pair of sensitivity/specificity values in Table 5. Using logistic regression as an example, calibration was poor (Hosmer-Lemeshow statistic = 32.2). Discrimination was reasonable, but not sufficient for performance criteria to be met (AUC = 0.78 for logistic regression, AUC = 0.78 for neural network).

4.2. In-hospital prediction of death

Of the 4,014 blunt trauma cases in the RMH dataset, 2,059 (51.3%) cases had complete data for the creation of prognostic models for death. The prevalence of death was 60/2,059 (2.9%). The final logistic regression model was

$$\begin{aligned}
 \widehat{X\beta} = & -10.388 \\
 & -1.676 * I(\text{Funding status} = \text{TAC}) \\
 & +1.172 * I(\text{Place of injury} = \text{Road/street/highway}) \\
 & -1.152 * I(\text{Place of injury} = \text{workplace}) \\
 & -0.912 * I(\text{Place of injury} = \text{other place of injury including sport/recreation area}) \\
 & +1.676 * I(\text{Presence of complications} = \text{Yes}) \\
 & -0.970 * I(\text{Triage category} = \text{emergency}) \\
 & -0.664 * I(\text{Triage category} = \text{urgent}) \\
 & -1.475 * I(\text{Triage category} = \text{semi- or nonurgent}) \\
 & +0.947 * I(\text{GCS motor response} = \text{abnormal}) \\
 & +0.989 * I(\text{GCS eye opening response} = \text{abnormal}) \\
 & +1.447 * I(\text{ISS} > 15) \\
 & +0.0864 * \text{age} \\
 & -0.0205 * (\text{SBP} - 136) \\
 & +0.00037 * (\text{SBP} - 136)^2 \\
 & +0.0232 * (\text{Pulse rate} - 84.5)
 \end{aligned}$$

The ROC plot shown in Fig. 2 was used as the basis for choosing a cutoff of > 0.02 on the predicted probability scale yielding 95% sensitivity and 85% specificity for the RMH data.

Table 6 outlines the seven terminal nodes of the classification tree that arose when false positive and false negative predictions were considered to be of equal importance. This model had 93% sensitivity and 85% specificity for the RMH data.

A final neural network was selected that was a “jump connection” network; three hidden layers, all layers using logistic activation function, consisting of 23, 18, 18, 18, and 1 neurons for input, three hidden layers, and output layer respectively, momentum 0.1, learning rate 0.1, trained

Table 4

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) at different cutoffs on the predicted probability scale for an artificial neural network model for intensive care unit stay using model development data (area under the receiver operating characteristic curve = 0.83)

Cutoff	Sensitivity %	Specificity %	PPV %	NPV %
0.10	88	46	33	93
0.125	85	57	37	93
0.15	81	68	43	93
0.20	75	79	52	92
0.25	70	86	59	91

Table 5

A summary of the fit of different statistical models to a validation dataset on at-scene predictors of an intensive care unit stay: sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)

Fitting criteria	Sensitivity %	Specificity %	PPV %	NPV %
Classification tree (High PPV, high sensitivity)	70	68	44	86
Logistic regression: (predicted probability > 0.10)	90	46	37	92
Neural network (predicted probability > 0.125)	84	52	39	90

using mean squared error with a random 10% test set, calibration interval of 200, and trained until 100,000 events without improvement. From Fig. 2 we see that the performance criteria were met in the RMH data and a cutoff of >0.05 was chosen which yielded 95% sensitivity and 50% specificity in these data.

Table 7 summarizes the performance in the VSTR dataset of the three models developed using the different statistical methods. One thousand, three hundred eighty-seven patients from the VSTR dataset had complete data and were used in validation. The prevalence of death among these patients was 4.4%, slightly higher than in the model-development dataset. Using logistic regression as an example, calibration was poor (Hosmer-Lemeshow statistic = 53.5). Logistic regression performed slightly better than the other two methods, although for all three models the balance of sensitivity against specificity was inappropriate in light of the performance criteria.

5. Discussion

We pursued more in-depth analyses using logistic regression [44,45] partly for convenience but also because this methodology has been accepted into mainstream use within trauma research, and our results show that its

performance in our database was not inferior to the other two methods.

For ICU stay, differing prevalence in the RMH and VSTR datasets led to calibration problems for our developed models and was one reason for the models not satisfying the performance criteria. The cutoff on the predicted probability scale is a critical part of models for an ICU stay if they are to form the basis of a triage system at the scene of a trauma, that is, if an immediate binary decision is required. This contrasts with our modeling of mortality. The differing prevalence of mortality in the two datasets also led to calibration problems, as these models would be used for purposes such as benchmarking hospital performance and auditing patient outcomes. The models could, however, be recalibrated by changing the predicted probability cutoff on the basis of mortality prevalence in the VSTR data, or in future data to which we wished to apply the models.

In this report we made no attempt to deal with the problem of missing data on the assumption that it was unlikely to affect the comparisons of performance among models developed using different statistical methods. Our crude approach of including in model creation and validation only those patients with complete data on the variables required for a specific model is clearly unsatisfactory. Future research is needed to explore the performance of these methods when dealing with missing data under a range of assumptions.

5.1. Comparing methods for developing prediction models

For predicting ICU stay, which had moderate prevalence of 23% in the development dataset, the three methods gave almost identical results. For prediction of death, which was uncommon (prevalence 2.9% in development dataset), logistic regression had slightly better results than the other two methods, although it had calibration problems and it also failed to meet the performance criteria. This

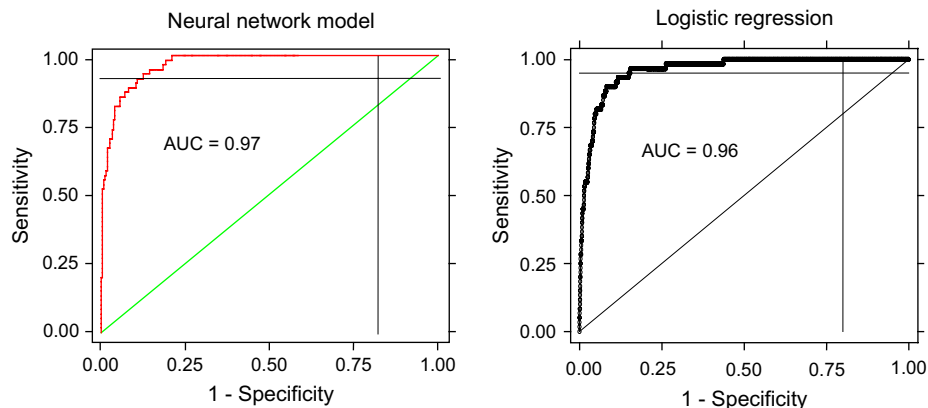


Fig. 2. ROC curve for prediction of death using neural networks and logistic regression with model development dataset. The performance criterion of sensitivity = 95% is marked with a horizontal line and the second criterion of specificity $>20\%$ is marked with a vertical line. The diagonal line represents worst possible discrimination ability and discrimination is summarized as area under the curve (AUC).

Table 6

A seven-node classification tree model for death from in-hospital information chosen for equal positive predictive value and negative predictive value costs; predicted outcome for individuals allocated to nodes, and split of patients from the model development dataset at each node

Node	Criteria for allocation of patient to node	Outcome for model development patients		Predicted outcome
		Death	Survive	
1	Complications = None; GCS ^a motor = Normal	4	1419	Survive
2	Complications = None; GCS Motor = Abnormal	8	55	Death
3	Complications = At least one; Age ≤27.125	0	135	Survive
4	Complications = At least one; Age > 27.125 and ≤60.83; GCS eye = Normal; Pulse rate ≤99	0	144	Survive
5	Complications = At least one; Age > 27.125 and ≤60.83; GCS eye = Normal; Pulse rate > 99	3	45	Death
6	Complications = At least one; Age > 27.125 and ≤60.83; GCS eye = Abnormal	9	48	Death
7	Complications = At least one; Age > 60.83	36	153	Death

^a GCS = Glasgow Coma Scale.

comparison of the three statistical methods suggests no one method is preferable in terms of the performance of prognostic models that are created. This is in broad agreement with other comparisons of the three methods, although no consensus exists. In particular, one study comparing these three methods with slightly different estimation methods from ours for classification trees and neural networks, with simulated data, and in the absence of performance criteria, concluded that logistic regression with piecewise-linear and quadratic functions of predictor variables did best, although having the largest drop between representative and non-representative data [4].

CART [22] is able to output trees in the form of computer (C language) subroutines. Logistic regression has already been implemented for prediction of trauma outcomes through its use in the TRISS method for predicting death [46]. The models developed by neural networks can easily be put into practice, for example, NeuroShell 2 [41] is able to save networks as computer (BASIC or C) subroutines. Hence, there is no reason to prefer a method for reasons of implementation.

The crossvalidation aspect of CART and NeuroShell 2 provides protection against overfitting of the development dataset; however, the resulting models still need to be tested on “fresh” data [3,4]. Crossvalidation can be implemented

with maximum-likelihood estimation for logistic regression and, alongside the bootstrap and jackknife, is one of a number of approaches to overfitting that have been proposed. We checked our final logistic regression models using the bootstrap and confirmed that overfitting was not a problem [47].

Each method has a large number of parameters that can be altered to arrive at different “final” models; hence, comparison of methods is problematic. One approach is to strictly define sets of parameters as a “submethod” and to compare many different parameter sets [48]. Another approach, more in keeping with our comparison, is to compare broad methods and allow for the expertise of the individual in choosing from the myriad submethods. In this approach, to avoid bias, comparisons of different methods need to be done under experimental conditions with experts blinded to the validation data and to each others results during model development. Interexpert agreement in performance of final models using the “same” method should also be examined. Despite these restrictions, comparison of broad methods is preferable to comparison of tightly defined submethods because the results are more applicable to the day-to-day activities of applied statisticians.

5.2. Performance criteria

The use of performance criteria against which a newly developed prognostic model is assessed is not common in prognostic model development despite such criteria being an important part of a rational approach to possible uptake of the new models [2]. Two reasons for this seem likely: (1) concern about subjectivity in choosing indicators to represent minimum performance thresholds and, (2) a lack of consideration of the application context when developing new prognostic models. We have concern about the subjectivity of performance criteria for our mortality model and more generally for audit tools used to assess hospital care performance. One requirement of such models is correct prediction of patients with adverse outcomes, that is, maximum sensitivity. This criterion is of no use on its own

Table 7

A summary of the fit of different statistical models to a validation dataset on in-hospital predictors of death: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and area under the receiver operating characteristic curve (AUC)

Fitting criteria	Sensitivity %	Specificity %	PPV %	NPV %	AUC
Classification trees (PPV = NPV)	61	85	15	98	
Logistic regression (predicted probability > 0.02)	77	83	18	99	0.91
Neural network (predicted probability > 0.05)	70	80	14	98	0.83

because it can be achieved by a tool that predicts an adverse outcome for every patient! To satisfy the general goal of “good” overall calibration a second criterion is required, however, this is difficult to quantify. Our choices of sensitivity $>95\%$ and specificity $>20\%$ for predicting mortality were in keeping with the general spirit of these requirements for audit tools but the actual values chosen were somewhat arbitrary. Ideally, one could take into account the likely prevalence of the adverse outcome and scores for the utility of identification of unexpected adverse outcomes. Problems with subjectivity do not lessen the importance of specifying *a priori* performance criteria.

An interesting potential use of performance criteria is their incorporation into the statistical method used for model creation. We have only used the performance criteria in the estimation process for the classification tree method. Even then the exact values specified by the criteria were not used, rather the method was altered to reflect the general structure of the criteria. By definition, maximum likelihood estimation for logistic regression does not allow for the optimization of any criteria other than the likelihood function. Similarly, ready available routines for classification trees and neural networks do not accommodate the optimization of performance indicators such as sensitivity. It would be of practical interest then to explore alternative, novel estimation methods for logistic regression and other techniques.

6. Conclusion

We did not find an optimal statistical method for the development of prognostic models for binary outcomes. All of our developed models failed to meet prespecified performance criteria. One possible reason for this is because we did not consider important predictor variables. Our databases provided access to most previously reported predictors of mortality and ICU stay. Variables not considered, but potentially able to be collected in the prehospital setting, for example, the presence of coexisting conditions and specific anatomic injuries, might lead to an improved model for an ICU stay. However, the measurement accuracy for these variables has not been tested, for example, the accurate identification of anatomic injury in the field by paramedics is difficult due to the lack of imaging and diagnostic equipment [44]. For predicting death from information available from the patient's hospital stay the presence of comorbidity was not found to be an independent predictor, consistent with findings for the Charlson Comorbidity Index [45]. However, there have been reports that specific comorbidities, rather than a global comorbidity score, may be predictive of mortality. Another explanation for our models failing to meet our prespecified criteria is that the criteria are too onerous, that is, that there is a natural limit to the predictability of an ICU stay following blunt trauma and subsequent mortality that is below our specified criteria.

Acknowledgment

The work presented in this article was performed with the assistance of funding from the Victorian Trauma Foundation.

References

- [1] Harrell F, Lee K, Mark D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–7.
- [2] Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453–73.
- [3] Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol* 2003;56:826–32.
- [4] Terrin N, Schmid CH, Griffith JL, D'Agostino RB, Selker HP. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *J Clin Epidemiol* 2003;56:721–9.
- [5] Ripley B, Ripley R. Neural networks as statistical methods in survival analysis. In: Dybowski R, Gant V, editors. *Artificial neural networks: prospects for medicine*. Austin, TX: Landes Biosciences Publishers; 1998.
- [6] Kattan MW, Hess KR, Beck JR. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. *Comput Biomed Res* 1998;31:363–73.
- [7] Allore H, Tinetti ME, Araujo KLB, Hardy S, Peduzzi P. A case study found that a regression tree outperformed multiple linear regression in predicting the relationship between impairments and social and preventive activities scores. *J Clin Epidemiol* 2005;58:154–66.
- [8] Eftekhari B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC Med Inform Decision Making* 2005;5:3.
- [9] Becalick DC, Coats TJ. Comparison of artificial intelligence techniques with UKTRISS for estimating probability of survival after trauma: UK Trauma and Injury Severity Score. *J Trauma* 2001;51:123–33.
- [10] DiRusso SM, Chahine AA, Sullivan T, et al. Development of a model for prediction of survival in pediatric trauma patients: comparison of artificial neural networks and logistic regression. *J Pediatr Surg* 2002;37:1098–104.
- [11] Costanza MC, Paccaud F. Binary classification of dyslipidemia from the waist-to-hip ratio and body mass index: a comparison of linear, logistic, and CART models. *BMC Med Res Methodol* 2004;4:7–16.
- [12] Hastie T, Tibshirani RJ, Friedman J. *Elements of statistical learning: data mining, inference and prediction*. New York: Springer; 2001.
- [13] Belson WA. Matching and prediction on the principle of biological classification. *Appl Stat* 1959;8:65–75.
- [14] Rosenblatt F. The perceptron: a probabilistic model of information storage and organization in the brain. *Psychol Rev* 1958;65:386–408.
- [15] Altman DG. *Practical statistics for medical research*. Boca Raton, FL: CRC; 2000.
- [16] StataCorp. *Stata statistical software version 7*. College Station, TX: Stata Corporation; 2001.
- [17] Cleveland WS. *Visualizing data*. Summit, NJ: Hobart; 1993.
- [18] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Appl Stat* 1994;43:429–67.
- [19] Hosmer DW, Lemeshow S. *Applied logistic regression*. New York: Wiley; 2000.
- [20] Morgan JN, Sonquist JA. Problems in the analysis of survey data and a proposal. *J Am Stat Assoc* 1963;58:415–34.
- [21] McKenzie DP, McGorry PD, Wallace CS, Low LH, Copolov DL, Singh BS. Constructing a minimal diagnostic decision tree. *Methods Inform Med* 1993;32:161–6.

- [22] Salford Systems. CART for Windows, version 4. San Diego, CA: Salford Systems; 2001.
- [23] Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and regression trees. Belmont, CA: Wadsworth; 1984.
- [24] Rainer TH, Lam PKW, Wong EMC, Cocks RA. Derivation of a prediction rule for post-traumatic acute lung injury. *Resuscitation* 1999; 42:187–96.
- [25] Langdorf MI, Rudkin SE, Dellota K, Fox JC, Munden S. Decision rule and utility of routine urine toxicology screening of trauma patients. *Eur J Emerg Med* 2002;9:115–21.
- [26] Holmes JF, Sokolove PE, Brant WE, Kuppermann N. A clinical decision rule for identifying children with thoracic injuries after blunt torso trauma. *Ann Emerg Med* 2002;39:492–9.
- [27] Palchak MJ, Holmes JF, Vance CW, et al. A decision rule for identifying children at low risk for brain injuries after blunt head trauma. *Ann Emerg Med* 2003;42:492–506.
- [28] Cotton BA, Beckert BW, Smith MK, Burd RS. The utility of clinical and laboratory data for predicting intraabdominal injury among children. *J Trauma* 2004;56:1068–74.
- [29] Guldner G, Babbitt J, Boulton M, O'Callaghan T, Feleke R, Hargrove J. Deferral of the rectal examination in blunt trauma patients: a clinical decision rule. *Acad Emerg Med* 2004;11:635–41.
- [30] Rovlias A, Kotsou S. Classification and regression tree for prediction of outcome after severe head injury using simple clinical and laboratory variables. *J Neurotrauma* 2004;21:886–93.
- [31] Haydel MJ, Preston CA, Mills TJ, Luber S, Blaudeau E, DeBlieux PM. Indication for computed tomography in patients with minor head injury. *N Engl J Med* 2000;343:100–5.
- [32] Widrow B, Hoff ME. Adaptive switching circuits, vol. 4. New York: Institute of Radio Engineers, Western Electronics Convention (WESCON); 1960. p 96–104.
- [33] Rumelhart DE, McClelland JL. Parallel distributed processing: explorations in the microstructure of cognition. Cambridge, MA: MIT Press; 1986.
- [34] Werbos P. The roots of back propagation: from ordered derivatives to neural networks. New York: Wiley; 1994.
- [35] Haykin S. Neural networks: a comprehensive foundation. Upper Saddle River, NJ: Prentice-Hall; 1999.
- [36] Marble RP, Healy JC. A neural network approach to the diagnosis of morbidity outcomes in trauma care. *Artif Intell Med* 1999;15:299–307.
- [37] Hunter A, Kennedy L, Henry J, Ferguson I. Application of neural networks and sensitivity analysis to improved prediction of trauma survival. *Comput Methods Programs Biomed* 2000;62:11–9.
- [38] Sinha M, Kennedy CS, Ramundo ML. Artificial neural network predicts CT scan abnormalities in pediatric patients with closed head injury. *J Trauma* 2001;50:308–12.
- [39] Estahbanati HK, Bouduhi N. Role of artificial neural networks in prediction of survival of burn patients—a new approach. *Burns* 2002; 28:579–86.
- [40] Lammers RL, Hudson DL, Seaman ME. Prediction of traumatic wound infection with a neural network-derived decision model. *Am J Emerg Med* 2003;21:1–7.
- [41] Ward Systems Group. NeuroShell 2, Release 4. Frederick, MD: Ward Systems Group; 2000.
- [42] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143: 29–36.
- [43] Senkowski C, McKenney M. Trauma scoring systems: a review. *J Am Coll Surg* 1999;189:491–503.
- [44] Gabbe BJ, Cameron PA, Wolfe R, Simpson P, Smith KL, McNeil JJ. Pre-hospital prediction of intensive care unit stay and mortality in blunt trauma patients. *J Trauma*, in press.
- [45] Gabbe BJ, Cameron PA, Wolfe R, Simpson P, Smith KL, McNeil JJ. Predictors of mortality, length of stay and discharge destination in blunt trauma. *Aust N Z J Surg* 2005;75:650–6.
- [46] Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. *J Trauma* 1987;27:370–8.
- [47] Austin PC, Tu JV. Bootstrap methods for developing predictive models. *Am Stat* 2004;58:131–7.
- [48] Lim T-S, Loh W-Y, Shih Y-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learn* 2000;40:203–29.

7. DISCUSSION

The final chapter summarises and integrates the results of the five applications included in Chapters Two to Six of this thesis, examines the contribution made by CART in these analyses, and then looks at specific ways in which this procedure could be improved with regard to screening, diagnosis and subgroup analysis.

7.1. Subgroups of Australian Gulf War veterans at high risk of hazardous or harmful alcohol consumption

In a study of hazardous or harmful alcohol consumption in male Royal Australian Navy Gulf War veterans the following risk factors were found by logistic regression to be statistically significant - current smoking; past smoking; enlisted military rank; single or never married; separated, divorced or widowed, current (12 month) DSM-IV major depression and current (12 month) DSM-IV PTSD. Several potential confounders, including age, current (12 month) DSM-IV other anxiety disorder (not including PTSD), level of schooling, and number of active deployments, were controlled for.

Although all of the above variables were made available to CART, the procedure generated a decision tree employing all of the above statistically significant risk factors apart from current DSM-IV PTSD and current DSM-IV other anxiety disorder. High risk subgroups consisted of 1/ current smokers; 2/ nonsmokers or former smokers, of non-commissioned officer (NCO) or enlisted rank, who are single, never married, separated or divorced; and 3/ nonsmokers or former smokers, of NCO or enlisted rank, who are married , and have current (12 month) DSM-IV major depression. While logistic regression analysis allowed the overall relationships between risk factors and the outcome to be readily ascertained, the CART analysis showed that there are specific combinations of risk factors that should be further examined. The CART results suggest that intervention and prevention strategies could usefully be targeted at current smokers, as well as married veterans with current major depression.

7.2. Temporal progression of psychiatric disorders in Australian Gulf War veterans

In male Royal Australian Navy Gulf War veterans with no prior psychiatric disorders, discrete time survival analyses performed using logistic regression found that affective, alcohol use and anxiety disorders peaked in the first two years (1990 – 1991) following the beginning of the Gulf War. Alcohol use disorders were the most likely to appear first. CART analyses showed that the risk of developing the above disorders, particularly within the first few years, was higher if veterans had been exposed to a high number of potential psychological stressors during their military service. In addition, lower military rank was associated with increased risk of alcohol use disorders, particularly in the first two years. In this analysis, CART was able to detect statistical interactions between time phase since the Gulf War, and other variables such as rank. These findings suggest that intervention and prevention strategies need to be put in place in the first few years post-War, particularly for those veterans who have encountered greater numbers of potential stressors, or in the case of alcohol use disorders, those of lower military rank.

7.3. Subtypes of DSM-IV major depression in the hospitalised medically ill

In hospitalised medically ill patients, logistic regression found three key symptoms of demoralization (helplessness – hopelessness), consisting of pessimism, worthlessness and thoughts of death, to be highly associated with current (past month) DSM-IV major depression. In addition, one key symptom of anhedonia (loss of interest or pleasure), namely less interest in activities with others, was also highly associated with major depression. Unlike the other three symptoms listed above, pessimism is not one of the DSM-IV criteria for major depression and yet it exhibited the strongest relationship with this diagnosis.

CART found that two combinations of symptoms – 1/ pessimism and worthlessness; and 2/ pessimism, loss of interest in others, and thoughts of death; were highly associated with major depression, suggesting that there may be at least two subtypes of this disorder, as reflected in different combinations of

symptoms. Such combinations of symptoms would have been more difficult to detect if only the logistic regression results were examined. A screening rule based upon both these combinations exhibited reasonable performance, at least for this sample. These results suggest that medical staff and family members should pay particular attention to feelings of pessimism, especially when accompanied by feelings of worthlessness, or by a loss of interest in others and thoughts of death, in medically ill patients.

7.4. Psychological and somatic symptoms of depression in young adults

The recently developed Somatic and Psychological Health Report (SPHERE) ³⁶³ was applied to young adults aged 22 to 34 years who were presumed to have a low prevalence of comorbid physical illness due to their youth, in order to test the performance of this instrument in a community sample. The presence of any current (past month), or any lifetime, DSM-IV psychiatric disorder was examined, as was any current DSM-IV anxiety disorder, and any current DSM-IV depressive disorder. Meeting criteria for probable caseness on either the psychological scale or on the somatic scale of the SPHERE, exhibited the best screening performance. Meeting criteria for caseness on both scales exhibited the highest specificity. CART did not improve upon the performance of the above screening rules at detecting anxiety disorder.

With regard to detecting depressive disorder however, CART constructed a combination of simple screening rules involving caseness on both scales, and caseness on the psychological scale. Such a combination could be expressed as ‘depressive disorder is present if there is caseness on both subscales or, if not, if there is caseness on the psychological subscale alone’. These results again suggest that CART can be useful in finding specific combinations of variables which together exhibit better screening performance than does each variable separately.

The CART screening rule performed better than the other rules, but still showed only low positive predictive value, suggesting that many persons with a positive result on the SPHERE may not actually have major depression. The

positive predictive value of the CART rule was far lower than that observed for the four item CART-generated rule for classifying past month DSM-IV depression described in Chapter Four. The two samples, however, are very different in composition (young community sample versus older hospitalised medically ill).

The above SPHERE results, as well as those reported earlier by Clarke and McKenzie in general practice patients³⁶⁵ indicate that further research into the actual composition of the SPHERE may be required, before determining which combinations of items have the best screening performance. Such an indication is further supported by the recent finding³⁶² that the specificity of the SPHERE is increased if used in conjunction with the Mental Component Summary score of the 12 item version of the Short Form Health Survey (SF-12)³⁸¹, a measure of health-related quality of life.

7.5. Prediction of binary trauma outcomes

Three computational procedures, logistic regression, including quadratic transformations; CART, and an artificial neural network, were employed to create models on metropolitan hospital trauma data and then test them on statewide trauma registry data. Two health outcomes were employed – admission to intensive care unit, and mortality.

The variables chosen by each procedure were listed in the published paper appearing in Chapter Six, but were not the focus of the discussion in that paper, as this was primarily a methodological study concerned with the performance of the different techniques. Logistic regression appeared to exhibit the best performance, followed by CART weighted so as to decrease the possibility of false negatives.

None of the techniques met the pre-specified performance criteria, based upon those commonly employed by emergency medicine physicians, indicating that important predictors of intensive care unit admission and mortality were either unknown, or not able to be measured, and therefore were not available for statistical analysis.

It has been suggested that models be chosen on the basis of factors such as ease of interpretation, as well as performance³¹⁰. The CART model presented

in Chapter Six was arguably easier to interpret than the logistic regression equation, and far more interpretable than the artificial neural networks. Artificial neural networks often provide very good performance, but have structures which are generally difficult to interpret.³⁸²

7.6. Ways in which CART could be improved

7.6.1. Permutation testing of splits and final subgroups

Logistic regression was used in Chapters Two, Three and Four to calculate crude odds ratios, as well as those adjusted for potential confounders such as age, for the comparison of each CART terminal node or end subgroup with a reference group. The latter consisted of that subgroup with the lowest frequency of the outcome variable (e.g., the lowest frequency of major depression).

The use of logistic regression to compare CART subgroups has been recommended by various researchers^{7,237,303}. What has rarely been discussed in the tree-building literature, however, is that the significance levels associated with the above odds ratios may be very misleading. As described in the Introduction (Chapter One), statistical comparisons based upon searching for maximally different subgroups, such as those generated by a recursive partitioning procedure, are often overly liberal.

Although CART can take the extensive searching for models into account when choosing the best sized tree, based upon the cross-validation process described in Chapter One, the procedure does not provide statistical significance levels for each individual split, or for comparisons of the final subgroups with each other.

A possible method of determining the statistical significance of splits or subgroup comparisons would be to use Monte Carlo permutation tests, that were described in Chapter One and which offer an alternative to Bonferroni type multiple comparisons¹⁷³. Permutation tests have recently been applied to the merging of contingency tables¹⁵⁶, as well as the generation of classification trees^{383,384}. In order for such an approach to be incorporated into CART, a large

number, say 10,000, of random permutations or shuffles of the dataset would be generated. For each random permutation, as well as the original, unshuffled, dataset, CART would be applied and the best decision tree chosen, according to the usual CART cross-validation criteria. The odds ratio for the difference between, for example, the reference group (the lowest frequency of the outcome) and whatever end subgroup has the highest frequency of the outcome, is then calculated ¹⁵⁶.

The number of times that the odds ratio obtained for the original dataset is exceeded by those obtained for the random permutations of the dataset, would then be computed. This number, expressed as a proportion of the number of random permutations plus one ¹⁶⁷, provides an estimate of the probability that an odds ratio as large as, or larger than, the original odds ratio would be obtained by chance. As outlined in Chapter One, the permutation testing of decision trees is a burgeoning area, although there is as yet a dearth of studies concerned with the comparison of particular end subgroups within decision trees. In order to obtain adjusted odds ratios, logistic regression could be used along with CART (e.g., regression routines added to the CART computer program), on each permutation, in order to control for possible confounders. The above permutation procedure for obtaining crude and adjusted odds ratios cannot currently be directly implemented by CART but would require extension of the actual software.

A far more expanded, but far more expensive version of CART, CART ProEX (www.salford-systems.com, accessed 30 November 2008), not used in this thesis, offers permutation testing. Such testing allows the statistical significance of the final tree to be ascertained, but does not allow comparisons at each stage of the tree, or specific comparisons between the reference group and the other end subgroups. An extended version of CHAID, commercially available as Optimus RP (www.goldenhelix.com, accessed 30 November 2008) employs permutation testing to ascertain the statistical significance of a particular split, which, unlike CART, may be binary or multi-way. The end subgroups generated by Optimus RP are not compared by permutation tests however, while merging

of categories is still solely determined by statistical significance testing rather than other measures such as effect size.

Rightly or wrongly, researchers often wish to report the statistical significance of their results. The proposed combination of CART, logistic regression and permutation testing would allow this reporting. Such a combination would require the incorporation of logistic regression routines into CART, if odds ratios adjusted for potential confounders were to be computed. Further research needs to be performed into the use of such hybrid techniques to obtain valid statistical significance levels, and ideally, valid confidence intervals.

Permutation testing has also been shown ^{175,385} to reduce the variable selection bias exhibited by CART in selecting variables with large numbers of categories, as described in Chapter One. Such a potential bias was not viewed as a major problem in the analyses presented in this thesis however, as the variables employed had only small numbers of categories. The use of a measure of cost (model performance) and complexity (in this case the number of categories of a particular predictor) such as the Akaike Information Criterion (AIC) ^{258,260,386}, Bayesian Information Criteria (BIC) ³⁸⁷ or a similar statistic ^{257,388}, along with a suitable search procedure ^{137,389,390}, could be incorporated into CART so as to find the best k-way split, or the best binary split, if the latter was desired.

7.6.2. Probabilistic assignment to subgroups, and more control over variables used in the growing of decision trees

Probabilistic assignment to subgroups, in a similar fashion to that employed by latent class cluster and regression analysis ⁷³ and related techniques, could be added to CART. Such assignment is available in various recursive partitioning algorithms ²⁹⁸ including ones developed very recently ²⁹⁹, which are not yet widely known in psychiatric research. Probabilistic assignment would allow the researcher to avoid the use of simple cut-points with dimensional variables such as age and psychological distress, if this was desired. The ultimate goal of such a method would be to maximise the homogeneity of

observed outcomes within subgroups, as well as of the coefficients of the regression equation developed for each subgroup. This would combine the benefits of latent class regression techniques and recursive partitioning. Care must however be taken to ensure that the resulting algorithm does not become too complex, or make unnecessary assumptions of the data, or of the researcher.

The new, limited facility of CART which allows the researcher to specify which variables are to be used in the top levels of a decision tree was employed in the analyses described in Chapter Three. This was done so that categorised number of years elapsed since the Gulf War could be manually specified as the initial splitting variable. It should be noted that the facility for limited variable selection does not override the tree-pruning criteria employed by CART. Even if the initial splitting variables are specified by the researcher, no tree will be generated unless the cross-validated performance is sufficiently high.

The ability to interactively select variables on the basis of theoretical interest is an important one, but is a feature in which CART is still somewhat lacking. The expanded CART ProEX allows the use of 'structured trees', whereby the researcher is able to specify at which stage of the tree-growing process a particular variable or group of variables should be used. This feature can be likened to one offered by the early AID procedure that allowed the researcher to assign priorities to each variable. AID also allowed an option for 'symmetry', whereby the researcher would specify that if a particular variable was chosen at one branch of the tree, that same variable must be chosen at the other branch, provided that the difference in model performance did not exceed a certain value ¹¹². In terms of the CART analysis presented in Chapter Two for example, such a procedure would estimate the effects of splitting the subgroup consisting of veterans who were not married, by the presence of current major depression, as well as splitting the subgroup consisting of married veterans by this variable, as was done in Chapter Two.

Structured trees and symmetry within the CART framework could also be extended to allow the cross-validated performance of particular variable combinations as well as individual variables to be assessed, building upon earlier

²⁰⁸, as well as more recent ²⁷⁴, approaches. Symmetry is not available in CART however, while only the high-end expanded CART ProEx includes the facility for structured trees. Even the latter version of CART does not actually allow the researcher to interactively select variables as the tree is being constructed, although this facility was offered in early procedures such as IDEA ¹³⁸, as well as the more recently developed KnowledgeSEEKER ¹⁶⁵.

7.6.3. Boolean rules within each split

In the study presented in Chapter Six, CART was employed with pre-defined simple Boolean screening rules such as caseness on psychological scale, or caseness on somatic scale. As such combinations are rarely pre-defined it would be very useful if CART could generate its own Boolean combinations, involving logical AND's and OR's, within a particular node of a tree. As described in Chapter One, the original CART monograph ⁸⁰ mentioned such a facility, but it has not yet been implemented.

A simple method of constructing, and assessing the performance of, Boolean rules would be to compare the cross-validated performance of pairs of binary items such as symptoms or caseness thresholds, linked by logical AND or logical OR (whichever gives the best results) with the cross-validated performance of single items ²⁷⁷. This method could also be extended to allow lists of three or more variables, perhaps using an approach such as that described in the recently developed logic regression procedure ^{282,285}, which employs both cross-validation and permutation testing to determine the optimally sized Boolean rule.

Further research into methods of including Boolean rules within CART is required, as is comparison of the performance of those trees that include Boolean rules within the branches, with those that do not.

7.6.4. Specification of performance criteria

CART allows the weighting of errors in order to maximise specificity or sensitivity, this facility being employed in Chapters Five and Six. However, determination of which weights to use is very much a trial and error process.

CART, as with many, but not all ²⁷⁷ statistical and machine learning procedures, is not able to directly maximise sensitivity or specificity, while minimum values for these parameters cannot be specified.

It is thought to be a comparatively simple matter to expand CART by adding features such as permutation testing of splits and end subgroups, splitting based upon cost-complexity criteria, the generation of simple Boolean rules within nodes, and the direct maximising of sensitivity or specificity. The incorporation of probabilistic assignment to subgroups would appear to be more complex, however various such algorithms already exist ^{298,299,302} and so may be able to be adapted. Such modification and expansions would need to be performed by the CART program developers however, which may be reliant on a perceived demand for such features.

7.7. Conclusions

This thesis has shown that the identification of salient combinations of variables by CART can aid in the identification of subgroups at risk of a particular outcome, such as hazardous or harmful drinking; as well as improving screening for depression, and aiding in the diagnosis of this disorder by uncovering possible subtypes. Rather than adopting a 'one size fits all' approach, CART allows treatment and prevention strategies to be targeted for particular subgroups ^{79,391}, such as those military personnel that are highly likely, or highly unlikely, to seek treatment for PTSD ⁷⁹, or adolescents at risk of becoming experimental, or regular, smokers ²³³. It is suggested that CART be further used in the identification of such at risk subgroups.

CART could also be further employed in general studies of resource utilisation, identifying those specific combinations of risk factors associated with higher costs of psychiatric health resources ²⁴¹⁻²⁴³. For example, CART was recently used to develop a model predicting daily cost of inpatient care, based upon combinations of risk factors such as age, psychiatric diagnosis, and deficits in daily activities ²⁴³. This model performed better than one based upon diagnosis-related-groups, which, as mentioned earlier, were originally created

using early recursive partitioning methods ¹²⁰ that were far less sophisticated than CART.

CART also has the potential to develop different screening instruments for different subtypes of a diagnosis such as depression, as illustrated in Chapter Four. Such instruments could augment diagnostic procedures such as the more theoretically driven Mood Assessment Program (MAP) ³⁹², a computer program that is able to classify depression into melancholic depression (characterised by psychomotor disturbance such as severe retardation of physical movement) and non-melancholic depression ³⁹³.

Caution must always be exercised when performing screening and diagnosis however. Clinicians, being only human, can sometimes make biased and inaccurate diagnoses. Mistakes can also be made using simple instruments such as the GHQ and SPHERE, or more sophisticated concept driven computer-based approaches such as the MAP ³⁹², and earlier techniques ³⁹⁴ or computer-based instruments developed using data driven procedures such as CART.

On the one hand, psychiatric illness such as depression may not be recognised when it is actually present. This is often the case in veterans, hospitalised medically ill and general practice patients, as was discussed in Chapters Two through Five. If depression is not detected, an individual may experience unnecessary suffering and anguish. On the other hand, if psychiatric illness is thought to have been detected when it is not actually present, then an individual may also experience unnecessary suffering and anguish. Such false alarms may be particularly troubling in a military setting ³⁹⁵, where the presumption of being mentally ill can be associated with even greater stigma than it is in a general community setting.

An advantage of computer-based instruments, including those developed using CART, is that there are clearly observable grounds for assigning caseness or making a diagnosis – ‘diagnosis x has been judged to be present because symptoms a, b and c are present’, for example. Widespread psychiatric screening, whether it be of military personnel, primary practice patients or other groups can be problematic if an individual’s work, family, social or medical

context is not taken into account. It can be difficult to incorporate knowledge of the individual's broader life setting into screening and diagnostic rules, whether the latter are derived using simple linear decision methods, logistic regression, latent class regression, machine learning techniques, or theoretical constructs. Within an actual clinical psychiatric interview, the presence of particular symptoms that would otherwise 'trigger' particular decision rules may lead the clinician to a different interpretation of symptoms, if such a context was able to be taken into account^{367,396,397}. The application of more 'mechanistic' screening and diagnostic rules should therefore be followed by actual clinical interviews, if at all possible.

Statistical methods such as logistic regression and machine learning methods such as CART should be used together wherever applicable, as has been done in this thesis. Such an approach allows researchers to ascertain global relationships, especially for individual symptoms or risk factors; as well as local relationships, especially for combinations of symptoms or risk factors, with particular outcomes.

Further methodological work involving empirical comparison via simulation studies of CART and other methods of subgroup detection, such as latent class and similar techniques, are areas for future research. This will aid understanding of how the techniques can best be employed in conjunction with each other, for example in allowing probabilistic assignment to subgroups, as well as identifying which specific patterns of risk factors are associated with which subgroups. The increased use of Rasch modelling, including Rasch latent class analysis and Rasch recursive partitioning methods, may increase clinical understanding of the symptoms of psychiatric illness, as well as aid in the detection of subgroups of observations with similar levels of illness.

The addition of permutation testing, probabilistic assignment to subgroups, Boolean decision rules, and the other extensions to CART that have been suggested above may improve the performance, as well as the interpretability, of the technique. As discussed in Chapter One, studies have variously shown that recursive partitioning techniques perform worse^{270,304-307}, about as well^{87,308-310},

or better^{31,47,235,268,311-313} than other techniques such as logistic regression. Further simulation studies are, however, needed to identify those conditions under which CART, and other techniques, perform well, and those under which they perform poorly. In particular, the performance of various techniques when applied to very limited information in the form of predictor variables or risk factors, as may have been the case with the datasets employed in Chapter Six, needs to be assessed, as does the effect of including risk factors and/or outcome variables of low validity.

Any computational method can only ever be as good as the information with which it is presented. In closing, it must be emphasised that even instruments created using the most sophisticated computational methods, and which have been well validated on a large number of datasets, should be used carefully and as an adjunct to good clinical practice. This latter point is particularly relevant if diagnostic results are to be provided to individuals, rather than being employed in groups in prevalence studies.

Notwithstanding all of the above caveats, it is concluded that CART, especially if it is extended to include the suggestions made above, should be employed widely, but prudently, in psychiatric screening, diagnosis and subgroup analysis.

REFERENCES

1. Huffman JC, Smith FA, Blais MA, Januzzi JL, Fricchione GL. Anxiety, independent of depressive symptoms, is associated with in-hospital cardiac complications after acute myocardial infarction. *Journal of Psychosomatic Research* 2008;**65**(6):557-63.
2. Kubzansky LD, Davidson KW, Rozanski A. The clinical impact of negative psychological states: expanding the spectrum of risk for coronary artery disease. *Psychosomatic Medicine* 2005;**67** **Supplement 1**:S10-4.
3. Kraemer HC. Evaluating medical tests : objective and quantitative guidelines. Newbury Park, California: Sage, 1992.
4. Kendell RE. The role of diagnosis in psychiatry. Oxford: Blackwell, 1975.
5. Schmidt NB, Kotov R, Joiner TE, eds. Taxometrics: towards a new diagnostic scheme for psychopathology. Washington, D.C.: American Psychological Association, 2004.
6. Hand DJ. Construction and assessment of classification rules. Chichester, UK: Wiley, 1997.
7. Zhang H, Singer B. Recursive partitioning in the health sciences. New York: Springer-Verlag, 1999.
8. Shannon WD, Faifer M, Province MA, Rao DC. Tree-based models for fitting stratified linear regression models. *Journal of Classification* 2002;**19**:113-130.
9. Belson WA. Matching and prediction on the principle of biological classification. *Applied Statistics* 1959;**8**:65-75.
10. Hunt EB, Marin J, Stone P. Experiments in induction. New York: Academic Press, 1966.
11. Hand DJ. Artificial Intelligence and psychiatry. Cambridge, UK: Cambridge University Press, 1985.
12. Einhorn H. Alchemy in the behavioral sciences. *Public Opinion Quarterly* 1972;**36**:367-378.
13. Babor T, Fuente J, Saunders J, Grant M. The Alcohol Use Disorders Identification Test: Guidelines for use in primary health care. Geneva: Division of Mental Health, World Health Organization, 1989.
14. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Archives of General Psychiatry* 1961;**4**:561-571.
15. Goldberg DP. The detection of psychiatric illness by questionnaire. Maudsley Monograph No. 21 ed. London: Oxford University Press, 1972.
16. Clarke DM, McKenzie DP. A caution on the use of cut-points applied to screening instruments or diagnostic criteria. *Journal of Psychiatric Research* 1994;**28**:185-188.
17. Friedman HP, Rubin J. On a method to aid in the analysis and classification of psychiatric data. In: Kline NS, Laska E, eds. Computers and electronic devices in psychiatry. New York: Grune & Stratton, 1968: 81-99.

18. Kaaya SF, Lee B, Mbwapbo JK, Smith-Fawzi MC, Leshabari MT. Detecting depressive disorder with a 19-item local instrument in Tanzania. *International Journal of Social Psychiatry* 2008;**54**(1):21-33.
19. Melrose JP, Stroebel CF, Glueck BC. Diagnosis of psychopathology using stepwise multiple discriminant analysis. *Comprehensive Psychiatry* 1970;**11**:43-50.
20. Pardo PJ, Georgopoulos AP, Kenny JT, Stuve TA, Findling RL, Schulz SC. Classification of adolescent psychotic disorders using linear discriminant analysis. *Schizophrenia Research* 2006;**87**:297-306.
21. Chouinard G, Annable L, Ross-Chouinard A, Nestoros JN. Factors related to tardive dyskinesia. *American Journal of Psychiatry* 1979;**137**:79-82.
22. Dunn G, Everitt BS. The natural history of depression in general practice: stochastic models. *Psychological Medicine* 1981;**11**:755-764.
23. Liu A, Tan H, Zhou J, et al. A short DSM-IV screening scale to detect posttraumatic stress disorder after a natural disaster in a Chinese population. *Psychiatry Research* 2008;**159**:376-81.
24. Muhwezi WW, Agren H, Musisi S. Detection of major depression in Ugandan primary health care settings using simple questions from a subjective well-being (SWB) subscale. *Social Psychiatry and Psychiatric Epidemiology* 2007;**42**:61-9.
25. Gruenewald TL, Mroczek DK, Ryff CD, Singer BH. Diverse pathways to positive and negative affect in adulthood and later life: an integrative approach using recursive partitioning. *Developmental Psychology* 2008;**44**(2):330-43.
26. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer, 2001.
27. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley, 2000.
28. Rodgers B, Parslow R, Degenhardt L. Affective disorders, anxiety disorder and psychological distress in non-drinkers. *Journal of Affective Disorders* 2007;**99**:165-172.
29. Isaacowitz DM, Seligman ME. Is pessimism a risk factor for depressive mood among community-dwelling older adults? *Behaviour Research and Therapy* 2001;**39**:255-72.
30. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. 2nd ed. Hoboken, New Jersey: Wiley, 2003.
31. Feldman S, Klein DF, Honigfeld G. A comparison of successive screening and discriminant function techniques in successive taxonomy. *Biometrics* 1969;**25**:725-734.
32. Alpaydin E. Introduction to machine learning. Cambridge, Massachusetts: MIT Press, 2004.
33. Samuel AL. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 1959;**3**:210-229.
34. Scharff R. The how and why wonder book of robots and electronic brains. New York: Wonder Books, 1963.

35. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco, California: Morgan Kaufmann, 2005.
36. Raynor WJ. The international dictionary of artificial intelligence. Chicago, Illinois: Glenlake, 1999.
37. Luger GF. Artificial intelligence: structure and strategies for complex problem solving. 5th ed. San Francisco, California: Addison-Wesley, 2005.
38. Bishop CM. Pattern recognition and machine learning. New York: Springer, 2006.
39. Werbos PJ. The roots of backpropagation: from ordered derivatives to neural networks and political forecasting. New York: Wiley, 1994.
40. Negnevitsky M. Artificial intelligence: a guide to intelligent systems. Harlow, England: Pearson, 2002.
41. Rosenblatt F. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological Review* 1958;**65**:386-408.
42. Specht DF. Vectorcardiographic diagnosis utilizing adaptive pattern-recognition techniques. Stanford, California: Systems Theory Laboratory, Stanford Electronics Laboratories, Stanford University, 1964.
43. Widrow B, Hoff ME. Adaptive switching circuits. Institute of Radio Engineers, WESCON (Western Electronics Convention) Convention Record, 1960: 96-104.
44. Belson WA. A series of four lectures on mass media research. Sydney: Market Research Society of Australia, 1961.
45. Belson WA. Matching and prediction on the principle of biological classification. *International Journal of Market Research* 1996;**38**:293-305.
46. Feinstein AR. Multivariable analysis: an introduction. New Haven, Connecticut: Yale University Press, 1996.
47. Kiernan M, Kraemer HC, Winkleby MA, King AC, Taylor CB. Do logistic regression and signal detection identify different subgroups at risk? Implications for the design of tailored interventions. *Psychological Methods* 2001;**6**:35-48.
48. Everitt BS, Landau S, Leese M. Cluster analysis. 4th ed. London: Arnold, 2001.
49. Hartigan JA. Clustering algorithms. New York: Wiley, 1975.
50. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. New York: Wiley, 1990.
51. Sokal RR, Sneath PHA. Principles of numerical taxonomy. San Francisco: WH Freeman, 1963.
52. Vermunt JK, Magidson J. Latent class cluster analysis. In: Hagenaaars JA, McCutcheon AL, eds. Applied latent class analysis. Cambridge, UK: Cambridge University Press, 2002: 89-106.
53. McQueen JB. Some methods of classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, eds. Proceedings : fifth Berkley Symposium 1965-1966 on mathematical statistics and probability. Berkeley, CA: University of California Press, 1967: 281-297.
54. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974;**61**:215-231.

55. Lazarsfeld PF. The logical and mathematical foundation of latent structure analysis. In: Stouffer SA, Guttman L, Suchman EA, Lazarsfeld PF, Star SA, Clausen JA, eds. *Measurement and prediction*. Princeton, New Jersey: Princeton University Press, 1950: 362-412.
56. Lazarsfeld PF, Henry NW. *Latent structure analysis*. Boston, Massachusetts: Houghton-Mifflin, 1968.
57. Manton KG, Woodbury MA, Tolley HD. *Statistical applications using fuzzy sets*. New York: Wiley, 1994.
58. Woodbury MA, Clive J. Clinical pure types as a fuzzy partition. *Journal of Cybernetics* 1974;**4**:111-121.
59. Woodbury MA, Clive J, Garson A. Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and biomedical research* 1978;**11**:277-298.
60. Hasselblad V. Estimation of parameters for a mixture of normal distributions. *Technometrics* 1966;**8**:431-444.
61. McLachlan GJ, Peel D. *Finite mixture models*. New York: Wiley, 2000.
62. Wolfe JH. A computer program for the maximum likelihood analysis of types. San Diego, California: US Naval Personnel Research Activity, 1965.
63. Clarke DM, Smith GC, Dowe DL, McKenzie DP. An empirically derived taxonomy of common distress syndromes in the medically ill. *Journal of Psychosomatic Research* 2003;**54**:323-330.
64. Davidson J, Woodbury MA, Pelton S, Krishnan R. A study of depressive typologies using grade of membership analysis. *Psychological Medicine* 1988;**18**(1):179-89.
65. Lincoln KD, Chatters LM, Taylor RJ, Jackson JS. Profiles of depressive symptoms among African Americans and Caribbean Blacks. *Social Science in Medicine* 2007;**65**(2):200-13.
66. Parker G, Hadzi-Pavlovic D, Boyce P, et al. Classifying depression by mental state signs. *British Journal of Psychiatry* 1990;**157**:55-65.
67. Parker G, Hadzi-Pavlovic D, Roussos J, et al. Non-melancholic depression: the contribution of personality, anxiety and life events to subclassification. *Psychological Medicine* 1998;**28**(5):1209-19.
68. Pilowsky I, Levine S, Boulton DM. The classification of depression by numerical taxonomy. *British Journal of Psychiatry* 1969;**115**:937-945.
69. Sullivan PF, Prescott CA, Kendler KS. The subtypes of major depression in a twin registry. *Journal of Affective Disorders* 2002;**68**:273-284.
70. Gibson WA. Three multivariate models: factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika* 1959;**24**:229-252.
71. DeSarbo WS, Cron WL. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* 1988;**5**:249-282.
72. Wedel M, DeSarbo WS. Mixture regression models. In: Hagenaars JA, McCutcheon AL, eds. *Applied latent class analysis*. Cambridge, UK: Cambridge University Press, 2002: 366-382.
73. Magidson J, Vermunt JK. Latent class models. In: Kaplan D, ed. *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, California: Sage, 2004: 175-198.

74. Wedel M, DeSarbo WS. A review of recent developments in latent class regression models. In: Bagozzi RP, ed. *Advanced methods of marketing research*. Cambridge, UK: Blackwell, 1994: 352-388.
75. Quandt RE, Ramsey JB. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association* 1978;**73**:730-738.
76. Muthen B, Muthen LK. Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research* 2000;**24**(6):882-91.
77. Stoolmiller M, Kim HK, Capaldi DM. The course of depressive symptoms in men from early adolescence to young adulthood: identifying latent trajectories and early predictors. *Journal of Abnormal Psychology* 2005;**114**(3):331-45.
78. Smits F, Smits N, Schoevers R, Deeg D, Beekman A, Cuijpers P. An epidemiological approach to depression prevention in old age. *American Journal of Geriatric Psychiatry* 2008;**16**:444-53.
79. Fikretoglu D, Brunet A, Schmitz N, Guay S, Pedlar D. Posttraumatic stress disorder and treatment seeking in a nationally representative Canadian military sample. *Journal of Traumatic Stress* 2006;**19**:847-58.
80. Breiman L, Friedman J, Olshen RA, Stone CJ. *Classification and regression trees*. Belmont, California: Wadsworth, 1984.
81. Salford Systems. *CART for Windows*, version 6 [computer software]. San Diego, California: Salford Systems, 2006.
82. SPSS for Windows, version 15.0 [computer software] [program]. Chicago, Illinois: SPSS Inc., 2006.
83. Gini C. Measurement of inequality of income. *The Economic Journal* 1921;**31**:124-126.
84. Light RJ, Margolin BH. An analysis of variance for categorical data. *Journal of the American Statistical Association* 1971;**66**:534-544.
85. Sonquist JA, Morgan JN. *The detection of interaction effects: a report on a computer program for the selection of optimal combinations of exploratory variables*. Ann Arbor, Michigan: Institute for Social Research, University of Michigan, 1964.
86. Chambers R, Hentges A, Zhao X. Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society, Series A* 2004;**167**:323-339.
87. Harper PR. A review and comparison of classification algorithms for medical decision making. *Health Policy* 2005;**71**:315-331.
88. McKenzie DP, Low LH. The construction of computerized classification systems using machine learning methods : an overview. *Computers in Human Behavior* 1992;**8**:155-167.
89. Morgan JN. History and potential of binary segmentation for exploratory data analysis. *Journal of Data Science* 2005;**3**:123-136.
90. Zhang H. Recursive partitioning and tree-based methods. In: Gentle JE, Hardle W, Mori Y, eds. *Handbook of computational statistics*. Berlin: Springer, 2004.

91. Gehrke J. Decision trees. In: Ne Y, ed. The handbook of data mining. Matwah, New Jersey: Erlbaum, 2003: 3-24.
92. Loh WT. Classification and regression tree methods. In: Ruggeri F, Kenett R, Faltin FW, eds. Encyclopedia of statistics in quality and reliability. Chichester, UK: Wiley, 2008: 315-323.
93. Murthy SK. Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Mining and Knowledge Discovery* 1998;**2**:345-389.
94. Ripley BD. Pattern recognition and neural networks. Cambridge, UK: University Press, 1996.
95. Rokach L, Maimon O. Data mining with decision trees: theory and applications. Singapore: World Scientific, 2008.
96. Goodall DW. Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. *Australian Journal of Botany* 1953;**1**:39-63.
97. Williams WT, Lance GN. Automatic subdivision of associated populations [letter to the editor]. *Nature* 1958;**182**:1755.
98. Williams WT, Lambert JM. Multivariate methods in plant ecology. I. Association-analysis in plant communities. *Journal of Ecology* 1959.
99. Fielding A, O'Muircheartaigh CA. Binary segmentation in survey analysis with particular reference to AID. *The Statistician* 1977;**26**:17-28.
100. Agostini JM. A method of market segmentation. *Advertiser's Weekly* 1966(January 7th):22-24.
101. Belson WA. The impact of television methods and findings in program research. Melbourne: F.W. Cheshire, 1967.
102. Belson WA. Investigating causal hypotheses concerning delinquent behavior, with special reference to new strategies in data collection and analysis. *The Statistician* 1978;**27**:1-25.
103. Thompson VR. Sequential dichotomisation : two techniques. *The Statistician* 1972;**21**:181-195.
104. de Ville B. Applying statistical knowledge to database analysis and knowledge base construction. Sixth Conference on Artificial Intelligence Applications 1990, Santa Barbara, California: 30-36.
105. Hunt EB. Concept learning: an information processing problem. New York: Wiley, 1962.
106. Shneidman E, Farberow NL. Clues to suicide. New York: McGraw-Hill, 1957.
107. Morgan JN, Sonquist JA. Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association* 1963;**58**:415-434.
108. Andrews FM. Multiple classification analysis: a report on a computer program for multiple regression using categorical predictors. Ann Arbor, Michigan: Survey Research Center, University of Michigan, 1967.
109. Yates F. The analysis of variance with unequal numbers in the different classes. *Journal of the American Statistical Association* 1934;**29**:51-66.

110. Sonquist JA. Multivariate model building : the validation of a search strategy. Ann Arbor, Michigan: Institute for Social Research, University of Michigan, 1970.
111. Maxwell SE, Delaney HD. Designing experiments and analyzing data: a model comparison perspective. 2nd ed. Mahwah, New Jersey: Erlbaum, 2004.
112. Sonquist JA, Baker EL, Morgan JN. Searching for structure. revised ed. Ann Arbor, Michigan: Institute for Social Research, University of Michigan, 1973.
113. Calahan D. Problem drinkers. San Francisco, California: Jossey-Bass, 1970.
114. Keil IJ. Sex role variations and women's drinking: results from a household survey in Pennsylvania. *Journal of Studies on Alcohol* 1978;**39**:859-868.
115. Reich T, Robins LN, Woodruff RA, Taibleson M, Rich C, Cunningham L. Computer-assisted derivation of a screening interview for alcoholism. *Archives of General Psychiatry* 1975;**32**:847-852.
116. Woodruff RA, Robins LN, Taibleson M, Reich T, Schwin R, Frost N. A computer assisted derivation of a screening instrument of hysteria. *Archives of General Psychiatry* 1973;**29**:450-454.
117. Robins LN, Marcus SC. The diagnostic screening procedure writer: a tool to develop individualized screening procedures. *Medical Care* 1987;**25**, **supplement**:S106-S112.
118. Mills R, Fetter RB, Riedel DC, Averill R. AUTOGRP: an interactive computer system for the analysis of health care data. *Medical Care* 1976;**14**(7):603-15.
119. Thompson JD, Fetter RB, Mross CD. Case mix and resource use. *Inquiry* 1975;**12**:300-312.
120. Young WW, Swinkola RB, Hutton MA. Assessment of the AUTOGRP patient classification system. *Medical Care* 1980;**18**(2):228-44.
121. Krupinski J, Alexander L, Carson N. Patterns of psychiatric care in Victoria: 1961-1978. Melbourne: Mental Health Research Institute, Mental Health Commission of Victoria, 1982.
122. Messenger R, Mandell L. A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association* 1972;**67**:768-772.
123. Morgan JN, Messenger RC. THAID - a sequential analysis program for the analysis of nominal scale dependent variables. Ann Arbor, Michigan: Institute for Social Research, University of Michigan, 1973.
124. Andrews FM, Messenger RC. Multivariate nominal scale analysis. Ann Arbor, Michigan: Institute for Social Research, University of Michigan, 1973.
125. Goodman LA, Kruskal WH. Measures of association for cross-classifications. *Journal of the American Statistical Association* 1954;**49**:732-764.
126. Fielding A. Binary segmentation : the automatic interaction detector and related techniques for exploring data structure. In: O'Muircheartaigh CA,

- Payne C, eds. The analysis of survey data : exploring data structures. Chichester, UK: Wiley, 1977: 221-256.
127. Freih L, Wall S. Quantity and variation in morbidity: THAID-analysis of the occurrence of gastroenteritis among Ethiopian children. *International Journal of Epidemiology* 1979;**8**:313-325.
 128. Cellard JC, Labbe B, Savitsky G. Le programme E.L.I.S.E.E. - presentation et applications. *METRA* 1967;**6**:503-520.
 129. Bouroche J-M, Tenenhaus M. Some segmentation methods. *METRA* 1970;**9**:407-418.
 130. Macnaughton-Smith P. The classification of individuals by the possession of attributes associated with a criterion. *Biometrics* 1963;**19**:364-366.
 131. Macnaughton-Smith P. Some statistical and other numerical techniques for classifying individuals. London: Her Majesty's Stationery Office, 1965.
 132. Feinstein AR, Landis JR. A computer program for finding multivariate prognostic clusters. *Clinical Research* 1973;**21**:725.
 133. Finifter BM. ERIV : a computer program for evaluating relative importance of variables in the analysis of interaction effects. *Behavioral Science* 1971;**16**:511-512.
 134. Gillo MW. MAID, A Honeywell 600 program for an automatized survey analysis. *Behavioral Science* 1972;**17**:251-252.
 135. Gillo MW, Shelly MW. Predictive modeling of multivariable and multivariate data. *Journal of the American Statistical Association* 1974;**69**:646-653.
 136. Henrichon EG, Jr., Fu KS. A nonparametric partitioning procedure for pattern classification. *IEEE Transactions on Computers* 1969;**18**:614-624.
 137. Press LI. IDEA : a technique for inductive data exploration and analysis.: University of California, 1967.
 138. Press LI, Rogers MS, Shure GH. An interactive technique for the analysis of multivariate data. *Behavioral Science* 1969;**14**:364-370.
 139. Stone PJ, Dunphy DC, Smith MS, Ogilvie DM. The General Inquirer : a computer approach to content analysis in the behavioral sciences. New York: McGraw Hill, 1966.
 140. Bardini T. Bootstrapping: Douglas Engelbart, co-evolution and the origins of personal computing. Stanford, California: Stanford University Press, 2000.
 141. Arthur Yates & Co. PL. Yates garden guide for the Australian gardener. 26th ed. Sydney: Arthur Yates & Co., Pty. Ltd, 1956.
 142. Efroymson MA. Multiple regression analysis. In: Ralston A, Wilf HS, eds. Mathematical methods for digital computers. New York: Wiley, 1960: 191-203.
 143. Miller AJ. Subset selection in regression. 2nd ed. Boca Raton, Florida: Chapman & Hall / CRC, 2002.
 144. Doyle P. The use of automatic interaction detector and similar search procedures. *Operational Research Quarterly* 1973;**24**:465-467.
 145. Hoffman R, Carpenter BK, Minkin VI. Ockham's razor and chemistry. *Hyle : an International Journal for the Philosophy of Chemistry* 1997;**3**:3-28.
 146. Russell B. A history of Western philosophy. London: Allen and Unwin, 1965.

147. Calaprice A. The expanded quotable Einstein. Princeton, NJ: Princeton University Press, 2000.
148. Holte RC. Very simple classification rules perform well on most commonly used data sets. *Machine Learning* 1993;**11**:63-91.
149. Murphy PM, Pazzani MJ. Exploring the decision tree forest. *Journal of Artificial Intelligence Research* 1994;**1**:257-275.
150. Kass GV. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 1980;**29**:119-127.
151. Stallones RA. The use and abuse of subgroup analysis in epidemiological research. *Preventive Medicine* 1987;**16**:183-194.
152. McKenzie DP, McGorry PD, Wallace CS, Low LH, Copolov DL, Singh BS. Constructing a minimal diagnostic decision tree. *Methods of Information in Medicine* 1993;**32**:161-166.
153. Dunn OJ. Multiple comparisons among means. *Journal of the American Statistical Association* 1961;**56**:52-64.
154. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979;**6**:65-70.
155. Perneger TV. What's wrong with Bonferroni adjustment. *British Medical Journal* 1998;**316**:1236-1238.
156. Austin PC, Goldwasser MA. Pisces did not have increased heart failure: data driven comparison of binary proportions between levels of a categorical variable can result in increased significance levels. *Journal of Clinical Epidemiology* 2008;**61**:295-300.
157. Miller R, Siegmund D. Maximally selected chi square statistics. *Biometrics* 1982;**38**:1011-1016.
158. Scott AJ, Knott M. An approximate test for use with AID. *Applied Statistics* 1976;**25**:103-106.
159. Worsley KJ. A nonparametric extension of a cluster analysis method by Scott and Knott. *Biometrics* 1977;**33**:532-535.
160. Kass GV. Significance testing in automatic interaction detection. *Applied Statistics* 1975;**24**:178-189.
161. Hawkins DM. Fitting multiple change-point models to data. *Computational Statistics and Data Analysis* 2001;**37**:323-341.
162. Jandhyala VK, Fotopoulos SB, Hawkins DM. Detection and estimation of abrupt changes in the variability of a process. *Computational Statistics and Data Analysis* 2002;**40**:1-19.
163. Young SS, Hawkins DM. Using recursive partitioning analysis to evaluate compound selection methods. In: Bajorath J, ed. Chemoinformatics: concepts, methods, and tools for drug discovery. Totowa, New Jersey: Humana Press, 2004: 317-334.
164. Feller WD. An introduction to probability theory and its application. New York: Wiley, 1968.
165. Biggs D, de Ville B, Suen E. A method of choosing multi-way partitions for classification and decision trees. *Journal of Applied Statistics* 1991;**18**:49-62.

166. Manly BFJ. Randomization, bootstrap and monte carlo methods in biology. 3rd ed. Boca Raton, Florida: Chapman & Hall / CRC, 2006.
167. McKenzie DP, Mackinnon AJ, Peladeau N, et al. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? *Journal of Psychiatric Research* 1996;**30**:483-492.
168. Mehta CR. The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research* 1994;**3**:135-156.
169. Fisher RA. The design of experiments. Edinburgh, Scotland: Oliver & Boyd, 1935.
170. Pitman EJR. Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society, Series B* 1937;**4**:119-130.
171. Chung JH, Fraser DAS. Randomization tests for a multivariate two-sample problem. *Journal of the American Statistical Association* 1958;**53**:729-735.
172. Dwass M. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 1957;**28**:181-187.
173. Westfall PH, Young SS. Resampling-based multiple testing. New York: Wiley, 1993.
174. Frank E, Witten IH. Using a permutation test for attribute selection in decision trees. Fifteenth International Conference on Machine Learning 1998: 152-160.
175. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics* 2006;**15**:651-674.
176. Rabinowitz J, Levine SZ, Hafner H. A population based elaboration of the role of age of onset on the course of schizophrenia. *Schizophrenia Research* 2006;**88**:96-101.
177. KnowledgeSEEKER [program]. Toronto, Canada: Angoss Software Corporation, 2003.
178. Barton CA, McKenzie DP, Walters EH, Abramson MJ, Victorian Asthma Mortality Study Group. Interactions between psychosocial problems and management of asthma : who is at risk of dying? *Journal of Asthma* 2005;**42**:249-256.
179. Keks NA, McKenzie DP, Low LH, et al. Multidiagnostic evaluation of prolactin response to haloperidol challenge in schizophrenia: maximal blunting in Kraepelinian patients. *Biological Psychiatry* 1992;**32**:426-437.
180. Bailey SL. The measurement of problem drinking in young adulthood. *Journal of Studies on Alcohol* 1999;**60**:234-244.
181. O'Connell J, Novins DK, Beals J, et al. The relationship between patterns of alcohol use and mental and physical health disorders in two American Indian populations. *Addiction* 2006;**101**:69-83.
182. Witbrodt J, Kaskutas LA. Does diagnosis matter? differential effects of 12-step participation and social networks in abstinence. *American Journal of Drug and Alcohol Abuse* 2005;**31**:685-707.

183. Yen CF, Cheng CP, Tsai JL, Hsu SY. Family, peer and individual factors related to methylenedioxymethamphetamine use in Taiwanese adolescents. *Psychiatry and Clinical Neuroscience* 2007;**61**(5):552-7.
184. Kissane DW, Bloch S, Dowe DL, et al. The Melbourne family grief study, I: perceptions of family functioning in bereavement. *American Journal of Psychiatry* 1996;**153**:650-658.
185. Welte JW, Barnes GM, Wieczorek WF, Tidwell MC. Gambling participation and pathology in the United States--a sociodemographic analysis using classification trees. *Addictive Behaviors* 2004;**29**:983-989.
186. Boscarino JA, Galea S, Adams RE, Ahern J, Resnick H, Vlahov D. Mental health service and medication use in New York City after the September 11, 2001, terrorist attack. *Psychiatric Services* 2004;**55**(3):274-83.
187. Handrinos D, McKenzie D, Smith G. Timing of referral to a consultation-liaison psychiatry unit. *Psychosomatics* 1998;**39**:311-317.
188. Marino R, Stuart GW, Wright FAC, Minas IH, Klimidis S. Acculturation and dental health among Vietnamese living in Melbourne, Australia. *Community Dentistry and Oral Epidemiology* 2001;**29**:107-119.
189. Levkoff SE, Safran C, Cleary PD, Gallop J, Phillips RS. Identification of factors associated with the diagnosis of delirium in elderly hospitalized patients. *Journal of the American Geriatrics Society* 1988;**36**:1099-1104.
190. Shah A, Chiu E, Ames D, Harrigan S, McKenzie D. The characteristics of aggressive subjects in Australian (Melbourne) nursing homes. *International Psychogeriatrics* 2000;**12**:145-161.
191. Tunis SL, Edell WS, Adams BE, Kennedy JS. Characterizing behavioral and psychological symptoms of dementia (BPSD) among geropsychiatric patients. *Journal of the American Medical Directors Association* 2002;**3**:146-151.
192. Saltini A, Mazzi MA, Del Piccolo L, Zimmerman C. Decisional strategies for the attribution of emotional distress in primary care. *Psychological Medicine* 2004;**34**:729-739.
193. Huang HC, Lin TK, Ngui PW. Analysing a mental health survey by chi-squared automatic interaction detection. *Annals of the Academy of Medicine, Singapore* 1993;**22**:332-337.
194. LaForge R. Confidence intervals or tests of significance in scientific research. *Psychological Bulletin* 1967;**68**:446-447.
195. Nunnally J. The place of statistics in psychology. *Educational and Psychological Measurement* 1960;**20**:641-650.
196. Rozenboom WW. The fallacy of the null-hypothesis significance test. *Psychological Bulletin* 1960;**57**:416-428.
197. Cohen J. The earth is round ($p < .05$). *American Psychologist* 1994;**49**:997-1003.
198. Harlow LL, Mulaik SA, Steiger JH. What if there were no significance tests. Mahwah, New Jersey: Erlbaum, 1997.
199. Grissom RJ, Kim JJ. Effect sizes for research: a broad practical approach. Mahwah, NJ: Erlbaum, 2005.

200. Kline RB. Beyond significance testing: reforming data analysis methods in behavioral research. Washington, D.C.: American Psychological Association, 2004.
201. Vacha-Haase T, Nilsson JE, Reetz DR, Lance TS, Thompson B. Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology* 2000;**10**:413-425.
202. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;**1**:43-46.
203. Geisser S. The predictive sample reuse method with applications. *Journal of the American Statistical Association* 1975;**70**:320-328.
204. Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* 1974;**39**:44-47.
205. Lachenbruch PA, Mickey MR. Estimation of error rates in discriminant analysis. *Technometrics* 1968;**10**:1-11.
206. Allen DM. Mean square error of prediction as a criterion for selecting variables. *Technometrics* 1971;**13**:469-475.
207. Mosteller F, Tukey JW. Data analysis, including statistics. In: Lindzey G, Aranson E, eds. Handbook of social psychology. Reading, Massachusetts: Addison-Wesley, 1968: 80-203.
208. Mabbett A, Stone M, Washbrook J. Cross-validators selection of binary variables in differential diagnosis. *Applied Statistics* 1980;**1980**:198-204.
209. Christensen R, Johnson W. A conversation with Seymour Geisser. *Statistical Science* 2007;**22**:621-636.
210. Stone M. Cross-validators choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B* 1974;**36**:111-147.
211. Friedman JH. A recursive partitioning decision rule for nonparametric classifiers. *IEEE Transactions on Computers* 1977;**C-26**:404-408.
212. Elder J, Pregibon D. A statistical perspective on knowledge discovery in databases. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, eds. Advances in knowledge discovery and data mining. Menlo Park, California: AAAI Press, 1996: 83-113.
213. Weiss SM, Kulikowski CA. Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. San Mateo, California: Morgan Kaufman, 1991.
214. Efron B. Computers and the theory of statistics: thinking the unthinkable. *SIAM Review* 1979;**21**:460-480.
215. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall, 1993.
216. Simon JL. Basic research methods in social science. New York: Random House, 1969.
217. Efron B. Estimating the error rate of a prediction rule: some improvements on cross-validation. *Journal of the American Statistical Association* 1983;**78**:316-331.

218. Efron B, Tibshirani R. Improvements on cross-validation : the .632+ bootstrap method. *Journal of the American Statistical Association* 1997;**92**:548-560.
219. Crawford S. Extensions to the CART algorithm. *International Journal of Man-Machine Studies* 1989;**31**:197-207.
220. Merler S, Furlanello C. Selection of tree-based classifiers with the bootstrap .632 rule. *Biometrical Journal* 1997;**3**:369-382.
221. Kullback S. Information theory and statistics. New York: Wiley, 1959.
222. Shannon C. A mathematical theory of communication. *The Bell System Technical Journal* 1948;**27**:379-423.
223. Shannon C, Weaver W. The mathematical theory of communication. Urbana, Illinois: University of Illinois Press, 1949.
224. Agresti A. Categorical data analysis. 2nd ed. Hoboken, New Jersey: Wiley, 2002.
225. Fisher RA. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* 1922;**222**:309-368.
226. Fisher RA. The conditions under which X^2 measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society* 1924;**87**:442-450.
227. Quinlan JR. Discovering rules by induction from large collections of examples. In: Michie D, ed. Expert systems in the micro electronic age. Edinburgh: University Press, 1979: 168-201.
228. Mingers J. An empirical comparison of selection measures for decision tree induction. *Machine Learning* 1989;**3**:319-342.
229. Povalec P, Lenic M, Zorman M, Kokol P, Dinevski D. Accuracy of intelligent medical systems. *Computer Methods and Programs in Biomedicine* 2005;**80**(Supplement 1):S95-S105.
230. Ericson WA. A note on partitioning for maximum between sum of squares. In: Sonquist JN, Morgan JA, eds. The detection of interaction effects : a report on a computer program for the selection of optimal combinations of exploratory variables. Ann Arbor, Michigan: Institute for Social Research, University of Michigan, 1964: 149-157.
231. Fisher WD. On grouping for maximum homogeneity. *Journal of the American Statistical Association* 1958;**53**:442-450.
232. Barnes GM, Welte JW, Dintcheff B. Drinking among subgroups in the adult population of New York State: a classification analysis using CART. *Journal of Studies on Alcohol* 1991;**52**:338-344.
233. Dierker LC, Avenevoli S, Goldberg A, Glantz M. Defining subgroups of adolescents at risk for experimental and regular smoking. *Prevention Science* 2004;**5**:169-183.
234. Elkashef A, Holmes TH, Bloch DA, et al. Retrospective analyses of pooled data from CREST 1 and CREST II trials for treatment of cocaine dependence. *Addiction* 2005;**100**(Supplement 1):91-101.
235. Sullivan PF, Kovalenko P, York TP, Prescott CA, Kendler KS. Fatigue in a community sample of twins. *Psychological Medicine* 2003;**33**:197-201.

236. D'Alisa S, Miscio G, Baudo S, Simone A, Tesio L, Mauro A. Depression is the main determinant of quality of life in multiple sclerosis: a classification-regression (CART) study. *Disability and Rehabilitation* 2006;**28**:307-314.
237. Schmitz N, Kugler J, Rollnik J. On the relationship between neuroticism, self-esteem, and depression: results from the National Comorbidity survey. *Comprehensive Psychiatry* 2003;**44**:169-176.
238. Biederman J, Petty C, Faraone SV, et al. Childhood antecedents to panic disorder in referred and nonreferred adults. *Journal of Child and Adolescent Psychopharmacology* 2005;**15**:549-561.
239. Biederman J, Petty CR, Faraone SV, et al. Antecedents to panic disorder in nonreferred adults. *Journal of Clinical Psychiatry* 2006;**67**:1179-1186.
240. Knable MB, Barci BM, Bartko JJ, Webster MJ, Torrey EF. Molecular abnormalities in the major psychiatric illnesses: classification and regression tree (CRT) analysis of post-mortem prefrontal markers. *Molecular Psychiatry* 2002;**7**:392-404.
241. Boerstler H, de Figuieredo JM. Prediction of use of psychiatric services: application of the CART algorithm. *Journal of Mental Health Administration* 1991;**18**:27-34.
242. Cromwell J, Drozd EM, Gage B, Maier J, Richter E, Goldman HH. Variation in patient routine costliness in U.S. psychiatric facilities. *Journal of Mental Health Policy and Economics* 2005;**8**:15-28.
243. Drozd EM, Gage B, Maier J, Greenwald LM, Goldman HK. Patient casemix classification for Medicare psychiatric prospective payment. *American Journal of Psychiatry* 2006;**163**:724-732.
244. Belle SH, Mendelsohn AB, Seaberg EC, Ratcliff G. A brief cognitive screening battery for dementia in the community. *Neuroepidemiology* 2000;**19**:53-56.
245. Royall DR, Palmer R, Mulroy AR, et al. Pathological determinants of the transition to clinical dementia in Alzheimer's disease. *Experimental Aging Research* 2002;**28**(2):143-62.
246. Craig TJ, Siegel C, Hopper K, Lin S, Sartorius N. Outcome in schizophrenia and related disorders compared between developing and developed countries: a recursive partitioning reanalysis of the WHO DOSMD data. *British Journal of Psychiatry* 1997;**170**:229-233.
247. Subotnik KL, Nuechterlein KH, Irzhevsky V, Kitchen CM, Woos SM, Mintz J. Is unawareness of psychotic disorder a neurocognitive or psychological defensiveness problem? *Schizophrenia Research* 2005;**75**:147-157.
248. Mann JJ, Ellis SP, Waternaux CM, et al. Classification trees distinguish suicide attempters in major psychiatric disorders: a model of clinical decision making. *Journal of Clinical Psychiatry* 2008:e1-e9.
249. Quinlan JR. C4.5: programs for machine learning. San Mateo, California: Morgan Kaufmann, 1993.
250. Quinlan JR. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research* 1996;**4**:77-90.
251. Quinlan JR. Learning efficient classification procedures and their application to chess endgames. In: Michalski RS, Carbonell JG, Mitchell TM, eds.

- Machine learning: an AI approach. Palo Alto, California: Tioga, 1983: 463-482.
252. Quinlan JR. Decision trees and multi-valued attributes. In: Hayes JE, Michie D, Richards J, eds. Machine intelligence 11. Oxford, UK: Clarendon Press, 1988: 305-318.
 253. Quinlan JR. Simplifying decision trees. *International Journal of Man-Machine Studies* 1987;**27**:221-234.
 254. Stone M. Quantification of Ockham's razor by violation of the likelihood principle. Madison, Wisconsin: Department of Statistics, University of Wisconsin, 1966.
 255. Wallace CS, Freeman PR. Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B* 1987;**4**:223-265.
 256. Quinlan JR, Rivest RL. Inferring decision trees using the minimum description length principle. *Information Processing and Computation* 1989;**80**:227-248.
 257. Wallace CS, Patrick JD. Coding decision trees. *Machine Learning* 1993;**11**:7-22.
 258. Ciampi A, Hogg SA, McKinney S, Thiffault J. RECPAM : a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. methods and program features. *Computer Methods and Programs in Biomedicine* 1988;**26**:239-256.
 259. Ciampi A, Thiffault J, Nakache J-P, Asselain B. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival analysis with covariates. *Computational Statistics and Data Analysis* 1986;**4**:185-204.
 260. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974;**19**:716-723.
 261. Hastie T, Tibshirani RJ, Friedman JH. Elements of statistical learning : data mining, inference and prediction. New York: Springer, 2001.
 262. Lanterman AD. Schwarz, Wallace and Rissanen: intertwining themes in theories of model order estimation. *International Statistical Review* 2001;**69**:185-212.
 263. Rissanen J. Modeling by shortest data description. *Automatica* 1978;**14**:465-471.
 264. Wallace CS, Boulton DM. An information measure for classification. *Computer Journal* 1968;**11**:185-194.
 265. Wallace CS, Dowe DL. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics & Computing* 2000;**10**:75-83.
 266. Carinci F, Nicolucci N, Ciampi A, et al. Role of interactions between psychological and clinical factors in determining 6-month mortality among patients with acute myocardial infarction. Application of recursive partitioning techniques to the GISSI-2 database. Gruppo Italiano per lo Studio della Sopravvivenza nell' Infarto Miocardico. *European Heart Journal* 1997;**18**:835-845.

267. Kedia S, Williams C. Predictors of substance abuse treatment outcomes in Tennessee. *Journal of Drug Education* 2003;**33**:25-47.
268. Li L, Huang J, Sun S, et al. Selecting pre-screening items for early intervention trials of dementia - a case study. *Statistics in Medicine* 2004;**23**:271-283.
269. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. *Journal of Clinical Epidemiology* 2003;**56**:826-832.
270. Terrin N, Schmid CH, Griffith JL, D'Agostino RB, Selker HP. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *Journal of Clinical Epidemiology* 2003;**56**(8):721-9.
271. Kass GV. Automatic interaction detection (AID) techniques. In: Kotz S, Johnson NL, Read CB, eds. Encyclopedia of statistical sciences. New York: Wiley, 1982: 148-152.
272. Loh W-Y, Shih Y-S. Split selection methods for classification trees. *Statistica Sinica* 1997;**7**:815-840.
273. Kim H, Loh W-Y. Classification trees with unbiased multiway splits. *Journal of the American Statistical Association* 2001;**96**:589-604.
274. Kim H, Loh W-Y. Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics* 2003;**12**:512-530.
275. Chan K-Y, Loh W-Y. An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics* 2004;**13**:826-852.
276. Loh W-Y. Logistic regression tree analysis. In: Pham H, ed. Handbook of engineering statistics. New York: Springer, 2006: 537-549.
277. McKenzie DP, Clarke DM, Low LH. A method of constructing parsimonious diagnostic and screening tests. *International Journal of Methods in Psychiatric Research* 1992;**2**:71-79.
278. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). 4th ed. Washington, DC: American Psychiatric Association, 1994.
279. Cendrowska J. PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies* 1987;**27**:349-370.
280. Marshall RJ. The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology* 2001;**54**:603-609.
281. Forsyth R. BEAGLE: A Darwinian approach to pattern recognition. *Kybernetes* 1981;**10**:159-166.
282. Janes H, Pepe M, Kooperberg C, Newcomb P. Identifying target populations for screening or not screening using logic regression. *Statistics in Medicine* 2005;**24**:1321-1328.
283. Leenen I, Van Mechelen I. A branch-and-bound algorithm for Boolean regression. In: Balderjahn I, Mathar R, Schader M, eds. Classification, data analysis, and data highways. Berlin: Springer, 1998: 164-171.

284. Marshall RJ. A program to implement a search method for identification of clinical subgroups. *Statistics in Medicine* 1995;**14**:2645-2659.
285. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. *Journal of Computational and Graphical Statistics* 2003;**12**:475-511.
286. Van Mechelen I. Prediction of a dichotomous criterion variable by means of a logical combination of dichotomous predictors. *Mathematiques et Sciences Humaines* 1988;**102**:47-54.
287. Weiss SM, Indurkha N. Optimized rule induction. *IEEE Expert* 1993;**8**:61-69.
288. Pagallo G, Haussler D. Boolean feature discovery in empirical learning. *Machine Learning* 1990;**5**:71-100.
289. van der Linden WJ, Hambleton RK. Handbook of modern item response theory. New York: Springer, 1997.
290. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. New York: Oxford University Press, 2003.
291. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute of Educational Research, 1960.
292. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology* 2007;**46**(Pt 1):1-18.
293. Andrich D. Rating formulation for ordered response categories. *Psychometrika* 1978;**43**:561-573.
294. Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982;**47**:149-174.
295. Clarke DM, Mackinnon AJ, Smith GC, McKenzie DP, Herrman HE. Dimensions of psychopathology in the medically ill: a latent trait approach. *Psychosomatics* 2000;**41**:418-425.
296. Goldberg DP, Bridges K, Duncan-Jones P, Grayson D. Dimensions of neuroses seen in primary-care settings. *Psychological Medicine* 1987;**17**:461-470.
297. Grayson DA, Henderson AS, Kay DW. Diagnoses of dementia and depression: a latent trait analysis of their performance. *Psychological Medicine* 1987;**17**(3):667-75.
298. Jordan M, Jacobs R. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 1994;**6**:181-214.
299. Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 2008;**17**:492-514.
300. Formann AK. Linear logistic latent class analysis and the Rasch model. In: Fischer GH, Molenaar IW, eds. Rasch models: foundations, recent developments and applications. New York: Springer, 1995: 239-255.
301. Formann AK, Kohlmann T. Three-parameter linear logistic latent class analysis. In: Hagenars JA, McCutcheon AL, eds. Applied latent class analysis. Cambridge, UK: Cambridge University Press, 2002: 183-210.

302. Strobl C, Wickelmaier F, Zeileis A. Accounting for individual differences in Bradley-Terry models by recursive partitioning [technical report 54]. Munich: Department of Statistics, University of Munich, 2009.
303. Steinberg D, Cardell NS. The hybrid CART-logit model in classification and data mining. Keystone, Colorado: Eighth Annual Advanced Research Techniques Forum, American Marketing Association, 1998.
304. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AML mortality. *Statistics in Medicine* 2007;**26**(15):2937-57.
305. Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of Biomedical Informatics* 2001;**34**:28-36.
306. Sabbagh A, Darlu P. Data-mining methods as useful tools for predicting individual drug response: application to CYP2D6 data. *Human Heredity* 2006;**62**:119-134.
307. Thomas S, Leese M, Walsh E, et al. A comparison of statistical models in predicting violence in psychotic illness. *Comprehensive Psychiatry* 2005;**46**:296-303.
308. Costanza MC, Paccaud F. Binary classification of dyslipidemia from the waist-to-hip ratio and body mass index: a comparison of linear, logistic, and CART models. *BMC Medical Research Methodology* 2004;**4**:7.
309. Holt RN, Scarpello V, Carroll RJ. Toward understanding the contents of the "black box" for predicting complex decision-making outcomes. *Decision Sciences* 1983;**14**:253-239.
310. Selker HP, Griffith JL, Pair S, Long WJ, D'Agostino RB. A comparison of performance of mathematic predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. *Journal of Investigative Medicine* 1995;**43**:468-476.
311. Delen D, Walker G, Kadam A. Predicting breast cancer survivability. *Artificial Intelligence in Medicine* 2005:113-127.
312. Huang J, Lin A, Narasimhan B, et al. Tree-structured supervised learning and the genetics of hypertension. *Proceedings of the National Academy of Sciences, USA* 2004;**101**:10529-10534.
313. Lacher DA. Comparison of nonparametric recursive partitioning to parametric discriminant analyses in laboratory differentiation of hypercalcemia. *Clinica Chimica Acta* 1991;**204**:199-207.
314. Lim T-S, Loh W-Y, Shih Y-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning* 2000;**40**:203-228.
315. Kooperberg C, Bose S, Stone CJ. Polychotomous regression. *Journal of the American Statistical Association* 1997;**92**:117-127.
316. Houghton D, Oulabi S. Direct marketing modeling with CART and CHAID. *Journal of Interactive Marketing* 1997;**11**:42-52.
317. Hawkins DM, McKenzie DP. A data-based comparison of some recursive partitioning procedures. Statistical Computing Section, American

- Statistical Association. Raleigh, North Carolina: American Statistical Association, 1995: 245-252.
318. Bloemer JMM, Brijs T, Vanhoof K, Swinnen G. Comparing complete and partial classification for identifying customers at risk. *International Journal of Research in Marketing* 2003;**20**:117-131.
 319. Wolpert DH. The relationship between PAC, the statistical physics framework, the Bayesian framework and the VC framework. In: Wolpert DH, ed. *The mathematics of generalization*. Reading, Massachusetts: Addison-Wesley, 1995.
 320. Groot M, Tiffen R, eds. *Australia's Gulf War*. Melbourne: Melbourne University Press, 1992.
 321. Robins LN, Wing J, Wittchen HU, et al. The Composite International Diagnostic Interview : an epidemiological instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry* 1988;**45**:1069-1077.
 322. World Health Organization Collaborating Centre for Mental Health and Substance Abuse. *Composite International Diagnostic Interview: CIDI-Auto 2.1 - Administrator's guide and reference*. Sydney: World Health Organization Collaborating Centre for Mental Health and Substance Abuse, 1997.
 323. Harig PT. Substance abuse programs in military settings. In: Gal R, Mangelsdorff AD, eds. *Handbook of military psychology*. Chichester, UK: Wiley, 1991: 635-655.
 324. Micklewright S. Problem drinking in the Naval Service: a study of personnel identified as alcohol abusers. *Journal of the Royal Naval Medical Service* 1996;**82**:34-40.
 325. Hankin CS, Spiro A, Miller DR, Kazis L. Mental disorders and mental health treatment among U.S. Department of Veterans outpatients: the Veterans Health Study. *American Journal of Psychiatry* 1999;**156**:1924-1930.
 326. Branchey L, Davis W, Lieber C. Alcoholism in Vietnam and Korea veterans: a long term follow-up. *Alcoholism, Clinical and Experimental Research* 1984;**8**(6):572-575.
 327. O'Toole BI, Marshall RP, Grayson DA, et al. The Australian Vietnam Veterans Health Study: III. Psychological health of Australian Vietnam veterans and its relationship to combat. *International Journal of Epidemiology* 1996;**25**:331-340.
 328. Hotopf M. Treating Gulf War Veterans' illnesses - Are more focused studies needed? *Journal of the American Medical Association* 2003;**289**(11):1436-1437.
 329. Ikin JF, Sim MR, Creamer MC, et al. War-related psychological stressors and risk of psychological disorders in Australian veterans of the 1991 Gulf War. *British Journal of Psychiatry* 2004;**185**:116-126.
 330. The Iowa Persian Gulf Study Group. Self-reported illness and health status among Gulf War veterans: A population-based study. *Journal of the American Medical Association* 1997;**277**(3):238-245.

331. Sim M, Abramson M, Forbes A, et al. Australian Gulf War Veterans' Health Study Volumes 1-3. Canberra: Monash University for the Commonwealth of Australia, 2002.
332. Ismail K, Kent K, Brugha T, et al. The mental health of UK Gulf War veterans: phase 2 of a two phase cohort study. *British Medical Journal* 2002;**325**:525-576.
333. Forbes AB, McKenzie DP, Mackinnon AJ, et al. The health of Australian veterans of the 1991 Gulf War: factor analysis of self-reported symptoms. *Occupational and Environmental Medicine* 2004;**61**:1014-1020.
334. Kelsall H, Macdonell R, Sim M, et al. Neurological status of Australian veterans of the 1991 Gulf War and the effect of medical and chemical exposures. *International Journal of Epidemiology* 2005;**34**:810-819.
335. Kelsall HL, Sim MR, Forbes AB, et al. Symptoms and medical conditions in Australian veterans of the 1991 Gulf War: relationship to immunisations and other Gulf War exposures. *Occupational and Environmental Medicine* 2004;**61**:1006-1013.
336. Kelsall HL, Sim MR, Forbes AB, et al. Respiratory health status of Australian veterans of the 1991 Gulf War and the effects of exposure to oil fire smoke and dust storms. *Thorax* 2004;**59**:897-903.
337. Kelsall HL, Sim MR, Ikin JF, et al. Reproductive health of male Australian veterans of the 1991 Gulf War. *BMC Public Health* 2007;**7**:79.
338. Kelsall HL, Sim MR, McKenzie DP, et al. Medically evaluated psychological and physical health of Australian Gulf War veterans with chronic fatigue. *Journal of Psychosomatic Research* 2006;**60**:575-584.
339. McKenzie DP, Ikin JF, McFarlane AC, et al. Psychological health of Australian veterans of the 1991 Gulf War: an assessment using the SF-12, GHQ-12 and PCL-S. *Psychological Medicine* 2004;**34**:1419-1430.
340. Goss Gilroy Inc. Health study of Canadian Forces personnel involved in the 1991 conflict in the Persian Gulf. Ottawa, Canada, 1998.
341. Ishoy T, Suadican P, Guldager B, Appleyard M, Hein HO, Gyntelberg F. State of health after deployment in the Persian Gulf. The Danish Gulf War Study. *Danish Medical Bulletin* 1999;**46**(5):416-419.
342. Salamon R, Verret C, Jutand MA, et al. Health consequences of the first Persian Gulf War on French troops. *International Journal of Epidemiology* 2006;**35**:479-487.
343. Wessely S. The long aftermath of the 1991 Gulf War. *Annals of Internal Medicine* 2004;**141**(2):155-156.
344. Friedman MJ. Veterans' mental health in the wake of war. *New England Journal of Medicine* 2005;**352**(13):1287-1290.
345. Ikin J, McKenzie D, Creamer M, et al. War zone stress without direct combat: the Australian naval experience of the Gulf War. *Journal of Traumatic Stress* 2005;**18**:193-204.
346. Mateczun JM, Holmes EK. Return, readjustment, and reintegration: the three R's of family reunion. . In: Ursano RJ, Norwood AE, eds. Emotional aftermath of the Persian Gulf War: Veterans, Families, Communities and Nations. Washington, DC: American Psychiatric Press, 1996: 369-392.

347. Jones E, Wessely S. Shell shock to PTSD: military psychiatry from 1900 to the Gulf War. New York: Psychology Press, 2005.
348. Singer JD, Willett JB. Applied longitudinal data analysis : modeling change and event occurrence. New York: Oxford University Press, 2003.
349. Steptoe A. Depression and physical illness. Cambridge, UK: Cambridge University Press, 2007.
350. Clarke DM. Psychological factors in illness and recovery. *New Zealand Medical Journal* 1998;**111**:410-412.
351. Grassi L, Mangelli L, Fava GA, et al. Psychosomatic characterization of adjustment disorders in the medical setting: Some suggestions for DSM-V. *Journal of Affective Disorders* 2007;**101**:251-254.
352. Frank JD, Frank JB. Persuasion and healing: a comparative study of psychotherapy. 2nd ed. Baltimore, Maryland: John Hopkins University Press, 1973.
353. Willner P. Anhedonia. In: Costello CG, ed. Symptoms of depression. New York: Wiley, 1993: 63-84.
354. Clarke D, Smith G, Herrman H, McKenzie D. The Monash Interview for Liaison Psychiatry (MILP): development, reliability and procedural validity. *Psychosomatics*. 1998;**39**:318-328.
355. Goldberg DP, Williams P. A user's guide to the General Health Questionnaire. Windsor, England: NFER-Nelson, 1988.
356. Siegel B, Vukicevic J, Spitzer RL. Using signal detection methodology to revise DSM-III-R: reanalysis of the DSM-III-R national field trials for autistic disorders. *Journal of Psychiatric Research* 1990;**24**:293-311.
357. Zimmerman M, Chelminski I, McGlinchey JB, Young D. Diagnosing major depressive disorder X: can the utility of the DSM-IV symptom criteria be improved? *Journal of Nervous and Mental Disease* 2006;**194**:893-7.
358. Nair J, Nair SS, Kashani JH, Reid JC, Mistry SI, Vargas VG. Analysis of the symptoms of depression--a neural network approach. *Psychiatry Research* 1999;**87**:193-201.
359. Costello CG, ed. Symptoms of depression. New York: Wiley, 1993.
360. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. Washington, DC: American Psychiatric Association, 1994.
361. Williams JBW, Noel PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? *Journal of the American Medical Association* 2002;**287**:1160-1170.
362. Wilhelm KA, Finch AW, Davenport TA, Hickie IB. What can alert the general practitioner to people whose common mental health problems are unrecognised. *Medical Journal of Australia* 2008;**188**(supplement 12):S114-S118.
363. Hickie IB, Davenport TA, Hadzi-Pavlovic D, et al. Development of a simple screening tool for common mental disorders in general practice. *Medical Journal of Australia* 2001;**175 Suppl**:S10-7.
364. Hickie IB, Davenport TA, Scott EM, Hadzi-Pavlovic D, Naismith SL, Koschera A. Unmet need for recognition of common mental disorders in

- Australian general practice. *Medical Journal of Australia* 2001;**175** Suppl:S18-24.
365. Clarke DM, McKenzie DP. An examination of the efficiency of the 12-item SPHERE questionnaire as a screening instrument for common mental disorders in primary care. *Australian and New Zealand Journal of Psychiatry* 2003;**37**:236-239.
 366. Baker R. Beating the blues: mental health takes the industry pills. *The Age* 2006 8 August 1,6.
 367. Horwitz AV, Wakefield JC. The loss of sadness: how psychiatry transformed normal sorrow into depressive disorder. New York: Oxford University Press, 2007.
 368. Rief W, Hessel A, Braehler E. Somatization symptoms and hypochondriacal features in the general population. *Psychosomatic Medicine* 2001;**63**:595-602.
 369. Collins P. Burn: the epic story of bushfire in Australia. Sydney: Allen & Unwin, 2006.
 370. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837-845.
 371. McKenzie DP, Mackinnon AJ, Clarke DM. KAPCOM : A program for the comparison of kappa coefficients obtained from the same sample of observations. *Perceptual and Motor Skills* 1997;**85**:899-902.
 372. Gilbody S, Richards D, Barkham M. Diagnosing depression in primary care using self-completed instruments: UK validation of PHQ-9 and CORE-CM. *British Journal of General Practice* 2007;**57**:650-652.
 373. Clarke DM, McKenzie DP, Marshall RJ, Smith GC. The construction of a brief case-finding instrument for depression in the physically ill. *Integrative Psychiatry* 1996;**10**:117-123.
 374. Clarke DM, Smith GC, Herrman HE. A comparative study of screening instruments for mental disorders in general hospital patients. *International Journal of Psychiatry in Medicine* 1993;**23**:323-337.
 375. Fleminger S. Long-term psychiatric disorders after traumatic brain injury. *European Journal of Anaesthesiology Supplement* 2008;**42**:123-130.
 376. Hoge CW, McGurk D, Thomas JL, Cox AL, Engel CC, Castro CA. Mild traumatic brain injury in U.S. soldiers returning from Iraq. *New England Journal of Medicine* 2008;**358**:453-63.
 377. Mushkudani NA, Hukkelhoven CW, Hernandez AV, et al. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. *Journal of Clinical Epidemiology* 2008;**61**:331-343.
 378. Stiell IG, Clement CM, McKnight RD, et al. The Canadian C-spine rule versus the NEXUS low-risk criteria in patients with trauma. *New England Journal of Medicine* 2003;**349**:2510-8.
 379. Stiell JG, Wells GA, Vandemheen KL, et al. The Canadian C-spine rule for radiography in alert and stable trauma patients. *Journal of the American Medical Association* 2001;**286**:1841-1848.

380. Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. *Lancet* 1974;**2**:81-84.
381. Ware JE, Kosinski M, Keller SD. A 12-item Short-Form Health Survey. Construction of scales and preliminary tests of reliability and validity. *Medical Care* 1996;**34**:220-233.
382. Silipo R. Neural networks. In: Berthold M, Hand DJ, eds. Intelligent data analysis : an introduction. 2 ed. Berlin: Springer, 2003: 269-320.
383. Boulesteix A-L. Maximally selected chi-square statistics for ordinal variables. *Biometrical Journal* 2006;**48**:451-461.
384. Boulesteix A-L. Maximally selected chi-square statistics and binary splits of nominal variables. *Biometrical Journal* 2006;**48**:838-848.
385. Strobl C, Boulesteix A-L, Augustin T. Unbiased split selection for classification trees based on the Gini index. *Computational Statistics & Data Analysis* 2006.
386. Sakamoto Y, Akaike H. Analysis of cross-classified data by AIC. *Annals of the Institute of Statistical Mathematics* 1978;**30**:187-197.
387. Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 1978;**6**:461-464.
388. Mehta M, Rissanen J, Agrawal R. MDL-based decision tree pruning. Proceedings of Knowledge Discovery in Databases 1995, Montreal, Canada: 216-221.
389. Dayton CM. SUBSET: Best subsets using information criteria. *Journal of Statistical Software* 2001;**6**(2).
390. Eberhart RC, Shi Y. Computational intelligence: concepts to implementation. Burlington, Massachusetts: Morgan Kaufmann, 2007.
391. Andreescu C, Mulsant BH, Houck PR, et al. Empirically derived decision trees for the treatment of late-life depression. *American Journal of Psychiatry* 2008;**165**:855-62.
392. Parker GB, Fletcher K, Hyett MP. The Mood Assessment Program: a computerised diagnostic tool for deriving management plans for mood disorders. *Medical Journal of Australia* 2008;**188**(supplement 12):S126-S128.
393. Parker G, Manicavasagar V. Modelling and managing the depressive disorders. New York: Cambridge University Press, 2005.
394. Wing JK, Sturt E. The PSE-ID-CATEGO system supplementary manual. London: Medical Research Council Social Psychiatry Unit, 1978.
395. Wessely S. The role of screening in the prevention of psychological disorders arising after major trauma : pros and cons. In: Ursano RJ, Fullerton CS, Norwood AE, eds. Terrorism and disaster : Individual and community mental health interventions. Cambridge: Cambridge University Press, 2003: 121-145.
396. Dworkin RW. Artificial happiness: the dark side of the happy new class. New York: Carroll and Graf, 2006.
397. Leader D. The new black: mourning, melancholia and depression. London: Hamish Hamilton, 2008.

APPENDIX A - AUDIT (ALCOHOL USE DISORDERS IDENTIFICATION TEST) QUESTIONNAIRE

Please circle the answer that is correct for you

1. How often do you have a drink containing alcohol?

- Never
- Monthly or less
- 2–4 times a month
- 2–3 times a week
- 4 or more times a week

2. How many standard drinks containing alcohol do you have on a typical day when drinking?

- 1 or 2
- 3 or 4
- 5 or 6
- 7 to 9
- 10 or more

3. How often do you have six or more drinks on one occasion?

- Never
- Less than monthly
- Monthly
- Weekly
- Daily or almost daily

4. During the past year, how often have you found that you were not able to stop drinking once you had started?

- Never
- Less than monthly
- Monthly
- Weekly
- Daily or almost daily

5. During the past year, how often have you failed to do what was normally expected of you because of drinking?

- Never
- Less than monthly
- Monthly
- Weekly
- Daily or almost daily

6. During the past year, how often have you needed a drink in the morning to get

yourself going after a heavy drinking session?

- Never
- Less than monthly
- Monthly
- Weekly
- Daily or almost daily

7. During the past year, how often have you had a feeling of guilt or remorse after drinking?

- Never
- Less than monthly
- Monthly
- Weekly
- Daily or almost daily

8. During the past year, have you been unable to remember what happened the night before because you had been drinking?

- Never
- Less than monthly
- Monthly
- Weekly
- Daily or almost daily

9. Have you or someone else been injured as a result of your drinking?

- No
- Yes, but not in the past year
- Yes, during the past year

10. Has a relative or friend, doctor or other health worker been concerned about your drinking or suggested you cut down?

- No
- Yes, but not in the past year
- Yes, during the past year

APPENDIX B - SPHERE (SOMATIC AND PSYCHOLOGICAL HEALTH REPORT) QUESTIONNAIRE

For more than TWO WEEKS have you:

1. Felt sad, down or miserable most of the time?
2. Lost interest or pleasure in most of your usual activities?

If you answered "YES" to either of these questions, complete the symptom checklist below

Behaviours	
1.	<input type="checkbox"/> Stopped going out
2.	<input type="checkbox"/> Not getting things done at work
3.	<input type="checkbox"/> Withdrawn from close family and friends
4.	<input type="checkbox"/> Relying on alcohol and sedatives
5.	<input type="checkbox"/> Stopped doing things you enjoy
6.	<input type="checkbox"/> Unable to concentrate
Thoughts	
7.	<input type="checkbox"/> "I'm a failure"
8.	<input type="checkbox"/> "It's my fault"
9.	<input type="checkbox"/> "Nothing good ever happens to me"
10.	<input type="checkbox"/> "I'm worthless"
11.	<input type="checkbox"/> "Life is not worth living"
Feelings	
12.	<input type="checkbox"/> Overwhelmed
13.	<input type="checkbox"/> Unhappy, depressed
14.	<input type="checkbox"/> Irritable
15.	<input type="checkbox"/> Frustrated
16.	<input type="checkbox"/> No confidence
17.	<input type="checkbox"/> Guilty
18.	<input type="checkbox"/> Indecisive
19.	<input type="checkbox"/> Disappointed

20.	<input type="checkbox"/>	Miserable
21.	<input type="checkbox"/>	Sad
Physical		
22.	<input type="checkbox"/>	Tired all the time
23.	<input type="checkbox"/>	Sick and run down
24.	<input type="checkbox"/>	Headaches and muscle pains
25.	<input type="checkbox"/>	Churning gut
26.	<input type="checkbox"/>	Can't sleep
27.	<input type="checkbox"/>	Poor appetite/weight loss