Part III

Clinical Studies

Chapter 5

DIAGNOSIS OF MELANCHOLIA

5.1 The Clinical Decision Making Problem: Diagnosis of Melancholia

Melancholia is a subtype of depression, which is usually diagnosed by the presence of a set of endogeneity symptoms, such as diurnal variation in mood, sleep disturbance and change in appetite, in addition to symptoms of low mood. The Mood Disorders Unit of the School of Psychiatry at the University of NSW has undertaken a series of studies (Parker et al, 1990; Parker et al, 1994; Parker & Hadzi-Pavlovic, 1996) investigating the hypothesis that melancholia is a disorder of movement as well as of mood. They have also investigated whether the movement disorder symptoms, unlike the traditional endogeneity features used to diagnose melancholia, are specific to the melancholia subtype.

Research efforts have focused upon the development of a clinician-rated behaviourally focussed measure of the presence and severity of the psychomotor disturbance (PMD). The 18-item CORE measure assesses clinical features, which are hypothesised to be the surface manifestations of underlying neuropathological processes. Parker et al [1994] describe CORE-defined PMD as a biological marker implicating likely underlying neurobiological disturbances, which are associated with both PMD and depression.

In developing the CORE instrument, the Mood Research Unit group has investigated the hypothesis that PMD is both necessary and sufficient to the definition of melancholia (Parker et al, 1995a). In the final development study (Parker et al, 1994) Linear Discriminant Function Analysis (LDFA) and Logistic Regression (LR) were used to examine the capacity of the CORE scale and of the traditional endogeneity symptoms to predict 'melancholia' assignment.

In this study, we reanalyse the same data set as Parker et al [1994]. This provides the opportunity to compare the results of analysis with a Multi-layer Perceptron neural network to the previous analyses with linear statistical techniques. Specifically we investigate the hypothesis that a non-linear classification of depressed cases into melancholic and non-melancholic sub-types is more accurate than the linear classification carried out by Parker et al. [1994].

To test this hypothesis, we investigate relationships between three sets of predictor variables and three separate diagnostic criteria for melancholia using both linear and non-linear models. If non-linear models are found to fit the data better than linear models, then the hypothesis is supported.

5.2 METHOD

Subjects

Relevant details are reproduced from Florio, Parker, Austin, Hickie, Mitchell & Wilhelm [1998], which is a published account of the current study. Further details can also be found in the original study: Parker et al, [1994].

"We enrolled a heterogeneous sample of depressed patients, recruiting in-patients and out-patients from a number of Sydney psychiatric hospitals as well as from our tertiary referral Mood Disorders Unit (MDU), subject to patients having a primary clinical diagnosis of a depressive episode present for at least two weeks. Research psychiatrists undertook a comprehensive intake interview, obtaining data generating DSM-III-R (APA, 1987) diagnoses and scores on the Newcastle Index (Carney et al 1965).

Symptom data (coded '0' if absent or '1','2', or '3' if present and of increasing severity) considered in this paper involved the following 17 features held to have some specificity to melancholia: appetite loss, weight loss, slowed thinking, indecisiveness, unpleasant thoughts, slowed physically, suicidal thoughts, loss of interest, anticipatory anhedonia, consumatory anhedonia, non-reactivity to pleasant events, non-reactivity to social support, mood worse in morning, energy worse in the morning, terminal insomnia, non-variable mood, and constipation.

CORE scores were generated by the research psychiatrists, all trained in rating PMD by that strategy. The research psychiatrists were also required to assign an MDU 'clinical diagnosis'. In essence, subjects were assigned a diagnosis of 'endogenous depression' or ED if they had classical features of melancholia (Nelson & Charney, 1981), including significant psychomotor disturbance, vegetative features, pervasive anhedonia and non-reactive mood) as well as absence of delusions and hallucinations. A diagnosis of 'psychotic depression' PD was made if they had such features and delusions and/or hallucinations. Diagnosis of 'neurotic depression' (ND) or 'reactive depression' (RD) required 'classical' melancholic features to be few or absent, with ND requiring evidence of a pre-morbid neurotic style and RD being diagnosed when depression appeared related clearly and principally to a significant antecedent life event. In our analyses we combine PD and ED, as well as ND and RD, and regard the two groups as reflecting melancholic and non-melancholic depression respectively.

Two other diagnostic systems were used to distinguish melancholic and non-melancholic sub-groups: the DSM-III-R system (APA, 1987) with depressed patients with delusions and/or hallucinations being here assigned to the melancholic (vs non-melancholic) group; and the Newcastle Scale (Carney et al, 1965), with a score of 6 or more being the cut-off for melancholia.

Thus we had three estimates of melancholia vs non-melancholia depression (now termed 'Clinical', 'DSM' and 'Newcastle'). In addition we had three sets of predictors: (1) a set of 18 items comprising the CORE scale (Parker et al, 1994), which we will hereafter

refer to as the CORE set; (2) a set of 17 items measuring symptoms held to be over represented in melancholia, also taken from the earlier study Parker et al. (1994), which we will hereafter refer to as the SYMPTOM set; and (3) a set of 35 items comprising the combined 18-item CORE and 17-item SYMPTOM sets, which we will hereafter refer to as the CORE+SYMPTOM set." Florio et al [1998].

The mean age of the 407 patients was 51 years, half being inpatients and with 66% female. Clinical diagnosis allocated 12% as Psychotic Depression, 27% as Endogenous Depression (thus allocating 39% as having clinically diagnosed 'melancholia'), 54% as Neurotic Depression and 7% as Reactive Depression. DSM-III-R criteria assigned 57% as having melancholia. Finally 29% scored 6 or more on the Newcastle Scale and where thus allocated a Newcastle diagnosis of melancholia. The three diagnostic systems thus resulted in three distinctly differing percentages of the sample assigned to 'melancholic' classes.

Analyses

Classification Problems

Nine classification problems were generated by examining three sets of predictor variables (CORE – 18 items, SYMPTOM – 17 items, and CORE+SYMPTOM – 35 items) against the three criteria for melancholia diagnosis (Clinical, DSM and Newcastle).

MLP Neural Networks & Linear Discriminant (LD)

In order to compare an MLP non-linear model with a LD linear model for classification accuracy, eight MLP neural network (with 2 to 9 hidden units) was trained for each of the nine classification problems. In addition we also trained an MLP without a hidden layer and without any hidden units as a Logistic Discriminant (LD), for each of the nine classification problems we studied. Thus, in total 81 individual LD or MLP models were examined.

In summary the design is:

3 diagnostic outcome criteria **(Clinical, DSM and Newcastle)**	**X**	**3 Predictor Item Sets** **(CORE, SYMPTOM & CORE+SYMPTOM)**	**X**	**9 Models** **(LD + MLP2 to MLP9)**	**=**	**81 Models** **9 LD Models** **72 MLP Models**

All models (LDs and MLPs) were trained with QuickProp optimisation, Early Stopping with a 25% holdout, and Weight Decay (-0.01). These technical details are discussed in Chapter 4 and in Appendix 2.

MLP Model Selection

As outlined in Chapter 4, the Akaike Information Criterion (AIC) was used to select one MLP model from amongst eight (MLP2 to MLP9) as the MLP model which will be directly compared to the LD model. AIC values are calculated using data from the training dataset. For each of the nine classification problems, the MLP model with the lowest AIC value was selected as the MLP model to be used in comparison with the LD model.

Measurement of Classification Accuracy

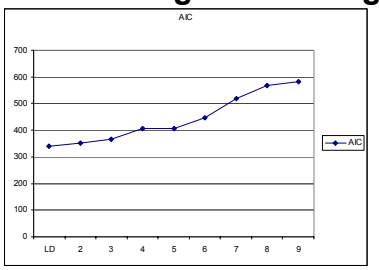
The classification accuracy, of both the LD model and the selected MLP model, was measured using Area Under the ROC Curve ($A_{Z)}$, calculated according to the method outlined by Harrel et al [1984]). In order to obtain a measure of classification accuracy that can be generalised to the entire population from which the sample was drawn, bootstrapping (Efron & Tibshirani, 1993) was used to produce an estimate of A_Z corrected for optimistic bias. One hundred (100) bootstraps were used in each analysis.

For each of the nine classification problems, the A_Z value of the LD model and the selected MLP model are statistically compared using Hanley & McNeil [1983]'s formula for comparing two A_Z values, as outlined in Chapter 4.

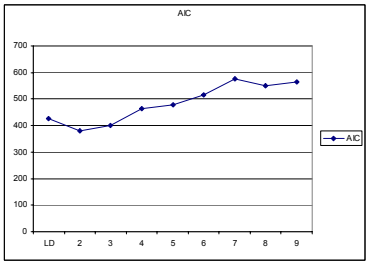
5.3 RESULTS

Model Selection

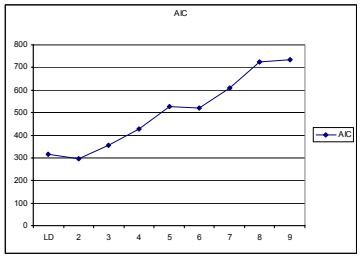
Clinical Diagnosis – Diagnostic Criteria



Core Predictor Set

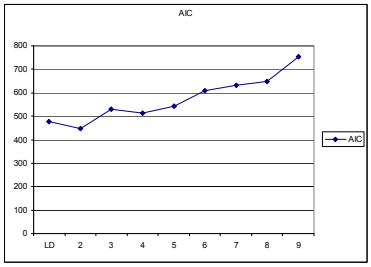


Symptom Predictor Set

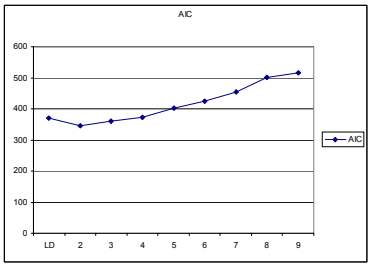


Core+Symptom Predictor Set

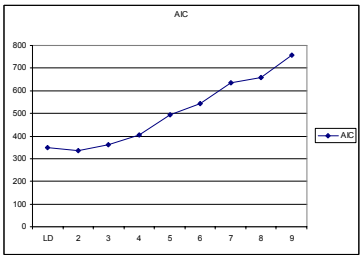
DSM– Diagnostic Criteria



Core Predictor Set

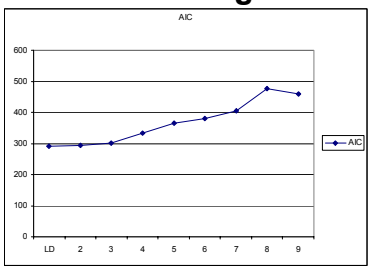


Symptom Predictor Set

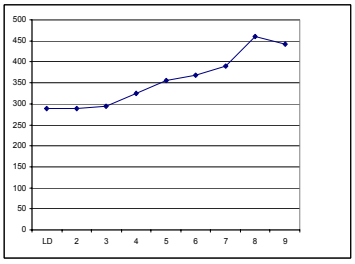


Core+Symptom Predictor Set

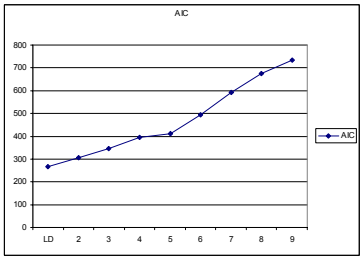
Newcastle– Diagnostic Criteria



Core Predictor Set



Symptom Predictor Set



Core+Symptom Predictor Set

Figure 5.1 AIC values for LD and eight MLP (2 to 9 hidden units) models in each of nine classification problems (3 diagnostic criteria X 3 predictor sets).

In all nine graphs above the MLP2 model is the MLP model with the lowest AIC value amongst the MLP models. Thus for all nine classification problems the MLP2 model will be used for comparison to the LD model.

The results presented in Tables 5.1, 5.2 and 5.3 are Training dataset derived Areas under the ROC Curve (A_Z), corresponding Bootstrap (100 bootstraps) corrected $A_{Z,}$ the standard deviation of the Bootstrap A_Z and Shrinkage, which is an estimate of the optimistic bias contained in the Training Dataset derived value of $A_{Z.}$

The hypothesis under investigation is that the "selected" non-linear model (in all cases MLP-2) will more accurately classify subjects into diagnostic classes than an LD linear model. This hypothesis is tested by comparing the Bootstrap corrected A_Z of the MLP-2 with the Bootstrap corrected A_Z of the LD.

All significance tests for these differences use Hanley & McNeil's [1983] z-test, as outlined in Chapter 4, which takes into account the degree of correlation between two classifiers.

To take account of fact that we are carrying out a large number of significance tests, our criteria for significance of any one difference will be a p value of .01 or less.

Diagnostic Criteria: Clinical Diagnosis

Item Set & Model	Classification Accuracy: Training Dataset A_Z	Bootstrap Corrected A_Z	Std Dev of Bootstrap A_Z	Shrinkage A_Z
CORE (ns)				
LD	.887	.863	.018	.024
MLP 2	.886	.865	.025	.021
SYMPTOM (ns)				
LD	.845	.816	.021	.029
MLP 2	.868	.824	.021	.044
CORE + SYMPTOM (ns)				
LD	.936	.893	.011	.043
MLP 2	.981	.918	.019	.049

Table 5.1 Training Dataset and Bootstrap Corrected A_Z (Area under the ROC Curve) and their standard deviations, for a Logistic Discriminant (LD) and the MLP2 non-linear model (2 hidden units), for three predictor sets: CORE, SYMPTOM and CORE + SYMPTOM combined, using the criterion of a Clinical Diagnosis of Melancholia as the output.
(ns) = difference between LD and MLP2 is not significantly different
(*) = difference between LD and MLP2 is significantly different

By applying significance tests to the differences between the LD and MLP2 models in Table 5.1, we find the following. Firstly, for the CORE item predictor set that assesses PMD, the non-linear model (MLP2) was not significantly more predictive of Clinical Diagnosis than the LD linear model (LD vs MLP2 = .863 vs .865, z = 0.56 p = .290, r_{neg} = .830, r_{pos} = .904). For the SYMPTOM predictor set. The linear solution was also as accurate a classifier as the non-linear solution (LD vs MLP2 = .816 vs .824, z = ..57, p = .283, r_{neg} = .845, r_{pos} = .841). And the same was also true for the CORE + SYMPTOM item set (LD vs MLP2 = .893 vs .918, z = 1.92, p = 0.028, r_{neg} = .764, r_{pos} = .819).

Diagnostic Criterion: DSM Diagnosis

Item Set & Model	Classification Accuracy: Training Dataset A_Z	Bootstrap Corrected A_Z	Std Dev of Bootstrap A_Z	Shrinkage A_Z
CORE (*)				
LD	.800	.763	.023	.037
MLP 2	.846	.801	.036	.045
SYMPTOM (ns)				
LD	.892	.868	.017	.024
MLP 2	.918	.872	.016	.046
CORE + SYMPTOM (ns)				
LD	.926	.886	.011	.040
MLP 2	.957	.911	.011	.046

Table 5.2 Training Dataset and Bootstrap Corrected A_Z (Area under the ROC Curve) and their standard deviations, for a Logistic Discriminant (LD) and the MLP2 non-linear model (2 hidden units), for three predictor sets: CORE, SYMPTOM and CORE + SYMPTOM combined, using the criterion of a DSM Melancholia Diagnosis as the output.
(ns) = difference between LD and MLP2 is not significantly different
(*) = difference between LD and MLP2 is significantly different

The pattern of results for a criterion diagnosis of DSM-III-R Melancholia, are depicted in Table 5.2 above. For the CORE only dataset, the non-linear MLP2 model classified significantly better than the LD model (LD vs MLP2 = .763 vs .801, z = 2.24, p = 0.013, r_{neg} = .776, r_{pos} = .828), but the absolute size of the difference was small. For SYMPTOM only dataset the difference between the LD and MLP2 models was not significantly different (LD vs MLP2 = .868 vs .872, z = 0.27, p = 0.395, r_{neg} = .737, r_{pos} = .793). In respect of the CORE + SYMPTOM dataset, the MLP2 did not classify significantly better than the LD model (LD vs MLP2 = .886 vs .911, z = 1.67, p = 0.047, r_{neg} = .696, r_{pos} = .757).

Diagnostic Criteria: Newcastle Diagnosis

Item Set & Model	Classification Accuracy: Training Dataset A_Z	Bootstrap Corrected A_Z	Std Dev of Bootstrap A_Z	Shrinkage A_Z
CORE (ns)				
LD	.924	.904	.014	.020
MLP 2	.925	.897	.020	.028
SYMPTOM (ns)				
LD	.882	.856	.018	.026
MLP 2	.929	.881	.016	.048
CORE + SYMPTOM (ns)				
LD	.954	.914	.011	.040
MLP 2	.957	.923	.018	.034

Table 5.3 Training Dataset and Bootstrap Corrected A_Z (Area under the ROC Curve) and their standard deviations, for a Logistic Discriminant (LD) and the MLP2 non-linear model (2 hidden units), for three predictor sets: CORE, SYMPTOM and CORE + SYMPTOM combined, using the criterion of a Newcastle Diagnosis of Melancholia as the output.
(ns) = difference between LD and MLP2 is not significantly different
(*) = difference between LD and MLP2 is significantly different

Table 5.3 depicts the results using a Newcastle scale score greater than 6 as a criteria to classify subjects into the Melancholia class. For all three datasets the difference between LD and the non-linear MLP2 model were not significant. For the CORE only dataset (LD vs MLP2 = .904 vs .897, z =0.64, p = 0.262, r_{neg} = ..886, r_{pos} = .822). For the SYMPTOM only dataset (LD vs MLP2 =.856 vs .881, z = 1.63, p = 0.052, r_{neg} = .817, r_{pos} = .711). For the CORE + SYMPTOM dataset (LD vs MLP2 = .914 vs .923, z = 0.72, p = 0.235, r_{neg} = .764, r_{pos} = .739).

Shrinkage

The greater the magnitude of the shrinkage, the less optimal the training, in terms of producing a model which generalises well to future cases (see discussion of the Bias – Variance Trade Off in Chapter 2).

The average shrinkage for the LD models was .031, which was significantly less than the average shrinkage of the MLP-2 model of .040 (paired samples $t = 2.74$, $p = .025$, $df = 8$). This is expected because the sample size was the same for both models, but the MLP2 would have more error due to Variance. This indicates that a larger training dataset sample size would produce more accurate models in both cases but possibly more so with the more complex MLP-2 models.

The absolute values of the shrinkage were small, indicating that though there was overfitting, the degree of overfitting was not large.

5.4 CONCLUSIONS

In one out of nine of the classification problems studied in this chapter the non-linear model generated by an MLP neural networks classified more accurately than linear model generated by a Linear Discriminant. In the remaining eight classification problems the two types of models were found to classify at an equivalent level. However, in the comparison which yielded a significant difference, the magnitude of the difference (less than 4 A_z units) was relatively small.