Prediction accuracy and stability of regression with optimal scaling transformations

Kooij, A.J. van der

Chapter 5

Application of the CATREG-Lasso: Severity of Bulimia Nervosa — Components of the Syndrome and Definition of Therapy Outcome

To identify the most important components of syndrome severity of bulimia nervosa, as well as to identify clinical cases, we explored the relation between dimensional and categorical assessment, by studying the performance of a standard battery of outcome questionnaires in predicting an expert rating of syndrome severity. From a consecutive treatment sample, 213 cases were selected covering the whole range of severity of bulimia. We applied Regression with Optimal Scaling to model nonlinear relations in the data, and the Lasso-method with bootstrapping for predictor selection. The sensitivity and specificity of case classification using the obtained model was determined with ROC curve analysis. We found the best set of predictors to be a combination of an expert rating of binge frequency, and the Eating Disorder Inventory (EDI)-factors ”bulimia” and ”drive for thinness”. The model can effectively predict

This chapter is based on Hartmann A.H., Zeeck, A. & Van der Kooij, A.J. (submitted for publication). Severity of Bulimia Nervosa — Components of the Syndrome and Definition of Therapy Outcome.

being a clinical case at a rate of 88%. The presented statistical methods are innovative and promising approaches that can help researchers to better define sets of outcome variables for treatment studies, reviews or meta-analyses.

5.1. Introduction

Reliable and valid operationalizations of severity of a disorder, as well as thresholds separating pathological states of health, are necessary for the evaluation of treatment outcome. This is a generic problem that also pertains to the evaluation of eating disorders. Clausen (2004a,b) found that recovery rates of outcome studies vary from 24% to 74% and that a great amount of this variance is accounted for by the methods of the assessment (Clausen 2004a) and its time point (Clausen 2004b). In a similar study Olmsted, Kaplan, and Rockert (2005) showed that varying definitions of relapse and remission in bulimia nervosa produce relapse rates ranging from 21% to 55% in the same sample. Obviously, defining outcome of bulimia nervosa is not a trivial problem.

For bulimia nervosa, the DSM-IV (American Psychiatric Association 1994) lists criteria for eating behavior and weight control as well as the abnormal relation between self esteem and weight or shape, but no criterion of severity or amount of impairment. The criterion defining the borderline between pathology and normality (an average of 2 binges per week over 3 months) is not based on empirical investigations, but rather on expert's consensus.

Many patients suffering from bulimia nervosa suffer from more than abnormal eating behavior and weight concern. Very often we find a co-morbidity with depression, with anxiety disorders (Casper 1998; Cooper and Fairburn 1986; Levy, Dixon, and Stern 1989; Tobin and Griffing 1995) and with personality disorders (Cassin and von Ranson 2005), contributing to the severity of overall impairment. Furthermore, the course of bulimia nervosa is often fluctuating with periods of symptom abstinence followed by relapses.

Severity or status of illness or health can be determined categorical or dimensional (Rounsaville et al. 2002). To determine diagnoses is an ostensible simple, categorical solution. But the absence of an intake diagnosis at termination of treatment is not necessarily a sign of good outcome. Patients who suffered from bulimia nervosa may not fulfill the DSM (or ICD) criteria at termination, without good outcome. For example, there are patients who have ceased bingeing completely, but they still vomit after regular meals, keeping their weight low and suffering from cognitive fixations on food or body shape. Of course, these persons could be diagnosed as EDNOS (Eating Disorder Not Otherwise Specified) preventing a misclassification as healthy.

Also, there is a problem with clinically relevant improvement. For a patient having started with many binges and purges a day, a reduction down to two purges a week would be an excellent treatment outcome, while the patient still has to be diagnosed as bulimic. For a patient having started with many binge/purges a day, a reduction down to two purges a week would be an excellent treatment outcome, while according to the DSM criteria the patient still has to be diagnosed as bulimic. The clinical evaluation of "good outcome" (with respect to number of binge-purges) is contradicted by the diagnostic label "still ill".

In psychotherapy research the concept of clinical significance (Jacobson, Follette, and Revenstorf 1984; Jacobson, Roberts, Berns, and McGlinchey 1999; Jacobson and Truax 1991) has been developed to operationalize outcome. Methods of clinical significance provide two measures: The cut-off value "C", drawing a line between healthy and ill on a continuous variable, and the "RCI" (Reliable Change Index), defining the difference between two values over time where the probability of random change is $p < 0.05$. The calculation of C is based on distribution characteristics of samples of ill and healthy persons, determining a cut off where the minimum of misclassifications is to be expected (based on the assumption of normal distributions of the outcome variable in both samples). The calculation of RCI is based on the retest reliability of a measure (sometimes on Crohnbach's alphas) and estimates the minimal non-random difference.

The use of the cut-off value C suffers from the same flaw as diagnostic categories: relevant improvement within the pathological range is not covered. Some authors combine C and RCI, forming a set of six outcome classes: [healthy | ill] × [reliably improved | unchanged | reliably deteriorated]. This is clearly a better solution, but the RCI is a very lenient index for improvement. For example: The subscale Bulimia from the Eating Disorders Inventory (EDI) scores on a range from 0 to 21 (Garner 1991; Meermann, Napierski, and Schulenkorf 1987). For this scale $C = 4.5$ and $RCI = 2.4$, meaning that a change of 2.4 or more points on this scale is regarded as a non-random relevant change. But a change of 3 scoring points may mean something very different, depending on the starting point: Improving from 6 to 3 means to cross the line between clinical and normal ($C = 4.5$), improving from 18 to 15 means reliable improvement but is clinically irrelevant; whereas, improving from 18 to 6 would be reliable and clinically relevant.

The question is how to link clinically relevant categories to the scores of questionnaires or interviews, and how to interpret differences in the light of clinical judgment. Another issue is multi-measure, multi-source outcome measurement. Many authors claim that it is necessary to measure many facets of a disorder and to use multiple sources of information (e.g. Fichter and Quadflieg 1997). This leads to the unfortunate situation where the investigations produce a lot of data, where the relevance of the variables, their independence and the significance levels of comparisons are open to question. In order to avoid the problems of multiple testing, researchers then have to reduce the complexity of their data. Some choose the solution to define one variable as final endpoint — the "rest" of multi-measure outcome data entering exploratory analyses. Others factorize their outcome data set, either using factor scores to define outcome or selecting the highest loading variables.

All these considerations show the complexity of determination of outcome or status (for bulimia nervosa) using categorical, dimensional, multi-variate and multi-source approaches. We would argue that it is most relevant to know which variables contribute to the overall severity of a disorder and how variables are related to severity. This seems to be a necessary condition to determine whether therapy is successful. Empirical investigations on this issue would be valuable for researchers leading their decisions on primary outcome selection as well as for reviewers leading their selection of specific outcome variables from the sets of variables reported in studies.

Our research questions are: Firstly, how can we investigate which facets of the eating disorder and of general psychopathology determine expert rating of overall syndrome severity? In other words: How do we select an optimal subset of variables from a standard battery of outcome questionnaires to predict severity of bulimia nervosa as rated by experts? Secondly, can we empirically determine a cut-off score between health and pathological states? And finally, can we determine the score region of the predictor variables where discrimination is best? In answering these questions we will take into account the non-interval scale of the variables and possible nonlinear relations between them.

5.2. Method

5.2.1 Sample and instruments

We draw our sample from all consecutive admissions of patients suffering from bulimia nervosa (ICD-10) who were seen in our outpatient clinic at the Department for Psychosomatic Medicine and Psychotherapy of the University Medical Center in Freiburg, Germany. Our first selection criteria are being diagnosed with the Structured Interview for Anorexia and Bulimia nervosa, a standardized and comprehensive interview for eating disorders (SIAB-EX, Fichter (1991) and Fichter and Quadflieg (2001)), and having filled out a battery of psychometric tests within 3 weeks before or after the interviews. The aim of the case selection was to obtain a sufficient number of cases for all outcomes.

Between January 1990 and December 2005 a total of 862 patients have been seen in our department. Depending on the therapist's capacity and the funding of research projects, 290 patients were diagnosed with the SIAB-EX, some of them more than once, resulting in 535 available interviews. Sorting out the cases with psychometric data and caring for sufficient frequencies for extremes ("no eating disorder" vs. "most severe eating disorder") resulted in a final data set containing $N = 213$ independent cases. The interviews were obtained mainly from a treatment sample, so there were relatively few people who were considered to show no pathology and who had no need for a treatment at intake. The majority of these cases were found after treatment, at discharge or at follow-up assessment for 3 months to 3 years after termination of therapy (see Table 5.2).

The interviewers were all trained and checked for their reliability. The training procedure included the lecture of the manual and related articles, viewing at least 3 videotapes of experienced interviewers, a rating of one tape and in case of deviations in the trainee's ratings discussion of reasons with an expert; viewing and rating of 8 videotapes of expert interviews, and finally the computation of kappas and intra class correlations between the trainee's ratings and the Gold Standard defined by the experts for the 8 tapes. In case the trainees ratings did not reach a kappa or an ICC$>$ 0.7, the respective variables had to be retrained with a reliable rater.

In our investigation we use the therapists global severity rating of the bulimic eating disorder of the SIAB-EX (Fichter (1991) and Fichter and Quadflieg (2001)). The interviewers give the global severity rating at the end of the interview, which takes one hour on average. The instruction says that the global severity rating shall summarize all symptoms found in the inter-

Table 5.1. Rating of the severiy of the eating disorder (SIAB item no. 84).

Rating	Instructions, meaning
0	No eating disorder
1	Slight eating disorder
2	Marked eating disorder; outpatient treatment sufficient
3	Severe eating disorder; inpatient treatment or intense outpatient treatment necessary
4	Very severe eating disorder, inpatient treatment necessary

Table 5.2. Severity by Time of measurement.

	Ratings of Severity					
	0	1	2	3	4	
Time	None	Slight	Marked	Severe	Very severe	Total
Intake	3	4	35	58	34	134
Post treatment	1	6	4	0	0	11
Follow-up	24	21	17	6	0	68
Total	28	31	56	64	34	213

view. The rating represents an overall clinical impression of the interviewer and may explicitly contain symptoms not only related to the eating disorder, except for isolated obsessions or anxieties. The therapists ranked the severity into five categories according to Fichter, Herpertz, Quadflieg, and Herpertz-Dahlmann (1998) (see Table 5.1). We selected the severity ratings from the interview. There has been a change of SIAB-Version, but it only changed the item number, not the content (old version: item no. 53; latest version: item no. 84).

We explored the relationship of the severity of the disorder to a set of outcome variables containing self report measures SCL90R (Derogatis 1977; Franke 2002) and the EDI-F2 (Garner 1991) and an expert rating of binge frequency from the SIAB-EX. This can be considered as a typical multi source, multi measure data set containing both variables measuring cognitive and behavioral aspects of the disorder as well as general psychopathology. About 63% of the observations were obtained in the context of intake diagnostic procedures; about 32% were obtained during follow-up, the rest at the end of treatment (see Table 5.2). As to be expected, no or mild symptoms are mostly found at follow up, while most of the higher ratings are found at intake. This distribution also mirrors daily practice and period of observation of the interview. The SIABs often ask for a three month period of observation for the items; therefore, it is not recommended to use the SIAB as an outcome measure immediately after treatment, if treatment lasts only a few weeks or months. But this is not the research question here, the time point of measurement (pre, post, or follow-up) is not relevant for the operationalization of severity.

5.2.2 Self and Expert Ratings

We are also interested in the relation of self and expert ratings of severity. The SIAB-S is a questionnaire strictly parallelized with the interview. Since 1999 we use it as a self report instrument for eating disorders. For all interviews later than 1999 we calculated the correlation of the experts' ratings of severity and the self report of "impairment caused by the eating disorder" as rated by the patients. Both items are highly correlated ($r = 0.78$, $rho = 0.77$, $N = 68$), supporting our decision to use the larger sample of experts ratings.

5.2.3 Statistical Analysis

We choose to use CATREG, which is a method for regression with (unordered or ordered) categorical variables using optimal scaling (Gifi 1990; Van der Kooij and Meulman 2004). The method is implemented in the program CATREG in the Categories module of SPSS (Meulman et al. 1999, 2004). Using CATREG we neither need to assume that the variables are interval scaled, nor that they are linearly related, nor have normality of residuals. The CATREG model is the classical linear regression model applied to transformed variables. The variables are transformed by replacing categories with optimal values, called category quantifications, using the optimal scaling methodology (Gifi 1990). The transformed variables have all numeric properties of interval variables. The regression coefficients and the quantifications are estimated simultaneously in an iterative process.

In the CATREG approach all variables are considered as categorical, regardless of their measurement level. Each value of a variable can be considered as a category; the categories being labels for nominal variables, rank numbers for ordinal variables, and values for continuous variables. Thus, a continuous variable for example is considered as a categorical variable with N categories (values). In the optimal scaling process, properties of the observed variables are respected to a varying degree, depending on the optimal scaling level. We distinguish the property of containing grouping information (nominal variables), grouping and ordering information (ordinal variables), and grouping, ordering, and interval information (numeric variables).

The researcher has to specify an optimal scaling level for each variable. This decision can be based upon the measurement level of the variable, but need not be. For example, applying numeric scaling level to a continuous variable will respect all numeric properties of the "category" values, thus all information in the observed variable is retained in the transformed variable. But the researcher might not be interested in interval information and only want to retain the rank order. Then ordinal scaling level is appropriate; with this scaling level the interval information is ignored and only grouping and ordering information is retained in the quantified variable. Nominal scaling level also ignores the ordering information and thus only retains grouping information in the quantifications. When applying numeric scaling level to all variables CATREG is equal to standard linear regression. When the researcher wants to explore nonlinear relations in the data, non-numeric scaling levels should be chosen, allowing for nonlinear transformations (monotonic for ordinal scaling level, non-monotonic for nominal scaling level). If variables are nonlinearly related, nonlinear transformations will result in a higher R-squared (proportion of variance explained) than can be obtained with linear regression.

The target variable, severity of syndrome, is clearly rank ordered. All the predictor variables (see Table 5.3: the EDI factors, T-GSI, and Freq-Binges) were treated as rank ordered variables too, although from test theory they might be considered to have interval character. Treating them as rank ordered allows the CATREG analysis to tell from the data if this holds. For rank ordered variables with a small number of categories CATREG provides the ordinal scaling level; for variables with a large number of categories the monotonic spline scaling level is more suited. With both these scaling levels the transformation is a monotonically nonincreasing function; the ordinal scaling level results in a step function (degrees of freedom equal to the number of categories that have received different quantified values), while the transformation using monotonic spline scaling shows a smooth curve and is more restrictive if the number of categories is large (degrees of freedom equal to number of interior knots and degree of the spline). In our analysis we applied ordinal scaling level to severity of syndrome and the predictors Freq-Binge and T-GSI (see table3), and monotonic spline scaling level with two interior knots and two degrees to the EDI predictors.

For selecting a subset of stable predictors that well predict severity of syndrome, we used the Lasso (Least Absolute Shrinkage and Selection Operator; Tibshirani (1996)), which was recently incorporated into the CATREG method (Van der Kooij and Meulman 2006b). The Lasso combines the improved prediction accuracy of Ridge regression in the presence of multicolinearity (Hoerl and Kennard 1970b,a) with the better interpretability of subset selection. Ridge regression improves prediction accuracy by reducing the variability of the estimates of the regression coefficients by shrinking them. The shrinkage is accomplished by adding a penalty term to the regression model, penalizing the sum of the squared regression coefficients. With ridge regression the coefficients are shrunken towards zero, but never become exactly zero, thus all predictors remain in the model. Subset selection on the other hand

Table 5.3. Patients characteristics.

Variable	Range[1]	M	(SD)
Age at intake	16 – 54	26.02	(6.56)
Age at onset (self report)	3 – 34	16.96	(4.09)
BMI at intake	17.51 – 35.70	21.88	(3.22)
EDI 1 Drive for thinness	0 – 21	9.90	(6.15)
EDI 2 Bulimia	0 – 21	7.47	(5.57)
EDI 3 Body Dissatisfaction	0 – 27	15.15	(8.42)
EDI 4 Ineffectiveness	0 – 28	9.30	(6.91)
EDI 5 Perfectionism	0 – 18	5.98	(4.47)
EDI 6 Interpersonal Distrust	0 – 17	6.10	(3.81)
EDI 7 Interocept. Awareness	0 – 30	9.30	(7.70)
EDI 8 Maturity Fears	0 – 23	4.72	(3.30)
SCL GSI T-Score	34.4 – 144.4	73.80	(24.60)
SCL GSI raw score	0.01 – 2.84	1.05	(0.66)
	Category label	N	(%)
Gender	Male	6	(2.8)
	Female	207	(97.2)
SCL GSI T-Score-Classes (further referred to as: T-GSI)	T $\leq$ 60	73	(35.1)
	60 $<$ T $\leq$ 70	28	(13.5)
	70 $<$ T $\leq$ 80	36	(17.3)
	80 $<$ T $\leq$ 90	21	(10.1)
	90 $<$ T	50	(24.0)
Frequency of binge eating (rated by interviewer based on overall evidence, further referred to as: Freq-Binges)	0 No marked binge eating	58	(27.2)
	1 Rarely	18	(8.5)
	2 Occasionally (average at least twice weekly)	40	(18.8)
	3 Frequently (up to once per day)	46	(21.6)
	4 Very frequently (more than once per day)	51	(23.9)

[1] The ranges of most of the variables cover the full span of values, which is due to including patients in the sample with no pathology to most severe pathology.

provides sparse models but does not reduce the estimation variability of the coefficients. The Lasso combines sparse models and improved prediction accuracy by shrinking the coefficients towards zero and some of them to exactly zero. The Lasso applies a penalty to the sum of the absolute values of the coefficients. With a penalty value of zero, the model including all predictors (unshrunken) is obtained and with a high penalty value all predictors are shrunken to zero. Penalty values in between zero and some high value result in models with different shrunken coefficients with some of them shrunken to zero. Thus, different values of the penalty result in different models.

The optimal value of the penalty can be determined using resampling techniques like cross validation or the bootstrap. With resampling techniques the error in predicting future observations is estimated (expected prediction error). The expected prediction error is a corrected estimate of the prediction error for the data set at hand (the apparent expected prediction error), which is usually too optimistic. We used the .632 bootstrap (Efron 1983, for details of using the .632 bootstrap procedure with CATREG see Van der Kooij and Meulman (2006a)), which is essentially a smoothed version of leave-one-out cross validation to estimate expected prediction error (and its standard error). The .632 bootstrap procedure has to be applied to each Lasso model(starting with a penalty value of zero and increasing the penalty value in small steps until all predictors are shrunken to zero). To select the optimal Lasso model usually the one-standard-error rule is applied: select the most parsimonious Lasso model within one standard error of the Lasso model with minimum expected prediction error.

For sake of simplicity we did not want to use a penalized model; we only used the Lasso in an explorative way to select a subset of predictors, which were then analyzed with CATREG without shrinkage. We selected predictors according to the Lasso results for models with one predictor to the model including all predictors. The expected prediction error for each unshrunken model was estimated with the .632 bootstrap, and the most parsimonious model with expected prediction error within one standard error of the model with minimum expected prediction error was selected. In the CATREG-Lasso analysis we excluded the 10 observations that have a missing value on one or more predictors, so $N = 203$. For the unshrunken models with a subset of predictors we only excluded the observations with a missing value on the selected predictors.

CATREG saves the predicted values (for each observation the sum over transformed variables times their regression coefficient) as a variable and we used this prediction variable as discriminant scores to evaluate the discriminative power of the model for classification into health and illness. For this purpose we divided the severity ratings into 0–1 (no and slight) and 2–4 (marked through very severe). Cut-off values for clinical caseness are determined and there sensitivity (SENS; proportion of true positives) and specificity (SPEC; proportion of true negatives) are determined by Receiver Operating Curve analyses (ROC; Altman and Bland (1994)). ROC analysis is commonly used in research on medical diagnostic testing, where a parameter (e.g. creatinine) serves as an indicator for pathology (e.g. myocardial infarction). ROC curves show how sensitivity and specificity of a classification vary with different cut-off values for the test variable.

5.3. Results

5.3.1 Model selection

In Figure 5.1 the CATREG-Lasso paths are displayed. On the right hand side of the plot ($s = 1$, no penalty) we see the regression coefficients for the model including all predictors unshrunken, and on the left ($s = 0$, high penalty) all predictors are shrunken out. The vertical line ($s = .51$) represents the shrunken model selected with the .632 bootstrap applying the one-standard-error rule. Four predictors are included in this model: EDI-F2-Bulimia, Freq-Binges, EDI-F1-Drive for thinness, and EDI-7-Interoceptive Awareness. The expected prediction error for this model is 0.386 (apparent error 0.335).

Figure 5.2 displays the error estimates with their standard errors for unshrunken models with number of predictors from one to ten and predictors entering according to the CATREG-Lasso analysis. The model with the minimum error estimate is the four-predictor model. Applying the one-standard-error rule results in selecting the three-predictor model. The error results for these two models are given in Table 5.4. Thus, the best subset of predictors (in terms of expected prediction accuracy and parsimony) of symptom severity contains the frequency of binges and the EDI factors EDI-F2 “Bulimia” and EDI-F1 “Slimness Ideal”. The apparent proportion of explained variance (R-squared; is 1− apparent error) is 0.67 ($F = 45.3; df = 9; p < 0.0001$). Table 5.5 displays the regression coefficients of the predictors and their Relative Importance (Pratt 1987).

The regression with optimal scaling procedure shows that there are nonlinear relations between the independent variables and severity rating of eating disorder, indicated by the nonlinear curves in the transformation plots (Figure 5.3), especially for EDI-F1 and EDI-F2. The categories “No” and “Slight” of the severity rating variable have negative quantified values and the “Marked”, “Severe”, and “Very severe” categories have positive quantifications. As in

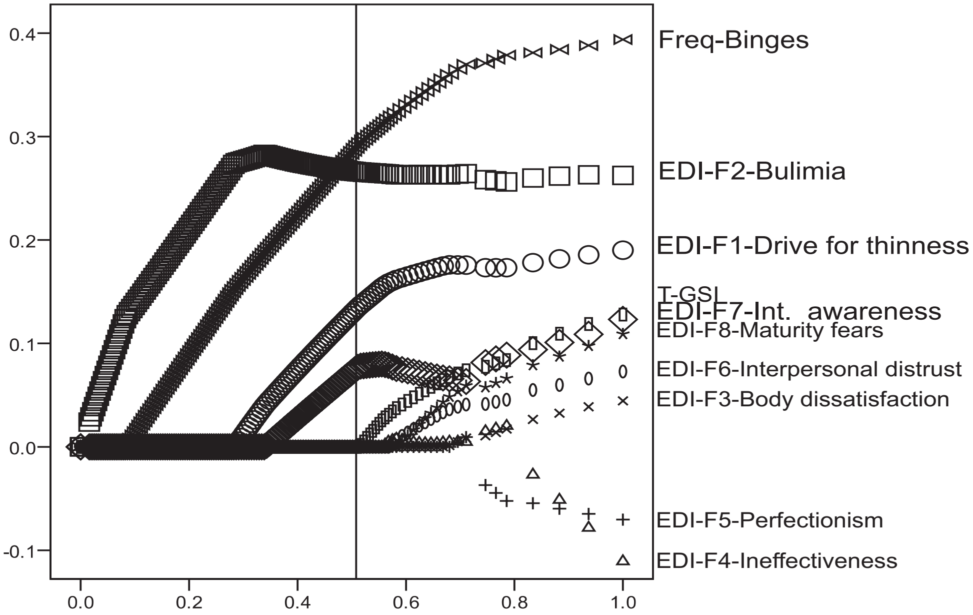

Figure 5.1. CATREG-Lasso paths. The vertical axis represents the regression coefficients β. The horizontal axis represents $s = \sum_{j=1}^{J}|\beta_j^{l}/\sum_{j=1}^{J}|\beta_j^{us}|$, the sum of the absolute values of the $\beta's$ at a particular location on the horizontal axis divided by this sum for the full unshrunken (us) model. The vertical line represents the optimal model selected with the $.632$ bootstrap.

standard linear regression, the contribution of a predictor variable to the predicted value for a person is the person's value on the predictor variable multiplied with the regression coefficient; with CATREG the value on the predictor variable is not the score on the original variable, but the quantification of that score. So, the categories of the predictor variables with negative quantified values (below the "Zero line" in the transformation plot) contribute negatively to the prediction and thus are related to "no" and "slight" eating disorder, while the categories with positive quantifications contribute positively and thus are related to "marked", "severe", and "very severe" eating disorder.

For the expert rating of binge frequency the categories 0 (no) and 1 (slight) get negative quantification values that are markedly different, meaning that they are to a varying degree related to health. The categories 2 to 4 (marked to very severe binge eating) do all indicate pathology, but patients in categories 2 to 4 can not clearly be separated into the eating disorder categories 2, 3, and 4, as is indicated by the small differences between the positive quantifications. So, the binge rating variable distinguishes 3 groups: persons without

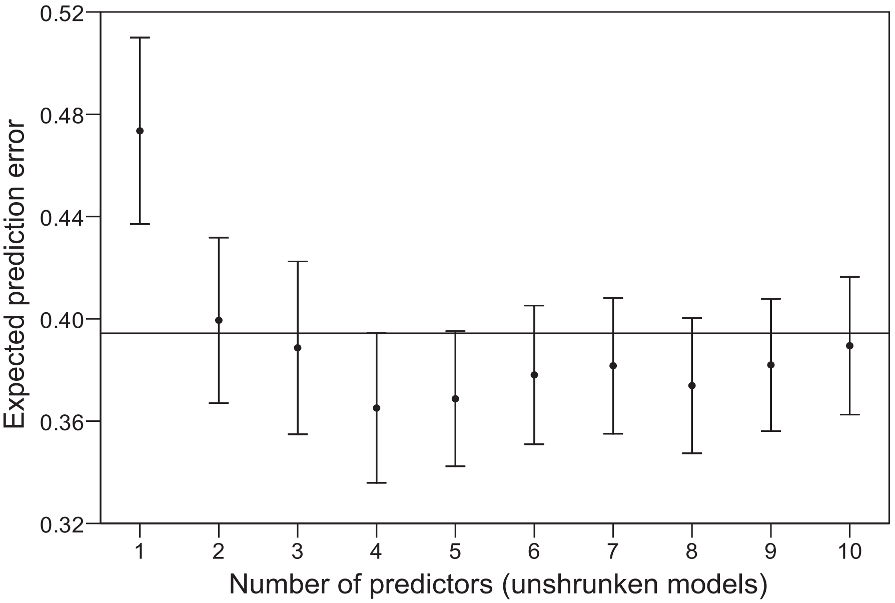

Figure 5.2. Expected prediction error estimated with the .632 bootstrap for unshrunken models, with predictors as suggested by CATREG-Lasso. The bars around the error points represent the standard error. The horizontal line represents the mininum error estimate (at p = 4) plus one standard error.

Table 5.4. Error results for two candidate models.

	Exp. prediction error (Std. Error)	App. prediction error	N
model 1			
EDI-F2-Bulimia, Freq-Binges, EDI-F1-Drive for thinness, EDI-7-Interoceptive Awareness	0.365 (0.0292)	0.303	210
model 2			
EDI-F2-Bulimia, Freq-Binges, EDI-F1-Drive for thinness	0.389 (0.0338)	0.330	211

Table 5.5. Regression coefficients for the selected model.

	Beta	SE	df	F	Sig	Importance
Freq-Binges	0.355	0.052	3	47.437	0.000	0.365
EDI-F1 Drive for thinness	0.225	0.058	3	14.982	0.000	0.221
EDI-F2 Bulimia	0.370	0.065	3	32.155	0.000	0.415

binge eating, with slight binge eating, and with marked to very severe binge eating. In the transformation plots for EDI-F1 "Drive for thinness" and EDI-F2 "Bulimia" we see the same trend: For EDI-F1 "Drive for thinness" the transformation curve starts to level off at category 6, and at category 5 for EDI-F2 "Bulimia", thus the lower categories (related to health) are well separated from each other and from the higher categories (related to pathology), but within the pathology range there is little discrimination. The findings with respect to distinguishing health from pathological states are confirmed if we apply a clinical significance measure: the cut-off value between normal and pathological population C; for EDI-F1 "Drive for thinness" C=6 and for EDI-F2 "Bulimia" C=4.5.

5.3.2 Discrimination between health and pathology

CATREG with nominal scaling level for the dependent variable is equivalent to Discriminant Analysis. Because the severity rating is ordinally related to the predictors, the nominal quantifications are equal to the ordinal quantifications. Thus, the predicted values are discriminant scores, that are optimal to discriminate between the five severity classes. To evaluate how well these scores discriminate cases in the two health classes from cases in the three pathological classes, we used them as the test variable in an ROC curve analysis, showing very good separation of cases (AUC = 0.93, SE = 0.019, CI95 [0.892|0.968]). For a classification with maximum sensitivity (SENS; true positive rate) and efficiency (EFF; overall correct classification rate) we chose 1.50 as the cut-off value for the severity rating recoded to a two class variable with values 1 and 2, which corresponds to a cut-off value of −0.49 for standardized two-class severity (because CATREG standardizes the transformed variables, the predicted values estimate the standardized dependent variable). We also used the ROC curve analysis results to identify the cut-off value for maximum specificity (SPEC; true negative rate) and the cut-off value that maximizes both sensitivity and specificity. The results are displayed in Table 5.6, also displaying the predicitive value of the positive test (PVP), the

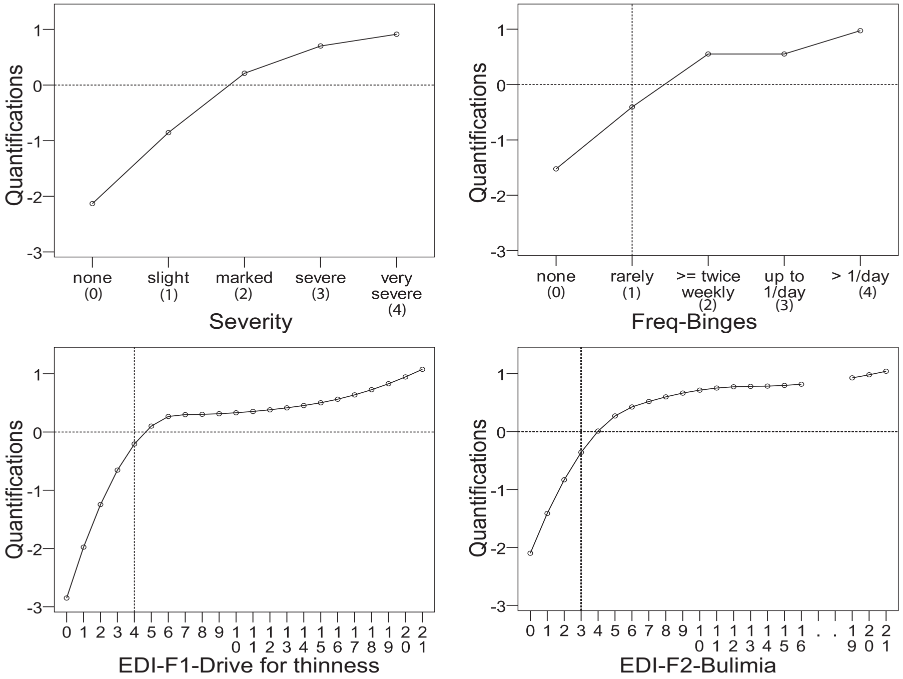

Figure 5.3. Transformations Severity and predictors in selected model.

predicitive value of the negative test (PVN), and Cohen's Kappa, a measure of agreement.

5.4. Discussion

Severity of disorder

From our results we conclude that the best operationalization of syndrome severity of bulimia nervosa combines a behavioural measure of binge frequency and measures of cognitive-emotional aspects of the disorder, as measured by the scales "Bulimia" and "Drive for thinness" of the EDI. Drive for thinness is an important predictor of outcome in bulimia, as patients aiming to reduce weight stay in the circulus vitiosus of restricted eating and craving for food. To include a measure of cognitive occupation with thoughts on eating clinically makes sense, as patients are impaired by fixation on thoughts on eating as by eating behaviour itself. Surprisingly, and in contrast to previous findings (Cooper and Fairburn 1986; Nollett and Button 2005), the variable covering general psychopathology (T-GSI) did not remain in the final model, which

Table 5.6. Classification into clinical and healthy/subclinical cases.

	Predicted							
Observed	Clinical	Healthy/ Subclin.	SENS	SPEC	EFF	PVP	PVN	Kappa
Max. SENS			0.95	0.70	0.88	0.89	0.84	0.68
Clinical	144	8						
Healthy/Subclin.	18	41						
Max. SPEC			0.70	0.95	0.77	0.97	0.55	0.54
Clinical	107	45						
Healthy/Subclin.	3	56						
Max. SENS & SPEC			0.83	0.83	0.83	0.93	0.65	0.61
Clinical	126	26						
Healthy/Subclin.	10	49						

may be due to multicollinearity.

We recommend that the measurement of syndrome severity should include both the behavioural and the cognitive/emotional aspects of the disorder. A severity index can be computed with the regression model presented above.

Discrimination between health and pathology

Using results of regression with optimal scaling it is possible to separate remitted and sub clinical cases from marked and severe cases with a high efficieny (EFF=0.88). With this classification quality the model might be used as a tool for the support of clinical decision making. The labels of the overall severity rating include indication for psychotherapy, thus the expert rating also implied determination of being a clinical case with consequences for further clinical action. Clinical action is very much dependent on the national / local health care system, for example, very severe bulimia can only lead to inpatient treatment in health care systems providing this treatment option; on the other hand, clinicians from many cultures might well consent to our classification of two classes: “to treat or not to treat”.

Depending on the aim of clinical decision making, either high sensitivity or high specificity may be important. For example, a high specificity would allow for an economic selection of pathological cases, where only a few false positives might get an unnecessary treatment. On the other hand a high sensitivity is necessary to correctly identify the clinical cases, thus not withholding treatment from patients who need it, at the cost of more false positives. We showed that our results allow for either high sensitivity with acceptable specificity or high specificity with acceptable sensitivity.

Discrimination within the range of pathology

Accurate prediction of severity ranking within the range of pathology would be useful for decision making in stepped care. The predictors included in our model show lower accuracy in distinguishing the pathological severity ratings (2 to 4) from each other. This was also observed when including other and/or more predictors in the model. So, for the purpose of discrimination within the pathological range, we would recommend to add another set of variables, adding further aspects of pathology like presence or severity of personality disorder, chronicity, parasuicidal behaviour, anxiety or depression.

Regression with Optimal Scaling

The results of the regression with optimal scaling clearly show nonlinear (monotone) relationships between the predictors and the severity rating of bulimia nervosa. With linear regression we obtained roughly the same results for the binary classification into health and illness. But without further investigation using other analyses methods, the lack of discriminative power in the pathological range of especially EDI-F1 and EDI-F2, implying that persons with less severe pathological ratings of eating disorder can not be well distinguished from persons with more severe ratings, is not revealed. Also, the good discrimination between none and slight bulimia nervosa is obscured by the linear requirement. Using regression with optimal scaling, the discriminative characteristics of the predictors become immediately clear by inspecting the transformations plots. Furthermore, by loosening the restriction of linear relations to monotonic relations, the apparent explained variance with our model increased from 58.2% to 67.0%. Therefore, CATREG is more suited for our data than linear regression, and we prefer CATREG above standard Nonlinear Regression because of the freedom in modelling nonlinear relations (Nonlinear Regression requires a predefined mathematical function).

Lasso

The Lasso provides a solution for the problem of selecting a sparse set of predictors with good prediction accuracy from a collection of intercorrelated predictors, where prediction accuracy pertains to the prediction of future observations. Stepwise regression methods, like forward and backward selection, do not provide models with stable estimates of the regression coefficients (among other problems (Judd and McClelland 1989)), and are therefore not preferred. Implementation of the Lasso and the .632 bootstrap in CATREG is scheduled for SPSS Categories version 16. The Lasso for linear regression is available in the LARS software for R and S-PLUS and can be downloaded from Tibshirani's Lasso page (http://www-stat.stanford.edu/~tibs/lasso.html).

Answers to the research questions

Using regression with optimal scaling to find nonlinear transformations, the Lasso to select a sparse model with stable predictors, and the .632 bootstrap to assess the prediction accuracy, we identify a subset of three variables that predicts severity rating of bulimia nervosa best (namely: the frequency of binges and the EDI-factors 1 "drive for thinness" and 2 "bulimia"). Determining cut-offs between clinical severity and normal range, we showed that the predicted values could be used for case or outcome classification with high sensitivity and specificity. We also showed that the lower score regions of the predictors do a better job in discrimination than the higher score regions. Especially the higher scores of the EDI factors do not discriminate patients with marked, severe, and very severe bulimia nervosa well. It is open to further investigation to replicate or challenge the results with other instruments or other samples.

Authors' note

The authors express their appreciation to the patients who invested much time in completing questionnaires, and to the therapists of the ward and the day clinic for their continuous support of this research. We are grateful as well to Elvira Bozkaya for help in monitoring the data, and to Thomas Herzog who helped to establish some important foundations for this work.

Leiden University holds the copyright of the procedures in the SPSS Package Categories, and the Department of Data Theory receives the royalties.