A chance for change : building an outcome monitoring feedback system for outpatient mental health care

Jong, K. de

Author: Jong, Kim de
Title: A chance for change : building an outcome monitoring feedback system for outpatient mental health care
Date: 2012-04-17

Chapter 4

Risk models for negative treatment outcomes in psychiatric outpatients: predicting end state functioning and rate of change using classification and regression trees (CART) and multilevel modeling

De Jong, K., Nugter, M.A., Ninaber, C., Lutz, W.,
Van Ginkel, J.R., Heiser, W.J. & Spinhoven, P.

Manuscript submitted for publication.

Abstract

Objective. Risk models that aimed to identify consistent predictors for negative outcomes have encountered several challenges. This study uses state of the art statistical techniques to handle these problems, by using multilevel analysis combined with multiple imputation to predict the rate of change and classification and regression tree (CART) analysis to predict end state functioning.

Method. A naturalistic sample of 1540 outpatients (63% female; age range 17-67 years, M = 37.5, SD = 11.7) was collected in three mental health care organizations in the Netherlands. Patients completed the Outcome Questionnaire (OQ-45; Lambert et al., 2004) regularly during treatment. In addition, several potential predictor variables were collected.

Results. Initial severity, educational level, expectancies, Global Assessment of Functioning (GAF) and the working alliance were significant predictors for end state functioning. In predicting rate of change, the same predictors were found, except for educational level and expectancies. In addition, previous treatment, comorbidity and having a personality disorder as main diagnosis were significant predictors for rate of change as well.

Conclusions. Although there was overlap in predictors of negative outcomes with regard to end state functioning and rate of change as outcome variables, both analyses provide different information. By combining the prediction models, patients that may need to be monitored more closely during treatment can be identified so that negative outcomes may be prevented. By using CART and multilevel analysis combined with multiple imputation substantially more data could be used in analysis than would otherwise have been the case, thus reducing the selection bias and improving generalization.

Introduction

Risk models that aim to predict the future course and outcome of disease processes are common in medical and health research. Good risk models are valuable for a wide variety of purposes, including policy making, adjusting for differences in patient case-mix between institutions, and assisting patients and clinicians to make informed decisions about treatment (Ambler, Omar & Royston, 2007). In medical situations, knowing the risk factors for negative treatment outcomes might mean the difference between life and death. Although psychotherapy is usually not involved in life and death situations, negative treatment outcomes can have a large impact on patients' quality of life and potentially constitute an increased risk of long-term psychiatric complaints and higher costs for mental health care. Hansen, Lambert and Forman (2002) have shown that in naturalistic settings a lack of treatment success is common: 3-14% of patients deteriorate and 45-60% show no clinically significant change during treatment. By comparison, in clinical trials results are much better: 67% of patients improve significantly. Similar results were found by Barkham et al. (2008), who showed that approximately 18% more patients were clinically significantly improved and effect sizes were more than twice as large in randomized trials than in practice-based studies. This suggests that there is much room for improvement in clinical practice. Although not all patients are likely to achieve treatment success and some might be on a progressive decline that cannot be stopped, evidence suggests that at least some patients might worsen as a result of therapy (Lambert & Ogles, 2004). Therefore, knowing risk factors for negative treatment outcomes in psychotherapy can be very valuable for treatment selection.

In psychotherapy research, risk models are best known from the line of research that is referred to as *patient-focused research* and health services research such as quality assurance programs. Such research aims to prevent negative treatment outcomes by making predictions about the patient's progress using variables that have previously been associated with positive or negative treatment outcomes (e.g. Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lueger et al., 2001; Lutz, 2003). These models usually attempt to forecast the course of treatment or rate of change, rather than end state functioning, and are often referred to as expected treatment recovery (ETR) curves. Most models use multilevel modeling or similar techniques to make these predictions. Although these models are very valuable, it has been a challenge to identify reliable predictors of change beyond initial severity. Some authors have found additional predictors, including patients' expectancies of outcome and the Global Assessment of Functioning score, but in these studies initial severity still explained the highest proportion of outcome variance, compared with all other predictors (Lutz et al., 2005; Lutz, Lowry, Kopta, Einstein, & Howard, 2001; Lutz, Martinovich, & Howard, 1999).

Moreover, most factors have not been consistently replicated by others. For outcome in terms of functioning at the end of treatment, initial severity is an important predictor too. In the review by Clarkin & Levy (2004), a high initial severity of symptoms was related with poor treatment outcomes, especially in depression and addiction populations. The results on other client variables were less consistent: comorbid personality disorders for instance could have a positive or negative effect on outcome and client demographics were usually not consistently found as relevant predictors (Clarkin & Levy, 2004).

One reason for not finding many predictors for end state functioning and rate of change might be that we oversimplify the relationship between the predictor variable and outcome. Most prediction models suppose a linear relationship between the predictor and outcome, whereas the true relation might be much more complex. Kendler (2008) states that we need to start looking at more complex interactions in explanatory models for psychiatric illness – models that consider predictors at different levels: micro and macro, within and outside the individual – to better understand what risk factors are relevant in psychiatry. Assuming simple linear relationships between a predictor and outcome may be misleading. Take for example a predictor like the working alliance. The working alliance is one of the more robust predictors of outcome that has been identified (e.g. Horvath & Symonds, 1991; Klein et al., 2003; Martin, Garske, & Davis, 2000). It is often assumed that a strong alliance leads to better outcomes, yet not all patients with a high working alliance have good outcomes. Suppose that the true relationship would only hold for those patients with average to low severity of symptoms and not for patients with a high level of symptom distress. Standard regression models, would only find working alliance to be a predictor and might miss the interaction.

Another challenge for developing accurate risk models is ensuring sufficiently complete data. Most risk models for health outcomes are estimated using data routinely collected in clinical practice. The advantage of that approach is that there is usually a large dataset available, with high external validity. The disadvantage is that those datasets typically have many missing values. A review of more than 800 articles published in three leading personality journals showed that almost half of the articles reported missing data problems (Van Ginkel, Sijtsma, van der Ark, & Vermunt, 2010). It is not unusual for some important predictors to be missing for over 50% of patients (e.g. Ambler et al., 2005). Moreover, in many studies patients have missing data on several variables simultaneously. Combinations of missing data on different predictor variables may result in percentages of complete data as low as 32% (Ambler, Omar, & Royston, 2007). Although missing values are common, information on the percentage of missing data and how missing data are handled statistically is often not reported. A survey of 46 papers in a counseling psychology journal showed that only a little over one third of the papers reported missing data percentages (Schlomer, Bauman, & Card, 2010). A review of 100 articles using longitudinal data in three leading journals in developmental psychology showed that 82% of the studies that did report having missing data, used missing data techniques that are problematic (Jelicic, Phelps, & Lerner, 2009). In addition, anecdotes from other psychotherapy researchers tell us that it is not uncommon for more than half of the data to be "cleaned up" prior to final analysis (see Hatfield, McCullough, Frantz, & Krieger, 2010 for an example) although this is seldom reported in the published articles. Complete case analysis can result in substantially smaller sample sizes, biased regression coefficients and reduced reliability for predicting future observations as a result (Ambler, Omar & Royston, 2007). Moreover, researchers have demonstrated that serious violations of statistical assumptions occur when missing data are ignored (e.g. Allison, 2003; Graham & Hofer, 2000; Wothke, 2000).

Several statistical techniques have been developed that handle missing data problems well, including repeated measures multilevel analysis (Hedeker & Gibbons, 2006; Raudenbush & Bryk, 2002; Singer & Willet, 2003) and classification and regression trees (CART; Breiman, 1984). Multilevel analysis handles missing data on the dependent variable by estimating an individual change trajectory that depends on the observed variables for each person and as a result can handle missing data on the dependent variable very well, but it does need complete values on predictor variables. CART is a regression based datadriven technique that automatically searches for moderators in the data and calculates optimal splits for predictor variables. The missing data are handled by the use of surrogate splits. These type of split points are created as backup for the main split, meaning that if the split variable being evaluated is missing, a surrogate predictor is used (Breiman, 1984). Contrary to most techniques it can handle missings on the predictor variables very well, but needs complete cases on the dependent variable. CART has the additional advantage of being able to select important predictors automatically, especially when there are many variables (categorical and/or interval), and has the ability to uncover complex interaction effects between them. Furthermore, the method is robust for non-normally distributed data (Briand, Ducharme, Parache, & Mercat-Rommens, 2009). The resulting trees are easy interpretable by lay persons and can be used to create decision rules, which can be generalized to a non-research setting (Lewis, 2000). CART and multilevel analysis address different aspects of outcome: multilevel analysis assesses factors that are related to the rate of change, whereas CART assesses factors related to end of treatment functioning.

Another approach to handling missing data is multiple imputation. Multiple imputation can be used to address missing data in both independent and dependent variables and restores the dataset to the full size. Many authors have illustrated that multiple imputation gives better results in statistical analyses than listwise deletion (e.g. Ambler, et al., 2007; Little & Rubin, 2002; Rubin, 1987; Schafer, 1997; Schlomer, et al., 2010). It decreases estimation bias in both the parameters and the standard errors and when the imputation model is correct and the data are missing at random, the precision of the parameter estimates will approach the precision that would have been achieved with complete data (Huang et al., 2009). However, clinical researchers seem to be reluctant to use it (Jelicic, et al., 2009). Although multiple imputation is a very useful method to deal with missing data and has been rapidly developed for many types of data, it still has trouble handling unbalanced designs like those that are common in naturalistic settings in clinical psychology. For example, patients' progress is often measured multiple times during the course of treatment, yet the frequency of sessions and assessments may differ considerately among patients - one patient might have 5 sessions of treatment with one measurement missing whereas another patient has 15 sessions with 3 measurements missing. Even with complete data the number of measurements would differ between patients. In the currently available software packages, the full data matrix is filled up, rather than just the true missing values. In this situation, multiple imputation can be used to impute the predictor variables, but not the outcome variable.

In summary, developing prediction model for negative treatment outcomes is important, since many patients in clinical practice do not experience sufficient change. Yet, handling the missing data that are typical in the naturalistic data used for these models poses a challenge. Applying techniques that are more flexible in handling missing data seem a good solution, but these methods have their limitations as well. In the current article, we will use a naturalistic dataset collected in three community outpatient facilities in the Netherlands to predict patients at risk for negative treatment outcomes. Data on the predictor variables will be multiply imputed and multilevel analysis will be used to predict factors related to the rate of change. Factors directly related to negative outcomes will be assessed using CART on the original data. The predictor variables in the multilevel analysis and CART analysis will be identical and include clinical variables (e.g. initial severity, diagnosis, duration of complaints), demographic variables (e.g. sex, age, educational level) and process variables (expectancies, working alliance). Differences in the results and the performance of both models will be compared.

Method

Participants

Data were collected between June 2006 and June 2009 in eight treatment departments of three mental health care organizations in the Netherlands. Subjects were outpatients who were seen for psychological or psychiatric treatment. Data were collected as part of routine care, but patients were offered the option to refuse participation. The research proposal was evaluated by the local ethical committees of the participating institutions. There were 4447 patients in the original sample, including 1689 males (38%) , 2752 females (62%) and 6 persons with unknown gender. The age of the patients ranged from 17 to 71, with a mean of 37.7 years (*SD* = 11.8). Patients with psychotic disorders, mental retardation or in a current crisis at the time of referral (*n* = 92), patients who received non-verbal treatments (*n* = 151), patients who did not have a sufficient level of understanding of Dutch (*n* = 121) and patients who did not receive more than one treatment session (*n* = 148) or were unable to participate in the study for other reasons (*n* = 91) were excluded from the study. A total of 503 patients actively refused to participate in the study, and an additional 1801 patients completed less than two questionnaires during treatment - 773 of which never completed a single questionnaire. The remaining sample of 1540 consisted of 561 men (36%) and 976 women (63%) and 3 people of unknown gender. The age of the subjects ranged from 17 to 67 with a mean age of 37.5 (*SD* = 11.7) (see Table 1).

A subset of 541 patients was used for the complete case analysis in the multilevel model, to demonstrate the effect of the multiple imputation on the analysis. Since only 19% of patients (*n* = 295) had complete values on all predictor variables, an optimal set of predictor variables was selected in such a way that the most relevant predictor variables could be included, without losing too many other cases (see Table 2).

Table 1 shows the baseline characteristics of the sample that met inclusion criteria, the final selection and the complete case sample. There were no significant differences found between the samples in sex, age, main diagnosis for treatment following the *Diagnostic and Statistical Manual of Mental Disorders IV* (DSM-IV) and Outcome Questionnaire – 45 (OQ-45; Lambert et al., 2004) scores at intake. The most common main diagnosis in our sample was mood disorder, followed by anxiety disorder and adjustment disorder.

Instruments

Outcome Questionnaire-45 item version (OQ-45)

The Dutch translation of the Outcome Questionnaire-45 item version (OQ-45; Lambert et al., 2004) was used to measure patient progress during treatment. The OQ-45 is

Table 1 Patient characteristics of the sample meeting inclusion criteria, the sample selected for main analyses and sample selected for the complete case analysis

		Sample meeting inclusion criteria (n = 3844)	*Selected for main analyses (n = 1540)*	*Complete case analysis (n = 541)*
Sex				
-	Male	1419 (37%)	561 (36%)	197 (36%)
-	Female	2419 (63%)	976 (63%)	344 (64%)
-	Unknown	6 (0.2%)	3 (0.2%)	
Age		*M* = 37.6	*M* = 37.5	*M* = 38.5
		SD = 11.8	*SD* = 11.7	*SD* = 11.6
Main DSM-IV diagnosis				
-	Mood disorder	1149 (30%)	466 (30%)	170 (31%)
-	Anxiety disorder	658 (17%)	302 (20%)	106 (20%)
-	Adjustment disorder	580 (15%)	269 (18%)	93 (17%)
-	Disorders usually first diagnosed in childhood	215 (6%)	78 (5%)	24 (4%)
-	Personality disorder (Axis II)	236 (6%)	102 (6%)	33 (6%)
-	Cognitive disorder	136 (4%)	71 (5%)	47 (9%)
-	Substance related disorders	86 (2%)	19 (1%)	8 (2%)
-	Other	486 (13%)	193 (13%)	59 (11%)
-	No DSM-IV Axis I or II diagnosis	8 (0.2%)	6 (0.4%)	1 (0.2%)
-	Unknown or missing	290 (8%)	34 (2%)	-
OQ-45 intake score		M = 77.5[a]	M = 77.4 [b]	M = 76.6
		SD = 23.9	SD = 23.0	SD = 23.0

Note. DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, 4th Edition; OQ-45 = Outcome Questionnaire-45 item version
[a] n = 2452, [b] n = 1419

a self-report instrument and has 45 items, 9 of which are reversed, asking how the respondent has felt over the last week on a 5 point rating scale, ranging from 0 (never) to 4 (almost always). The OQ-45 consists of three subscales that assess different domains of client functioning: Symptom Distress, Interpersonal Relations and Social Role. The Symptom Distress domain consists of 25 items relating to psychological symptoms that are common in highly prevalent mental disorders. The Interpersonal Relations domain consists of 9 items that assess the functioning of the patient in interpersonal relationships, and the Social Role domain assesses the patients functioning in social roles, such as work and school. The Dutch OQ-45 has satisfactory psychometric properties. The internal consistency for the Total score ranges between 0.92 and 0.96 in university, community, patients and community and patients combined samples. For the subscales the international consistency is 0.90-0.95 for the Symptom Distress (SD) scale, 0.74-0.84 for the Interpersonal Relations (IR) subscale and 0.53-0.72 for the Social Role (SR) subscale (De Jong, Nugter, Lambert, & Burlingame, 2009).

Working Alliance Inventory (WAI)

The therapeutic relationship between therapist and patient was assessed using the Working Alliance Inventory Client's Form (WAI; Horvath & Greenberg, 1989). The Dutch version of the WAI is called the Werkalliantie Vragenlijst (WAV; Vervaeke & Vertommen, 1996) and consists of 36 items that are scored on a 5-point rating scale, ranging from 1 (never) to 5 (always). An example of an item from the WAI is 'I believe ____ (therapist's name) likes me'. The WAI has three subscales that consist of 12 items each: The Bond, Task and Goal subscales. The Bond subscale assesses the therapeutic bond between the patient and therapist, the Goal subscale measures the level of agreement in therapy goals between patient and therapist, and the Task subscale assesses the level of agreement between therapist and patient on who has to do what in the treatment. The Dutch version of the WAI had internal consistencies for the Bond, Task and Goal subscales of 0.85, 0.88 and 0.88 respectively (Vervaeke & Vertommen, 1993).

Treatment Credibility Questionnaire (TCQ)

The patient's expectations of the treatment results were measured with a questionnaire derived from the Treatment Credibility Questionnaire (TCQ; Borkovec & Nau, 1972). Our version was based on the adaptation by Addis et al. (2004), with one additional item ("How much improvement in your symptoms do you think will occur") from the new version of the TCQ, the Credibility Expectancy Questionnaire (Devilly & Borkovec, 2000). The TCQ version that was used in this study consisted of 7 items that are scored on a 7-point rating scale, ranging from 1 (not at all) to 7 (extremely). The TCQ consists of two factors, Expectancies and Credibility. The Expectancies subscale consists of 4 items and assesses patients' expectations of therapy outcome. The Credibility subscale consists of 3 items and measures the degree to which the patient thinks the therapy is credible. The version that was used in this study has an internal consistency of 0.89 for the Expectancies subscale and 0.84 for the Credibility subscale (Hüpscher, 2007).

Patient characteristics

A variety of patient characteristics were extracted from the electronic registration systems of the participating mental health care institutions. Patient characteristics included demographic variables such as age, gender, education, and clinical variables such as DSM-IV diagnosis and prior treatment (see Table 2 for the full list). The information that was retrievable differed between the participating institutions. In addition, information from intake forms and electronic patient files was used in order to complete missing data when possible.

Procedure

Patients were informed about the study through a letter, before the intake. Participation in the study was on a voluntary basis and patients could object to participate by filling out the enclosed rejection form, but were automatically included if they did not reject. Patients were asked to report to the reception desk fifteen minutes before their treatment session started. The receptionist provided them with an OQ-45. The therapist handed out the TCQ and WAI to the patient after session 2. The patient could hand in the completed forms the next session at the reception desk. Patients were asked to complete the OQ-45 at intake, the first five sessions of treatment and subsequently every fifth session of treatment (the tenth, fifteenth, etc.), for a maximum period of one year. If patients were still in treatment after one year, the final measure was administered at that time; this applied to 445 patients (n = 1494; 30%).

Analysis

Missing Data

In the final sample, only 295 (19%) of the 1540 patients had completely observed responses on all the relevant predictor variables. The amount of missing values ranged between 0 and 20, with a mean of 5.3 (SD = 3.8). Missing data on the predictor variables were handled using Multiple Imputation using Chained Equations (MICE; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). In general, multiple imputation works as follows: Missing data are estimated multiple times using one or more statistical models. The procedure then generates plausible random values for the missing data that resemble the observed data as much as possible. By estimating the missing data multiple times several complete versions of the incomplete dataset are created, which are analyzed by standard statistical procedures. The results of these analyses are then pooled into one final analysis. Multiple imputation was carried out using PASW 17.0 (SPSS, 2009).

The MICE procedure uses linear regression to estimate the missing values on continuous variables, using the other variables as predictors, and logistic regression to estimate the missing values on categorical variables. Van Buuren, Boshuizen & Knook (1999) recommend using a maximum of 25 variables. Thus, a selection of variables was used which were necessary for carrying out the statistical analyses of interest. The variables that were used in the imputation procedure are given in Table 2. Based on recommendations by Graham (Graham, 2009; Graham, Olchowski, & Gilreath, 2007) data were imputed 20 times and the results of the 20 imputed databases were combined using Rubin's (1987) rules for multiple imputation. PASW 17.0 (SPSS, 2008) automatically combines the results of multiply imputed datasets into one pooled analysis.

Definition of negative outcome

The OQ-45 was used to define treatment outcome, based on the score at the end of treatment, or, if that score was missing, the last available score for that patient. Negative treatment outcome was defined as a patient scoring in the clinical range (55 or higher on the OQ-45) at the end of treatment and having experienced *no change* (less than 14 points decrease in score on the OQ-45) or *deterioration* (more than 14 points increase in score) according to the criteria for reliable change within the concept of clinical significance by Jacobson & Truax (1991). This definition is slightly different from that used by others. Usually negative outcomes are defined as deterioration or no change, regardless of the level of functioning at the end of treatment. We think that people who are functioning in the normal range at the end of treatment should not be considered as having experienced a negative outcome, so we excluded this group from the negative outcomes group. Based on the definition that we used, 727 (49%) patients in our sample had negative treatment outcomes, of which 130 patients deteriorated and 597 showed no (reliable) change.

Classification And Regression Trees (CART)

The Classification And Regression Trees (CART) method to acquire a valid model can be broken down in a couple of steps. The CART algorithm searches through the predictor values for a split point. This way the data are split up in subsets, also called nodes, and the homogeneity of the outcome variable in the new subgroups has increased. The best split is selected, based on the criterion that there is an optimal division of the outcome variable. This partitioning of data is then repeated for the new (sub)sets. In other words, we start with the full dataset, which is split up based on an optimal cut point value on a particular predictor and then the same procedure is applied to the new sets, also called child nodes. Repeating this process many times will lead to a situation where the number of subjects in a child node becomes one or they all have the same outcome. In case of the latter, the partitioning process can be stopped. It can be reasoned that the large model constructed may be poorly generalizable, in other words an overfitted model, because it is tailored for a particular dataset. To overcome this problem Breiman (1984) proposed a regularization method, which aims to find the optimal trade-off between the size of a tree and its predictive power. We start with a very large tree, which has overfitted the data. Each branch causes some amount of homogeneity gain in their end nodes. But this reduction needs to be viewed in relation to the size of the branch, for example a branch early in the tree probable leads to a larger reduction than one almost at the end. This ratio, often referred to as the cost-complexity ratio, is used to select the branch which benefits the model the least and pruned (removed) out of the tree. This pruning procedure is then sequentially repeated and the optimal size of the final tree is determined based on cross-validation.

Table 2 Predictor variables prior to multiple imputation (n = 1540) and correlations with change and the last available OQ-45 score (end).

	r (change)	*r* (end)	% missing	In complete case analysis	Values
Sex	-0.03	0.06	0%	Yes	See Table 1
Age	-0.04	0.04	0%	Yes	See Table 1
Main DSM-IV diagnosis category			2%	Yes	See Table 1
- Mood disorder	-0.12	0.25			
- Anxiety disorder	0.06	-0.11			
- Adjustment disorder	-0.05	-0.04			
- Disorders usually first diagnosed in childhood	0.04	-0.05			
- Personality disorder	0.06	0.03			
- Cognitive disorder	0.05	0.02			
- Substance related disorders	0.01	0.08			
- Other	0.02	-0.16			
Educational level	-0.06	-0.03	36%	No	31% Low, 32% middle, 33% high, 4% other
Having a paid job	-0.02	-0.15	32%	No	64% Yes
Previous treatment	0.04	0.12	31%	No	55% Yes
Using psychiatric medication at intake*	-0.03	0.23	34%	No	49% Yes
Duration of complaints	0.05	0.01	34%	No	52% Longer than one year
Diagnosis on Axis I	-0.05	0.07	2%	Yes	98% Yes
Comorbidity on Axis I	0.01	0.18	2%	Yes	34% Yes
Diagnosis on Axis II	0.03	0.10	2%	Yes	17% Yes
Comorbidity Axis I and II	0.02	0.12	2%	Yes	16% Yes
Problems with primary support group	-0.05	0.11	13%	No	46% Yes
Problems related to social environment	0.01	0.06	13%	No	18% Yes
Occupational problems	-0.05	0.10	13%	No	33% Yes
Legal problems / crime	0.00	-0.09	13%	No	1% Yes
GAF score (Axis V)	0.01	-0.22	6%	Yes	M=59.6, SD=8.6
OQ-45 SD subscale	-0.33	0.89	8%	Yes	M=48.3, SD=15.3
OQ-45 IR subscale	-0.22	0.71	8%	Yes	M=16.1, SD=6.5
OQ-45 SR subscale	-0.26	0.64	10%	Yes	M=13.0, SD=5.2
TCQ Expectancies subscale	-0.05	-0.13	47%	Yes	M=21.2, SD=4.2
TCQ Credibility subscale	-0.03	-0.11	47%	Yes	M=15.1, SD=3.2
WAI Bond subscale	-0.03	-0.15	56%	Yes	M=49.6, SD=6.4
WAI Task subscale	-0.02	-0.24	56%	Yes	M=48.7, SD=6.8
WAI Goal subscale	-0.02	-0.23	56%	Yes	M=48.1, SD=6.7

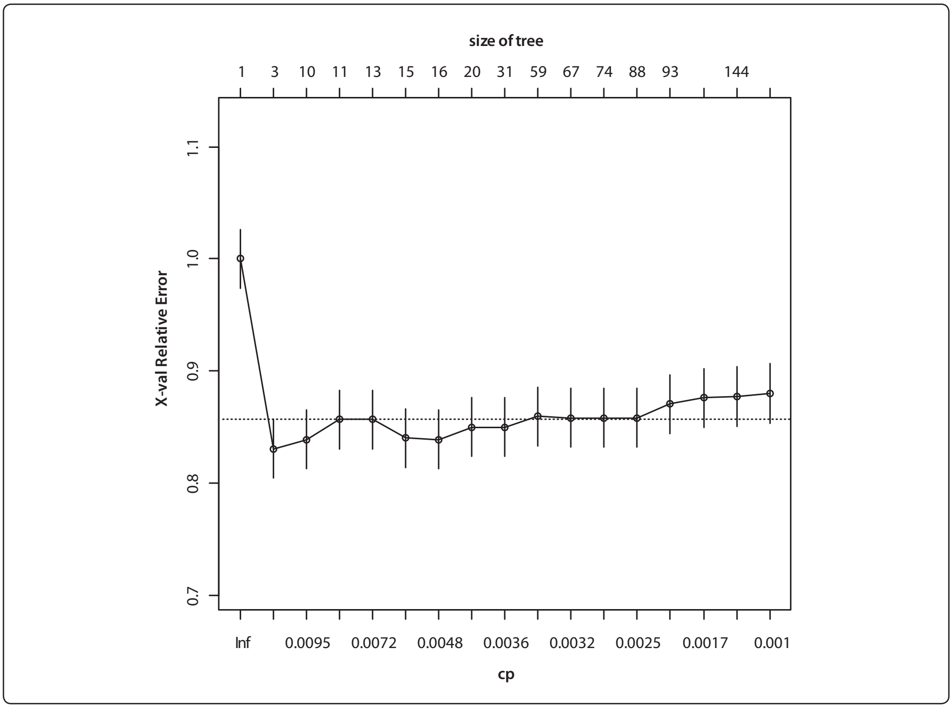
Note. Correlations between the predictor variables and change and the last available measurement of the OQ-45 (end) are pooled estimates, based on the imputed dataset. * Psychiatric medication had already been prescribed by the general practitioner for these patients

For the building of the initial models based on the CART method, the R package Rpart (Therneau & Atkinson, 1997) was chosen, because it closely follows the work and propositions of Breiman (1984). To acquire a large tree, which could later on be pruned back, the following model parameters were used. The maximum depth possible for a tree was 30 layers. The minimum number of cases in a terminal node was set to 10 cases. The complexity parameter was set to 0.001. The Gini splitting criterion was used because it in general is more favorable (Breiman, 1984). The optimal classification threshold was set at 0.50, based on ROC curve analyses for the two final models.

Multilevel Analysis

Two multilevel analyses were performed, on the complete case sample and on the multiply imputed datasets. Both models were two-level random intercept random slope multilevel models, using maximum likelihood estimation and with an unstructured covariance structure. The analyses were performed in PASW 17.0 (SPSS, 2009). Time was set as the logarithm of the session number. A backwards procedure was applied, starting with a full model and removing non-significant predictors (based on the Wald test for fixed effects) one by one until a parsimonious model was reached that was not significantly worse than the full model.

Figure 1 Cost-complexity versus relative error per tree size (number of nodes)



Results

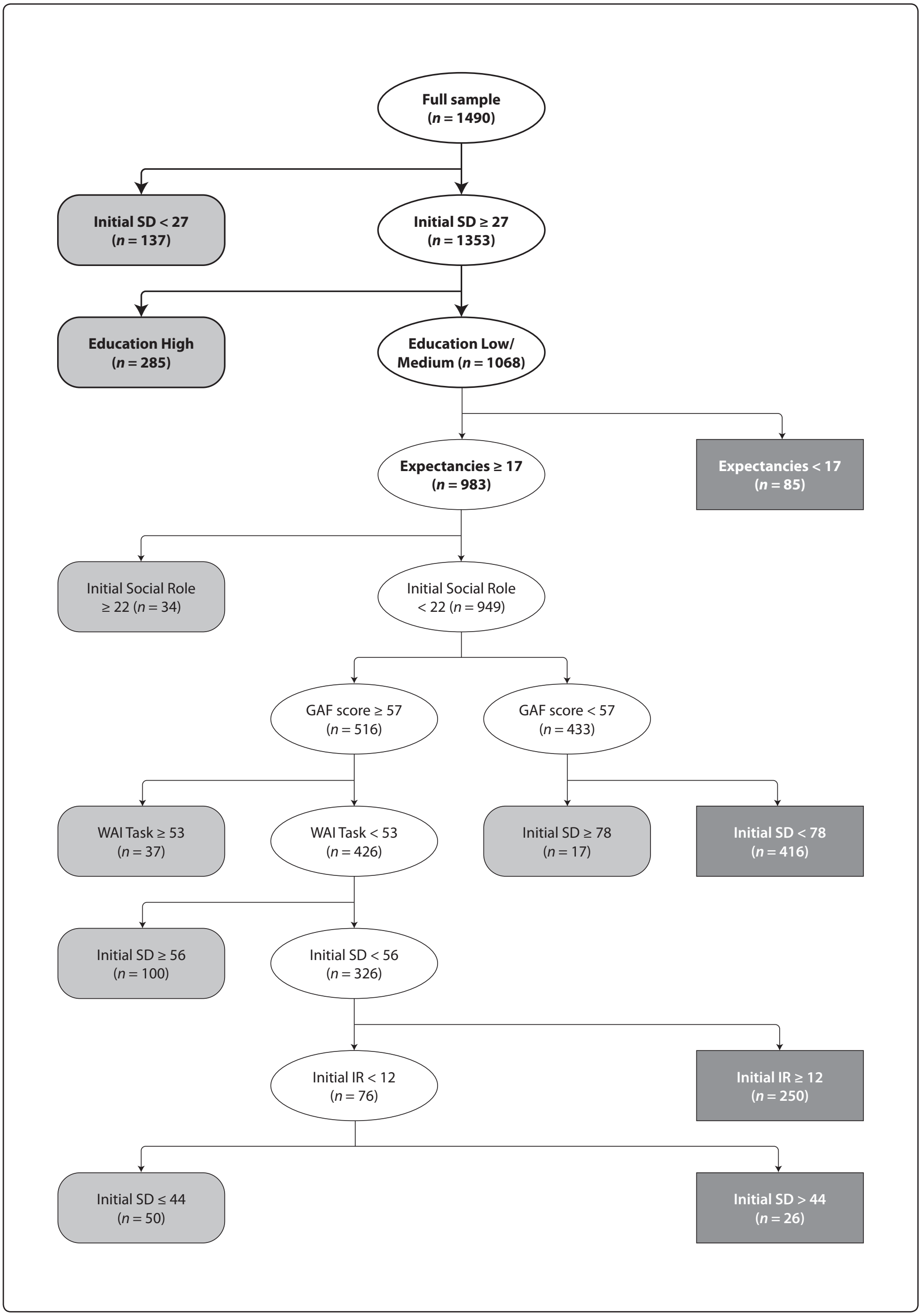
Predicting negative outcomes at end of treatment

A classification and regression tree was modelled to predict negative treatment outcomes. First, a tree that is too large was constructed. The pruning step followed by using the cross-validated error rates for different cost-complexities (see Figure 1). Even though a cross validation of 10-fold is a generally accepted as a satisfactory amount, we applied 250-fold to attain a more stabilized model. Having acquired our estimations, we selected the complexity value that was within the 1 *SE* range of the lowest error rate (three nodes). With this parameter set, the tree was pruned upward and the final model was acquired. In addition to the three nodes model, a second tree with ten nodes that also fitted the cost-complexity criterion and had a similar fit was selected (the second tree under the dotted line in Figure 1). The second model should be considered an exploratory model and was used to study more complex interactions between predictors. Figure 2 shows both models, nested within each other (the three node model in bold). As can be seen in Figure 2, the first model shows that patients who have low pre-treatment scores on the SD subscale of the OQ-45 have favorable outcomes, as do patients who have a high level of education (bachelor degree or higher). Patients with a high score on the SD subscale and a low or medium level of education are most at risk for negative treatment outcomes. It also shows that patients with low expectations of treatment outcomes – given higher initial symptom severity and lower education – have an increased risk for negative outcomes. Model 2 further explores the risk factors for negative outcomes. Patients with higher expectancies, but with more problems on social role functioning on the other hand have more favorable outcomes. From here on the model becomes more complex: Given low social role problems, high expectancies, low to medium educational level, a medium to high symptom severity, in patients with a initial GAF score below 57 (poor overall functioning), but a score on the initial SD subscale between 27 and 78 negative outcomes are more likely. Those patients who have an initial GAF above 57 and a good working alliance (WAI above 53) are predicted to have favorable outcomes, whereas a low working alliance, combined with an initial severity on the SD scale of 27-55 and high initial problems on interpersonal relationships are predicted to have negative outcomes. That is also true for patients with an initial severity on the SD scale between 44 and 55 and low initial problems on interpersonal relationships.

Predicting rate of change

Variables predicting the rate of change were investigated by using a two level multilevel analysis, with the repeated measures within patients at level 1 and differences

Figure 2 Nested CART models for negative treatment outcomes



Note: The two classification trees are nested. The smallest model is in bold, the extended model consists of all the branches that are shown. Negative outcomes are dark grey and square, positive outcomes light grey and square with rounded corners.

Table 3 The unconditional growth model and final model predicting the rate of change for the complete case analysis sample (*n* = 541) and the imputed data (*n* = 540)

		Complete case analysis	Imputed data		
	Parameter	*Estimate (SE)*	*Estimate (SE)*	*Fraction of missing information*	*Relative increase variance*
Fixed effects					
Initial status	Intercept	78.37 *** (1.03)	79.40 (0.62)	0.002	0.00
Rate of change	Intercept	-21.51 * (9.64)	-15.52 (9.42)	0.577	1.29
	Initial SD	0.80 *** (0.07)	0.38 *** (0.08)	0.631	1.61
	Initial IR	0.72 *** (0.15)	0.44 *** (0.12)	0.244	0.32
	Initial SR	0.76 *** (0.18)	0.31 * (0.14)	0.231	0.29
	GAF score	-0.25 *** (0.10)	-0.20 ** (0.07)	0.238	0.30
	WAI Task scale	-0.71 *** (0.13)			
	WAI Goal scale		-0.41 (0.12)	0.492	0.92
	Mood disorder	-3.88 * (1.83)			
	Adjustment disorder	-5.70 * (2.29)			
	Comorbidity on Axis 1		2.61 * (1.21)	0.090	0.10
	Previous treatment		3.63 ** (1.25)	0.202	0.25
	Personality disorder		4.65 * (2.22)	0.076	0.08
Variance components					
Level 1	Within-person	94.71 (3.46)	94.23 (2.41)	0.002	0.00
Level 2	Intercept	481.00 (34.69)	467.38 (21.13)	0.004	0.00
	Covariance	-550.29 (86.95)	-296.35 (42.72)	0.331	0.48
	Slope	878.10 (190.27)	467.97 (66.57)	0.339	0.50
Goodness of fit	Deviance	20900	49929		
	AIC	20926	49956		

Note: Time is modeled as the 10log of the session number. *p<0.05 *** p<0.001
Due to different sample sizes, the deviance and AIC values of the two analyses cannot be compared directly.

between patients at level 2. First, the complete case sample was analyzed, followed by the imputed data set. The final models for both analyses are presented in Table 3. As can been seen in Table 3, some predictor variables were significant in both models, but some differences could be observed as well. For pre-treatment scores on the SD, IR and SR subscales of the OQ-45 higher scores slow down the rate of change. A higher GAF score, indicating better functioning, is positively related to the rate of change. The working alliance Task scale has a positive relationship with the rate of change in the complete case sample, whereas the Goal scale is positively related to the rate of change in the imputed data set. The complete case sample emphazises the relationship with Axis I disorders, showing a more positive rate of change for patients that have mood disorders or adjustment disorders as their main diagnosis. The imputed dataset shows a stronger emphasis on the complexity of the presented problems, demonstrating that having comorbid Axis I disorders, having had previous treatment for psychological complaints and having a personality disorder as main diagnosis slow down the rate of change. Since the imputed data set most likely provides the most valid base for prediction, further analyses and interpretations were only performed for this model.

Model performance

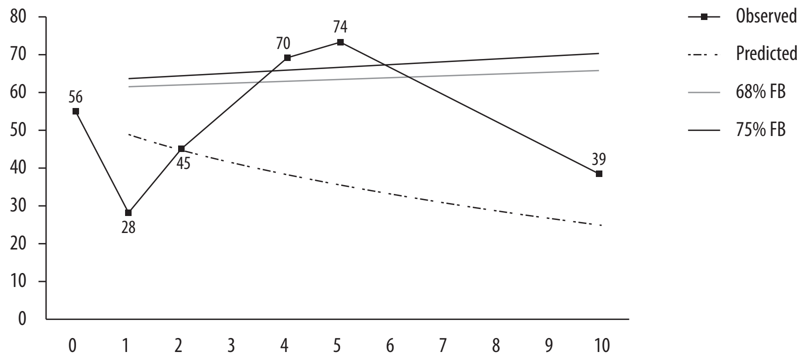
To test model performance, the sensitivity, specificity and the positive and negative predicted values were computed for both CART models (see Table 4). Model 1 had better performance in sensitivity (0.82), which means that most cases with negative outcomes are picked up, but had poor specificity (0.39), so many people that would be detected by the model as having a high risk of negative outcomes do not actually

Table 4 Model performance for the CART models and multilevel model with failure boundaries (n=1540)

	Negative outcome				Deterioration	
	CART model 1	*CART model 2*	*ML 68% FB*	*ML 75% FB*	*ML 68% FB*	*ML 75% FB*
Sensitivity	0.82 [0.80-0.84]	0.68 [0.66-0.70]	0.40 [0.38-0.42]	0.34 [0.32-0.36]	0.89 [0.87-0.91]	0.85 [0.83-0.87]
Specificity	0.39 [0.37-0.41]	0.63 [0.61-0.65]	0.78 [0.76-0.80]	0.82 [0.80-0.84]	0.75 [0.73-0.77]	0.80 [0.78-0.82]
Positive predicted value	0.55 [0.53-0.57]	0.64 [0.62-0.66]	0.62 [0.60-0.64]	0.64 [0.62-0.66]	0.25 [0.23-0.27]	0.29 [0.27-0.31]
Negative predicted value	0.70 [0.68-0.72]	0.67 [0.65-0.69]	0.58 [0.56-0.60]	0.57 [0.55-0.59]	0.99 [0.97-1.00]	0.98 [0.96-1.00]

Note: CART= Classification and Regression Tree; FB = Failure bound; ML = Multilevel analysis.
The 95% confidence intervals for the values are reported between brackets.

Figure 3 Example of the multilevel model for an individual patient



Note. FB = Failure bound

have negative outcomes. Model 2 has a better balance between sensitivity (0.68) and specificity (0.63), although both values were lower than ideally would be the case.

For the multilevel model based on the multiply imputed data sets, two confidence interval based failure bounds were computed in order to determine if deviation from the predicted treatment course (based on the multilevel model), would result in a higher risk for negative outcomes. If the failure bound was crossed the patient was considered to be at risk for negative outcomes (see Figure 3). As can be seen in Table 4, both the 68% and 75% failure bound had reasonably good specificity values (0.78 and 0.82 respectively), meaning that patients who do not cross the interval have a good chance of having positive treatment outcomes. However, specificity is poor (0.40 and 0.34), so patients at risk for negative outcomes are not picked up very well. Since we defined negative outcomes as still scoring in the clinical range and either no change or deterioration, secondary analyses were performed to test how well the multilevel model performs in identifying patients at risk for deterioration. After all, for some patients, (reliable) change may not be feasible. The multilevel model does predict deterioration well. The 68% failure bound has somewhat better sensitivity (0.89) but the 75% failure band had better specificity (0.80) while maintaining satisfying sensitivity (0.85).

Discussion

In this article, we aimed to predict risk factors for negative treatment outcomes at the end of treatment using CART and for rate of change using multilevel modeling (combined with multiple imputation). Fifty-one per cent of the patients in our sample improved and had scores on the OQ-45 outside the clinical range at the end of treatment. In the CART analyses we found that patients with relatively low pre-treatment scores for symptom distress, and patients with high education and positive expectancies have a better chance of favorable outcomes. The extended model showed the complexity of the relation between predictors and outcome and showed how pre-treatment expectancies, social role problems and GAF scores and the working alliance at the beginning of treatment (Task subscale) interacted in different ways to predict negative outcomes at the end of treatment. The multilevel analyses showed that initial severity, the working alliance (Task or Goal subscale) and GAF score were significant predictors for the rate of change in patients. In the complete case sample, having a mood or adjustment disorder as main diagnosis had a positive relationship with the rate of change, whereas in the imputed sample previous treatment, having comorbid Axis I disorders and having a personality disorder as main diagnosis had a negative relationship with the rate of change. The model based on the multiply imputed data was considered the most reliable model, and further analyses were computed only for this model. The CART models and multilevel models differed in their sensitivity to detect negative outcomes. The first CART model had high sensitivity, but low specificity, whereas the multilevel model had high specificity and low sensitivity. The multilevel model was good at picking up deterioration, but not at identifying the no change group. The extended CART model had the best balance between sensitivity and specificity. Although sensitivity and specificity values were not as high as we are used to for instance in diagnostic tests (e.g. Gilbody, Richards, Brealey, & Hewitt, 2007), in predicting outcomes, those values are quite good, considering that there are many variables that influence the course and outcome of therapy and that most of the predictors in our study were measured pre-treatment. Our results match earlier multilevel prediction models (Lutz et al., 1999; 2001), in which the GAF score, expectancies and initial severity were found to be significant predictors, except we found some additional predictors to be significant and did not find an effect of expectancies in the multilevel analyses.

The factors that influence outcome and rate of change are in part overlapping, but not exactly the same. Expectancies and educational level were relatively strong predictors in the CART models, but not significant for predicting outcome. Having had prior treatment and having a personality disorder as main diagnosis was predictive for the rate of change, but not for outcome. Although they are highly related – people that progress fast, usually have a better chance of favorable outcomes – there are differences too. For instance, patients who have low initial severity (in the non-clinical range) in terms of symptom distress, usually have a low rate of change, but a better chance of positive outcome according to our definition. And patients who start with high initial severity and have an average progress may still have poor outcomes. We have defined negative outcomes as being in the clinical range at the end of treatment and having experienced no (reliable) change or deterioration. This is a slightly different definition than that used by others (e.g. Lutz et al., 2006), that combined deterioration and no change as negative outcomes, including patients who functioned in the normal range at the end of treatment. We chose to define people who were functioning in the normal range as positive outcomes; since functioning for this group is comparable to people that do not have psychiatric complaints. As a result, our results differed as well. Lutz et al. (2006) found their model to be more sensitive than ours, but less specific, probably because they had a more homogeneous sample, resulting in smaller confidence intervals. The definition of negative outcome also influences the CART models, in the sense that patients who function in the normal range at the beginning of treatment usually still do so at the end of treatment. So it is no real surprise that patients with low initial severity are in the first branch of the regression tree. Patients who start just above the cut-off score for normal functioning (55) and end up just below it, and who have in fact not changed a lot, are also considered a treatment success according to our definition. However, every definition of negative outcomes has its drawbacks, including the Jacobson & Truax (1991) criteria, which are considered very conservative (Jacobson, Follette, & Revenstorf, 1984). Slight variations in the definition of negative outcomes could lead to different results in the prediction models. Sensitivity analysis could provide more insight in the extent to which these variations have impact on the results. It should be noted that the term negative outcomes has the connotation that patients could do better, but for some patients no change may be the best obtainable result. The no change group is a complicated group to start with, as it probably consists of several subgroups, including patients who have a more chronic course and are not expected to change, patients who have somewhat improved, but not enough to meet the criteria for reliable change of Jacobson & Truax (1991) and patients who come to therapy for other reasons than symptom reduction (e.g. insight or life phase problems) (Watson, 2011). This may be the reason that the multilevel model performed better in predicting deterioration than deterioration and no change combined.

In clinical practice, a combination of the CART models and multilevel model could be used to identify patients at risk for negative outcomes and to develop measures to prevent them. For instance, patients who are having high symptom distress and low or medium education may be monitored more closely during treatment, as they have an increased risk of having negative outcomes. The multilevel model can then be used to predict change. We can distinguish four groups: (1) patients who are at risk for negative outcomes and have an expected change treatment course that shows no change or deterioration; (2) patients who are at risk, but have an expected positive treatment course; (3) patients who are predicted to have positive outcomes, but have a negative or no change expected treatment course and (4) patients who have predicted positive outcomes and also a positive expected treatment course. The first group probably has the highest risk of having actual negative outcomes. As these patients are *expected* to have a negative treatment course, the patient may not cross the failure boundary, but still have a negative treatment result. In these cases, the multilevel model might not be very effective in preventing the negative outcomes from happening, but intensive monitoring is still advisable. Other treatment options, such as seeing patients more frequently, or combining individual therapy and group therapy should be considered as well. The extended CART model could be used to identify subgroups of patients with better chances of favorable outcomes within this high risk group, and factors that also influence the rate of change, such as expectancies and the agreement on what needs to be done in therapy (Task) should be actively addressed by the therapist. For the second group, that is at risk for negative outcomes but with a favorable predicted treatment course, the multilevel model could be combined with intensive monitoring to assure that patients who go off track (and thus have a fair risk of deterioration) are identified early on in treatment. The third group, with positive predicted outcomes, but a non-positive expected treatment course probably includes patients who start with low symptom severity and are not expected to improve much. Another options is that they have favorable characteristics (e.g. low symptom severity, high education), combined with negative expectancies and a low (early) working alliance. Again, the extended CART model could provide more insight in which patients are more likely to have negative outcomes within this group. The fourth group probably has the best chances of favorable outcomes. In this group, the frequency of measurements could be decreased. We know from studies by Lambert that 30-50% of patients go off the expected track during treatment (e.g. Harmon et al., 2007; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004). His feedback system greatly improves outcomes for those off track patients (Lambert, 2007; Shimokawa, Lambert, & Smart, 2010), but has the consequence that patient progress has to be monitored for all patients on a session by session basis, which is a big investment of time and money. If we identify at risk patients prior to treatment, we could focus on monitoring those patients who are at risk more intensely and relax the measurement frequency for patients who have a low risk of negative outcomes and thereby improve the cost-effectiveness of the tracking system.

Using state of the art statistical techniques in analyzing our data enabled us to use all available data, and provided us with enough power to detect relevant predictor variables. The complete case sample shows that had we not used multiple imputation, results would have been different, as we would have been forced to drop some of the predictor variables, including several of the ones we were interested in (e.g.

educational level, previous treatment and duration of complaints), and drop cases without complete values on the left-over predictor variables. Interesting as these methods are, like all other methods, they have their limitations as well. For instance, one of the criticisms on CART is that is not a hypothesis driven technique and therefore strongly depends on the data. Another issue is that the models usually are not very stable. That is why the models need to be cross-validated. The imputation mechanism of using surrogate splits in CART is more limited than multiple imputation, as it uses a binary value as a surrogate for the missing value (higher or lower than split value X on predictor Y), whereas multiple imputation provides a plausible value for the missing value. A major drawback is that CART cannot be performed on a multiply imputed dataset, as it would create a different tree for each imputed dataset and results could not be pooled. As a result, when outcome variables are missing, only single imputation or the last observation carried forward method can be used, both of which are not ideal. It should be noted that all methods that deal with missing data can only fix the problem to the extent that the data are Missing At Random (MAR), in other words: There should not be selective missingness that depends on unobserved variables. Although the methods applied here perform much better in Missing Not at Random (MNAR) situations than most other methods (e.g. Graham, 2009), results are still likely to be somewhat biased. Like in most naturalistic datasets, it is probable that at least part of our data was MNAR.

The main objective of this study was to develop a good prediction model that is useful in clinical practice and could be used to help prevent negative treatment outcomes. In our aim to improve outcomes for this group, an important consideration is how much outcome is to be expected. Some of the patients may simply never achieve positive outcomes. However, comparing results from randomized trials and clinical practice (Barkham, et al., 2008; Hansen, et al., 2002) suggest that in clinical practice there is still room for improvement, even though patients who participate in clinical trials may not be entirely representative of patients who are seen in everyday practice. Using prediction models in clinical practice has mainly been successful in reducing deterioration rates, but less effective in improving the outcomes for the no change group. As a field, we need to continue looking for better prediction models to help improve outcomes for this group as well and combining models that aim to predict outcomes as well as rate of change. Searching for interactions between predictive factors might be helpful in building more complex predictive models. Fortunately, statistical developments are progressing fast to help us to develop more complex prediction models suitable for analysing data from naturalistic settings.

Acknowledgements

The authors would like to thank patients and staff from GGZ Noord-Holland-Noord, GGZ Dijk en Duin and PsyQ Haaglanden that have participated in this study, in particular Patricia van Sluis, Ed Berretty and Kosse Jonker, as well as all the students that have assisted in the data collection. We would also like to thank John Ogrodniczuk for his feedback on the draft of this article.