

The Probability of Treatment Success, Failure and Duration— What Can Be Learned from Empirical Data to Support Decision Making in Clinical Practice?

Wolfgang Lutz,^{1*} Michael J. Lambert,² S. Cory Harmon,²
Armita Tschitsaz,¹ Eva Schürch¹ and Niklaus Stulz¹

¹University of Berne, Switzerland

²Brigham Young University, Provo, UT, USA

Empirical methods have been found to be superior to clinical judgment for the purpose of correctly identifying patients at risk for treatment failure and, hence, to enhance psychotherapy outcomes. The development and evaluation of an empirical approach aimed at supporting clinical decisions during the course of psychotherapy is described. The tool provides predictions based on a patient-specific sampling strategy called the nearest neighbors method and on growth curve approaches to model an expected treatment course for each patient. Using session-by-session data from an outpatient center in the US ($N = 4365$), this new empirically derived decision model was evaluated and compared with a clinically based approach loosely based on an adaptation of clinically significant change concepts. The empirically derived decision system was found to be superior to the rational clinically based one in almost all measures of prediction accuracy, indicating its potential to identify patients at risk for treatment failure. Copyright © 2006 John Wiley & Sons, Ltd.

While there is evidence for the efficacy and effectiveness of psychotherapy for a wide variety of psychiatric disorders and psychological problems, the so-called 'scientist–practitioner gap' and the lack of practical implications of psychotherapy research are chronic problems that continue to hinder the impact of psychotherapy (see, e.g., Barkham et al., 2001; Borkovec, Echemendia,

Ragusea, & Ruiz, 2001; Goldfried & Wolfe, 1996; Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lambert, Hansen, & Finch, 2001; Newman & Tejada, 1996). Clinicians often criticize researchers for not producing clinically relevant research results, while researchers complain that few research findings and empirically supported treatment alternatives find their way into clinical practice. Among other reasons, this might be due to the fact that researchers often represent a nomothetic way of thinking, i.e., they ask about the *average* efficacy of a treatment. Therapists, on the other hand, are confronted with the question of whether this treatment is working for this *specific* patient.

A general method of dealing with this issue has recently emerged as a new research paradigm,

*Correspondence to: Wolfgang Lutz, Department of Psychology, University of Berne, Muesmattstrasse 45, CH-3012 Berne 9, Switzerland.

E-mail: wolfgang.lutz@psy.unibe.ch

Contract/grant sponsor: Swiss National Science Foundation (SNF), PP01-102651/1.

which is called *patient-focused psychotherapy research* (see, e.g., Howard et al., 1996; Lambert, 2001; Lutz, 2002). Patient-focused psychotherapy research aims to use empirical information on the course of individual patients' treatment to forecast the ongoing course of treatment and to feed back this information to the therapist in a timely manner. Several specific approaches have been developed to predict and feed back an individual patient's response to therapy in order to enhance patients' outcomes. These approaches use different versions of decision rules based on a patient's expected level of progress, which can be summarized into two broad classes: *rationally derived methods* and *empirically derived methods* for identifying patients at risk for treatment failure. Rationally derived methods are based on predefined clinical judgments about poor progress (see, e.g., Lambert et al., 2002) and first applications of them have already found their way into specific computer programs and into routine care. Empirically derived rules, on the other hand, use statistically derived expected recovery curves, with which the actual treatment course of a patient can be compared to allow for adaptive treatment planning (see, e.g., Lutz, Martinovich, & Howard, 1999).

Lambert and his colleagues (2003) demonstrated beneficial effects of giving therapists feedback on their clients' progress in a meta-analysis of feedback studies. They specifically showed that feedback based on rationally derived identification procedures is helpful in identifying patients at risk for treatment failure and reducing negative outcomes. The rationally derived decision rules used in these studies were based on information about early response to treatment, the dose-response relationship and the concept of clinically significant change. The authors created decision matrices for different treatment phases (e.g., sessions 2-4), which enabled clinicians to classify every patient into one of four feedback categories based on her/his initial score on the Outcome Questionnaire (OQ-45) (vertical axis of the matrix) and on her/his change on this instrument at the session of interest (horizontal axis). The areas of the matrix, i.e., the feedback categories, were color coded and resulted in different recommendations to therapists. White feedback, for example, was given if the patient was within the range of normal functioning and included the recommendation 'consider termination of therapy'.

Much previous work on empirically derived rules used an expected treatment response (ETR) method (e.g., Finch, Lambert, & Schaalje, 2001;

Lueger et al., 2001; Lutz et al., 1999), which evolved from the dose-effect model (Hansen, Lambert, & Forman, 2002; Howard, Kopta, Krause, & Orlinsky, 1986), which assumes a linear relationship between the logarithm of the session number and the benefit of patients due to psychotherapy. Growth curve analyses of large, longitudinal data sets were used to generate the expected response to treatment of each individual client. For example, in one study, individual differences in change were predicted using the following seven indices of client status measured at intake: well-being, symptom distress, social functioning, previous therapeutic experience, duration of problem, client's expectation of improvement in psychotherapy and evaluation of client's functioning status by the therapist (Lutz et al., 1999). Each client's expected change over treatment was then modeled as a log-linear function of the session number. That is, an ETR curve was generated using client information from the seven predictors and the actual client course was compared with the predicted course to evaluate treatment progress. Further work using the ETR model found that predictions for change in later sessions were enhanced by incorporating information about the change clients experienced during early sessions (e.g., Haas, Hill, Lambert, & Morell, 2002; Lutz, Rafaeli, Howard, & Martinovich, 2002).

In one study, Lambert et al. (2002) compared the rational method described above for predicting patient treatment outcome with an empirically derived rule based on an ETR technique and showed essential equivalence, though the ETR approach was marginally more accurate in predicting patient treatment failure. However, the rationally derived method was faster in identifying patients at risk for treatment failure.

By further refining an ETR method, Lutz et al. (2005) found that an approach based on the nearest neighbors (NN) method was superior to the ETR model in predicting rate of change. Therefore, in this demonstrational study, we went beyond the ETR approach to consider case-specific predictions of session-to-session changes in clients' impairment using a new clinical sample and a different measure of progress and outcome. Hypothesizing that particular predictors may work best for restricted subsets of clients, the NN approach identifies those previously treated clients who most closely match the target client (hence 'nearest neighbors') on intake variables. It then uses this homogeneous subgroup to generate predictions of treatment progress for the target client. This strat-

egy has been used to estimate the probabilities of alpine avalanches occurring (see, e.g., Brabec & Meister, 2001). In developing NN prediction models for avalanches, researchers used a large database of days with many kinds of potentially relevant parameter e.g., temperature, barometric pressure, depth of snow, recent rainfall. To make a prediction for a given day, they chose the most similar days and used the relative frequency of avalanches among those nearest neighbors as the prediction (Brabec & Meister, 2001).

In adapting the NN approach to predict clients' progress in psychotherapy, our aim was to make forecasts based on the course of similar already-treated patients. That is, we aimed to select subsamples of clients similar to an incoming client to predict his/her course, based on the assumption that the incoming client's course will resemble the course of most similar already-treated patients. This approach mirrors the way clinicians often talk about how they use their clinical experience (Lutz et al., 2006).

Although the ETR and NN approaches may use the same variables, they differ in the ways they model change. The ETR approach is essentially an advanced regression (level 1) model, which uses the intake variables for the full sample to predict individual change (see, e.g., Raudenbush, 2001). Individual slopes form the input for the regression (level 2) model and weights show the influence of the intake variables on the prediction of individual change. In the NN approach, however, the intake variables are used only to find the nearest neighbors. Individual change is then predicted using an unconditional growth model, i.e., using the average growth (slope) for the nearest neighbors as the prediction.

In this development and evaluation study, we used the NN approach to demonstrate the development of a statistical decision rule to evaluate clinical progress. This statistical model to predict final treatment outcome was then evaluated and compared with a rational decision rule using a large clinical data set drawn from an outpatient setting. The rational approach used in this study was derived by adapting the above mentioned procedure to the Outcome Questionnaire-30 (OQ-30, Lambert, Hatfield, Vermeersch, & Burlingame, 2001).

We addressed the following research questions: (a) how well do the rationally derived and the empirically derived decision rules predict treatment outcome?; (b) how early in treatment can the decision rules identify negative developments and which system is better at doing this? and (c) is the

number of warning signals provided from these decision rules predictive of therapy outcome?

METHODS

Participants

Participants included $N = 3968$ patients, who were treated for a total of 4365 therapy episodes at a university counseling center between June 1996 and May 2004. They were predominately female (63.2%), US citizens (92.0%) and single (67.3%; married, 30.8%; divorced, 1.0%; other marital status or no information, 0.8%).

At the beginning of treatment 24.5% of the participants fulfilled criteria for an affective disorder according to DSM-IV (American Psychiatric Association, 1994). Additional diagnoses included 9.7% anxiety disorder, 4.7% eating disorder, 9.8% adjustment disorder, 8.4% other diagnosis (e.g., personality disorder), 30.2% V-diagnosis and 12.7% of the participants had no diagnosis or no diagnostic information was available.

All of the 162 therapists were doctoral level students in training or doctoral licensed mental health professionals. They had a variety of treatment orientations, with most integrating two or more theoretical systems (e.g., cognitive and behavioral). The average duration of therapy was 9.3 sessions (SD = 8.9, range 3–129).

Measures

Participants completed the Outcome Questionnaire-45 (OQ-45, Lambert et al., 2004) at the beginning of each therapy session. This instrument measures three fundamental aspects of the patient's progress: (a) subjective discomfort (intrapsychic functioning), (b) interpersonal relationships and (c) social role performance. The 45 items address commonly occurring problems across a wide variety of disorders and tap symptoms most likely to occur. Each item is scored on a five-point scale (0 = never, 1 = rarely, 2 = sometimes, 3 = frequently, 4 = almost always).

In this study, only the 30 items of the short form called the OQ-30 (OQ-30, Lambert et al., 2001b) were further analyzed, since completion of the OQ-30 takes less than five minutes and is therefore well suited to assess data on a session-by-session basis. Although the OQ-30 provides three subscale scores, only the total score, which provides a global assessment of patient functioning, was used.

Internal consistency of the OQ-30 is high ($r = 0.93$) and the three-week test-retest value is satisfactory ($r = 0.84$). Correlations between the total score of the OQ-30 on the one hand, and the GSI of the SCL-90-R (Derogatis, 1977) and the total scores of the BDI (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the IIP (Horowitz, Rosenberg, Baer, Ureño, & Villasenor, 1988) and the SAS (Weissman & Bothwell, 1976) on the other hand, range from 0.59 to 0.70. Furthermore, the OQ-30 proved to be change sensitive for psychotherapy patients (Lambert et al., 2001b).

The average pretreatment OQ-30 score of the $N = 4367$ therapy episodes was 53.0 (SD = 12.5); the average posttreatment score was 43.3 (SD = 12.5).

Procedure

Rationally Derived Method

First, the decision matrices were adapted for the OQ-30 by adapting the boundaries between the feedback categories so that feedback frequencies were approximately identical to the OQ-45 (Lambert et al., 2002). We used three decision matrices for sessions 2–4, 5–9 and 10 and above. Each of them contained four possible feedback messages, two indicating that the patient is on track and two indicating that he or she is not.

Empirically Derived Method

The nearest neighbors (NN) method was used to model the expected treatment response curves. In a first step, for every patient, the 10 most similar cases were selected based on the initial scores of the 15 items of the OQ-45 not included in the OQ-30 (Table 1). These homogenous subgroups were calculated based on Euclidian distances. Since all 4367 episodes were already finished and session-by-session reports were available in the database, growth curve slopes were calculated separately for each client as a linear function of the logarithm of session number based on the 10 nearest neighbors (using the SAS procedure 'Mixed', Little, Milliken, Stroup, & Wolfinger, 1996; Raudenbush, 2001). In this way, the average slopes of the 10 nearest neighbors were used as a change parameter to model each client's expected change over treatment. For every individual prediction, the 67, 75, 84, 90, 95, 97.5 and 99.5% prediction intervals were determined (see Figure 1). If the observed course deviated from the expected one by more than the negative border of the confidence interval, this was treated as a warning signal indicating risk for neg-

Table 1. The 15 items used to define the nearest neighbors for each patient

-
1. I get along well with others
 2. I tire quickly
 3. I feel unhappy in my marriage/significant relationship
 4. I work/study too much
 5. I have a unfulfilling sex life
 6. I feel loved and wanted
 7. I enjoy my spare time
 8. I like myself
 9. My heart pounds too much
 10. I have sore muscles
 11. I feel afraid of open spaces, of driving or of being on buses, subways and so forth
 12. I feel my love relationships are full and complete
 13. I have too many disagreements at work/school
 14. I feel angry enough at work/school to do something I might regret
 15. I have headaches
-

ative treatment outcome. If the observed course deviated from the expected one by more than the positive border of the confidence interval, this was treated as a positive signal, indicating a positive development and a higher chance for positive outcome.

RESULTS

To evaluate the two decision rules under examination, treatment outcome was classified into four categories based on the concept of clinically significant change (see Jacobson & Truax, 1991): (a) clinically significant change, (b) reliable improvement, but not in the range of the functional population at the end of therapy, (c) no (reliable) change and (d) reliable negative change. Subsequently, (a) and (b) will often be treated as positive treatment outcomes, which do not require an alert or warning signal during treatment, while (c) and (d) represent negative outcomes.¹ According to these conservative change criteria, 1460 (33.5%) episodes showed

¹A clinically significant change indicates that the patient changed reliably and improved from the clinical to the functional population. Reliable change (RC) was defined via the Jacobson and Truax (1991) formula: OQ-30 change-score equal to at least 9. The cut-off score between the functional and the clinical population is 44 (Lambert et al., 2001b). A reliable negative change indicates a negative change of 9 OQ-30 raw scores from pre- to post-score.

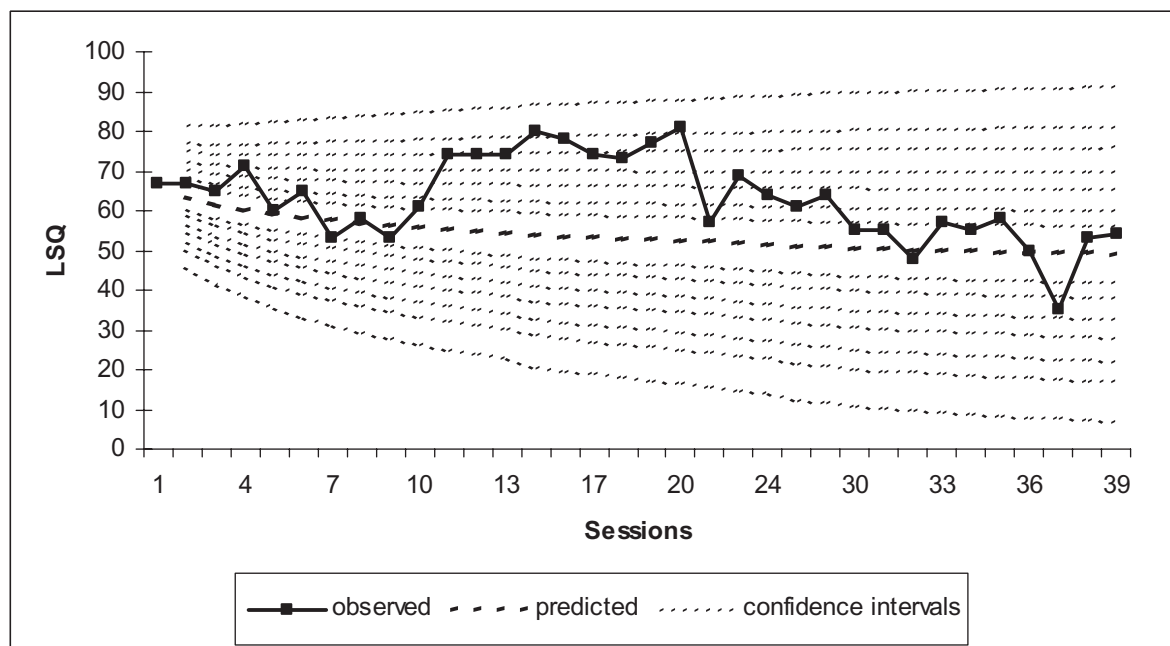


Figure 1. Predicted individual treatment response and confidence intervals (example)

Table 2. Statistical differences between different empirical decision rules and the rational decision rule

	Rational	Empirical						
		67% CI	75% CI	84% CI	90% CI	95% CI	97.5% CI	99.5% CI
Pos. pred. value	60.4	63.2***	65.8***	69.9***	73.1***	75.7***	77.5***	78.8***
Neg. pred. value	63.1	75.4***	70.3***	65.1***	61.8***	58.3***	56.1***	53.9***
Sensitivity	57.4	79.0***	68.4***	53.1***	41.1***	27.5***	17.8***	7.3***
Specificity	65.9	58.3***	67.8	79.3***	86.3***	92.0***	95.3***	98.2***
Correct identified	61.9	68.2***	68.1***	66.8***	64.8***	61.4	58.5***	55.0***

* $p \leq 0.05$ (McNemar's χ^2 -test, two-tailed significance).

*** $p \leq 0.001$ (McNemar's χ^2 -test, two-tailed significance).

a clinically significant change and an additional 830 (19.0%) showed at least a reliable improvement. 1864 (42.7%) episodes indicated no reliable change and 211 (4.8%) indicated negative change.

In order to adapt the analyses to feedback conditions in daily practice, we combined reliable negative change and no change into a *negative outcomes group*, which require warning signals for the therapists, and reliable and clinically significant improvement into a *positive outcomes group*, which does not require an alert. We also compared all different confidence bounds of the statistical rule with the rational rule. As can be seen in Table 2, the statistical decision rule with a 75% confidence

interval is an empirically derived system that is superior to the rational one in all indices of detection accuracy. Feedback based on a 67% confidence interval, for instance, is characterized by a higher sensitivity (79.0%). However, the problem here would be a lower specificity (58.3%), which is even lower than the one of the rational model (see Table 2). Following this result, we used the 75% rule for further analyses of the empirically derived decision rule.

Besides a high detection accuracy, a system to monitor psychotherapy should also be able to identify patients at risk for treatment failure early in treatment. Table 3 compares the two models in

Table 3. Percentage of patients with negative outcome ($n = 2075$), who would have received at least one warning signal by sessions 2-8

Rule	Session						
	2*	3*	4*	5*	6*	7*	8*
Empirically ¹	33.5	45.1	53.2	58.9	62.0	63.6	64.9
Rationally	29.7	38.1	42.6	47.5	49.6	51.6	52.8

¹75% negative prediction interval.

*A statistically significant difference between the empirically derived and the rationally derived decision rule at the $p < 0.05$ level.

Table 4. Probabilities (and numbers) of positive and negative treatment outcomes depending on the number of warning signals generated by the rational and the 75% statistical decision rule

	Rational		Empirical		p^1
	Positive	Negative	Positive	Negative	
No signal	0.63 (1510)	0.37 (874)	0.70 (1552)	0.30 (655)	<0.001
1 signal	0.49 (388)	0.51 (411)	0.47 (347)	0.53 (396)	n.s.
2 signals	0.38 (158)	0.62 (255)	0.30 (137)	0.70 (314)	<0.05
>2 signals	0.31 (234)	0.69 (525)	0.26 (254)	0.74 (710)	<0.05

¹Tests for a statistical significant difference between the percentages of correct identified patients by the empirically derived and the rationally derived decision rules.

their capacity to identify patients at risk early in treatment. The rational model would have generated an alert for 29.7% of the non-responders (i.e., patients without at least reliable improvement) in the second session and almost 50% of them would have received at least one warning signal by session eight. The empirical rule with a prediction interval of 75% is not only more accurate, it is also faster in identifying patients at risk for treatment failure, with more than half of the non-responders detected by session 4 (see Table 3).

Furthermore, as can be seen in Table 4, an increasing number of warning signals during treatment is associated with a higher probability of an actual negative treatment outcome in both the rational and the 75% statistical model. For example, 396 out of the 743 patients with exactly one empirically derived warning signal do not benefit from the treatment (positive predictive value = 0.53).² This probability for a negative treat-

ment outcome increases to 70% given two signals, and to 74% if a patient receives three or more alerts during treatment. The probability of being a treatment non-responder is also increasing with an incremental number of warning-signals provided by the rational rule. However, the predictive values of this method are on a lower level than the statistical rule (see Table 4).

In order to further explore the empirical decision system, different prediction intervals were considered. Figure 2 shows the positive and negative predictive values for different confidence intervals (67, 75, 84, 90, 95, 97.5 and 99.5%) and for different numbers of signals (0, 1, 2, and 3 and more signals).

As can be seen in Figure 2(a), the higher the negative prediction interval, the higher the percentage of a negative treatment outcome, if one or more warning signals were reported (see differences between the prediction intervals for each amount of negative feedback on the left half of Figure 2(a)). That is, the less probable a deviation of the actual course from the predicted one is due to chance, the more indicative it is for a negative treatment outcome. As can also be seen in this example, the cost of higher prediction intervals are lower

²The positive predictive value indicates the true positives (those signals under investigation) divided by all positives (in this case negative outcome).

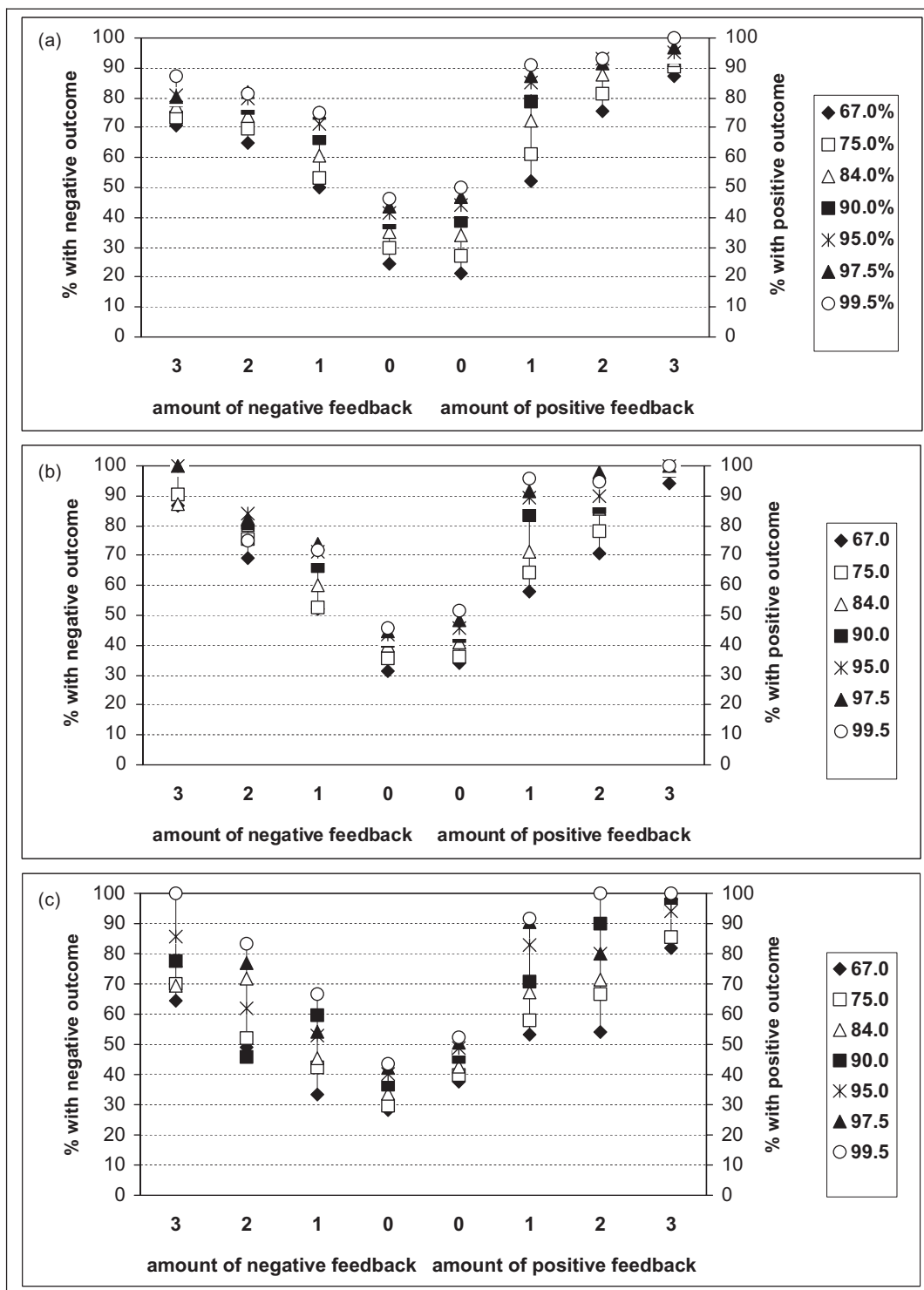


Figure 2. Positive and negative predictive values against the amount of feedback and different confidence intervals: (a) all patients and all sessions ($N = 4395$); (b) feedbacks between session 2 and 4, outcomes between session 5 and 8 ($n = 1423$); (c) feedbacks between session 2 and 8, outcomes between session 17 and 28 ($n = 389$)

sensitivities, i.e., using a high prediction interval might give rise to the problem of missing negative outcome cases since with high prediction intervals a signal is only generated when the patient strongly deviates from expected course.

The same holds for positive feedback, i.e., feedback that is generated if the observed course falls below the positive border of the confidence interval. As can be seen in the right half of Figure 2(a), the probability of at least a reliable improvement increases with higher positive prediction intervals and with an increasing amount of positive feedback.

Figure 2(b) shows the predictive validities of early feedback—i.e., of feedback during sessions 2–4—for a sub-sample of patients with short treatments (five to eight sessions, $n = 1423$). All of these patients, who had a score outside of the range of the 90% confidence interval in sessions 2, 3 and 4 terminated treatment reliably improved (in the case of three times falling below the positive confidence interval) or did not improve (when all times exceeded the negative confidence interval). Again, as expected, there is a tendency that the more scores outside the confidence intervals a patient receives and the higher the chosen prediction interval, the higher is the predictive validity. For example, if a patient exceeds the negative 67% confidence interval only once during sessions 2–4, then the percentage of an actual negative outcome is only 52.0% (compared with 74.0% when exceeding the 97.5% confidence interval once). Although this value is not significantly better than the base rate (46.5% of the patients with five to eight sessions have a negative outcome, $p = \text{n.s.}$), it is still significantly higher than for patients who never received a warning signal during this period of time (31.5% show negative outcome, $p < 0.05$).

However, the association between higher confidence intervals and increasing predictive values is not always consistent. For example, among patients with short treatments (five to eight sessions), two warning signals indicate a probability of 82.4% for a negative treatment outcome given a confidence interval of 97.5%. This predictive value tends to decrease with a higher confidence interval of 99.5% and falls to 75% (see Figure 2(b)). However, due to small n (only 17 received two signals when applying the 97.5% border, and only four when using the 99.5% border), this difference is not significant.

A quite similar picture is shown for another sub-sample of patients ($n = 389$) with longer therapies (17–28 sessions), in which measurements in ses-

sions 2–8 were used for the prediction of final therapy outcome. The predictive values again depend on the confidence intervals and on the number of positive and negative feedback messages, respectively (see Figure 2(c)).

DISCUSSION

Clinical decision support tools are helpful for clinicians to identify patients at risk for treatment failure and, hence, can enhance the rate of positive treatment outcomes. Therefore, the aim of this study was to develop and evaluate an empirically derived model to forecast treatment outcome in psychotherapy and to compare it with a more rationally derived prediction system, which proved to be effective in enhancing treatment outcomes (Lambert et al., 2003). The main idea of the statistical model used here is to select homogenous sub-samples of the most similar patients for a given patient based on intake variables. In a second step, the estimated growth curve slope of these nearest neighbors was then used to predict the expected course of the patient under consideration. This procedure mirrors what therapists often say they do in using their clinical experience: Confronted with a new patient, they are looking for some similar cases and use the experiences they have with past cases to develop an optimal treatment for the incoming patient. Unfortunately, decisions based on clinical experience are typically less accurate than statistical predictions and, in the case of predicting eventual treatment failure, woefully underpredict such outcomes (Hannan et al., 2005).

The prediction accuracy of the statistical system used here was good (68.1% of the patients were correctly identified, positive predictive value = 70.3%) and superior to that of the adapted rationally derived decision rule (61.9% correctly identified, positive predictive value = 60.4%). Therefore, the NN method seems to be an interesting alternative to forecast individual treatment courses and to build clinical decision support tools. However, since the calculation of the nearest neighbors in a large data set and the growth curve modeling require a software system for routine application, which is not yet developed, the decision matrices of the rational system may be the best method for clinical practice.

Despite the good prediction accuracy of the empirically derived decision system, there still may be potential to enhance forecasts made by the NN method. For example, additional or other

intake variables (e.g., disorder-specific measurements) and/or other similarity measures than Euclidian distances (e.g., Gower similarity coefficient, which can include dichotomous and continuous variables) may enhance prediction accuracy and should therefore be investigated. Furthermore, the predictions may also be enhanced by selecting more or less than the 10 nearest neighbors (Lutz et al., 2005). For example, one could try to define cut-off criteria for each of the intake variables and then to select for the patient under consideration all cases that fulfill these criteria (e.g., all patients with the same gender and diagnosis, which do not differ by more than five OQ-30 points from the target patient in the initial measurement). If such methods were used, the number of selected nearest neighbors would not necessarily have to be the same for all patients. Furthermore, a re-definition of the nearest neighbors could be made over the course of therapy, with predictive information being available at each specific point in time during treatment and based on a new subset of neighbors. Also, given the fact that the decision rules can be differentially conservative, it is not clear when to give specific feedback to a therapist about a potential negative development or whether it should be an open system with the possibility for a therapist to get access and to get updated information at each point during treatment whenever he or she wishes.

Several investigators (e.g., Gibbons et al., 1993; Hansen et al., 2002; Howard et al., 1986) described a log-linear dose–effect relationship between the amount of treatment and the benefit due to psychotherapy. According to this, we modeled the individual growth curves as linear functions of the logarithm of session number. However, the exact shapes of change have been the subject of discussion in the literature for quite some time (e.g., Barkham, Stiles, & Shapiro, 1993; Krause, Howard, & Lutz, 1998) and indeed for some patients other change functions might be more appropriate to describe treatment progress and should therefore be considered.

Besides the potential to refine a clinical decision support system based on the NN method, this study has limitations: Whenever patients were classified into positive and negative outcomes, those with no (or no reliable) improvement were treated as negative outcomes since a goal of psychological treatment should be the reduction of psychological distress. However, under certain circumstances, the prevention of further exacerbation of psychological impairment may already be a

success of psychotherapy and the lack of reliable improvement is, therefore, not necessarily an instance of negative treatment outcome. Furthermore, not all cells in the outcome matrix have an equal number of patients. Therefore, as can be seen in Figure 2(b), not all of the calculated percentages have the same kind of validity based on large N values. For example, for some cases there might be not enough patients available in the database to define the probability of success for a longer term treatment.

Further research should try to refine the presented clinical decision support systems and to evaluate them with other datasets. In addition, the effects of giving feedback based on these decision rules to therapists should be evaluated with respect to the reduction of negative treatment outcomes.

ACKNOWLEDGMENT

This work was partially supported by a grant from the Swiss National Science Foundation (SNF), No. PP01-102651/1 (Wolfgang Lutz).

REFERENCES

- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (4th edition)—DSM IV*. Washington, DC: Author.
- Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., et al. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Clinical Outcomes in Routine Evaluation—Outcome Measures. Journal of Consulting and Clinical Psychology, 69*(2), 184–196.
- Barkham, M., Stiles, W.B., & Shapiro, D.A. (1993). The shape of change in psychotherapy: Longitudinal assessment of personal problems. *Journal of Consulting and Clinical Psychology, 61*, 667–677.
- Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychology, 4*, 53–63.
- Borkovec, T.D., Echemendia, R.J., Ragusea, S.A., & Ruiz, M. (2001). The Pennsylvania Practice Research Network and future possibilities for clinically meaningful and scientifically rigorous psychotherapy effectiveness research. *Clinical Psychology: Science and Practice, 8*, 155–167.
- Brabec, B., & Meister, R. (2001). A nearest-neighbor model for regional avalanche forecasting. *Annals of Glaciology, 32*, 130–134.
- Derogatis, C.R. (1977). *SCL-90, administration, scoring, and procedures. Manual 1 for the R(evised) version and other instruments of the Psychopathology Rating Scale Series*. Baltimore: Johns Hopkins University School of Medicine.

- Finch, A.E., Lambert, M.J., & Schaalje, B.G. (2001). Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology and Psychotherapy*, 8, 231–242.
- Gibbons, R.D., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H.C., Greenhouse, J.B., et al. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data. Application to the NIMH Treatment of Depression Collaborative Research Program dataset. *Archives of General Psychiatry*, 50, 739–750.
- Goldfried, M.R., & Wolfe, W. (1996). Psychotherapy practice and research: Repairing a strained alliance. *American Psychologist*, 51, 1007–1016.
- Haas, E., Hill, R.D., Lambert, M.J., & Morell, B. (2002). Do early responders to psychotherapy maintain treatment gains? *Journal of Clinical Psychology*, 58(9), 1157–1172.
- Hannan, C., Lambert, M.J., Harmon, C., Nielsen, S.L., Smart, D.W., Shimokawa, K., et al. (2005). A lab test and an algorithm for identifying patients at risk for treatment failure. *Journal of Clinical Psychology*, 61(2), 155–163.
- Hansen, N.B., Lambert, M.J., & Forman, E.M. (2002). The psychotherapy dose–response effect and its implication for treatment delivery systems. *Clinical Psychology: Science and Practice*, 9, 329–343.
- Horowitz, L.M., Rosenberg, S.E., Baer, B.A., Ureño, G., & Villasenor, V.S. (1988). Inventory of Interpersonal Problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology*, 56, 885–892.
- Howard, K.I., Kopta, M., Krause, M.S., & Orlinsky, D.E. (1986). The dose–effect relationship in psychotherapy. *American Psychologist*, 41, 159–164.
- Howard, K.I., Moras, K., Brill, P., Martinovich, Z., & Lutz, W. (1996). The evaluation of psychotherapy. *American Psychologist*, 52, 1059–1064.
- Jacobson, N.S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12–19.
- Krause, M.S., Howard, K.I., & Lutz, W. (1998). Exploring individual change. *Journal of Consulting and Clinical Psychology*, 66(5), 838–845.
- Lambert, M.J. (2001). Psychotherapy outcome and quality improvement: Introduction to the special section on patient-focused research [special section]. *Journal of Consulting and Clinical Psychology*, 69, 159–172.
- Lambert, M.J., Hansen, N.B., & Finch, A.E. (2001a). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology*, 69(2), 159–172.
- Lambert, M.J., Hatfield, D.R., Vermeersch, D.A., & Burlingame, G.M. (2001b). *Administration and scoring manual for the Life Status Questionnaire (LSQ)* [draft version]. Orem, UT: American Professional Credentialing Services L.L.C.
- Lambert, M.J., Morton, J.J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R.C., et al. (2004). *Administration and scoring manual for the OQ-45.2 (Outcome Questionnaire)*. Orem, UT: American Professional Credentialing Services L.L.C.
- Lambert, M.J., Whipple, J.L., Bishop, M.J., Vermeersch, D.A., Gray, G.V., & Finch, A.E. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy*, 9, 149–164.
- Lambert, M.J., Whipple, J.L., Hawkins, E.J., Vermeersch, D.A., Nielsen, S.L., & Smart, D.W. (2003). Is it time to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice*, 10, 288–301.
- Little, R.C., Milliken, G.A., Stroup, W.W., & Wolfinger, R.D. (1996). *SAS system for Mixed Models*. Cary, NC: SAS.
- Lueger, R.J., Howard, K.I., Martinovich, Z., Lutz, W., Anderson, E.E., & Grissom, G. (2001). Assessing treatment progress with individualized models of predicted response. *Journal of Consulting and Clinical Psychology*, 69, 150–158.
- Lutz, W. (2002). Patient-focused psychotherapy research and individual treatment progress as scientific groundwork for an empirical based clinical practice. *Psychotherapy Research*, 12, 251–273.
- Lutz, W., Leach, C., Barkham, M., Luccock, M., Stiles, W.B., Evans, C., et al. (2005). Predicting change for individual psychotherapy clients based on their nearest neighbors. *Journal of Consulting and Clinical Psychology*, 73(5), 904–913.
- Lutz, W., Martinovich, Z., & Howard, K.I. (1999). Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *Journal of Consulting and Clinical Psychology*, 67, 571–577.
- Lutz, W., Rafaeli, E., Howard, K.I., & Martinovich, Z. (2002). Adaptive modeling of progress in outpatient psychotherapy. *Psychotherapy Research*, 12, 305–327.
- Lutz, W., Saunders, S.M., Leon, S.C., Martinovich, Z., Kosfelder, J., Schulte, D., Grawe, K., & Tholen, S. (2006). Empirical and clinical useful decision making in psychotherapy: Differential predictions with treatment response models. *Psychological Assessment*, 18(2).
- Newman, F.L., & Tejada, M.J. (1996). The need for research that is designed to support decisions in the delivery of mental health services. *American Psychologist*, 51, 1040–1049.
- Raudenbush, S.W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, 52, 501–525.
- Weissman, M.M., & Bothwell, S. (1976). Assessment of social adjustment by patient self-report. *Archives of General Psychiatry*, 33, 1111–1115.