



Predicting individual change during the course of treatment

Edward A. Wise, David L. Streiner & Robert J. Gallop

To cite this article: Edward A. Wise, David L. Streiner & Robert J. Gallop (2016) Predicting individual change during the course of treatment, *Psychotherapy Research*, 26:5, 623-631, DOI: [10.1080/10503307.2015.1104421](https://doi.org/10.1080/10503307.2015.1104421)

To link to this article: <http://dx.doi.org/10.1080/10503307.2015.1104421>



Published online: 12 Nov 2015.



Submit your article to this journal [↗](#)



Article views: 127



View related articles [↗](#)



View Crossmark data [↗](#)

METHOD PAPER

Predicting individual change during the course of treatment

EDWARD A. WISE¹, DAVID L. STREINER^{2,3}, & ROBERT J. GALLOP⁴

¹*Mental Health Resources, PLLC, Memphis, TN, USA;* ²*Department of Psychiatry, McMaster University, St. Joseph's Healthcare, Hamilton, Ontario, Canada;* ³*Department of Psychiatry, University of Toronto, ON, Canada &* ⁴*Applied Statistics Program, Department of Mathematics, West Chester University, West Chester, PA, USA*

(Received 10 January 2015; revised 15 September 2015; accepted 22 September 2015)

Abstract

Objective: An empirically derived prediction model was developed in a private practice setting to monitor on-track and off-track weekly treatment progress in an intensive outpatient program (IOP). **Method:** The predictive equation was derived as a function of the baseline measure and time. The formulae for the predictive equations were derived from two groups of psychiatric patients ($N = 400$ each) in an IOP diagnosed with major depression. Each equation was cross-validated between these two psychiatric IOP samples and a dual diagnosis sample ($N = 198$) using κ , the reliable change index (RCI), receiver operating characteristic curves, and Youden's J . **Results:** Using varying RCI classifications, approximately 66–75% of both samples reliably improved, 23–24% were indeterminant, and only 1–3% deteriorated. Of patients identified as off-track, which included patients classified as indeterminant and deteriorated, 83% were correctly identified. Of those identified as on-track, 85% were correctly classified. Those identified as on-track (85%) are highly likely to respond to treatment as expected. **Conclusions:** The overall efficiency index (hit rate) for the correct classification of all patients was 85%. Implications for using this predictive model as a clinical support decision tool with relatively homogeneous populations in other practice settings are discussed.

Keywords: routine outcomes monitoring; psychotherapy outcomes; predicting change; intensive outpatient

Previous research has demonstrated that depressed patients with symptom severity levels comparable to, and in many cases exceeding inpatients, can be effectively treated in a private practice-based intensive outpatient program (IOP; Wise, 2003). These findings were based on pre- and post-treatment measures, as well as weekly symptom data, using reliable change indices (RCIs) and clinically significant (CS) change methodology (Jacobson & Truax, 1991). Furthermore, a significant and predictable dose–response relationship was seen between depressive symptoms and amount of treatment, as well as a Cohen's $d = 1.68^1$ for the amount of change in pre- to post-treatment depressive symptoms (Wise, 2005). However, moving beyond *ex post facto* patient outcome for individuals and groups to monitoring and predicting outcomes for individuals in real time has proved challenging.

The identification of patients at risk for deterioration is perhaps best illustrated by Lambert (2010). Lambert, Hansen, and Finch (2001) provided a conceptual model involving nearly 12,000 patients divided into 50 groups, from which they created estimated recovery curves that were used to predict individual outcomes. Fifteen and ten percent of patients were identified as making less than adequate progress and deviating significantly from the expected trajectory of change. These patients became identified as “signal alarm” and “off-track” cases, respectively. Finch, Lambert, and Schaalje (2001) explained the rationale for determining that the patient's intake score was the most robust predictive variable to base an expected recovery curve and also defined a signal alarm or early warning signal to detect individuals at risk for deterioration. They also noted, however, that “any desired proportion of a patient

Correspondence concerning this article should be addressed to Edward A. Wise, Mental Health Resources, PLLC, PLLC, 1037 Cresthaven Road, Memphis, TN 38119, USA. Email: Wise@MHRMemphis.com

population, such as 15 or 20%” could be used to identify signal alarm and off-track patients (p. 233). Lambert et al. (2002) and Spielmans, Masters, and Lambert (2006) compared rational and empirical methods to predict psychotherapy outcomes and found the latter to outperform rational or clinical predictions.

Unfortunately, the use of this empirical system, and more importantly the methods, are limited to extremely large sample sizes such as that used in the development sample of the signal alarm parameters using the OQ-45 (Lambert et al., 1996). Due in large part to the extremely large sample size in Lambert et al.’s (2001) study, their statistical approach to predicting treatment outcomes is successful because it produces expected recovery curves depending on an individual’s baseline severity, which may vary over the range of the dependent measures. However, few clinicians or researchers have access to such large samples with repeated measures. The challenge of developing similar clinical decision support tools based on actuarial prediction models using smaller sample sizes, which could be applied to other instruments, has thus far been difficult. Other limitations to these and similar outcome monitoring systems relate to their generalizability to specific types of patients, in specific settings.

Thus, even in studies utilizing relatively smaller sample sizes, the vast majority of patients involved in outcome prediction studies tend to be those seen for traditional outpatient services in low acuity settings. For example, Crits-Christoph et al. (2001) studied 105 moderately depressed and 79 moderately anxious outpatients receiving manually guided individual therapy and found that early patterns of change in sessions 2–4 were associated with symptom remission. Similarly, Renaud et al. (1998) studied 100 depressed adolescents receiving supportive, family, or cognitive therapy and demonstrated that rapid responders had better outcomes and longer periods of time to recurrence. Delgado et al. (2014) studied low-intensity interventions (e.g., brief, self-help, psycho-educational; average and modal number of sessions = 4.8 and 2, respectively) provided to 2710 patients in a primary care delivery system and demonstrated that early treatment response predicted treatment outcomes.

Unlike patients treated in routine outpatient settings, however, patients admitted to the previously mentioned IOP are treated primarily in groups, tend to be acutely depressed, with suicidal ideation, multiple co-morbidities, and typically meet the criteria for inpatient hospitalization. A further complication arises when one considers the generalization of the predictive model to dual diagnosis substance abuse patients in a similar IOP. These patients are

seen in manually guided programs for an average of 9 hr per week over a course of treatment of approximately 6 weeks. The practical question is whether it is possible to use data already being collected to predict the outcomes of similar patients, in the same programs, as a foundation for a clinical decision support tool in the future. We believe this is a frequent question faced by many programs and that a practical, cost-effective solution would represent a significant contribution to monitoring treatment progress, providing real-time feedback and improving treatment outcomes. We were unable to locate any published studies utilizing routine outcome monitoring (ROM) procedures in IOP settings. Some barriers to the adoption of ROM in higher level of care programs are related to the acuity of these patients, the lack of normative data, and the limited sample sizes in private practice settings on which to develop predictive equations. Nonetheless, the development of a clinical decision support tool that predicts when these individual’s trajectory of change substantially deviates from an on-track profile would be highly desirable as these patients are potentially at risk for admission to a higher level of care (e.g., hospitalization) should they deteriorate.

Method

Clinical Setting

An IOP is a state licensed facility-based program that meets up to 3 hr per day and that patients can attend up to 5 days per week. It is a group-based program but may include individual or family therapy, as well as psychiatric management. The treatment team is composed of Master’s level therapists and psychiatrists. Our treatment programs are guided by treatment manuals. Broadly speaking, treatment consists of a traditional process group, in which dysfunctional repetitive relationship themes are addressed, cognitive behavioral groups, and skills training groups. Additionally, both cultures benefit from the use of a motivational interviewing (Miller & Rollnick, 2013) and stages of change theoretical framework (DiClemente & Prochaska, 1998), while the dual diagnosis program is based on an integrated treatment model and incorporates a harm reduction approach. (The interested reader is referred to Wise, 2003, and 2010 for further details.)

Participants

These studies were all based on analyses of data gathered for clinical purposes at a multidisciplinary private practice that operates two IOPs. All study

participants provided signed informed consent to use their de-identified data for research purposes. The primary sample of interest is the psychiatric IOP. Patients are typically employed or are a family member of an employee and are referred by employee assistance programs, employers, primary care physicians, and behavioral health providers. A semi-structured interview of approximately 90 min is used as the primary intake assessment tool to ensure the consistent and reliable collection of information by Master's level clinicians. Additionally, all patients complete a battery of self-administered assessment tools prior to being seen by the clinician, including various standardized symptom and functional rating scales. All patients are commercially insured and must pass the pre-authorization process instituted by their respective insurance plans to access this higher level of care. Pre-authorization by behavioral health managed care plans typically involves a review of symptoms, functional impairments, and acuity level to ensure that the medical necessity criteria for IOP have been met. Medical necessity for IOP typically requires that the patient has an acute exacerbation of a psychiatric condition; which may include passive or fleeting suicidal and/or homicidal ideation; that results in significant functional impairments (e.g., occupational, school, social, interpersonal, etc.); that cannot be treated in a lower level of care; can be expected to improve with treatment; and without which could result in admission to a higher level of care, such as an inpatient admission. This external review is usually conducted over the telephone and includes an agreement that the criteria for primary diagnosis have been met, functional impairments are present, and meet the specific medical necessity criteria of the insurer for admission to IOP. The behavioral health managed care organization typically requests an intermittent concurrent review of symptoms and response to treatment (as often as weekly) to ensure that the patient continues to meet the medical necessity criteria for continued care. Patients who are imminently suicidal, homicidal, or psychotic are not appropriate for this level of care.

In the mental health IOP, we had access to a total of $N = 1,956$. Of these, 744 cases were excluded (377 were seen only for intake and never started; 99 were considered dropouts [attended ≤ 3 visits]; and 268 had < 3 data points), resulting in 1212 patients available for the study from the mental health IOP. In the dual diagnosis IOP, we had access to a total of $N = 616$. Of these, 205 were excluded (133 were seen only for intake and never started; 40 were considered dropouts [attended ≤ 3 visits]; and 32 had < 3 data points), resulting in a total of 411 patients available for the study from the dual diagnosis IOP.

For study one, we obtained two non-overlapping random samples of 400 each from the sample of 1212 from the psychiatric IOP. In order to obtain randomized, non-overlapping samples, a similar procedure was followed with the dual diagnosis sample.² The study included all patients who completed at least three measures (e.g., intake baseline, T1, and T2), even if those patients later dropped out of treatment.

As can be seen in Table I, the primary population served in the psychiatric IOP are depressed females, with approximately 14.5 years of education, and with similar proportions of African-American and Caucasian participants. The comparison dual diagnosis IOP sample, used as a validation sample, is also described in Table I. These demographics are similar to those previously reported (Wise, 2003, 2005, 2010).

Measures

Intake scores on the depression scale serve as initial pre-treatment measures and patients complete the same measure on a weekly basis in a treatment plan/weekly update group. We used six common depression item stems (i.e., suicidal thoughts, loss of interest, hopelessness, worthlessness, loneliness, and depressed mood) rated on a scale of 0–4. The items are prefaced with rating instructions and then presented in a list. The depression stems were previously validated in an IOP sample with reference to the BSI-18 Depression subscale (Derogatis, 2001). In that study (Wise, 2005), patients were administered the SCL-90-R, from which we extracted the BSI-18 Depression Score. On the same day, they completed our six depression item stems. The obtained correlations were .79 for the pre-treatment group ($n = 30$) and .76 for the post-treatment group ($n = 50$), demonstrating reasonable equivalence

Table I. Demographic information.

Variable	Psychiatric IOP ($n = 1212$)	Dual diagnosis IOP ($n = 421$)
Race ^a		
Caucasian ^a	570 (47%)	294 (70%)
African-American ^a	618 (51%)	122 (29%)
Other/missing	24 (2%)	4 (1%)
Age ^a	41.99 (10.84)	38.47 (12.73)
Gender (% male) ^a	21	53
Education	14.50 (2.24)	14.23 (4.24)
Major depression ^{a,b}	1176 (97%)	302 (72%)
Substance abuse ^a	1 (0.1%)	411 (98%)
IOP days ^a	15.81 (4.09)	13.49 (2.12)

^a $p < .05$.

^bPrimary or secondary diagnosis.

between the two forms, particularly in light of the few items (Cronbach's $\alpha = .87$) When the pre-treatment reliability study was replicated ($n = 100$), we obtained an $r = .76$. All subsequent calculations that require a test-retest r were conducted using this value. Our depression measure performs as a parallel version of the BSI Depression Scale and in keeping with the SCL-90-R (Derogatis, 1983) and BSI (Derogatis, 1993), the item scores are averaged.

Statistical Analyses

Like Lambert et al. (2001), we believe that it is desirable for the recovery trajectory to be dependent on an individual's intake score or baseline measure, and to vary over the range of the scale, including missing data at different intervals with varying lengths of treatment. To accommodate these two features, we fit the following ordinary least-squares regression:

$$Y_{\Delta i} = b_1 \text{Week}_i + b_2 \text{Week}_i \times \text{Baseline} + e_i,$$

where $Y_{\Delta i}$ is the change score in the outcome from baseline to the last available score for the i th person, and Week_i is the elapsed time in weeks as described above for the i th person. As in all regression models, e_i is the error in fit of the regression model assumed to be normally distributed with a mean of 0 and variance $= \sigma^2$. The b_2 term corresponds to the regression coefficient for the interaction of the baseline measure and time, which will accommodate the two features of dependency on an individual's baseline and allowing for variability over the range of the scale. The regression line has no intercept term which refers to a regression line through the origin, corresponding to 0 change based on no change in time from baseline. A separate fitted line and error estimate was calculated for every patient, which accounted for missing data and allowed for the comparison of individuals, irrespective of number of sessions, time between sessions, and varying lengths of treatment. Confidence intervals provide predictions with regard to the average change in score over time, dependent on baseline score, which is similar to Lambert's "tolerance intervals" (Lambert, 2010, p. 93).

Table II. Regression beta coefficients.

Group	b (Week 1)	b (Week 1 \times depression)
Sample 1 \times Sample 2	-0.014883	0.097255
Sample 2 \times Sample 1	-0.052558	0.110737
Sample 1 \times Sample 3	-0.006482	0.116412
Sample 2 \times Sample 3	-0.149572	-.17449

Our main goal is the assessment of an individual being on-track or off-track during the course of treatment. We derived the predicted interval equation as a function of the baseline measure and time using the formula of Neter, Wasserman, and Kutner (1989, p. 246) for regression models. Power is the probability of rejecting the null hypothesis when it is in fact false. Common desired power levels are between 80% and 90%. In this setting, we are interested in the probability of correct classification of an off-track individual; therefore, power levels can be used as guidelines for level specification for the prediction interval. For example, when the prediction interval level is set to 80%, we want at least 80% correct classification of an off-track subject. In developing a predictive equation, validation is a necessity to ensure our recovery model is reproducible. We produced two random samples where we derived the recovery curves for each individual within each sample. To validate our model, we compared the classification of off-track versus on-track patients using the equation from the development sample and the validation data within our validation sample. We used the κ coefficient to measure the amount of agreement beyond chance (Landis & Koch, 1977). We expect a very high κ to show validation of the on-track and off-track classification groups. We took two additional validation steps: (i) we calculated RCI and calculated receiver operating characteristic (ROC) curves based on the on-track flag created by the predictive equation and (ii) we reported the absolute agreement obtained based on the RCI cut scores, based on Youden's (1950) index (\mathcal{J}), which is a summary statistic that defines the optimal cut point between the true-positive rate and the false-positive rate.

Results

In study 1, after obtaining two random samples of 400 each from the psychiatric IOP, we initially derived two prediction equations, one from each sample, to determine those who were on-track and not on-track (see Table II for b coefficients). To cross-validate the prediction equations, we applied the equation derived from the depressed derivation sample (Sample 1) to the depressed validation sample (Sample 2), and vice versa. When the equation derived from Sample 1 was run with the patient weekly depressive scores from Sample 2, we obtained a $\kappa^2 = .97$, 95% CI [.93, 1.00] or near perfect agreement of patients as being on or off-track (Landis & Koch, 1977).³ This equation flagged 70 (17.5%) patients as off-track and there were only four patients for whom the models did

not agree. Similarly, when the equation obtained from the second sample was applied to the first sample, $\kappa = .96$, 95% CI [.92, .99]; 68 (17%) patients were flagged as off-track and 5 were not agreed upon. It thus appears that both equations had very high agreement in classifying patients as on or off-track.

We next calculated RCIs (Jacobson & Truax, 1991) using the average depression scores [i.e., $(M_{\text{pre-treatment}} - M_{\text{post-treatment}})/\text{SEM}$; test-re-test $r = .76$]. Using $\text{RCI} \geq 1.96$, 66.3% ($n = 265$) and 65.3% ($n = 261$) of Samples 1 and 2, respectively, met or exceeded this cut-off for reliable improvement. When the RCI criterion for reliable improvement was lessened to ≥ 1.65 (90%) or ≥ 1.28 (80%) (Wise, 2004), 68% ($n = 272$) and 74.5% ($n = 298$) of Sample 1 met or exceeded these cut-offs and 66.3% ($n = 265$) and 73.5% ($n = 294$) of Sample 2 met or exceeded these cut-offs. Using the least stringent RCI deterioration cut-off of ≥ -1.28 , only 2.8% ($n = 11$) and 2.5% ($n = 10$) of Samples 1 and 2, respectively, were reliably worse following treatment. Finally, 22.8% ($n = 91$) of Sample 1 and 24% ($n = 96$) of Sample 2 were classified as indeterminants using the RCI criterion between ≤ -1.28 and ≥ 1.28 .

ROC analyses for the equations were then conducted on Samples 1 and 2, plotting continuous RCI scores and the on-track/off-track flag obtained from the equations mentioned above. We obtained areas under the curve (AUC) of .91 and .89 for using the on-track variable for Samples 1 and 2, respectively. The best cut-points to maximize sensitivity and specificity were $\text{RCI} = 1.60$ ($\mathcal{F} = .72$; sensitivity = .84, specificity = .88; 95% CI [.87, .94]) and 1.71 ($\mathcal{F} = .70$; sensitivity = .84, specificity = .86; 95% CI [.86, .93]) for Samples 1 and 2, respectively.

In study 2, we applied the equation derived from Sample 1 to a random group derived from the depressed dual diagnosis IOP described earlier (Sample 3; $n = 198$). These patients were also previously shown to be primarily depressed (Wise, 2010), but as is evident in Table I, they are engaged in addictive behavior(s), with 98% involved specifically in some form of substance abuse. In this comparison, $\kappa = .55$, 95% CI [.37, .73]; only 13 (7%) were flagged as off-track and 18 (9%) were inconsistently classified. When we applied the equation derived from Sample 2 to Sample 3, the results were identical to those obtained with equation from Sample 1 (i.e., $\kappa = .55$, 95% CI [.37, .73]); 13 (7%) flagged as off-track and 18 (9%) were inconsistently classified. It thus appeared that neither of the equations based on the depressed psychiatric patients generalized to the dually diagnosed group, despite their high comorbidity of depression.

For the IOP depressed patients, the equation derived from Sample 1 has a slightly higher κ and

Table III. Prediction of flag \times RCI ≥ 1.60 cut score.

Flag	RCI		Total
	Off-track	On-track	
Off-track	60 (15.0%)	12 (3.0%)	72
On-track	50 (12.5%)	278 (69.5%)	328
Total	110	290	400

AUC compared to the equation derived from Sample 2. Consequently, the equation derived from Sample 1 was chosen as the better model for future predictions of the depression scores for the psychiatric IOP patients. Since the best cut point to maximize sensitivity and specificity for this equation was an RCI score of 1.60, we used this as a cut-off and derived Table III. Table III shows that of those who are on-track ($\text{RCI} \geq 1.60$), 95.86% would be correctly classified. Similarly, the positive predictive value (PPV) of the flag for those who are classified as off-track was 83.33% and the negative predictive value (NPV) of the flag for those on-track was 84.75%. That is, of those identified as off-track ($n = 72$), 83.33% were correctly identified and of those identified as on-track ($n = 328$), 84.75% were indeed responding positively to treatment. However, of those who were off-track ($\text{RCI} < 1.60$), only 54.54% would be correctly identified. Fortunately, only 15% of the sample ($n = 72$) were identified as off-track. With an NPV of 85%, we can be fairly certain that those who are identified as on-track are responding to treatment as expected. However, the flag misclassifies 17% ($n = 12$) as off-track, which appears to be a relatively acceptable error rate, in light of the fact that it allows us to correctly identify 83% ($n = 60$) as off-track. In fact, the overall efficiency index (hit rate) for the correct classification of all patients is 85%.

Table IV. Depression improvement summary by samples.

	Sample		
	1	2	3
N	400	400	198
Mean (SD) depression score			
Pre-treatment	2.43 (1.02)	2.51 (1.04)	1.92 (1.24)
Post-treatment	1.02 (0.94)	1.11 (0.93)	0.76 (0.84)
Effect size	1.30	1.24	1.07
RCI number (percent)			
$\geq 1.28^a$	298 (74.5)	294 (73.5)	92 (46.5)
≤ -1.28	11 (2.8)	10 (2.5)	10 (5.0)

^aRCI ± 1.28 corresponds to an 80% CI of improvement or deterioration.

An inspection of **Table IV** provides some illumination on these results. As expected, the psychiatric samples showed higher average pre-treatment depression scores compared to the dual diagnosis sample. While we expect baseline severity to affect recovery rates, it should not influence the accuracy of the prediction model and classification accuracy rates. **Table IV** also shows that although all three samples had large effect sizes (Cohen, 1988), the dual diagnosis samples had lower RCI improvement rates. In fact, a one-way ANOVA of the change scores was significant ($F(2/995) = 3.828, p = .022$). The Newman-Keuls post hoc test showed that Samples 1 and 2 had similar changes ($M = 1.41, SD = 1.08$ and $M = 1.39, SD = 1.12$, respectively), and both were significantly higher than Sample 3 ($M = 1.15, SD = 1.08$).

Discussion

Two of the major findings from this study are: (i) it is possible to derive equations for relatively homogeneous patient groups that can identify patients who are off-track using relatively small sample sizes and (ii) equations derived from such groups may not generalize to other groups of patients. In contrast to Lambert's (2010) report that "... what matters most is how disturbed a patient is—not which disorder they have or if they have more than a single diagnosis—because these variables correlate with degree of disturbance" (p. 104), our findings suggest that with smaller, homogeneous sample sizes, diagnoses and comorbidities made a difference. This was evident in comparing the groups of depressed and dual diagnosis patients with respect to their means, standard deviations, RCIs, ESs, and change scores (**Table IV**).

Our findings are also consistent with Molenaar and Campbell (2009), who argue against attempts to predict change in heterogeneous psychiatric populations due to inter-individual variation and instead promote an idiographic approach to predicting change. In fact, they call for sample homogeneity to understand and ultimately predict intra-individual behavior. The non-generalizability of both equations from the depressed IOP patients to the depressed dual diagnosis patients, both relatively homogeneous samples, are consistent with their "person specific paradigm" (p. 112) and reflect a benefit of a sample-specific approach.

This study conceptually replicates the client-specific predictive model provided by Gallop, Conolly-Gibbons, Mack, and Crits-Christoph (2013). RCI scores were calculated on the depression symptom measure to provide an estimate of

improvement to validate the classifications of the predictive equation. RCI scores serve as a benchmark to not only facilitate the identification of improvement rates but also identify those who are not responding to treatment as expected. Using RCI classifications, approximately 66–75% of both samples reliably improved, 23–24% were indeterminant, and only 1–3% deteriorated. Using the RCI criteria, these studies demonstrate that relatively accurate treatment prediction equations can be developed for specific groups of patients with relatively small sample sizes, using intake level of symptom severity and weekly response to treatment as predictor variables.

Additionally, the predictive equations were statistically and clinically reliable within these relatively homogeneous samples. Using the most liberal RCI cut-off (1.28) allowed the identification of approximately 23–24% of the samples who were classified as indeterminant, and 2–3% were classified as reliably deteriorated. Hence, being able to identify the roughly 75% of patients who are reliably changing would allow us to deploy additional treatment resources to those in the lower quartile who are not responding as expected (including those classified as indeterminant) and those who are getting worse (deteriorating). While identifying approximately 25% of the sample as indeterminant or deteriorated may appear excessive, it should be noted that historically fewer than 5% of our patients have deteriorated, but 36–42% have been classified as indeterminant, based on RCI depression scores (Wise, 2003, 2005, 2010). In fact, the indeterminant classification group is typically the second largest RCI group and a primary challenge is to identify these and facilitate their movement into the improved group. Therefore, given the small number of patients from Sample 1 who deteriorated (3%), the majority of patients misclassified as off-track are actually from the indeterminant range (23%), not only the deteriorated group. By identifying patients in both the indeterminant and deterioration groups, we are able to expand the opportunity to potentially improve the outcomes of more patients. Our findings are consistent with Lambert's (2010, p. 104) review, which reported that the sensitivity for predicting those who deteriorate from treatment ranges from 80% to 100% and the overall correct classification rates for the OQ-45 and OQ-30 is approximately .80; both of these findings are consistent with our results. This indicates that the 80% cut-off tolerance interval will identify many of the individuals who are in the indeterminant range during the course of treatment, which in turn will aid providers in targeting these individuals for additional interventions.

It is also possible that there are multiple trajectories of change contained within these relatively

homogeneous samples. Stulz, Lutz, Leach, Lucock, and Barkham (2007) demonstrated that more severely impaired patients showed a distinct trajectory of change. Lutz, Stulz, and Köck (2009) analyzed a group patients diagnosed with Major Depression and found three distinct patterns of change. More recently, Nordberg, Castonguay, Fisher, Boswell, and Kraus (2014) replicated similar findings and highlighted the importance of both symptom severity and functional impairment on treatment outcomes. This growing body of research has identified different patterns of change associated with baseline severity symptom scores and functional impairments, in relatively homogeneous diagnostic groups, that were associated with different outcomes and lengths of treatment. This suggests, for example, that our indeterminate group could be composed of individuals with chronic or persistent symptoms, or of those who report lower initial symptom severity and functional impairment, and who subsequently experience only moderate improvement rates of change during treatment.

Similarly, Witkiewitz and Marlatt (2004) have enumerated a number of established variables that contribute to substance abuse outcomes, including depression, self-efficacy, expectancies, motivation, craving, quality of social support, and so forth. More recently, Zheng, Cleveland, Molenaar and Harris (2015) used an idiographic approach to study the roles of positive social experiences, craving, and negative affect in a comparatively homogeneous group of abstinent young adults treated for substance dependence in inpatient or addiction treatment centers for 3 or more months, engaged in 12-step recovery work. They analyzed the individual heterogeneity in the daily dynamic process of substance use and identified two subgroups with different recovery patterns. Given the comparatively lower baseline severity scores, accompanied by the comorbid substance abuse, one would expect to see different rates of change in the dual diagnosis sample compared to the depressed sample, as the comorbid sample is attempting to cope with two identified primary problems. Such findings emphasize the likelihood of multiple change trajectories among the dual diagnosis group that also suggest different intra-individual dynamic change variables (e.g., cravings, need for social support, coping with negative affect) compared to the depressed group, as evidenced by the effect size scores and significant differences between the change scores of the depressed and dual diagnosis groups.

Lambert's (2010) program of research has demonstrated the utility of providing feedback to clinicians and clients about their treatment status, based on

expected response to treatment. The statistical model presented here allows for the derivation, validation, and dissemination of individual predictive recovery curves derived from baseline ratings and helps to bridge the gap between science and practice by increasing the accessibility of predictive modeling. This preliminary study is essentially a proof of concept that demonstrates that the use of pre-post trajectory of change scores can be used to predict on-track or off-track status, in real time, week by week, based on patient end-point scores. Future study will be required to test the model in routine outcomes monitoring throughout treatment. In light of the fact that real-time feedback provided to clients and clinicians has been shown to improve treatment outcome, particularly among patients identified as off-track (e.g., Lambert & Shimokawa, 2011; Shimokawa, Lambert, & Smart, 2010), it is hoped that the development and dissemination of predictive models such as that presented here may facilitate the development of similar equations for specific practice settings as well as complimentary electronic clinical decision support tools for the provision of feedback to therapists and patients. We recognize that the current example is not quite as sophisticated as Lambert's (2010) exemplary multilevel model. The advantages of multilevel modeling are numerous and include accounting for the dependent nature of time-series data, the nested nature of clinical data, robustness for missing data, allowing for non-linear growth trajectories, and so on. However, as previously indicated, extremely large sample sizes are necessary to develop such models. We believe that multiple signal alarms (e.g., white, green, yellow, and red feedback) and additional data points could be added to the current model, if desired.

In light of the finding that the equations derived from the psychiatric IOP sample did not generalize to the dual diagnosis IOP sample, further work will be required to replicate this model with other disorders. Additionally, this could indicate that separate equations might be required for relatively homogeneous groups of patients. Further study involving the generation and cross-validation of predictive equations for dual diagnosis patients would be a logical extension of this work and could further clarify the need for population-specific predictive equations with relatively homogeneous samples. On the other hand, it could be that an equation for a group of depressed and anxious individual therapy patients, for example, might generalize to other similar settings. Nonetheless, while this statistical model can be tailored for specific patient populations, the fact that it requires a sample size large enough for a derivation and validation group could be barriers

for small practices, particularly if they are treating diverse patient groups. On the other hand, larger sample sizes, such as those found in group practices, could produce more generalizable equations. However, more diverse populations are more likely to demonstrate more heterogeneous trajectories of change, and may require additional sub-group specifications to delineate different patterns of change. Similarly, the model assumes linear change over time, which may not fit other clinical scenarios, such as longer time frames. Even within our sample, the model applies only to those patients who attended three or more IOP days and completed three or more repeated measures. Hence, the model does not inform us about dropouts or patients who choose to not start treatment. Additionally, the model does not account for different patterns of treatment response, but does provide a bridge to identify various trajectories of change, which could also be studied in the future. This model does demonstrate that actuarial treatment outcome prediction models and their corresponding expected trajectories of change are not limited to extremely large sample sizes. Specialty clinics, group practices, inter-disciplinary practices, and hospital-based programs, for example, could access and utilize this predictive equation, develop clinical decision support tools unique to their setting and population, and take the prediction of treatment outcomes one step closer to front line clinicians. The replication and validation of our conceptual model and predictive equation await further study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

- ¹ Cohen's $d = 1.68$ would be classified as a large effect size ($.20 =$ small, $.50 =$ medium, and $\geq .80 =$ large; Cohen, 1988).
- ² In the dual diagnosis sample, we originally drew a random sample of 200. However, we later discovered two cases in this sample that contained missing data and were eliminated resulting in one sample of 198 dual diagnosis IOP patients.
- ³ The κ coefficient is a correlation coefficient used to assess interrater reliability and has a range from 0 to 1.00, with larger values indicating better reliability. Landis and Koch (1977) consider values of 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect.

References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Crits-Christoph, P., Connolly, M. B., Gallop, R., Barber, J. P., Tu, X., Gladis, M. & Siqueland, L. (2001). Early improvement during manual-guided cognitive and dynamic psychotherapies predicts 16-week remission status. *Journal of Psychotherapy Practice and Research, 10*, 145–154.
- Delgado, J., McMillan, D., Lucock, M., Leach, C., Ali, S., & Gilbody, S. (2014). Early changes, attrition, and dose-response in low intensity psychological interventions. *British Journal of Clinical Psychology, 53*, 114–130.
- Derogatis, L. (2001). *Brief symptom inventory: Administration, scoring and procedures manual*. Minneapolis, MN: NCS Pearson.
- Derogatis, L. R. (1983). *SCL-90-R: Administration, scoring and procedures manual-II*. Towson, MD: Clinical Psychometric Research.
- Derogatis, L. R. (1993). *Brief Symptom Inventory (BSI): Administration, scoring and procedures manual* (4th ed.). Minneapolis: NCS Pearson, Inc.
- DiClemente, C., & Prochaska, J. (1998). Toward a comprehensive, transtheoretical model of change. In W. Miller & N. Heather (Eds.), *Treating addictive behaviours* (pp. 3–24). New York, NY: Plenum Press.
- 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.
- Gallop, R., Connolly-Gibbons, M. B., Mack, R., & Crits-Christoph, P. (October, 2013). *The development of client specific recovery curves for the community clinician feedback system*. Paper presented at Society for Psychotherapy Research, Memphis, TN.
- 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*, 12–19.
- Lambert, M. J. (2010). *Prevention of treatment failure*. Washington, DC: American Psychological Association.
- Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*, 159–172.
- Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okishi, J., & Burlingame, G. M. (1996). *Administration and scoring manual for the OQ-45.2*. Wilmington, DE: American Professional Credentialing Services.
- Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy, 48*, 72–79.
- Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G., & Finch, A. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy, 9*, 149–164.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
- Lutz, W., Stulz, N., & Köck, K. (2009). Patterns of early change and their relationship to outcome and follow-up among patients with major depressive disorders. *Journal of Affective Disorders, 118*, 60–68.
- Miller, W. R., & Rollnick, S. (2013). *Motivational interviewing: Helping people change* (3rd ed.). New York, NY: Guilford.
- Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychology, 18*, 112–117.
- Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models*. Homewood, IL: Irwin.
- Nordberg, S. S., Castonguay, L. G., Fisher, A. J., Boswell, J. F. & Kraus, D. (2014). Validating the rapid responder construct within a practice research network. *Journal of Clinical Psychology, 70*, 886–903.
- Renaud, J., Brent, D. A., Baugher, M., Birmaher, B., Kolko, D. J., & Bridge, J. (1998). Rapid response to psychosocial treatment for adolescent depression: A two-year follow-up. *Journal of the American Academy of Child and Adolescent Psychiatry, 37*, 1184–1190.

- Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and megaanalytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78*, 298–311.
- Spielmanns, G. I., Masters, K. S., & Lambert, M. J. (2006). A comparison of rational versus empirical methods in prediction of negative psychotherapy outcome. *Clinical Psychology & Psychotherapy, 13*, 202–214.
- Stulz, N., Lutz, W., Leach, C., Lucock, M., & Barkham, M. (2007). Shapes of early change in psychotherapy under routine outpatient conditions. *Journal of Consulting and Clinical Psychology, 75*, 864–874.
- Wise, E. A. (2003). Psychotherapy outcome and satisfaction methods applied to intensive outpatient programming in a private practice setting. *Psychotherapy: Theory, Research and Practice, 40*, 203–214.
- Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change and recommendations for future directions. *Journal of Personality Assessment, 82*, 50–59.
- Wise, E. A. (2005). Effectiveness of intensive outpatient programming in private practice: Integrating practice, outcomes and business. *American Psychologist, 60*, 885–895.
- Wise, E. A. (2010). Evidence based effectiveness of a private practice intensive outpatient program with dual diagnosis patients. *Journal of Dual Diagnosis, 6*, 25–45.
- Witkiewitz, K., & Marlatt, G. A. (2004). Relapse prevention for alcohol and drug problems: That was Zen, this is Tao. *American Psychologist, 59*, 224–235.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*, 32–35.
- Zheng, Y., Cleveland, H. H., Molenaar, P. C., & Harris, K. S. (2015). An alternative framework to investigating and understanding intraindividual processes in substance abuse recovery: An idiographic approach and demonstration. *Evaluation Review, 39*(2), 229–254. doi:10.1177/0193841X14567313