Journal of Consulting and Clinical Psychology

0022-006X/16/$12.00 http://dx.doi.org/10.1037/ccp0000083

Predicting Therapist Effectiveness From Their Own Practice-Based Evidence

David R. Kraus, Jordan H. Bentley, and Pamela C. Alexander
Outcome Referrals, Inc., Framingham, Massachusetts

James F. Boswell
University at Albany, SUNY

Michael J. Constantino
University of Massachusetts Amherst

Elizabeth E. Baxter
Outcome Referrals, Inc., Framingham, Massachusetts

Louis G. Castonguay
The Pennsylvania State University

Objective: Differences between therapists (therapist effect) are often larger than differences between treatments (treatment effect) in explaining client outcomes, and thus should be considered relevant to providing optimal treatment to clients. However, research on therapist effectiveness has focused largely on global measures of distress as opposed to a multidimensional assessment, and has failed to risk-adjust for client characteristics. The purpose of this study was to examine the stability and predictive validity of therapist effectiveness across multiple outcome domains using risk-adjusted outcomes. ***Method:*** Initial and follow-up outcome data on the Treatment Outcome Package (Kraus, Seligman, & Jordan, 2005) were collected on 3,540 clients who were treated in naturalistic settings by a sample of 59 therapists. After risk-adjusting outcomes based on case-mix variables using random forest models, outcome data from the first 30 clients of each therapist were used to classify each therapist's effectiveness on 12 outcome domains. These results were then compared with outcome data from the therapist's next 30 clients. ***Results:*** Results demonstrated that therapist effectiveness was relatively stable, although somewhat domain specific. Therapists classified as "exceptional" were significantly more likely to remain above average with future cases, suggesting that a therapist's past performance is an important predictor of their future performance. ***Conclusions:*** Clients are likely to experience differential benefit depending on the particular therapist and his or her strengths. Clinical outcomes may be improved by developing the best possible prediction model for each new client and then providing that client with referrals to therapists with well-matched strengths.

What is the public health significance of this article?
Therapist effectiveness in treating different domains of client functioning can be predicted from past performance, and using this actuarial information in clinical decision making holds promise for improving the percentage of clients who experience a positive treatment effect.

Keywords: therapist effectiveness, Treatment Outcome Package, risk adjustment, multilevel modeling, random forest

In their recent review of efficacy studies employing randomized controlled trial (RCT) designs, Baldwin and Imel (2013) estimated that approximately 5% of the variance in psychotherapy treatment outcomes is attributable to between-therapist differences. This effect is present despite the fact that therapists in most RCTs are required to adhere to specific treatment protocols and engage in

David R. Kraus, Jordan H. Bentley, and Pamela C. Alexander, Outcome Referrals, Inc., Framingham, Massachusetts; James F. Boswell, Department of Psychology, University at Albany, SUNY; Michael J. Constantino, Department of Psychological and Brain Sciences, University of Massachusetts Amherst; Elizabeth E. Baxter, Outcome Referrals, Inc.; Louis G. Castonguay, Department of Psychology, The Pennsylvania State University.

This research was funded in part by the Annie E. Casey Foundation (Grant No. 212.0369) and the Duke Endowment (Grant No. 1935-SP). We thank them for their support. We acknowledge that the findings and conclusions presented in this report are those of the authors alone and do not necessarily reflect the opinions of these foundations. David R. Kraus is the developer and owner of TOP.

Correspondence concerning this article should be addressed to David R. Kraus, Outcome Referrals, 1 Speen Street, Framingham, MA 01701. E-mail: dkraus@outcomereferrals.com

intensive uniform training and supervision (e.g., Blatt, Sanislow, Zuroff, & Pilkonis, 1996; Crits-Christoph & Mintz, 1991; Huppert et al., 2001; Kim, Wampold, & Bolt, 2006). Not surprisingly, the therapist effect is estimated to be even larger in naturalistic studies (7%), for which there is probably greater variability in therapist abilities and styles of treatment delivery. And across both highly controlled and naturalistic treatment settings, the estimate of therapist effect is larger than the percentage of outcome variance explained by between-treatment differences (i.e., differences that emerge when comparing protocol-driven treatment packages; Wampold & Imel, 2015). Thus, across populations of therapists and clients, the therapist effect has important implications for mental health care outcomes (Saxon & Barkham, 2012), especially if it can be shown that the effect is stable and predictable.

The fact that therapist effects have been found in both RCT and naturalistic research approaches should increase confidence in its reliability, as two different research approaches complement and compensate for each other's strengths and limitations (Castonguay, 2013). Naturalistic studies are more likely to contain threats to internal validity, yet can also yield more ecologically valid data. For example, in naturalistic outcome studies, complex comorbidity is the norm and treatment is applied idiographically (cf., Almlöv, Carlbring, Berger, Cuijpers, & Andersson, 2009; Almlöv et al., 2011; Cella, Stahl, Reme, & Chalder, 2011; Erickson, Tonigan, & Winhusen, 2012; Kim et al., 2006; Laska, Smith, Wislocki, Minami, & Wampold, 2013; Wiborg, Knoop, Wensing, & Bleijenberg, 2012). On the other hand, although there are typically fewer threats to internal validity in controlled research, there also are potential problems of generalizability and therapist allegiance to a given treatment despite being crossed by design (Falkenström, Markowitz, Jonker, Philips, & Holmqvist, 2013).

One potential concern for naturalistic outcome research is overreliance on a single global measure of distress as the outcome variable. In contrast, controlled trial research has a history of placing greater emphasis on multiple relevant treatment outcome indicators (Ogles, 2013), which allows researchers and therapists to examine differential patterns of response. For example, in a transdiagnostic treatment trial (Farchione et al., 2012), treatment effect sizes ranged between $g = 0.40$ and 1.39 depending on the outcome variable. This finding highlights the potential danger of relying solely on general indicators or a composite measure when examining outcomes or establishing benchmarks for therapist- or system-level performance (Dinger, Strack, Leichsenring, Wilmers, & Schauenburg, 2008; McAleavey, Nordberg, Kraus, & Castonguay, 2012; Saxon & Barkham, 2012; Wampold & Brown, 2005). A singular emphasis on global distress also ignores the recommendations of the Society for Psychotherapy Research (SPR) and American Psychological Association's (APA's) Core Battery Conference (Horowitz, Lambert, & Strupp, 1997), which recommended, among other things, that a core battery should (a) not be bound to specific theories, (b) show validity and appropriateness across all diagnostic groups, (c) measure subjective distress, (d) measure symptomatic states, (e) measure social and interpersonal functioning, (f) have general and clinical population norms to help discriminate between patients and nonpatients, (g) be highly sensitive to change, and (h) be easy to use and relevant to clinical needs.

A core battery with multidimensional assessment enhances the field's capacity to capture the complexity of clients' response to treatment, as well as the nuance of therapist effectiveness in treating different domains of client's presenting concerns. Therefore, it is not surprising that in one of the first cross-study reanalyses of RCTs exploring therapist effects, Crits-Christoph and Mintz (1991) found therapist effects as large as 39% on discrete, domain-specific outcome measures. Many, if not most, studies examining therapist effects have focused on global outcome indicators (e.g., distress total score) and, notably, have yielded smaller variance estimates when compared with Crits-Christoph and Mintz.

Consistent with the need for more naturalistic studies and domain-specific measures of outcome, Kraus, Castonguay, Boswell, Nordberg, and Hayes (2011) examined therapist effectiveness across a large sample of diverse clients seen in outpatient settings by therapists employing varied treatment approaches. The prevalence of "effective" and "harmful" therapists was estimated by analyzing the multidimensional pre–post treatment outcomes of nearly 700 therapists as measured by the Treatment Outcome Package (TOP; Kraus, Seligman, & Jordan, 2005). Outcomes were assessed across 12 domains, including quality of life, functional outcomes (e.g., work functioning), and symptomatic domains (e.g., depression). Therapists were classified based on whether, on average, clients reliably improved ("effective"), worsened ("harmful"), or showed no change ("unclassifiable or ineffective"). Results varied by problem/symptom domain, with widespread pervasiveness of unclassifiable, ineffective and harmful therapists. Although 96% of therapists had at least one area in which they were effective, this percentage evidenced at least one area in which they were ineffective or harmful. In fact, there was a range of 33% to 65% of therapists classified as ineffective or harmful across the 12 TOP domains. Relatively small correlations were observed between domains within the same therapist, suggesting that therapists may possess domain-specific competence rather than a general competence. On the other hand, general and domain-specific competencies are not necessarily mutually exclusive. In fact, many in the field take this for granted, as exemplified by therapists who have a general credential but then also pursue credentialing in specific practice domains.

Irrespective of the type of outcome used, however, the current state of research strongly suggests that therapist differences are highly relevant to any effort aimed at providing clients with optimal treatment (Boswell, Constantino, Kraus, Bugatti, & Oswald, 2015; Lambert, 2010). The practical implications of therapist differences have, thus far, been largely unaddressed. The field's acknowledgment of the prevalence and meaningfulness of therapist differences raises the critical question of how knowledge of such differences translates into improving patient care, professional training, and service delivery systems. For example, providing the public access to therapist effectiveness data (i.e., performance "report cards") might be a highly impactful practice implication of assessing, valuing, and attending to therapist differences (Boswell et al., 2015). This notion would seem to fit within the current health care climate of accountability. Health care systems are increasingly emphasizing outcome assessment and performance-based payment models (Bremer, Scholle, Keyser, Knox Houtsinger, & Pincus, 2008; Institute of Medicine Committee on Redesigning Health Insurance Performance Measures, Payment, and Performance Improvement Programs, 2007; Scanlon, Lindrooth, & Christianson, 2008). The fairness and utility of systems- and therapist-level performance data-driven decision making, however, rests on the stability and predictive capacity of performance estimates. For example, stakeholders need to be confident that a therapist who has been labeled "effective" based on a demonstrated outcome track record is likely to demonstrate positive outcomes with future clients (Wampold & Brown, 2005).

Outcome simulations have suggested that high-performing therapists would achieve good outcomes on 80% of their cases, whereas underperforming therapists would achieve good outcomes on only 20% of their cases (Wampold & Imel, 2015). However, only a few studies have prospectively examined the stability of a therapist's observed effectiveness across subsequent cases. Wampold and Brown (2005) used a unidimensional measure to examine the stability of therapists' client outcomes. When treating successive cases, therapists in the top quartile (as benchmarked against their peers) had between 7% and 13% more of their clients reliably improve than did therapists in the bottom quartile. These results provide some evidence that past therapist performance predicts future performance. What has not yet been studied, however, is the stability of a therapist's domain-specific outcomes. With such information, referral systems could be developed that match each client's needs to a subset of therapists who have an empirically documented successful track record at treating those issues.

In order to improve estimation precision and enhance decision-making utility, another critical consideration is risk adjustment, which assumes that certain client (case-mix) characteristics may themselves predict outcomes. Therefore, any assessment of therapists must control for the effects of these client characteristics on outcomes, especially because these characteristics are unlikely to be randomly distributed across therapists in real-world settings. Measures of risk adjustment for mental health populations have typically included sociodemographic data as well as baseline mental health severity (Rosen et al., 2010). Moreover, the inclusion of physical health data, stress indicators, and comorbid mental health disorders greatly increases the amount of variance explained in outcomes (Hermann, Rollins, & Chan, 2007; Jones et al., 2004; Raghavan, 2010). Risk adjustment is thus relevant to assessing therapist effectiveness, and the absence of this was a notable limitation of the Kraus et al. (2011) study as well as other large, naturalistic studies of therapist effects. In contrast, Saxon and Barkham (2012) incorporated client severity into their models and found that the between-therapist effect on global outcomes ranged from 1% at very low levels of initial client severity to more than 10% at higher levels of severity. As they described it, "the more severe a client's intake symptoms, the more his or her outcome depended on which therapist he or she saw" (p. 542).

Conceivably, statistically adjusting for initial client characteristics might lower the magnitude of the estimated therapist effect. This is because such risk adjustment could account for variance that would otherwise be attributed to differences in therapist skill but is actually related to the fact that some therapists see more complicated-to-treat clients. For example, in one non-risk-adjusted model, 7.8% of the variance was attributed to the therapist; however, when initial client risk-of-harm scores were added to the risk-adjustment model, the estimate dropped to 6.6% (Saxon & Barkham, 2012). In other words, more than 1% of the variance that would have been attributed to the therapist appeared to be an artifact of differences in clients seen by each therapist. An assessment of client characteristics, or case-mix risk adjustment, thus appears to be necessary to obtain accurate assessments of therapist effectiveness within and between therapists' caseloads.

The purpose of this study was to extend previous research by (a) examining the stability and predictive validity of therapist effectiveness across multiple outcome domains, and (b) using risk-adjusted outcomes (controlling for client characteristics) before analyzing the therapist effect. This extension of previous findings may move the field closer to models of therapist effectiveness that could be employed fairly and effectively in key areas of mental health care decision making and practice (e.g., client referrals, continuing education).

Method

Participants

Similar to Kraus et al. (2011), an archival data set of de-identified naturalistic outcomes was mined for analysis. All patients were assessed as part of routine care and provided informed consent, allowing their self-reported, de-identified data to be used in research. From this archival data, a sample of therapists was identified who had treated at least 60 adult clients with baseline and follow-up outcome data. The data set included 3,540 clients and 59 therapists. The client sample was predominantly female (55%), with an average age of 38.1 years ($SD = 12.2$), an average of 11.7 years of education ($SD = 3.3$), and race/ethnicity that was predominantly European American (80%), with an additional 5% African American, 7% Hispanic, 1% Asian American, and 1% other race/ethnicity (additional participants chose not to answer this item). Clients from low-income households were overrepresented in the sample, with 62% of clients having household incomes of $20,000 or less. Therapists were also predominantly female (59%), with an average age of 33.25 years ($SD = 10.7$), and race/ethnicity that was 68% European American, 5% African American, 14% Hispanic, 8% Asian American, and 5% other race/ethnicity. Therapists, who had an average of 10.2 years ($SD = 7.5$) of postlicensing experience, were social workers (46%), mental health counselors (26%), psychologists (20%), drug and alcohol counselors (5%), and those with another type of professional license (e.g., psychiatrist; 3%).

All treatment was individual psychotherapy. Twenty-nine (49%) of the therapists worked in traditional outpatient therapy service settings like community mental health centers and independent practice. The rest delivered services in milieu treatment settings like hospitalization, residential, and day-treatment programs. No differences in outcomes were detected by treatment setting.

Outcome Measure

Patients completed the TOP, a routine outcome assessment tool designed for clinical and research purposes in naturalistic settings (Kraus et al., 2005). Developed to meet the criteria established by the SPR- and APA-sponsored Core Battery Conference (Horowitz et al., 1997), TOP assesses a wide array of behavioral health symptoms and functioning, demographics, and risk-adjustment (case-mix) variables. The clinical scales consist of 58 items that assess 12 symptom and functional domains: Work Functioning, Sexual Functioning, Social Conflict, Depression, Panic (Somatic Anxiety), Psychosis, Suicidal Ideation, Violence, Mania, Sleep, Substance Abuse, and Quality of Life. Psychometric studies have provided support for the TOP's reliability and construct validity in both adults (Kraus et al., 2005) and children (Kraus, Boswell, Wright, Castonguay, & Pincus, 2010). In addition, the TOP has demonstrated excellent sensitivity to change, with 50% of clients demonstrating reliable improvement (Jacobson & Truax, 1991) on single subscales, 91% demonstrating reliable improvement on at least one of the 12 domains, and 67% demonstrating reliable deterioration on at least one subscale (Kraus et al., 2005). In addition to the 12 outcome domains, TOP assesses demographic, medical, and life-stress data, providing a resource for multidimensional outcome assessment with risk adjustment.

The TOP collects information on numerous case-mix variables, thus providing a unique opportunity to risk-adjust outcomes. In a 1999 project for The Joint Commission (the primary hospital-accrediting body in the United States), a sample of more than 14,000 clients was administered pre- and posttreatment TOP measurements. Items available for inclusion as independent variables in a regression-based model included any data collected by the TOP system on the client at intake (e.g., age, ethnicity, education level, income), information collected on the provider or service program upon registering to use the TOP data collection system (years of experience, treatment setting), and select information about the duration of treatment including length of treatment and number of sessions. Stepwise regression analyses were conducted to predict follow-up TOP scores for each domain. In all cases, each domain's initial severity score accounted for the largest amount of postoutcome score variance, typically followed by level of comorbid medical issues and life stress scores. This initial risk-adjustment model (1999 model) was used as a baseline in the current study to compare model improvements. The substance abuse domain had not been developed at this point and was not included in the model.

Procedure

Either the therapist or clinic involved in the data collection had previously contracted to collect assessment and outcome data on all clients as part of routine care. As part of their consent for services, clients were told that identifiable data would be used by their therapist to better understand their issues and needs for treatment, and that repeat assessments would be used to conjointly monitor progress toward mutually developed goals. Clients and therapists were also told that de-identified, aggregated data could be used for research.

Data Analyses

Analyses were based on a sample of 59 therapists who had each treated at least 60 clients with a TOP assessment at the beginning of care and a follow-up outcome assessment between 30 and 180 days later. We chose a 30-day minimum follow-up period in order to ensure a reasonable treatment dose and because early response research indicates that change in the first 4 to 5 weeks (or sessions) of treatment is a significant predictor of longer term outcomes (e.g., Lutz et al., 2014). We analyzed the outcome data from the first 30 clients of each therapist in the sample, and these scores were used to classify each therapist's relative effectiveness on each of the 12 TOP domains. We then compared and contrasted these results with outcome data from the next 30 clients of the same therapists. At each step, we used risk-adjusted outcome data that accounted for client variables, such as initial severity and life stress.

In an effort to build a more robust model than the 1999 model, we utilized a random forest (Breiman, 2001) machine-learning algorithm based on the case-mix variables collected on the TOP. Although random forests are not widely used in clinical psychology research, it has been recommended that they be used for forecasting outcomes (King & Resick, 2014). A random forest creates a large number of regression trees from randomly selected subsets of the "training" data and combines them into an ensemble model. This approach has several advantages over linear regression. Random forests are nonparametric, and, as such, are not limited by linear or even curvilinear relationships between the risk-adjustment variables and the independent outcome score. Random forests can also account for multiple interaction effects, which do not need to be specified beforehand. There is less risk of overfitting, and random forest methods do not suffer a loss of performance when weak predictors are included as inputs (Breiman, 2001; Strobl, Malley, & Tutz, 2009).

A risk-adjustment training sample of 27,045 clients with archived data was used to construct the model in R (R Core Team, 2012) using the random forest package (Liaw & Wiener, 2002). Because random forests contain an internal measure of variance explained (out-of-bag estimation), which is not prone to overfitting or inflated by degrees of freedom, a separate validation sample was unnecessary to estimate the variance explained in the training data. Similar to the findings of Saxon and Barkham (2012), we expected that the risk-adjustment models would reduce the variance attributed to the therapist.

We conducted three primary analyses (see analyses A, B, and C), and in each case the random forest model was used to create an expected outcome score for each client. If a therapist was able to achieve an outcome better than what would be expected based on risk adjustment for the client, this difference score was attributed to a positive therapist effect. If a therapist achieved an outcome worse than the prediction, this difference was attributed to a negative therapist effect. Analyses A, B, and C were performed on each therapist's risk-adjusted outcomes.

Analysis A: Classification analysis. Using the procedures outlined by Saxon and Barkham (2012), we calculated confidence intervals for therapists who were then classified as either *exceptional*, *average*, or *below average* based on whether the 86% CI of the mean difference of a therapist's actual outcomes and those predicted by the random forest risk-adjustment predictions crossed zero. That is to say, if the mean difference for a therapist was significantly positive, the therapist was considered to be *exceptional*; significantly negative mean differences indicated a *below average* therapist, and a lack of significance resulted in a designation of *average*. A chi-square test was used to assess whether the classification of therapists based on their first 30 cases remained consistent for their next 30 cases.

Analysis B: Mean difference analysis. Arguably, Analysis A (classification analysis) transforms a continuous variable—therapist effectiveness—into a categorical variable for the sake of simplicity. So as not to lose the specificity in the underlying construct, Analysis B assessed the correlation (Pearson's r) between the average risk-adjusted outcomes of each therapist's first

30 clients compared with the average risk-adjusted outcomes of their second 30 clients.

Analysis C: Hierarchical linear modeling (HLM) analysis. In order to account for the dependency typically found in nested data structures, as well as uneven spacing of measurement, Analysis C used HLM and a fixed slope assumption (see Wampold & Brown, 2005) to determine an intercept score for each therapist's first and second group of 30 clients. These scores were then compared using Pearson's *r*.

Results

Risk Adjustment

The first goal was to risk-adjust the data to minimize the effect of differential distribution of case mix variables across therapists. These risk-adjusted algorithms provided an estimate of the variance explained by client characteristics. The amount of variance explained in the training sample of 27,045 is presented in Table 1. The 1999 model rarely explained more variance than the initial domain score alone. In contrast, the newly built random forest model was able to explain substantially more variance, ranging from 1.8% more variance for work functioning to 11.1% more variance for quality of life. Furthermore, unlike the 1999 model, the random forest model was intentionally limited to only initial client characteristics and the duration of treatment. For example, the 1999 model included information on events involving the patient, such as additional life stressors that occurred after the initiation of treatment. It was decided to eliminate these midtreatment life stressors from the new model, as they do not help in a prospective, predictive analysis of therapist skill. By contrast, they may be important sources of additional, explainable variance in retrospective analyses of variance. As such, the random forest model was shown to be superior to the 1999 model.

Although random forest models excel at making use of predictors that are overlooked by other techniques, they can be difficult to interpret (King & Resick, 2014). Inspection of the node split purity, a measure of feature importance in random forest models, showed that the initial score for the domain under question was the most significant variable for all models. Other important features included age, health, employment, and severity on other problem domains. Gender and ethnicity were of low importance for all domains except substance abuse.

Table 1
Variance in Outcomes Explained by Client Characteristics When Treated by Average Therapist

TOP domain	Intake score[a]	1999 model[a]	Random forest model[b]
Sexual Functioning	28.6%	26.7%	30.8%
Work Functioning	16.6%	13.5%	18.4%
Violence	23.8%	22.4%	26.7%
Social Functioning	20.5%	24.9%	24.7%
Panic/Anxiety	34.0%	36.6%	40.4%
Substance Abuse	26.9%	N/A	30.7%
Psychosis	34.9%	34.2%	40.5%
Quality of Life	21.3%	20.3%	32.4%
Sleep	32.6%	32.2%	39.8%
Suicidality	26.9%	20.5%	32.1%
Depression	33.3%	38.0%	42.1%
Mania	18.6%	17.1%	21.5%
Total score	32.8%	N/A	44.2%

Note. TOP = Treatment Outcome Package; N/A = not applicable.
[a] Variance explained calculated by ANOVA R^2. [b] Variance explained calculated by built in out-of-bag random forest variance measure.

Analysis A (Classification Analysis)

Results from the final 30 × 30 (first 30 clients by next 30 clients) sample are presented in Table 2. Classifications demonstrated increased stability with a larger number of cases. All chi-square tests were highly significant except for the Mania and Psychosis domains (Psychosis, $p = .08$). For this analysis, *exceptional* was defined by a significant *t* test of the therapist's risk-adjusted outcomes compared with the *average* therapist. *Above average* was defined as being above the 50th percentile in sample rank. Table 3 presents the percentage of therapists who were classified as *exceptional* with their first 30 clients, which ranged from 8% for Mania to 46% for Depression. Table 3 also presents the percentage of those therapists who were classified as *exceptional* with their first 30 clients who remained at least above average (above the 50th percentile) with their next 30 clients. These results ranged from 40% for Mania to 91% for Substance Abuse.

Analyses B (Mean Difference Analysis) and C (HLM Analysis)

Significant correlations were found between the single-level risk-adjusted outcomes (Analysis B) and the results from the HLM analyses with risk adjustment (Analysis C), and are reported in Table 4. The results were very similar in both sets of analyses and generally provided support for past performance predicting future performance in a particular domain. A therapist's effectiveness was much less stable, however, in treating Mania and Violence than the other domains, with correlations ranging from $r = .53$ for Sexual Functioning (Analysis B) to .94 for Substance Abuse in the multilevel model analysis.

Other Results

Table 5 presents the primarily moderate, positive correlations between therapists' outcome rankings for each domain. Table 6 presents the frequency distributions of therapists' total above-average and below-average domains. Approximately two thirds of therapists had four or fewer above-average domains. If excluding the Mania domain, which exhibited difficulty in classifying therapists, 9% ($n = 5$) of therapists were above average on all 11 remaining domains. All five of these "super shrinks" remained above average on all domains in treating their next 30 clients. In contrast, no therapist was found to be below average on more than eight domains, and most (57%) had no below-average domains.

Finally, Table 7 presents an overall estimate of the variance explained at the client and therapist levels, contrasting the various methods presented by Wampold and Brown (2005), Saxon and Barkham (2012), and this study. The risk-adjusted variance percentage in each pair of columns was calculated using ANOVA to generate an R^2 value between the risk-adjusted projections and

Table 2
30 × 30 Results: Classification Stability of Therapists' Risk-Adjusted Outcomes

TOP domain	Validation sample classification (based on criterion sample)	E	A	B	χ^2
Sexual Functioning	E	10	4	0	14.81**
	A	7	28	2	
	B	1	6	0	
Work Functioning	E	9	8	0	10.99*
	A	7	21	4	
	B	0	7	2	
Violence	E	12	6	0	22.02**
	A	11	27	1	
	B	1	0	1	
Social Functioning	E	7	9	0	13.51**
	A	8	27	5	
	B	0	1	2	
Panic/Anxiety	E	10	8	0	11.51*
	A	10	22	3	
	B	0	4	2	
Substance Abuse	E	12	11	0	53.73**
	A	6	23	0	
	B	0	1	5	
Psychosis	E	9	8	0	8.41 (p = .08)
	A	9	25	3	
	B	0	5	0	
Quality of Life	E	14	4	1	24.97**
	A	11	16	4	
	B	0	3	6	
Sleep	E	13	5	0	25.34**
	A	8	26	4	
	B	0	1	2	
Suicidality	E	13	10	1	20.45**
	A	7	17	3	
	B	1	2	5	
Depression	E	19	7	1	31.01**
	A	4	13	3	
	B	0	5	7	
Mania	E	0	5	0	4.23 (p = .38)
	A	9	31	5	
	B	0	8	1	
Total score	E	14	6	1	26.48**
	A	8	17	4	
	B	0	6	8	

Note. Rows represent the first 30 patients; columns represent the next 30 patients. E = exceptional; A = average; B = below average.
* $p < .05$. ** $p < .01$.

actual outcomes. The therapist variance explained percentage was calculated as the R^2 between the predictions made by the HLM model and the actual outcomes, from which we then subtracted the risk-adjusted variance explained percentage. Risk adjustment with only intake score mirrored the analyses conducted by Wampold and Brown (2005), and the risk adjustment with intake and risk scores mirrored the analyses conducted by Saxon and Barkham (2012). The potential overestimation of the therapist effect using these methods can be seen by contrasting these results with the final set of columns using the full random forest model. In sum, these results indicate that a therapist's domain-specific performance is stable and predictive of future performance in that domain. In addition, the application of risk adjustment appears to result in more precise estimates of the variance accounted for by the therapist, yet the magnitude of the therapist effect remains significant and meaningful.

Table 3
Percentage of Exceptional Therapists That Remain Above Average

TOP domain	Therapists classified as exceptional (first 30)	Exceptional therapists who remained above average[a] (second 30)
Sexual Functioning	24%	79%
Work Functioning	29%	76%
Violence	31%	61%
Social Functioning	27%	81%
Panic/Anxiety	31%	83%
Substance Abuse	40%	91%
Psychosis	29%	71%
Quality of Life	32%	79%
Sleep	31%	89%
Suicidality	41%	79%
Depression	46%	78%
Mania	8%	40%
Total score	33%	76%

Note. TOP = Treatment Outcome Package.
[a] Above average = above the 50th percentile.

It should also be noted that no differences in outcomes were detected by treatment setting. And when comparing degrees of effectiveness across clinician licensing type and years of experience, no significant results were found after applying the Bonferroni correction. By contrast, length of treatment was a significant risk-adjustment variable and was accounted for in the random forest model. The mean length of treatment was 80 days, with a range from 30 to 180. Analyzed independently, length of treatment accounted for 1.14% of the outcome variance.

Discussion

The goals of this study were to examine the influence of risk adjustment on treatment outcome prediction, its implications for estimating therapist effects, and the stability of observed therapist

Table 4
Pearson's Correlation Between Risk-Adjusted Outcomes in Criterion and Validation Samples

TOP domain	Analysis B[a] (mean diff.)	Analysis C[b] (HLM)
Sexual Functioning	.531	.547
Work Functioning	.593	.575
Violence	.317	.374
Social Functioning	.643	.643
Panic/Anxiety	.579	.572
Substance Abuse	.924	.938
Psychosis	.593	.595
Quality of Life	.859	.863
Sleep	.687	.700
Suicidality	.682	.683
Depression	.806	.811
Mania	.259	.259
Total score	.847	.850

Note. TOP = Treatment Outcome Package; diff. = difference; HLM = hierarchical linear modeling.
[a] Analysis B = Pearson's r between average risk-adjusted outcomes of first and second groups of 30 clients. [b] Analysis C = hierarchical linear modeling-derived Pearson's r between intercept scores for first and second groups of 30 clients.

Table 5
Correlation (Kendall's Tau-B) Between Risk-Adjusted Therapist Rankings by TOP Domain

TOP Domain	SEXFN	WORKF	VIOLN	SCONF	PANIC	SA	PSYCS	LIFEQ	SLEEP	SUICD	DEPRS	MANIA	TOTAL
SEXFN	1.00												
WORKF	.39	1.00											
VIOLN	.28	.47	1.00										
SCONF	.54	.52	.40	1.00									
PANIC	.60	.43	.30	.51	1.00								
SA	.34	.30	.36	.33	.34	1.00							
PSYCS	.46	.56	.46	.49	.57	.42	1.00						
LIFEQ	.54	.47	.44	.55	.57	.50	.59	1.00					
SLEEP	.50	.39	.49	.50	.57	.51	.57	.68	1.00				
SUICD	.58	.50	.45	.55	.59	.46	.66	.68	.59	1.00			
DEPRS	.58	.46	.38	.56	.62	.47	.64	.76	.69	.75	1.00		
MANIA	.07	−.05	.07	.06	.06	−.04	−.09	−.08	−.01	−.07	−.09	1.00	
Total	.53	.43	.35	.49	.56	.45	.58	.71	.64	.66	.75	−.09	1.00

Note. $N = 59$. Based on 30 × 30 HLM rankings. TOP = Treatment Outcome Package; DEPRS = Depression; LIFEQ = Quality of Life; PSYCS = Psychosis; SA = Substance Abuse; SCONF = Social Conflict; SEXFN = Sexual Functioning; SUICD = Suicide; VIOLN = Violence; WORKF = Work Functioning; HLM = hierarchical linear modeling.

effectiveness (or ineffectiveness) over time. Results suggested that risk adjustment is clearly important when estimating therapist effects. This research supports the conclusion that without sufficient risk adjustment, the unequal distribution of client characteristics will lead to an overestimation of the therapist effect. In fact, enough variance is explained by using only initial outcome scores to risk-adjust outcomes that it could explain all or most of the "therapist effect" found in some other large naturalistic studies. Importantly, this study employed a relatively robust risk adjustment model that included multiple dimensions of functioning, symptom severity, and quality of life, yet we were still able to document a substantial therapist effect. Indeed, the therapist effect remained meaningful even after extensive risk adjustment and accounting for multiple sources of variance with multilevel models.

Our results also demonstrated that therapist effectiveness could be predicted from past performance. Exceptional therapists tended to remain in the top 50th percentile in their outcomes with subsequent clients. Predicting a therapist's future performance was robust in a number of TOP domains. For example, a therapist's past performance in treating substance abuse could be used to make an extremely accurate prediction of future clients' substance abuse outcomes. Use of past performance to estimate a therapist's future performance in treating depression and improving quality of life was also very high. For other domains (e.g., Mania), the relationship between previous and subsequent performance was less robust. However, the overall pattern of results shows that a therapist who was labeled as *exceptional* in a domain was likely to remain in the top 50th percentile with their next 30 cases, providing support for the idea that standardized outcome measures can be used to better match clients to therapists (Boswell et al., 2015). Consistent with the conclusions of Wampold and Imel (2015), a data-driven matching approach is likely to increase one's odds of benefiting from treatment.

Table 6
Therapists' Total Number of Above-Average and Below-Average Domains

	Above average		Below average	
Number of domains	Therapists with this many above-average domains (n)	Therapists above average	Therapists with this many below-average domains (n)	Therapists below average
0	10	18%	32	57%
1	13	23%	10	18%
2	4	7%	4	7%
3	5	9%	1	2%
4	5	9%	1	2%
5	3	5%	2	4%
6	3	5%	2	4%
7	2	4%	3	5%
8	3	5%	1	2%
9	1	2%	0	0%
10	2	4%	0	0%
11	5	9%	0	0%
12	0	0%	0	0%

The Mania scale did not perform as well as the other domains. For this subscale domain, there may be a bimodal relationship to health in which both extremes of the dimension are related to pathology (mania at one end and depression at the other), while the middle of the domain represents relatively healthy scores (Boswell, Kraus, Castonguay, & Youn, 2015). The use of the mania scale to predict therapist performance may require additional statistical manipulation before using HLM and other techniques that assume a linear relationship to health.

By contrast, the domains of Substance Abuse and Quality of Life demonstrated a much more pronounced therapist effect than the others. We speculate that the treatment of substance abuse may require a higher level of specialized training that may not be captured by type of degree and years of experience, which our overall exploration demonstrated had no significant impact.

Quality of life, on the other hand, is not often directly or explicitly targeted in disease-specific treatment approaches, as these approaches focus primarily on reducing symptoms. Adherence to empirically supported treatments is likely to reduce therapist variability and, therefore, may diminish the therapist effect in many focal disease-specific areas. By contrast, therapists may feel

Table 7
Estimates of Variance Explained at the Client and Therapist Levels by Risk-Adjustment Variables

TOP domain	RA with only intake score		RA with intake score and risk scores		RA with full random forest model	
	Client	Therapist	Client	Therapist	Client	Therapist
Sexual Functioning	22.87%	5.63%	25.19%	4.90%	28.79%	4.44%
Work Functioning	7.70%	6.97%	12.04%	5.58%	13.01%	6.59%
Violence	16.88%	9.04%	19.15%	8.77%	25.56%	5.85%
Social Functioning	13.39%	8.14%	20.29%	6.30%	24.07%	5.30%
Panic/Anxiety	29.96%	7.82%	35.20%	6.40%	40.81%	4.35%
Substance Abuse	28.90%	22.33%	29.04%	21.84%	33.78%	18.28%
Psychosis	27.57%	7.87%	36.05%	6.19%	41.50%	3.71%
Quality of Life	20.54%	22.39%	22.33%	21.44%	26.93%	18.72%
Sleep	32.05%	7.28%	32.37%	6.71%	36.30%	5.18%
Suicidality	23.88%	16.06%	26.82%	16.00%	32.83%	12.78%
Depression	23.07%	17.72%	31.97%	15.25%	38.64%	11.82%
Mania	10.86%	2.46%	14.90%	1.82%	18.27%	1.56%
Total score					40.16%	12.93%

Note. TOP = Treatment Outcome Package; RA = risk adjusted.

more freedom, whether in the context of naturalistic treatment or controlled trials, to practice intuitively with regard to existential and quality-of-life issues. The TOP assesses quality of life with four questions regarding satisfaction with relationships, daily responsibilities, general mood and feelings, and life in general. It is plausible that therapists, with their perceived freedom in how they address more global issues of life quality, vary widely in their focus on, assessment of, and skill in addressing patient's concerns in this area. If this speculation proved accurate, this would be unfortunate, given that low quality of life appears to be the primary driver that motivates patients to treatment (Kraus et al., 2005).

Our results are based on measured outcomes in specific domains, suggesting that a more granular assessment of a therapist's performance may both capture outcome complexity and yield greater predictive and/or decision-making utility than a global or composite outcome indicator. We believe the results are also encouraging for practicing therapists. More than half (57%) of therapists did not have a single area of underperformance (i.e., below-average outcomes), and 88% of therapists had at least one domain in which they were exceptional. Crediting Ricks (1974) for the term, a number of "super shrinks" were identified. These therapists made up 9% of the sample and were above average in 11 of 12 domains. We agree with Baldwin and Imel's (2013) conclusion that extremely valuable knowledge would likely be gained by studying the characteristics and in-session behaviors of these consistently high-performing therapists. Process research methods would help the field discover what it is about them that sets them apart from the average therapist. Previous research and the present results indicate that differences in client characteristics do not fully explain observed between-therapist differences.

A few words of caution are important when interpreting these results. The data were collected from naturalistic settings and there was no attempt to randomize clients to therapists. Despite the rather robust risk-adjustment methodology, some important client characteristics likely remain unaccounted for. For example, client motivation for treatment was not assessed. An RCT at a single site would help control for these types of variables. Furthermore, some clients in the sample may not have completed their last TOP at the formal termination of treatment. The duration of treatment was an important risk-adjustment variable and was accounted for in the Random Forest model. Nevertheless, the relationship between the dose of treatment and outcome remains complicated. How dose and rate of change affect the therapist effect in various lengths of treatments remains unknown (Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009). Future research needs to test these relationships more fully.

In addition, we do not have information on client dropout, which could have caused a therapist who easily loses clients to show an inflated measure of effectiveness. It is possible that the subset of clients who had their last TOP assessment closer to 30 days included both clients who rapidly received what they desired from treatment and others who dropped out. On the other hand, Lutz, Leon, Martinovich, Lyons, and Stiles (2007) showed that therapists who had higher rates of client dropout had poorer outcomes with the clients they retained, so this may be less of a concern. In addition, a subset of therapists appeared to perform better when treating clients with mild problems and others when treating clients with severe problems. Such therapists violate the HLM "fixed slope" assumption that was used in this analysis, an assumption that appears valid for the vast majority of therapists. Identifying therapists who have this nonstandard effect and testing whether it is a cross-problem-domain construct could improve predictive models. For example, although we controlled for initial severity, the current predictive model presented in this article assumed that a therapist would be similarly effective (compared with other therapists) for mild as well as severe pathology. For this subgroup of therapists who do not appear to have fixed slopes, the prediction of their outcomes would likely improve by allowing their slopes to vary.

Similarly, we suspect that some therapists may achieve better outcomes with certain demographic groups, and this could further improve predictive models. For example, a therapist with documented, exceptional outcomes working with gender identity issues may achieve better outcomes with clients struggling with these issues, independent of the therapist's domain-specific skills. Time is also a critical variable to explore. At what point in history (e.g., how many years back) does past performance stop being predictive of future performance? Because therapists may learn and evolve (and some may also deteriorate) in their therapeutic skills, it will be important to determine when previous outcomes have less predictive power. It may be possible to improve predictions by discounting or ignoring certain, too-old outcomes. In addition, Saxon and Barkham (2012) showed that a higher concentration of clients in a therapist's caseload who were at risk of harming themselves or others was predictive of poorer outcomes for all clients seen by that therapist. The authors speculated that this was related to therapist burnout. Understanding this relationship is critical to preventing otherwise effective therapists from performing poorly when overburdened with difficult referrals. Future research should continue the caseload analyses that Saxon and Barkham insightfully began. This might include integrating an assessment of therapist burnout in a longitudinal study of therapist effectiveness as well as deriving a measure of an individual therapist's sensitivity to the effects of burnout.

Research on routine outcome monitoring and feedback has demonstrated that the largest impact is on clients who have been labeled as "off track" and at risk for deterioration (Shimokawa, Lambert, & Smart, 2010). Similarly, when it comes to the stability and implications of therapist outcomes, potential harm may be more important to investigate than effectiveness. A therapist's ineffectiveness with an out-of-the-norm client-specific domain can bring down the outcome and effectiveness of the therapist on other domains, even those on which they are typically exceptional. For example, a therapist who has negative outcomes when treating sexual dysfunction, but is usually exceptional at treating depression, may underperform in treating a client with depression who also has a significant sexual dysfunction. A question for future research is whether steering clients away from potential harm is more important to maximizing outcomes than matching a client's most problematic domain to a therapist who is exceptional at treating that domain. In such cases, rather than steering clients toward a different therapist, a potentially fruitful alternative strategy might involve ensuring that the therapist receives additional supervision, consultation, and/or continuing education to improve his or her clients' outcomes. In either case, treating only those clients with symptoms for which one has an adequate level of expertise is a professional and ethical issue as well as a research concern.

The moderate correlations between therapist skill domains found in Table 5 are noteworthy. These results are significantly different from the small correlations found in the non-risk-adjusted outcome rankings reported by Kraus et al. (2011). They are, however, comparable with the correlations found between these same factors in previous TOP validation studies (e.g., Kraus et al., 2005) and probably denote that some domains are more closely related than others. As such, the conclusions made in 2011 do not appear to hold up with proper risk adjustment. Rather than concluding that there is little to no correlation between a therapist's skill in one area and another, there appears to be quite a bit of a relationship.

This, then, raises the question about the relative importance of each domain in a multidomain outcome model. Future research should explore the weighting of domains in future predictive models. Patients often present with elevation on multiple domains at intake. Which domain is of highest importance for matching skills to patient need? Is treating elevated suicidal ideation more urgent than a more out-of-the-norm sexual functioning? Does the number of domains on which a client is matched to a therapist with a positive track record in those domains moderate the potential utility of matching?

Finally, the ultimate goal of this research is to provide the best possible prediction model (i.e., set of therapist referral options) for each new client. Optimally, data-based algorithms could provide referral sources with a list of well-matched therapists in the geographic area who meet other standard requests, such as insurance coverage, distance to travel, and demographic characteristics. Choice seems critical in preparing such a list. Future research must, therefore, build more granular predictive models for the individual client (and not just for the next set of clients served by a therapist). In summary, therapist effectiveness appears to be relatively stable and can be demonstrated across clients. Although its prediction requires careful attention to risk-adjustment, case-mix, and problem domains, the identification of well-matched therapists for individual clients has the potential to greatly enhance mental health outcomes.

References

Almlöv, J., Carlbring, P., Berger, T., Cuijpers, P., & Andersson, G. (2009). Therapist factors in Internet-delivered cognitive behavioural therapy for major depressive disorder. *Cognitive Behaviour Therapy, 38,* 247–254. http://dx.doi.org/10.1080/16506070903116935

Almlöv, J., Carlbring, P., Källqvist, K., Paxling, B., Cuijpers, P., & Andersson, G. (2011). Therapist effects in guided internet-delivered CBT for anxiety disorders. *Behavioural and Cognitive Psychotherapy, 39,* 311–322. http://dx.doi.org/10.1017/S135246581000069X

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose-effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology, 77,* 203–211. http://dx.doi.org/10.1037/a0015235

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (pp. 258–297). Hoboken, NJ: Wiley.

Blatt, S. J., Sanislow, C. A., III, Zuroff, D. C., & Pilkonis, P. A. (1996). Characteristics of effective therapists: Further analyses of data from the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology, 64,* 1276–1284. http://dx.doi.org/10.1037/0022-006X.64.6.1276

Boswell, J. F., Constantino, M. J., Kraus, D. R., Bugatti, M., & Oswald, J. (2015). The expanding relevance of routinely collected outcome data for mental health care decision making. *Administration and Policy in Mental Health and Mental Health Services Research*. Advance online publication.

Boswell, J. F., Kraus, D. R., Castonguay, L. G., & Youn, S. (2015). Treatment outcome package: Measuring and facilitating multidimensional change. *Psychotherapy, 52,* 422–431.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32. http://dx.doi.org/10.1023/A:1010933404324

Bremer, R. W., Scholle, S. H., Keyser, D., Knox Houtsinger, J. V., & Pincus, H. A. (2008). Pay for performance in behavioral health. *Psychiatric Services, 59,* 1419–1429.

Castonguay, L. G. (2013). Psychotherapy outcome: A problem worth re-revisiting 50 years later. *Psychotherapy, 50,* 52–67.

Cella, M., Stahl, D., Reme, S. E., & Chalder, T. (2011). Therapist effects in routine psychotherapy practice: An account from chronic fatigue syndrome. *Psychotherapy Research, 21,* 168–178. http://dx.doi.org/10.1080/10503307.2010.535571

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology, 59,* 20–26. http://dx.doi.org/10.1037/0022-006X.59.1.20

Dinger, U., Strack, M., Leichsenring, F., Wilmers, F., & Schauenburg, H. (2008). Therapist effects on outcome and alliance in inpatient psychotherapy. *Journal of Clinical Psychology, 64,* 344–354. http://dx.doi.org/10.1002/jclp.20443

Erickson, S. J., Tonigan, J. S., & Winhusen, T. (2012). Therapist effects in a NIDA CTN intervention trial with pregnant substance abusing women: Findings from a RCT with MET and TAU conditions. *Alcoholism Treatment Quarterly, 30,* 224–237. http://dx.doi.org/10.1080/07347324.2012.663295

Falkenström, F., Markowitz, J. C., Jonker, H., Philips, B., & Holmqvist, R. (2013). Can psychotherapists function as their own controls? Meta-analysis of the crossed therapist design in comparative psychotherapy trials. *Journal of Clinical Psychiatry, 74,* 482–491. http://dx.doi.org/10.4088/JCP.12r07848

Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Thompson-Hollands, J., Carl, J., . . . Barlow, D. H. (2012). Unified protocol for transdiagnostic treatment of emotional disorders: A randomized controlled trial. *Behavior Therapy, 43,* 666–678. http://dx.doi.org/10.1016/j.beth.2012.01.001

Hermann, R. C., Rollins, C. K., & Chan, J. A. (2007). Risk-adjusting outcomes of mental health and substance-related care: A review of the literature. *Harvard Review of Psychiatry, 15,* 52–69. http://dx.doi.org/10.1080/10673220701307596

Horowitz, L. M., Lambert, M. J., & Strupp, H. H. (Eds.). (1997). *Measuring client change in mood, anxiety, and personality disorders: Toward a core battery*. Washington, DC: American Psychological Association Press.

Huppert, J. D., Bufka, L. F., Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2001). Therapists, therapist variables, and cognitive-behavioral therapy outcome in a multicenter trial for panic disorder. *Journal of Consulting and Clinical Psychology, 69,* 747–755. http://dx.doi.org/10.1037/0022-006X.69.5.747

Institute of Medicine Committee on Redesigning Health Insurance Performance Measures, Payment, and Performance Improvement Programs. (2007). *Rewarding provider performance: Aligning incentives in Medicare: Pathways to quality health care series*. Washington, DC: National Academy Press.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting & Clinical Psychology, 59,* 12–19.

Jones, D. R., Macias, C., Barreira, P. J., Fisher, W. H., Hargreaves, W. A., & Harding, C. M. (2004). Prevalence, severity, and co-occurrence of chronic physical health problems of persons with serious mental illness. *Psychiatric Services, 55,* 1250–1257. http://dx.doi.org/10.1176/appi.ps.55.11.1250

Kim, D.-M., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research, 16,* 161–172. http://dx.doi.org/10.1080/10503300500264911

King, M. W., & Resick, P. A. (2014). Data mining in psychological treatment research: A primer on classification and regression trees. *Journal of Consulting and Clinical Psychology, 82,* 895–905. http://dx.doi.org/10.1037/a0035886

Kraus, D. R., Boswell, J. F., Wright, A. G. C., Castonguay, L. G., & Pincus, A. L. (2010). Factor structure of the treatment outcome package for children. *Journal of Clinical Psychology, 66,* 627–640.

Kraus, D. R., Castonguay, L., Boswell, J. F., Nordberg, S. S., & Hayes, J. A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21,* 267–276. http://dx.doi.org/10.1080/10503307.2011.563249

Kraus, D. R., Seligman, D. A., & Jordan, J. R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The Treatment Outcome Package. *Journal of Clinical Psychology, 61,* 285–314. http://dx.doi.org/10.1002/jclp.20084

Lambert, M. J. (2010). *Prevention of treatment failure: The use of measuring, monitoring, and feedback in clinical practice*. Washington, DC: American Psychological Association Press.

Laska, K. M., Smith, T. L., Wislocki, A. P., Minami, T., & Wampold, B. E. (2013). Uniformity of evidence-based treatments in practice? Therapist effects in the delivery of cognitive processing therapy for PTSD. *Journal of Counseling Psychology, 60,* 31–41. http://dx.doi.org/10.1037/a0031294

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News, 2,* 18–22.

Lutz, W., Hofmann, S. G., Rubel, J., Boswell, J. F., Shear, M. K., Gorman, J. M., . . . Barlow, D. H. (2014). Patterns of early change and their relationship to outcome and early treatment termination in patients with panic disorder. *Journal of Consulting and Clinical Psychology, 82,* 287–297. http://dx.doi.org/10.1037/a0035535

Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Consulting and Clinical Psychology, 54,* 32–39. http://dx.doi.org/10.1037/0022-0167.54.1.32

McAleavey, A. A., Nordberg, S. S., Kraus, D., & Castonguay, L. G. (2012). Errors in treatment outcome monitoring: Implications for real-world psychotherapy. *Canadian Psychology, 53,* 105–114. http://dx.doi.org/10.1037/a0027833

Ogles, B. M. (2013). Measuring change in psychotherapy research. In M. J. Lambert (Ed.), *Bergin & Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 134–166). New York, NY: John Wiley & Sons.

R Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raghavan, R. (2010). Using risk adjustment approaches in child welfare performance measurement: Applications and insights from health and mental health settings. *Children and Youth Services Review, 32,* 103–112. http://dx.doi.org/10.1016/j.childyouth.2009.07.020

Ricks, D. F. (1974). Supershrink: Methods of a therapist judged successful on the basis of adult outcomes of adolescent patients. In D. F. Ricks, A. Thomas, & M. Roff (Eds.), *Life history research in psychopathology: III* (pp. 275–297). Minneapolis, MN: University of Minnesota Press.

Rosen, A. K., Chatterjee, S., Glickman, M. E., Spiro, A., III, Seal, P., & Eisen, S. V. (2010). Improving risk adjustment of self-reported mental health outcomes. *The Journal of Behavioral Health Services & Research, 37,* 291–306. http://dx.doi.org/10.1007/s11414-009-9196-9

Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology, 80,* 535–546. http://dx.doi.org/10.1037/a0028898

Scanlon, D. P., Lindrooth, R. C., & Christianson, J. B. (2008). Steering patients to safer hospitals? The effect of a tiered hospital network on hospital admissions. *Health Services Research, 43,* 1849–1868.

Shimokawa, K., Lambert, M. J., & Smart, D. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting & Clinical Psychology, 78,* 298–311. http://dx.doi.org/10.1037/a0019247

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14,* 323–348. http://dx.doi.org/10.1037/a0016973

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73,* 914–923. http://dx.doi.org/10.1037/0022-006X.73.5.914

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work* (2nd ed.). New York, NY: Routledge.

Wiborg, J. F., Knoop, H., Wensing, M., & Bleijenberg, G. (2012). Therapist effects and the dissemination of cognitive behavior therapy for chronic fatigue syndrome in community-based mental health care. *Behaviour Research and Therapy, 50,* 393–396. http://dx.doi.org/10.1016/j.brat.2012.03.002

Received August 12, 2015
Revision received December 18, 2015
Accepted December 21, 2015 ■