Psychotherapy Research, May 2011; 21(3): 267–276

Routledge
Taylor & Francis Group

Therapist effectiveness: Implications for accountability and patient care

DAVID R. KRAUS[1], LOUIS CASTONGUAY[2], JAMES F. BOSWELL[2], SAMUEL S. NORDBERG[2], & JEFFREY A. HAYES[2]

[1]*Behavioral Health Laboratories, Marlborough, MA, USA;* [2]*The Pennsylvania State University, Department of Psychology, Moore Building, University Park, PA, USA &* [3]*Pennsylvania State University, Department of Counselor Education, Counseling Psychology, and Rehabilitation Services, Cedar Building, University Park, PA, USA*

(Received 13 July 2010; revised 4 February 2011; accepted 9 February 2011)

Abstract
Significant therapist variability has been demonstrated in both psychotherapy outcomes and process (e.g., the working alliance). In an attempt to provide prevalence estimates of "effective" and "harmful" therapists, the outcomes of 6960 patients seen by 696 therapists in the context of naturalistic treatment were analyzed across multiple symptom and functioning domains. Therapists were defined based on whether their average client reliably improved, worsened, or neither improved nor worsened. Results varied by domain with the widespread pervasiveness of unclassifiable/ineffective and harmful therapists ranging from 33 to 65%. Harmful therapists demonstrated large, negative treatment effect sizes ($d = -0.91$ to -1.49) while effective therapists demonstrated large, positive treatment effect sizes ($d = 1.00$ to 1.52). Therapist domain-specific effectiveness correlated poorly across domains, suggesting that therapist competencies may be domain or disorder specific, rather than reflecting a core attribute or underlying therapeutic skill construct. Public policy and clinical implications of these findings are discussed, including the importance of integrating benchmarked outcome measurement into both routine care and training.

Keywords: outcome research; mental health services research

It has been well established that some patients do not benefit from psychotherapy (Lambert, 2007), with a substantial minority of patients reliably deteriorating (Lambert et al., 2001). Some patients may not be sufficiently motivated for change (Vogela, Hansen, Stiles, & Götestam, 2006). For others, the type of treatment offered may not match well with the specific patient's world view or personality style (e.g., insight oriented treatment offered to an externalizing patient; Beutler et al., 1991), or their treatment expectations (Greenberg, Constantino, & Bruce, 2006).

The nature and impact of potentially ineffective treatments (and therapists) has received some attention in the literature (e.g., Mays & Franks, 1985; Strupp, Hadley, & Gomez–Schwartz, 1977). In his article on proscription in psychotherapy, Mohr (1995) outlined potential indicators of treatment deterioration (e.g., client suspiciousness toward the therapist and underestimating the client's level of pathology). More recently, as reviewed elsewhere (Castonguay, Boswell, Constantino, Goldfried, & Hill, 2010), empirical evidence suggests that the therapist may contribute to treatment failures. Such potential contributions range from an inability to identify and repair budding alliance ruptures (Safran & Muran, 2000) to an overly hostile and dominant interpersonal style (Henry, Schacht, & Strupp, 1990; Henry, Strupp, Butler, Schact, & Binder, 1993), to serious ethical violations. In their comprehensive review of therapist variables, Beutler et al. (2004) reported that therapist well-being has a significant, yet modest relationship with outcome. They cite McCarthy and Frieze (1999), who found that therapist burnout was negatively associated with outcome. Beutler et al. (2004) also concluded that although variability exists across studies, therapist training, skill, experience, and style tended to be weak predictors of outcome ($r = .07$).

While serious ethical violations may be infrequent, survey data suggest that they may be more frequent than the field would like to believe. For example, on

Correspondence concerning this article should be addressed to David R. Kraus, Behavioral Health Laboratories, Marlborough, MA, USA. Email: dkraus@bhealthlabs.com

ISSN 1050-3307 print/ISSN 1468-4381 online
DOI: 10.1080/10503307.2011.563249

anonymous therapist surveys, 7–10% of respondents acknowledge having sexual contact with patients (Simon, 1999). In a survey conducted by Pope, Tabachnick, and Keith-Spiegel (1987) therapists were asked to report on the degree to which they engaged in a list of "unethical" behaviors. Approximately 2% of all respondents reported filing an ethics complaint against a colleague either "fairly" or "very often." Over 10% of surveyed therapists reported that they sometimes conducted therapy when too distressed to be effective. In addition, patients bring challenges and great variability to the treatment process (Barber, 2009), and it may well be that how therapists handle these challenges is critical to successful treatment outcomes (Gelso & Hayes, 2007; Hayes, 2004).

It has been shown that some therapists appear to provide little relief for their patients, even in situations where random case assignment distributed difficult clients across therapists (Luborsky, McLellan, Diguer, Woody, & Seligman, 1997). Most concerning is the evidence that some therapists might actually leave their average patient worse off than when he or she started treatment (Okiishi, Lambert, Nielsen, & Ogles, 2003). The Okiishi et al. study, however, was limited to one counseling clinic and the prevalence of these effects in other naturalistic settings is unknown. This study was further limited by the use of a total symptom distress score which may have masked deterioration effects in specific areas like substance abuse or suicidality.

On the other hand, since the 1970s, it has been documented that some therapists achieve consistently positive results, the outcome effects of which can still be measured years later (Miller, 1993; Ricks, 1974). As a group, these effective therapists achieve results that are exponentially greater than their average peers (Okiishi et al., 2003; Wampold & Brown, 2005). Nevertheless, the prevalence of these highly effective therapists is also unknown. Furthermore, it is not clear whether these superior skills generalize across disorders or problems treated.

The purpose of the current study was to answer two critical questions. First, what is the pervasiveness of both effective and harmful therapists in naturalistic settings; and second, are harmful therapists consistently harmful in treating various problems and functional domains, or are these negative effects domain/problem-specific?

Method

Participants

From a large, patient-de-identified archival dataset, a total of 15,217 adult patients in traditional outpatient care with outcome data at the first session of treatment and near the sixteenth week of treatment (within 14 days) were considered for inclusion in the study. The uniformity of data collection points was chosen to control for the effects of time. However, no attempt was made to limit the variability of patients by diagnosis or other factors in order to represent the conditions of patients as naturally occurring in treatment settings across the United States. The dose-response literature has demonstrated that change in psychotherapy is best modeled by a negatively accelerating relationship to number of sessions, such that each subsequent session evidences, on average, less change than the session before (Howard, Kopta, Krause & Orlinsky, 1986; Lutz, Lowry, Kopta, Eisnstein & Howard, 2001; Lutz, Martinovich & Howard, 1999). In addition, in a large sample of clients being seen in naturalistic settings, Lambert and colleagues (Lambert et al., 2001) demonstrated that 80% of clients had met the criteria for the reliable change index by session fifteen. Thus, given the relatively short-term nature of treatment in managed care, as well as the evidence that the majority of change takes place in early sessions, we limited our sample to 16 sessions of psychotherapy.

The data included outcomes for 3222 therapists, some of whom had only one patient who met the above criteria and others who had more than 300. In order to ensure a relatively consistent number of patients among therapists, the sample was further limited to those therapists with at least 10 patients and limited to each therapist's first 10 patients. A random sample of each therapist's caseload was considered, but with a mode of 11 cases per therapist, this approach would not have produced significantly different results. The final dataset included 6960 patients and 696 therapists, the characteristics of whom are summarized in Table I. The subset of therapists was quite similar to the larger group of 3222 clinicians, although the subset was slightly more ethnically diverse (71% white vs. 75%), a little less experienced (11.2 years vs. 14.2), but with very similar distribution of license types and age. Both patient and therapist populations were predominantly female (64% and 74%, respectively). The patient sample was ethnically diverse with the largest subgroup being European Americans (48%). Individual patients from low-income households were over-represented in the sample. The therapist sample consisted of experienced clinicians with an average of 11 years of post-licensing experience who were primarily social workers and mental health counselors, with less than 12% psychologists or psychiatrists.

Table I. Participant demographic and professional information

	Patients		Therapists	
	M	*SD*	*M*	*SD*
Age (mean/SD)	37.3	12.4	37.1	20.9
Education (years)	11.7	3.6		
Gender (% female)	64%		74%	
Race/ethnicity (%)				
White/Caucasian	48%		71%	
African-American	3%		6%	
Hispanic	24%		10%	
Asian	17%		6%	
Other	8%		6%	
Family income in thousands (%)				
0–10	51%			
10–20	18%			
20–30	9%			
30–40	6%			
40–50	5%			
50–75	6%			
75–100	3%			
Over 100	3%			
License types				
Social worker			43%	
Mental health counselor			35%	
Psychologist			10%	
Drug and alcohol counselor			5%	
Marital and family therapist			2%	
Psychiatrist			1%	
Clinical nurse			1%	
Other			3%	
Years of experience			11.2	8.1

Procedure

The data for this study came from a de-identified archival dataset which included patients seen in naturalistic settings. Either the clinician or clinic involved in the data collection had contracted with Behavioral Health Laboratories (BHL) to process assessment and outcome data on all patients as part of routine care. Patients were told that identifiable data would be used by the clinician to better understand patient issues and needs for treatment, and that repeat assessments would be used to conjointly monitor progress towards mutually developed goals. Patients and therapists were also told that de-identified data could be used for scientific studies.

Outcome Measure

The Treatment Outcome Package (TOP; Kraus, Seligman, & Jordan, 2005) is used by providers to assess patient strengths, psychopathology, and track patient improvement. The TOP is a behavioral health assessment and outcome battery designed for clinical and research purposes in naturalistic settings. Developed to meet the criteria established by the Society for Psychotherapy Research (SPR) and American Psychological Association (APA) sponsored Core Battery Conference (Horowitz, Lambert, & Strupp, 1997), it assesses a wide array of behavioral health symptoms and functioning, demographics, and 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.

The TOP possesses excellent confirmatory factor analysis (CFA) modeling (Brown, 2001) for both adults (Kraus, Seligman & Jordan, 2005) and children (Kraus, Boswell, Wright, Castonguay, & Pincus, 2010). In addition, the TOP has demonstrated excellent sensitivity to change with 50% of patients documenting reliable improvement (Jacobson & Truax, 1991) on single subscales, 91% documenting reliable improvement on at least one of the 12 domains, and 67% documenting reliable deterioration on at least one subscale (Kraus et al., 2005).

TOP feedback reports provide symptom severity for each of the 12 subscales in terms of standard deviations above or below the general population mean (Kraus & Castonguay, 2010). Additionally, the TOP assesses general health, substance use, stressful life events, treatment goals and satisfaction with treatment. It has demonstrated good test-retest reliability (see Table III) and high levels of convergent validity with scales such as the Beck Depression Inventory (BDI; Beck, Steer, & Ranieri, 1988), the Brief Symptom Inventory (BSI; Derogatis, 1975) and the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Graham, 1993; Hathaway & McKinley, 1989). The TOP requires approximately 8 minutes to complete.

Defining Effectiveness Categories and Ranking Therapists

This section describes a method for defining effective therapists in terms of their therapeutic impact. While the contribution of the therapist to treatment outcome has been established in the literature, ranking therapists on their observed effectiveness is a relatively new development in psychotherapy research, with few published studies and varied approaches (Brown, Jones, Lambert, & Minami, 2005; Luborsky et al., 1997; Okiishi et al., 2003; Wampold & Brown, 2005). Furthermore, none of these studies has attempted to determine the pervasiveness of therapists with effective or harmful outcomes. Considering the lack of accepted analytic methods to define types of therapists in terms of their treatment outcome, we decided to operationalize these definitions based on

established procedures to detect reliable change. Using the Reliable Change Index (RCI), we calculated whether patient change exceeded the measurement error of the scale (Jacobson & Truax, 1991). We used the formula:

$$\mathrm{RCI} = 1.96{*}\mathrm{SD}(1 - r)^{1/2} \quad (1)$$

where "SD" is the population standard deviation and "*r*" is the reliability estimate of the scale.

If the patient's change score exceeded the RCI for the specific scale, the change was determined to exceed the measurement error of the TOP and the patient was considered reliably changed. Change may reflect improvement or deterioration; improvement is defined as a reduction in scores on the TOP that exceeded the RCI and deterioration is defined as an increase in scores on the TOP that exceeded the RCI. As such, we propose the following:

An "effective" therapist is one whose average patient reliably improves.

An "unclassifiable/ineffective" therapist is one whose average patient neither reliably improves nor reliably deteriorates. The label "unclassifiable/ineffective" was chosen for this middle group due to the fact that it comprises a range of therapists who cannot be distinguished in the study. By definition these therapists have not witnessed enough change (deterioration or improvement) to be categorized as effective or harmful. For some this may be due to a low incidence of the defined pathology and for others because the therapist did not witness any significant change in their clients who had significant pathology. It was not the purpose of this study to further classify this group, and throughout the text, this group is referred to as "unclassifiable" for the sake of simplification.

A "harmful" therapist is one whose average patient reliably deteriorates.

These definitions ensured that the average level of patient change (for effective and harmful therapists) was meaningful, and by limiting each therapist caseload to 10 clients, that therapists with disproportionate caseloads (*n*s) were not advantaged or disadvantaged by different levels of statistical power.[1]

For the purposes of this study, which aimed to address the prevalence of effective and harmful therapists as defined above, we decided to use non-risk-adjusted change scores. While use of change scores has been controversial when comparing differences between empirically supported treatments (Cohen, Cohen, West, & Aiken, 1983), it is an acceptable standard of evaluating change at the individual level of the patient (Atkins, Bedics, McGlinchey, & Beauchaine, 2005) and has previously been used when evaluating and reporting therapist rankings (Luborsky et al., 1997).[2]

Results

We first endeavored to evaluate whether the level of patient improvement was comparable to other published TOP data by calculating RCIs for the patient sample and comparing them to the dataset reported by Kraus et al. (2005); Study 5, $n = 20{,}098$ from 511 different community samples). The data presented in Table II suggest that patients in the present community sample made considerable progress in treatment not dissimilar to other community samples. For example, the percent of patients demonstrating reliable improvement on the TOP Depression scale was 55% in the present study and 54% in the referenced comparison study. The largest differences were seen in the areas of sexual functioning (33% improved in the current study compared to 25% previously) and social functioning (46% in the current study compared to 38% previously). Taken

Table II. Reliable change by Treatment Outcome Package (TOP) domain

	% Reliably improved		% No change		% Reliably worsened	
TOP domain	Current study	2005 study*	Current study	2005 study	Current study	2005 study
Sexual functioning	33%	25%	44%	60%	23%	15%
Work functioning	34%	39%	43%	41%	23%	20%
Violence	24%	31%	60%	52%	16%	17%
Social functioning	46%	38%	22%	44%	32%	18%
Panic/anxiety	41%	41%	34%	42%	25%	17%
Substance abuse	32%	Not reported	50%	Not reported	19%	Not reported
Psychosis	41%	44%	36%	38%	23%	18%
Quality of life	46%	52%	30%	27%	24%	21%
Sleep	48%	47%	28%	33%	25%	20%
Suicidality	36%	42%	49%	44%	15%	14%
Depression	55%	54%	25%	32%	20%	14%
Mania	13%	10%	79%	84%	8%	6%

* Kraus, Seligman & Jordan (2005).

Table III. Initial and follow-up domain *Z* scores and reliability estimates

TOP* domain	Initial average *Z* score	Follow-up average *Z* score	Intraclass test-retest**
Sexual functioning	0.56	0.40	.92
Work functioning	0.13	−0.13	.90
Violence	1.17	0.81	.88
Social functioning	1.25	0.94	.93
Panic/anxiety	1.90	1.52	.88
Substance abuse	1.96	1.18	.89
Psychosis	1.73	1.27	.87
Quality of life	2.02	1.62	.93
Sleep	1.46	1.05	.94
Suicidality	1.91	1.19	.90
Depression	2.26	1.60	.93
Mania	−0.02	−0.13	.76

$n = 6960$.
* Treatment Outcome Package (TOP).
** Interclass test-retest and population standard deviations as calculated in Kraus, Seligman & Jordan (2005).

as a whole, however, this dataset appears to be consistent with other reported data, and appears to be reflective of naturalistic treatment in larger samples. Mean *Z* scores for the study participants at intake and follow-up are presented in Table III.

For each therapist, we calculated means and standard deviations across their caseload for each of the 12 TOP domains, and classified their effectiveness based on the reliable change criteria defined above. Results of the classification percentages by TOP domain are presented in Table IV. For all TOP domains there were large numbers of therapists whose average patient reliably improved or deteriorated with one exception—the Mania scale had few therapists establishing reliable change in either direction. The frequency of effective therapists ranged from a low of 29% in treating symptoms of sexual dysfunction to a high of 67% in treating symptoms of depression. The number of unclassifiable therapists ranged from a low of 30% in treating symptoms of depression to a high of 59% in treating symptoms of sexual dysfunction. Harmful therapists ranged from a low of 3% in treating symptoms of depression symptoms to a high of 16% in treating both symptoms of substance abuse and violence.

We then calculated the number of domains in which each therapist was effective, labeling this as a competency. Results are presented in Table V. The average, and modal, number of domain competencies was five. Ninety-six percent of therapists were identified as competent in at least one TOP domain. Only one therapist was competent in 11 domains and no therapist was competent on all domains.

Following this, we calculated the treatment effect sizes (Cohen's *d*) for all therapists and then for each of the effectiveness categories. Results are presented in Table VI. Effect sizes for all therapists ranged from small (0.27) for the treatment of sexual dysfunction symptoms to large (0.91) in treating depression symptoms. However, for the effective therapists, all effect sizes were large, ranging from 1.00 in treating psychotic symptoms to 1.52

Table IV. Pervasiveness of effective, harmful and unclassifiable therapists

TOP* domain	% Effective therapists	% Unclassifiable therapists	% Harmful therapists
Sexual functioning	29%	59%	12%
Work functioning	35%	58%	7%
Violence	38%	46%	16%
Social functioning	45%	41%	14%
Panic/anxiety	43%	47%	10%
Substance abuse	50%	34%	16%
Psychosis	46%	45%	9%
Quality of life	47%	48%	5%
Sleep	54%	37%	9%
Suicidality	58%	35%	7%
Depression	67%	30%	3%
Mania	0.7%	99%	0.3%

$n = 696$.
* Treatment Outcome Package (TOP).

Table V. Therapists with multiple competencies

Number of competencies	Number of therapists	Percent of therapists
0	31	4%
1	45	6%
2	64	9%
3	76	11%
4	85	12%
5	99	14%
6	85	12%
7	75	11%
8	61	9%
9	50	7%
10	22	3%
11	1	0%
12	0	0%

$n = 696$.

in treating work functioning. Conversely, harmful therapists yielded large, negative treatment effect sizes ranging from −0.91 (Psychosis) to −1.49 (Work Functioning).

Finally, we correlated the various rankings of therapists across each TOP domain, with these results presented in Table VII. Although mostly significant, the relatively low correlations demonstrate little common variance between the rankings of therapists across symptom domains. For example, the highest correlation (.33) was found between rankings in treating symptoms of depression with rankings of therapists treating symptoms of suicidality.

Discussion

On average, the findings from this study suggest that therapists in naturalistic settings tend to be quite effective with overall treatment effect sizes that range from 0.27 for the treatment of sexual dysfunction to 0.91 for the treatment of depression. However, these global findings mask tremendous variability in therapist skills and areas of competency. Due to the serious public policy implications of this data, we start with a discussion of the study's strengths and limitations.

This is the first known study that attempts to assess the pervasiveness of effective and harmful therapist effects in naturalistic settings. Although it improves on related research that has found wide variability in therapist effectiveness by virtue of the present study's larger sample sizes, number of practice setting and patient diversity, as well as the use of a multi-dimensional measurement approach, it is limited in three respects. First, it is limited by its reliance on a convenience sample of therapists and clinics that paid for the processing of outcome data. Lambert (2007) and others (Kraus, Castonguay, Boswell, & Nordberg, 2010) have demonstrated that outcome feedback improves the quality of care and outcomes. Therefore, using a convenience sample with therapists that has integrated real-time outcome feedback data (outcomes management) could overestimate effectiveness rates, given that outcomes management is still resisted by most therapists (Lipzin, 2009). Second, the study was further limited by not including measures for all disorder categories. For example, measurement of personality disorders, eating disorder issues or adult ADHD were not included and may have under-estimated the effectiveness of therapists who specialized in this type of work. Third, further study is needed to determine whether these multi-dimensional therapist effects are as stable as global measurements of effectiveness where as little as three clinical cases were needed to predict future performance (Wampold & Brown, 2005).

With these limitations noted, the results of this study indicate that the pervasiveness of harmful

Table VI. Treatment effect sizes (Cohen's *d*) for therapist categories

TOP* domain	Harmful therapists	Unclassifiable therapists	Effective therapists	All therapists
Sexual functioning	−1.36	.06	1.48	0.27
Work functioning	−1.49	0.10	1.52	0.44
Violence	−1.17	0.00	1.02	0.31
Social functioning	−1.31	0.09	1.46	0.48
Panic/anxiety	−0.97	0.12	1.17	0.42
Substance abuse	−0.98	0.04	1.14	0.47
Psychosis	−0.91	0.12	1.00	0.43
Quality of life	−0.95	0.16	1.51	0.68
Sleep	−0.87	0.08	1.20	0.57
Suicidality	−1.12	0.12	1.30	0.64
Depression	−1.05	0.04	1.41	0.91
Mania	Few data	Few data	Few data	Few data

$n = 696$.
* Treatment Outcome Package (TOP).

Table VII. Correlation (Kendall's tau-b) between therapist rankings by TOP domain

	DEPRS	LIFEQ	MANIA	PANIC	PSYCS	SA	SOCNF	SEXFN	SLEEP	SUICD	VIOLN
LIFEQ	.33										
MANIA	.23	.03									
PANIC	.35	.14	.16								
PSYCS	.30	.12	.23	.26							
SA	.18	.11	.14	.10	.16						
SOCNF	.24	.13	.10	.18	.25	.09					
SEXFN	.21	.11	.08	.15	.15	.06	.21				
SLEEP	.29	.14	.13	.27	.23	.10	.16	.11			
SUICD	.37	.18	.16	.22	.32	.24	.21	.14	.20		
VIOLN	.19	.07	.14	.16	.22	.16	.19	.15	.12	.29	
WORKF	.23	.09	.18	.18	.18	.10	.17	.16	.14	.18	.17

DEPRS, depression; LIFEQ, quality of Life; PSYCS, psychosis; SA, substance abuse; SCONF, social conflict; SEXFN, sexual functioning; SUICD, suicide; VIOLN, violence; WORKF, work functioning; $n = 696$.

therapists is more widespread than previously found. Rather than a small number of therapists that produce average negative outcomes on a measure's total score (cf. Okiishi et al., 2003), we found large numbers of therapists whose average patient ends treatment worse off than when they started (11–38%) depending on the TOP subscale, with 20% of therapists' average patients left more suicidal and 36% more violent. When we applied our stringent criteria for the label of "harmful" (meaning that this average patient worsening had to reach a certain threshhold) these numbers were significantly reduced, but still higher than expected (i.e., 0.3–16% were classified as harmful depending on the domian). We have chosen to label this subgroup of therapists "harmful" as their patients don't just end treatment deteriorated, they end treatment significantly worse. For example, 16% of therapists met the harmful criteria for treating signs of substance abuse and violence.

We also found preliminary evidence that therapist effectiveness is not a global construct. Therapists who are skilled in one domain may be harmful in another, and the correlations between rankings in various domains are relatively low. With only 1–9% of variance explained between ranking categories, it would be difficult to reliably infer a therapist's effectiveness in treating substance abuse, for example, from their effectiveness at treating psychosis ($r = 0.16$, Kendall's tau-b).

The widespread prevalence of negative treatment effects has significant public health and public policy implications. The large negative effect sizes associated with the work of domain-specific harmful therapists is very high across all domains measured (-0.91 to -1.49). Patients seen by these therapists leave treatment more suicidal, violent, psychotic and depressed than when they started treatment. The future implications for these patients, their families, co-workers and society may be of great significance and must be taken seriously by those setting public policy.

For therapists who are labeled effective within a specific domain (e.g., Depression), it is unknown how this effectiveness in improving symptoms of depression as measured by TOP translates into effectiveness with one or more depression diagnoses, and will require additional research to tease apart. However, since most clinicians treat problems and symptoms rather than diagnoses, this may be more of an academic exercise rather than a practical one. On the other hand, we have conducted preliminary analyses that do demonstrate that a specific therapist's skill at treating uncomplicated depressive symptoms can be compromised by the presence of co-morbid substance abuse, meaning that a therapist good at treating one population may not be good at treating another (Nordberg et al., 2010).

The implication of this study's findings must be evaluated within the current healthcare climate. Most notably, it is still not a routine aspect of standard practice to integrate outcome management into clinical care. Therapists often resist these demands, labeling them as intrusive, costly, unnecessary or poorly designed (Lipzin, 2009). Nevertheless, without routine measurement, many clinicians are probably not aware of the helpful or harmful consequences of their treatment decisions. Lambert has shown that most (or all) clinicians believe they are above average and they cannot predict patients who will have a negative treatment outcome (Hannan et al., 2005).

The findings from this study further emphasize the consequences of limited therapist predictive abilities and the problems with not evaluating and questioning one's professional abilities. As such, standards of ethical practice may require therapists to routinely measure their outcomes and focus their practices where they are most likely to succeed. The product of such standards would help all clinicians

improve their outcomes by helping them to avoid patients with whom they are less likely to succeed (until they receive further training and/or supervision), and patients, families, communities, and employers would benefit from greater productivity, quality of life, and lower healthcare costs.

Toward this goal, we find surprisingly hopeful results in these data. Nearly all therapists appear to have areas of strength where they consistently produce large positive effects. With assistance in finding the right therapist it is possible that the unique skills, strengths or competencies of each therapist can be more appropriately harnessed. Large clinics and community mental health centers may be best suited to realign how patients are assigned or referred to clinicians, basing these decisions on the inherent wealth of diversity and skill that large numbers of clinicians bring to the table. An ideal system may assist patients in finding therapists who are matched not only based on preference for gender, ethnicity or the variables currently used by prospective patients, but augmented with information about the therapist's prior track record of helping patients with similar issues.

The findings of the present investigation also have important training implications. As mentioned above with regard to practicing clinicians, this study points to the likely usefulness of providing regular and systematic feedback to graduate students, interns, and residents about the impact of their interventions on different aspects of their clients' functioning. As noted elsewhere, "Although research on such feedback has been conducted in different settings (counseling center, outpatient clinic), it stands to reason that routinely gathering outcome data would be particularly relevant and helpful in training clinics. What better way to help an inexperienced therapist learn what he/she is doing—or failing to do—that might facilitate or interfere with change than by monitoring, on a weekly basis, client change, positive or negative?" (Castonguay et al., 2010, p. 44).

Reliable and valid evidence that clients of a trainee tend to get worse on specific aspects of their functioning should led him/her, as well as his/her supervisor, to consider a number of strategies to remediate this less than optimal situation. With the recognition (by supervisors and trainees) that all therapists have their own vulnerabilities and weaknesses (Castonguay et al., 2010), such evidence could be viewed as a marker for introspection (e.g., "is there something in my past or current life that prevents me work well with depressed clients") and/or the adoption of Sullivan's (1953) participant-observer attitude during sessions (e.g., "are there issues frequently emerging when working with anxiety disorders clients that make me uncomfortable and/or unable to be attuned to their needs and resources?"). Additional strategies to address negative outcome feedback could include personal therapy, reading of empirical and clinical literature about the effective processes of change and treatment methods for particular clinical problems, extensive observations of videotaped sessions conducted by the trainees, and/or more frequent, specific, and expert training. The use of empirically based feedback over the course of a trainee's career can also help faculty members and supervisors with some of the most difficult but crucial decisions that they are required to make, such as when to significantly reduce the number of clients to be assigned to a trainee (while increasing the level or specificity of his/her supervision) until he/she demonstrates minimal competence, when to remove a trainee from clinical duties until he/she has addressed personal problems that might interfere with his/her clients' improvement and well-being, or when to encourage (or force) a trainee to abandon the clinical part of his/her graduate training. Such delicate and difficult decisions are intrinsic to all graduate trainers' responsibility to help the field meets its utmost ethnical duty: First, do no harm.

The results of the present study also clearly suggest that many trainees are more effective than others in treating clients with particular types of difficulty. If viewed as a friendly tool to improve one's work (Kraus, Wolfe & Castonguay, 2006), the assessment and regular monitoring of outcome can stimulate and guide a trainee's introspection and self-observation of his/her clinical skills, not only to develop a sharper and more articulated view of what he/she is doing well with particular types of problem, but also to help him/her generate strategies to be more effective in addressing other difficulties experienced by many clients. The close and systematic inspection of the students' outcome, in terms of both their strengths and limitations, can also help supervisors and faculty members to make two positive decisions: Who should be selected for funded positions that involve seeing a relatively large number of clients during an academic year? And who should be selected to serve as a co-supervisor for student with less experience?

As a final point, it is worth noting that this study failed to identify a single therapist who was effective in every clinical domain. With this finding and the low correlations between domain rankings, there is some preliminary evidence to suggest that there does not seem to be a core competency or skill that renders a great clinician good in all or most domains. For example, the clinician who was ranked best at treating depression was also very good at treating social conflict and panic. On the other hand, he/she

was one of the few with patients whose manic and violence symptoms reliably worsened. Armed with this knowledge, it might be prudent for this therapist to focus his/her practice around these core competencies and avoid patients with personal and family histories of violence or bi-polar illness. On the other hand, the clinician may wish to improve his/her outcomes in these deficient areas and find supervision or continuing education focused on improving these skills. Carl Whitaker once said that a therapist's right to practice is predicated on an undying pledge for personal and professional growth (Whitaker, 1989). For the benefit of patients, this appears to be prophetic.

Notes

[1] Obviously, the word "harmful" should be used cautiously. Nevertheless, if it was known that a therapist's patients were consistently ending treatment more violent, suicidal and depressed, using another, more euphemistic label may not draw the necessary attention to this public health dilemma, and therefore, we concluded that if such negative therapist effectiveness was detected it should be labeled "harmful."

[2] We considered using a risk-adjusted change score, which would control for client-level characteristics (case-mix variables) that may not be equally distributed across therapists. Residual gain scores (Wampold & Brown, 2005) and mixed-effect models (Okiishi et al., 2003) are examples that have been used for this purpose. These risk-adjusted models correlate highly (Kraus et al., 2009), and provide comparisons to expected values related to average therapist results. The reference point for these analyses becomes what is "normal" or average within a healthcare context and cohort. However, what is average or normal in many fields of medicine is not currently acceptable to the Institute of Medicine (2001), or others, with too many patients harmed by treatment as usual. If harm was the norm within a specific field of healthcare, relying exclusively on risk-adjusted data might mask these deplorable findings and label a provider or intervention as above average (sounding good) when it may more importantly be harmful. Although good for ranking, once risk-adjusted, results from these models lose their ability to provide useful reference points to which a definition of acceptability (some public policy or societal standard of cost effectiveness, value or worth) can be applied. We therefore conclude that a non-risk-adjusted analysis of the data is essential.

References

Atkins, D., Bedics, J. D., McGlinchey, J. B., & Beauchaine, T. P. (2005). Assessing clinical significance: Does it matter which method we use? *Journal of Consulting and Clinical Psychology, 73*, 982–989.

Barber, J. P. (2009). Towards a working through of some core conflicts in psychotherapy research. *Psychotherapy Research, 19*, 1–12.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8*, 77–100.

Beutler, L. E., Engle, D., Mohr, D., Daldrup, R. J., Bergan, J., Meredith, K., et al. (1991). Predictors of differential response to cognitive, experential, and self-directed psychotherapeutic procedures. *Journal of Consulting and Clinical Psychology, 59*, 333–340.

Beutler, L. E., Malik, M., Alimohamed, S., Harwood, T. M., Talebi, H., Noble, S., & Wong, E. (2004). Therapist variables. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 227–306). New York, NY: Wiley & Sons.

Blatt, S. J., Sanislow, C. A., 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.

Brown, G. S., Jones, E. R., Lambert, M. J., & Minami, T. (2005). Identifying highly effective psychotherapists in a managed care environment. *American Journal of Managed Care, 11*, 513–520.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford.

Castonguay, L. G., Boswell, J. F., Constantino, M. J., Goldfried, M. R., & Hill, C. E. (2010). Training implications of harmful effects of psychological treatments. *American Psychologist*, *65*, 34–49.

Cohen, P., Cohen, J., West, S. G., & Aiken, L. S. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

Derogatis, L. (1975). *Brief symptom inventory.* Baltimore, MD: Clinical Psychometric Research.

Domino, M. E., Burns, B. J., Silva, S. G., Kratochvil, C. J., Vitiello, B., Reinecke, M. A., et al. (2008). Cost-effectiveness of treatments for adolescent depression: Results from TADS. *American Journal of Psychiatry,165*, 588–596.

Finch, R. A., & Phillips, K. (2005). *An employer's guide to behavioral health services: A roadmap and recommendations for evaluating, designing, and implementing behavioral health services.* Washington DC: National Business Group on Health.

Gelso, C. J., & Hayes, J. A. (2007). *Countertransference and the therapist's inner experience: Perils and possibilities.* Mahwah, NJ: Erlbaum.

Graham, J. (1993). *MMPI-2: Assessing personality and psychopathology.* New York, NY: Oxford.

Greenberg, R. P., Constantino, M. J., & Bruce, N. (2006). Are expectations still relevant for psychotherapy process and outcome? *Clinical Psychology Review, 26*, 657–678.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., David W. Smart, D. W., Shimokawa, K., et al. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*, 155–163.

Hathaway, S., & McKinley, J. (1989). *Manual for administration and scoring.* Minneapolis.

Hayes, J. A. (2004). The inner world of the psychotherapist: A program of research on countertransference. *Psychotherapy Research,14*, 21–36.

Henry, W. P., Schacht, T. E., & Strupp, H. H. (1990). Patient and therapist introject, interpersonal process, and differential psychotherapy outcome. *Journal of Consulting and Clinical Psychology,58*, 768–774.

Henry, W. P., Strupp, H. H., Butler, S. F., Schacht, T. E., & Binder, J. L. (1993). Effects of training in time-limited dynamic psychotherapy: Changes in therapist behavior. *Journal of Consulting and Clinical Psychology, 61*, 434–440.

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

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose–effect relationship in psychotherapy. *American Psychologist*, *41*, 159–164.

Institute of Medicine. Committee on Quality of Health Care America. (2001). *Crossing the quality chasm: A new health system for the 21st century.* 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.

Kraus, D. R., & Castonguay, L. G. (2010). TOP: Development & use in naturalistic settings. In M. Barkham, G. Hardy, & J. Mellor-Clark (Eds.), *A CORE approach to delivering practice-based evidence in counseling and the psychological therapies.* London: Wiley Press.

Kraus, D. R., Boswell, J. F., Wright, A. G., 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., Castonguay, L., Boswell, J., & Nordberg, S. (2010). Empirically supported feedback in unstructured clinical settings: The utility of TOP client reports as an adjunct to treatment as usual. *Unpublished Manuscript*.

Kraus, D.R., Nordberg, S.S., Boswell, J.F., Castonguay, L.G., & Hayes, J., (2009). *Identifying effective therapists with the Treatment Outcome Package: The need for a multi-dimensional approach*. Unpublished manuscript.

Kraus, Nordberg, S. S., Boswell, J. F., Castonguay, L. G., Hayes, J. A., & Wampold, B. E. (2010). *Identifying effective therapists with the Treatment Outcome Package: The need for a multi-dimensional approach*. Unpublished manuscript.

Kraus, D. R., Seligman, D., & 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.

Kraus, D., Wolfe, A. & Castonguay, L.G. (2006). The outcome assistant: A kinder philosophy to the management of outcome. *Psychotherapy Bulletin, 41*, 23–31.

Lambert, M. J. (2007). Presidential address: What have we learned from a decade of research aiming at improving psychotherapy outcome in routine clinical care. *Psychotherapy Research,17*, 1–14.

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research,11*, 49–68.

Lipzin, B. (2009). Quality improvement, pay for performance, and "Outcomes Measurement": What makes sense? *Psychiatric Services, 60*, 108–111.

Luborsky, L., McLellan, A. T., Diguer, L., Woody, G., & Seligman, D. A. (1997). The psychotherapist matters: Comparison of outcomes across twenty-two therapists and seven patient samples. *Clinical Psychology: Science and Practice, 4*, 53–65.

Lutz, W., Lowry, J., Kopta, S. M., Einstein, D. A., & Howard, K. I. (2001). Prediction of dose–response relations based on patient characteristics. *Journal of Clinical Psychology*, 57, 889–900.

Lutz, W., Martinovich, Z., & Howard, K. I. (1999). Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *Journal of Consulting and Clinical Psychology*, *67*, 571–577.

Mays, D, & Frank, C. (Eds.). (1985). *Negative outcome in psychotherapy.* New York: Springer.

McCarthy, W.C., & Frieze, I.H. (1999). Negative aspects of therapy: Client perceptions of therapists' social influence, burnout, and quality of care. *Journal of Social Issues, 55,* 33-50.

Miller, L. (1993). Who are the best psychotherapists? Qualities of the effective practitioner. *Psychotherapy in Private Practice, 12*, 1–18.

Mohr, D. C. (1995). Negative outcome in psychotherapy: A critical review. *Clinical Psychology: Science and Practice*, *2*, 1–27.

Nordberg, S., Boswell, J., Kraus, D., Castonguay, L. G., Hayes, J. A., & Wampold, B. (2010, June). *Therapist effectiveness treating depression with and without co-morbid substance abuse*. Paper presented at the meeting of the Society for Psychotherapy Research, Asilomar, CA.

Okiishi, Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology and Psychotherapy, 10*, 361–373.

Pope, K.S., Tabachnick, B.G., & Keith-Spiegel, P. (1987). Ethics of practice: The beliefs and behaviors of psychologists as therapists. *American Psychologist*, *42*, 993–1006.

Ricks, D. F. (1974). Supershrink: Methods of a therapist judged successful on the basis of adult outcomes of adolescent patients. In D. F. Ricks, M. F. Roff, A. Thomas, D. F. Ricks, M. F. Roff, & A. Thomas (Eds.), *Life history research in psychopathology* (Vol. 3, pp. 275–297). Minneapolis, MN: University of Minneapolis.

Safran, J. D., & Muran, C. (2000). Resolving therapeutic alliance ruptures: Diversity and integration. *Journal of Clinical Psychology, 56*, 233–243.

Simon, R. I. (1999). Therapist–patient sex: From boundary violations to sexual misconduct. *Psychiatric Clinics of North America, 22*, 31–47.

Strupp, H.H., Hadley, S.W., & Gomes-Schwartz, B. (1977). *Negative effects in psychotherapy: Clinical, theoretical and research issues*. New York: Jason Aronson.

Sullivan, H. S. (1953). *The interpersonal theory of psychiatry.* New York, NY: Norton.

Vogela, P. A., Hansen, B., Stiles, T. C., & Götestam, K. G. (2006). Treatment motivation, treatment expectancy, and helping alliance as predictors of outcome in cognitive behavioral treatment of OCD. *Journal of Behavior Therapy and Experimental Psychiatry, 37*, 247–255.

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.

Whitaker, C. (1989). *Midnight musings of a family therapist.* New York, NY: W.W. Norton & Company.