
Article:
Saxon, D. orcid.org/0000-0002-9753-8477 and Barkham, M. orcid.org/0000-0003-1687-6376 (2012) Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. Journal of Consulting and Clinical Psychology, 80 (4). pp. 535-546. ISSN 0022-006X

https://doi.org/10.1037/a0028898

Patterns of therapist variability:

Therapist effects and the contribution of patient severity and risk

Dave Saxon and Michael Barkham

Centre for Psychological Services Research

University of Sheffield

To appear in: *Journal of Consulting and Clinical Psychology*

Abstract

Objectives: To investigate the size of therapist effects using multilevel modeling (MLM), to compare the outcomes of therapists identified as above and below average, and to consider how key variables, in particular patient severity and risk and therapist caseload, contribute to therapist variability and outcomes.

Method: We used a large practice-based data set comprising patients referred to the UK's National Health Service primary care counselling and psychological therapy services between 2000 and 2008. Patients were included if they had received ≥ 2 sessions of one-to-one therapy (including an assessment), had a planned ending to treatment and completed the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM) at pre- and post-treatment. The study sample comprised 119 therapists and 10,786 patients, whose mean age was 42.1 years, and 71.5% were female. MLM, including Markov chain Monte Carlo procedures, were used to derive estimates to produce therapist effects and to analyze therapist variability.

Results: The model yielded a therapist effect of 6.6% for average patient severity but it ranged from 1%-10% as patient non-risk scores increased. Recovery rates for individual therapists ranged from 23.5% to 95.6% and greater patient severity and greater levels of aggregated patient risk in a therapist's caseload were associated with poorer outcomes.

Conclusions: The size of therapist effect was similar to those found elsewhere but the effect was greater for more severe patients. Differences in patient outcomes between those therapists identified as above or below average were large and greater therapist risk caseload rather than non-risk caseload was associated with poorer patient outcomes.

Keywords: Therapist effects, multilevel modeling, severity, risk, CORE-OM

Patterns of therapist variability:
Therapist effects and the contribution of patient severity and risk

Randomized controlled trials (RCTs) of psychological therapies have primarily focused on addressing the effects of specific treatments for specific conditions (e.g., Elkin et al., 1989; Hollon et al., 1992). In contrast, the potential contribution of individual therapists (Crits-Christoph & Mintz, 1991) has been relatively neglected in study design and analyzes. Therapists' competence and their adherence to specific techniques have been studied, although invariably by post hoc analysis of trials designed for other purposes, and with mixed findings on their contribution to outcome (Shaw et al., 1999; Trepka, Rees, Shapiro, Hardy, & Barkham, 2004; Webb, DeRubeis, & Barber, 2010). However, systematic differences between therapists in their outcomes have been found, both in trials (Huppert et al., 2001; Luborsky et al., 1986) and routine clinical practice (Okiishi et al., 2006; Wampold & Brown, 2005) where, although most therapists have mixed outcomes, some achieve generally better or poorer results. This has important implications both for the interpretation of research results and in improving the outcomes of therapy services. Therapist effects can moderate the relationship between specific techniques and outcome. For example, an early report of a finding of the superiority of cognitive behaviour therapy over psychodynamic interpersonal therapy in the treatment of depression (Shapiro & Firth, 1987) was later found to be attributable to the relatively poorer outcomes of one therapist with the latter modality (Shapiro, Firth-Cozens, & Stiles, 1989).

Notwithstanding the focus on interventions, a degree of variability in patient outcome due to therapist effects has been identified in some treatment trials (e.g., Clark et al., 2006) although not in others (e.g., Wilson, Wilfley, Agras, & Bryson, 2011). Recent attempts to

revisit well-designed archived trial data sets in order to estimate the size of these therapist effects have also yielded equivocal results even when using the same dataset as provided by the National Institute for Mental Health Treatment of Depression Collaborative Research Project (NIMH TDCRP; see Elkin, Falconnier, Martinovitch, & Mahoney, 2006; Kim, Wampold, & Bolt, 2006). Accordingly, Elkin et al. (2006) suggested that therapist effects would be best investigated using (very) large samples drawn from managed care or practice-based networks.

Historically, attention to the importance of therapist effects originated with Martindale's (1978) observations on the nature of the effects and related design issues that were, in turn, extended both by Crits-Christoph and Mintz's (1991) literature review and the most recent and comprehensive review of therapist effects (Baldwin & Imel, in press). This literature has highlighted the problems with ignoring therapist effects (i.e., to assume that all therapists are equally effective), the main one being that treatment effects are overestimated as a result (see Wampold & Serlin, 2000). Given that therapists usually do vary in their outcomes to some degree, this should be reflected in study designs and explicit in their analyses. Such analyses should model the natural structure of therapists and patients in which patients are grouped within therapists and the outcomes of patients treated by the same therapist are likely similar in some way and different from the outcomes of patients seen by another therapist. Recent studies of therapist effects (e.g. Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Okiishi et al., 2006; Wampold & Brown, 2005) have increasingly turned to using methods, such as multilevel modeling, that better reflect this nested structure and allow for the partitioning of the total variance in patient outcomes between the patient level and the therapist level. The therapist effect is the proportion of the total variance that is at the therapist level (Snijders & Bosker, 2004, Wampold & Brown 2005).

The precision of estimates of therapist effects depends on the number of therapists and the number of patients per therapist in the sample. Large numbers of therapists, in the order of at least 50 or preferably 100, are necessary for best estimates (Maas & Hox, 2004) and in a commentary on the findings of the TDCRP re-analysis, Soltz (2006) recommended that researchers use a minimum N of 30 therapists with a minimum of 30 patients nested within each therapist. In general, it is unlikely that trials can yield such numbers for both patients and therapists. In addition to having a large enough sample of therapists and patients to produce reliable estimates of therapist effects, such estimates drawn from naturalistic settings will have enhanced external validity.

In two recent naturalistic studies using multilevel modeling and larger samples, albeit smaller than those recommended by Soldz (2006), therapist effects of 5% (Wampold & Brown, 2005) and 8% (Lutz, et al., 2007) have been reported. The size of these effects may appear small but they should be considered in the context of the overall effect of psychological therapy, estimated at 20%, which includes all the constructs of therapy such as therapist factors, adherence to protocol, and the working alliance (Baldwin & Imel, in press). Given this context, therapist effects of 5% or 8% are quite large and of major importance in explaining variation in patient outcomes.

Beyond the actual size of therapist effect, studies invariably report effect sizes as a single percentage figure representing the effect for average patient intake severity. As patient severity is a key factor in predicting patient outcome (e.g., Garfield, 1994), there may be differences in therapist effects as patient intake severity increases. Whether the size of therapist effect is consistent across all levels of patient severity or whether the size of the effect is a function of patient severity has not been studied to date.

In response to the methodological and sample size recommendations referred to above, particularly those of Soldz (2006), we used multilievel modeling with a large naturalistic

dataset from the UK to estimate the size of therapist effect for average patient severity. In addition, in order to assess whether the therapist effect varies with patient severity we also estimated the size of the therapist effect at different levels of initial patient symptom scores.

The Pattern of Variability in Therapist Effectiveness

Moving beyond establishing the extent of therapist effects, we sought to establish the range of effectiveness by which therapists might be viewed as more or less effective compared with their peers. In the psychological therapies, using methods such as the simple ranking of therapist outcomes may penalize those therapists who have not contributed sufficient data to make a reliable estimate of effectiveness or who see more patients that are difficult to treat. By contrast, in the fields of education and health, Goldstein and Spiegelhalter (1996) argued for the adoption of appropriate statistical models that take account of other significant variables and present outcomes with their degree of uncertainty quantified by confidence intervals. Such methods provide the fairest means of making comparisons between institutions or practitioners in terms of their relative effectiveness and also provide information on those factors that explain outcome variation. Studies in education research have ranked and plotted the differences in effectiveness of individual schools using confidence intervals, after controling for the intake attainment of students (Goldstein & Healy, 1995; Goldstein & Speigelhalter, 1996).

In our study, using similar methods, the variability in therapist effectiveness was represented by the degree to which a therapist's outcomes depart from those of the average therapist, while controling for other variables. Ranking and plotting this variability produces a graphical representation of the pattern of therapist variability in effectiveness. Given that all therapists will vary from the average to some extent, by plotting confidence intervals for the estimate for each therapist, therapists can more reliably be defined as within the average range or above or below the average range.

Case-mix

If comparisons of effectiveness are to be made between therapists, factors that are strongly associated with patient outcomes, that are likely to be unevenly distributed between therapists, need to be controlled in the analysis. Case-mix may be defined therefore as the characteristics, or profiles of the patients treated by a therapist. By including in the model measures of a therapist's case-mix that are predictive of outcome, not only are they controlled for but their relative impact on outcome can be estimated.

Initial patient severity is the leading case-mix variable associated with patient outcomes (Garfield, 1994; Kim et al., 2006). Okiishi et al. (2006), supporting earlier findings (cf. Luborsky, McLellan, Diguer, Woody, & Seligman, 1997), found that once initial severity was taken into account, other patient variables added relatively little value in predicting outcomes. However, another key patient variable that might contribute to therapist effectiveness is the level of patient risk. The risk of a patient harming themselves or others is of paramount concern to therapists and services and the risk level of patients is often monitored (Saxon, Ricketts, & Heywood, 2010). In responding to the presentation of patient risk, some therapists may, within a time-limited therapy, focus on addressing high patient risk at the expense of responding to other aspects of a patient's condition. Mindful of the priority for practitioners, we investigated the contribution of patient risk in addition to patient baseline severity.

The caseload burden of patient severity and risk may also have a significant effect on patient outcomes. There is a growing focus on caseload management in the helping professions. For example, Borkovec, Echemendia, Ragusea, and Ruiz (2001) found that the more patients a therapist had in their caseload, the poorer the average outcome of the caseload. Similarly, Vocisano et al. (2004) reported that therapist caseload was the second most important factor in determining treatment outcome. In a recent study of pediatric

community occupational therapists, Kolehmainen, MacLennan, Francis, and Duncan (2010) found that their caseload management behaviors were associated with children's length of treatment. Accordingly, we investigated risk and non-risk caseload as therapist variables.

In light of the above, we applied multilevel modeling to address the following three aims. First, to provide an estimate of the size of therapist effects in routine practice settings and to use the model to investigate whether the therapist effect is greater for more distressed patients. Second, to use reliable estimates of relative therapist effectiveness to identify and compare the outcomes of above and below average therapists. And third, to assess the individual contributions to outcome of patient intake severity and risk, as well as therapist severity and risk caseload.

Method

Original Data Set

The initial data set comprised data on 70,245 patients referred to UK primary care counseling and psychological therapy services between January 1999 and October 2008 and was named the Clinical Outcomes in Routine Evaluation Practice-Based Evidence National Database-2008. It represented data from 35 sites nationally and 1,059 therapists who saw between 1 and 1,084 patients each (M = 66.3; SD = 114.4). In most cases patients were allocated to the next available therapist and therapy was usually time-limited to 6 or 7 sessions (M = 5.9; SD = 3.0; *Median* = 6), including an assessment at the first session. This dataset was an updated version of earlier datasets used in studies by our research group (e.g., Stiles, Barkham, Connell, & Mellor-Clark, 2008a) and ethics approval for the study was covered by the UK National Health Service's Central Office for Research Ethics Committee, application 05/Q1206/128.

Study-specific Data Set

For the purposes of this study, patients were included if they were 18 or over, received two or more sessions comprising an initial assessment and one-to-one therapy, had a planned ending to treatment, and completed a common standardized outcome measure at the beginning and end of their treatment. Further, only therapists with 30 or more patients were included in order to satisfy the recommendations of Soltz (2006).

Patient demographics and assessment information were collected on all patients. However, the dataset contained therapists with a wide range of return rates of pre- and post-treatment patient outcome measures. For those patients meeting the other inclusion criteria, this ranged from 24.2% to 100%, despite all patients having a planned ending to treatment. Therefore, in order to address any bias due to possible case selection by therapists with particularly low return rates, a subset of those therapists with a pre-post measure return rate of 90% or more was selected, a return rate consistent with targets set by the UK's Department of Health in relation to its program on Improving Access to Psychological Therapies (Department of Health, 2008). Adopting this return rate resulted in a dataset of 10,786 patients seen by 119 therapists between September 2000 and July 2008. With only 22 sites and 10 sites having only 1 or 2 therapists, it was not possible to include site as a variable in the model.

Of the patients included, the mean age was 42.1 years (*SD* = 13.3), 71.5% were female, 94.4% were white British/European, and 50.2% were on medication, most commonly anti-depressants (44.8%). No formal diagnosis was recorded but therapists' assessments, derived from the CORE Assessment (Barkham, Gilbert, Connell, Marshall, & Twigg, 2005) indicated 77.2% to have some level of depression (44.0% rated as ranging between moderate and severe) and 84.6% had some level of anxiety (58.8% rated as ranging between moderate and severe)

Measurement: Assessment and Outcome

Our primary outcome measure was the CORE-OM (Barkham et al., 2001; Barkham, Mellor-Clark, Connell, & Cahill, 2006; Evans et al., 2002). The CORE-OM is a self-report measure comprising 34 items addressing the domains of subjective wellbeing (4 items: e.g., I have felt optimistic about my future), symptoms (12 items: e.g., I have felt totally lacking in energy and enthusiasm), functioning (12 items: e.g., I have felt able to cope when things go wrong), and risk (6 items). The risk domain captured both risk-to-self (4 items: e.g., I have made plans to end my life) and risk-to-others (2 items: e.g., I have been physically violent to others). The CORE-OM is reproduced in full elsewhere and is free to copy providing it is not altered in any way or used for financial gain (see Barkham et al., 2010a). Items are scored on a 5-point, 0-4 scale anchored by the following terms: *Not at all*, *Only occasionally*, *Sometimes*, *Often*, and *All or most of the time*. Forms are considered valid providing no more than three items are omitted (Evans et al., 2002). CORE-OM clinical scores are computed as the mean of all completed items, which is then multiplied by 10, so that clinically meaningful differences are represented by whole numbers. Thus, CORE-OM clinical scores can range from 0 to 40. The 34-item scale has a reported internal consistency of .94 (Barkham et al., 2001) and a one-month test-retest correlation of .88 (Barkham, Mullin, Leach, Stiles, & Lucock, 2007). Factor analysis indicates that the risk domain is measuring a different aspect of severity than the other 3 domains (Evans et al., 2002). Therefore mean risk items (n=6) and non-risk items (n=28) were scored separately to provide a risk and a non-risk score, each ranging from 0 – 40, for each patient. The risk and non-risk scales have internal consistencies of .79 and .94 respectively (Evans et al., 2002). Patients completed the CORE-OM prior to therapy and at the final treatment session. As measures of therapist caseload, therapist-level aggregated non-risk and risk scores were also calculated.

In addition, therapists' recovery rates were produced adopting procedures set out by Jacobson and Truax (1991) for determining reliable and clinically significant change in

patient outcome scores. Two criteria needed to be met. First, the change scores for patients needed to be greater than the reliable change index for the CORE-OM in order to take account of measurement error. We used a reliable change score of ±5 akin to the value used in other studies using the CORE-OM (e.g., Stiles et al., 2008). Hence a reduction of at least 5 points indicated reliable improvement while an increase of 5 points indicated reliable deterioration. Second, patients' scores had to change from being above the clinical cut-off at pre-treatment to being below the clinical cut-off at post-treatment. We used a clinical cut-off score of 10, which has reported sensitivity and specificity values of .87 and .88 respectively (for details, see Connell et al., 2007). Patients meeting both criteria (i.e., reliable improvement and moving from the clinical into the non-clinical population) were deemed to have made statistical recovery, a term we used to reflect the source of recovery being a statistical rather than a clinical procedure. The proportion of a therapist's patients who recovered statistically was considered a useful and meaningful measure of therapist effectiveness.

Analysis

The statistical concepts and methodology adopted in this study are fully described elsewhere (e.g., Kim et al., 2006; Rasbash, Steele, Browne, & Goldstein, 2009; Snijders & Bosker, 2004). A multilevel model was developed with patients at level 1 and therapists at level 2 and pre-treatment patient CORE scores were entered first, grand mean centered (Hoffmann & Gavin, 1998; Wampold & Brown, 2005). Other explanatory variables were added to the model, also grand mean centered, and were tested for significance by dividing the derived coefficients by their standard errors. Values greater than 1.96 were considered significant at the 5% level. Because patient outcome scores and patient intake risk scores were positively skewed, outcome scores and intake risk and non-risk scores were log-transformed for the model development.

Multilevel modeling software MLwiN v2.24 (Rasbash, Charlton, Browne, Healy, & Cameron, 2009) was used to estimate parameters, initially by Iterative Generalised Least Squares (IGLS) procedures. The multilevel model was developed from a single level regression model and improvements in the models judged by testing the difference in the -2*loglikelihood ratios produced by each model, against the chi squared distribution for the degrees of freedom of the additional parameters. Variation between therapists in the relationship between outcome and each explanatory variable was considered using random slope models.

The model produced by these IGLS procedures indicated a curvilinear relationship between the intake patient severity scores and outcome scores and also a cross-level interaction between a therapist variable and a patient variable. Such complexities can reduce the reliability of estimates produced by IGLS methods, therefore using the IGLS estimates as 'priors', Markov chain Monte Carlo (MCMC) estimation procedures, were run within MLwiN. This simulation approach uses the model to produce a large number of estimates of the unknown parameters that can be summarised to derive more reliable final estimates (Browne, 2009).

The therapist effect for the average patient severity was calculated by dividing the level 2 variance by the total variance in order to give the variance partition coefficient (VPC; Lewis et al., 2010; Rasbash, Steele et al., 2009). The VPC (akin to the intra-class correlation coefficient) is multiplied by 100 to give the therapist effect. In addition, the VPC and therapist effect were estimated for all levels of patient intake non-risk score.

The individual therapist residuals produced by the model represent the degree to which each therapist varies in effectiveness from the average therapist. This residual varies between therapists and is assumed to have a normal distribution and a mean of zero. In MLM, the intercept residual produced by the multilevel model represents the additional impact of

therapist on outcome, not explained by other variables contained in the model. Positively signed therapist residuals will have the effect of increasing outcome scores (i.e. worsen outcome), while negatively signed residuals will reduce outcome score. The size of the residuals can therefore be used to make comparisons between higher-level units, such as practitioners or institutions. (Goldstein & Speigelhalter, 1996; Rasbash, Steele et al., 2009; Wampold & Brown, 2005).

The therapist residuals were ranked and plotted with their confidence intervals (CIs). In education research the aim has been to provide a means of comparing the outcomes of pairs of schools, and CIs of 84% have been adopted (Goldstein & Healy, 1995). However our aim was not to compare pairs of therapists but rather to make more general comparisons between groups of therapists. Accordingly, the more usual 95% CI was used.

We constructed three groups of therapists based on the outcomes of their patients. Therapists whose residual CIs crossed the average therapist residual were identified as being of average effectiveness, while those therapists whose CI did not cross the average were considered either significantly above or below average effectiveness. In order to assess the differences between these three groups, patient and therapist outcomes and statistical recovery rate comparisons were made. Finally, using the estimates produced by the model, combinations of different levels of the included variables were plotted against predicted outcome scores to illustrate how the variables related to each other and to patient outcome.

Results

Initial analysis considered the data at the patient level in order to assess the data distributions and calculate overall effectiveness. Intake severity and outcomes were then calculated at both the patient and therapist level (Table 1) before development of the multilevel model.

For patients, the mean (*SD*) pre- to post-therapy change on the CORE-OM was 9.3 (*SD* = 6.3), with a range from -17.4 to +33.8 and yielded a pre- to post-therapy effect size of 1.55. Of patients scoring above the clinical cut-off (i.e., CORE-OM score ≥10 or more) at pre-therapy (N=9673), 61.6% met the criteria for reliable and clinically significant improvement (i.e., recovered statistically). For non-risk scores the mean change was 10.8 (*SD* = 7.3) with a range from -18.6 to +38.6. For risk scores 46% of patients had a risk score of zero (no risk) resulting in an overall small mean change of 2.5 (*SD* = 4.6), but there were extremes of -30.0 and +35.0. There were positive correlations between non-risk scores and outcome scores (*Pearson's r* = .428, $p < .001$), and between risk scores and outcome scores (*Pearson's r* = .292, $p < .001$).

For therapists, pre- to post-therapy change was normally distributed on all three indices of the measure (i.e., overall CORE-OM score, non-risk component, and risk component). For the CORE-OM the mean change was 8.9 (*SD* = 1.7) with a range from 4.5 to 13.5. For the non-risk items it was 10.3 (*SD* = 2.0) with a range 5.3 to 15.8 and for risk items the mean change was 2.5 (*SD* = 0.8) with a range from 0.9 to 4.6.

Therapist Effects

Multilevel modeling. IGLS methods were used to develop the model and provide estimates of the parameters for MCMC simulation procedures. Examination of the MCMC diagnostics and tests of convergence indicated a 'burn-in' of 500 followed by 25000 iterations to be adequate. Assumptions of Normality in the data were tested by plotting the patient level and therapist level residuals produced by the model to normal distribution curves (quantile-quantile plots). These were relatively linear ($x = y$), therefore Normality can be assumed. The final MCMC model is presented in Appendix A[(1)].

The MCMC model included patient non-risk and risk score and therapist risk caseload as significant predictors of outcome, with above average scores on each contributing to poorer

outcome. Therapist non-risk caseload and the interaction between patient non-risk score and therapist risk caseload, which had borderline significance in the IGLS model, were not significant following MCMC procedures.

The random slope for patient intake non-risk score indicates therapist variation in the relationship between patient intake non-risk score and outcome. The model also indicates a small positive covariance between therapist intercepts and the slopes (0.010, *SE* = 0.002), which describes a slight fanning out of the therapist regression lines. This would suggest that those therapists with poorer outcomes overall tended to be effected more negatively by increases in patient intake severity, than therapists with better outcomes overall.

The therapist effect for this final model was 6.6%. Considering the model without the therapist risk caseload variable produced a therapist effect of 7.8%, indicating that therapist risk caseload explained some of the variation between therapists. These therapist effects are slightly larger than those estimated by IGLS procedures (6.4% and 7.6% respectively).

Therapist effects and patient severity. The full MCMC model produced a VPC of 0.066, a therapist effect of 6.6%, for the average patient on all explanatory variables. Patient non-risk scores made the greatest contribution to outcomes and the VPCs were estimated for different patient intake non-risk scores (Rasbash, Steele et al., 2009). Figure 1, plots the VPCs and illustrates how the proportion of the unexplained difference in outcome between patients, attributable to therapists, varies with patient non-risk severity. It shows that with CORE non-risk scores of less than 3, there are differences in therapist effects of between 2% and 1%. However, as intake scores increase, the therapist effect rises to 10%.

Therapist Residuals and Effectiveness

In Figure 2, the therapist intercept residuals produced by the model are ranked and presented with their 95% confidence intervals. These represent how each therapist's outcomes differ from the average therapist outcome, controlling for the patient severity and

therapist caseload variables. Counterintuitively, but in common with the reporting of level 2 residuals elsewhere, better outcomes are presented to the bottom left with negative residuals while poorer outcomes have positive residuals (cf. Goldstein & Healy, 1995; Wampold & Brown, 2005). The plot indicates that for 79 (66.4%) therapists whose confidence intervals cross zero, their outcomes cannot be considered different from the average therapist. However, for 21 (17.7%) therapists their outcomes were better than average, while for 19 (16.0%) their outcomes were poorer than average (i.e., the CIs for these 40 therapists did not cross zero).

Although patient intake non-risk score is the main predictor of outcome score, the significant random slope in the model indicates that the relationship between patient intake non-risk score and outcome varied between therapists. The residuals for the slope of each therapist were highly correlated with the intercept residuals (*Pearson's r* = .996, $p < .001$), but the 95% CIs for the slope residuals indicated that only 17 therapists had a relationship between patient non-risk score and outcome that was significantly different than average. Eleven of these were amongst the 21 more effective therapists identified in Figure 2, for these 11 therapists, increases in patient severity had a less than average impact on their outcome scores. Six of the less effective therapists identified in Figure 2, also had a relationship between intake non-risk score and outcome that was significantly different to that of the average therapist. However, for these six therapists increases in patient intake score had a greater than average impact on their outcome scores.

Comparisons of Therapist Effectiveness

The mean (*SD*) recovery rate for all therapists was 58.8% (13.7), but the range across therapists varied from 23.5% to 95.6%. Tables 2 and 3 show the numbers of therapists and patients in each of the 3 groups of therapists, identified above as average or above or below average, and the group recovery rates. In Table 2, the proportion of patients scoring above the

clinical cut-off on CORE-OM at intake was similar across the three groups, while the patient recovery rate varied from 42.4% to 77.0%. Table 3 shows the pre- and post-therapy CORE-OM, risk and non-risk patient means for each therapist group. ANOVAs indicated no significant differences (all p values >0.05) between groups on intake measures but there were significant differences on all scores at outcome. Pre- to post-therapy change on the CORE-OM was 61% less for the below average group compared to the above average group.

Table 4 shows the aggregated therapist recovery rates and the range of individual therapist recovery rates within each group. When we considered the rate for reliable deterioration, the rate – albeit small – varied from 0.5% for the above average group, to 0.6% for the average group and 1.6% for the below average group. Table 4 indicates a considerable overlap of the recovery rate ranges due to the controling for intake scores and risk caseload in the model. Eight of the 19 therapists in our below average group, were not ranked in the bottom 19 therapists in terms of recovery rates, while eight therapists identified by our model as average were amongst those 19 therapists with the lowest recovery rates.

To assess the effect on patient outcomes of the 19 therapists identified as below average by the model, they and their 1947 patients were excluded from the dataset and the model development procedures repeated. The significant variables remained the same but the values of the coefficients changed and the therapist effect was reduced to 4.6%. The overall patient recovery rate increased from 61.6% to 64.9% while the aggregated, therapist mean recovery rate increased from 58.8% to 61.7%. If the 1704 clinical patients (Table 2) of the least effective therapists were treated by therapists with the average recovery rate (61.7%), then 1049 rather than 786 would have recovered, an additional 265 patients.

Graphical Representation of the Model

To illustrate how the different variables included in the model (Appendix A) relate and interact, predicted patient outcome scores were plotted for combinations of different levels of

patient non-risk and risk and therapist risk (Figure 3). Outcomes for the 5th, 50th and 95th percentile scores for patient intake risk (scores of 0, 1.7, 15.0), for therapist risk caseload (scores of 2.0, 3.5 and 5.4), were plotted for the 5th, 50th and 95th percentile scores of patient non-risk, (scores of 10.0, 20.7, 31.1) along the Y axis. Five of the 9 plots are shown in Figure 3, representing the average and the extremes of the range with the lines of other combinations located within this range.

The middle full line represents predicted outcomes for the 50th percentile therapist risk score and the 50th percentile patient risk score. Above this, the dashed line represents the 95th percentile therapist risk score and the 5th percentile patient risk score while the dotted line above represents the 95th percentile on both patient risk score therapist risk score. The lower dashed line is the predicted outcome for the 95th percentile patient risk score and the 5th percentile therapist risk score and the bottom dotted line represents the predicted outcome for the 5th percentile on both scores. Figure 3 illustrates how greater therapist risk caseload is associated with poorer patient outcomes with the poorest outcome predicted for a patient with a high risk score seen by a therapist with a high risk caseload. However, a patient with a high risk score seen by a therapist with a low risk caseload has a predicted outcome similar to a patient with median scores on both. The relationships between the variables are consistent across the levels of patient intake non-risk score, although as patient non-risk scores increase, the effect of risk increases slightly.

Discussion

In this practice-based study of primary care counseling and psychological therapy services in the UK, our aim was to establish the degree to which therapists contribute to variability in patient outcomes. In doing so, we used MLM and MCMC procedures to estimate the size of the therapist effect for different levels of patient intake severity and, adding to the evidence base for therapist variability, considered patient risk and therapist caseload as explanatory

variables. Using the multilevel model, we identified therapists that were either significantly more or significantly less effective than average therapists and compared their outcomes in terms of recovery rates. Our approach was in response to calls by commentators to adopt improved methods for the analyses of data sets such that for our analyzes we used a dataset meeting the most stringent recommended sample size of therapists and patients within therapists (Maas & Hox, 2004; Soldz, 2006), in which therapists were treated as random, assumptions of normality were tested, standard errors were reported, and the extremes of therapist variation considered (Crits-Christoph & Mintz, 1991; Elkin et al., 2006; Soldz, 2006).

In terms of the general effectiveness of the therapy delivered, the pre- to post-therapy effect size of 1.55 is broadly similar to outcomes reported in other independent datasets. For example, Richards and Suckling (2010) reported a pre-post effect size of 1.42 for the PHQ-9 on a completer sample of patients similar to that employed in the current study. Cahill, Barkham, and Stiles (2010) reported a slightly lower average pre-post effect size derived from 10 studies of 1.19 and a patient recovery rate of 56%. Our overall finding of 6.6% of variation in patient outcome due to therapist effects (7.8% when only pre-treatment patient scores were included in the model) lies between the 5% reported by Wampold and Brown (2005) in a study of managed care where therapy was more irregular and the 8.26% reported by Lutz et al. (2007), whose study included non-completers of treatment.

In other areas of healthcare, few studies have considered the practitioner as the grouping variable. Studies of surgery for colorectal cancer, found large differences in surgeon outcomes after controlling for known risk factors (McArdle, 2000; McArdle & Hole, 1991), while a study comparing treatments for back and neck pain found practitioner effects, derived from VPCs, of between 2.6% and 7.1% (Lewis et al., 2010).

The size of the therapist effect found in the current study and other naturalistic studies of psychological therapy are broadly consistent, although larger therapist effects may be found in the treatment of specific populations of patients. One study found a therapist effect of almost 29% in the treatment of racial and ethnic minority patients, although this finding was derived from a relatively small sample (Larrison & Schoppelrey, 2011).

In our study we found an increasing degree of variability between practitioners as the severity levels of patients became elevated (Figure 1). At very low levels of patient severity, where scores are similar to those found in the normative population (i.e. 0 to 5) the therapist effect is below 3% but this rises to 10% as patient intake severity increases. The sharp curve for very low scores may be partly due to the nature of these low-scoring patients and the reasons they are receiving therapy, but also the VPCs at the extremes of the non-risk score distribution may be less reliable due to the smaller sample size. For most of the pre-therapy non-risk distribution, as scores rose from 5 to 35 (out of a maximum of 40), therapist effects increased from about 3% to 9%. Therefore, the outcomes for less severe patients were more similar across therapists than outcomes for more severe patients. Put another way, the more severe a patient's intake symptoms, the more their outcome depended on which therapist they saw. Similar findings have been reported in a large naturalistic study of surgeon effects in adult cardiac surgery (Bridgewater et al., 2003).

Patient non-risk scores made the largest contribution to outcomes but the relationship between intake non-risk score and outcome score varied between therapists. Our results suggest that although greater intake severity may generally result in poorer outcomes, for some more effective therapists this had a less detrimental effect than average while for some less effective therapists the detrimental effect was greater than average. The relationship between patient risk score and outcome did not vary significantly between therapists and the difference between our above and below average therapists in the pre-post change on risk

score was proportionally less than the difference for non-risk score. The differences in the impact of patient risk and non-risk scores suggests some support for Kraus, Castonguay, Boswell, Nordberg, & Hayes (2011) who, using single level analyses, found that therapists varied in effectiveness on different aspects of the patient's condition, as measured by different domains of the outcome measure.

We found that at the therapist level, where patient risk and non-risk were each aggregated to produce measures of therapist caseload, a greater therapist risk caseload contributed to poorer patient outcomes, while therapist non-risk caseload was not predictive of patient outcome. We can only speculate as to why this may be. Therapists may feel more pressure to help patients at risk of harming themselves or others and this heightened pressure may be contributing to a reduction in their overall effectiveness. This may be linked to therapist burnout, which has been shown to have a negative effect on patient outcomes (McCarthy & Frieze, 1999). The issue of caseload has been identified as crucial in the management of the psychological therapies and there have been calls for this factor to have greater prominence due to its relevance to public health (Vocisano et al., 2004).

The shape of therapist variability found by ranking and plotting therapist residuals and their confidence intervals (recall Figure 2), is similar to profiles found in the comparison of health and education institutions (Goldstein & Healy, 1995; NHS Performance Indicators, 2002). However, only a few studies have considered psychological therapist variation using therapist residuals (e.g., Wampold & Brown, 2005). The plotted residuals show the extent of variation in performance after controlling for case-mix and caseload, with the most and least effective therapists being considered the tails of the distribution of therapist effectiveness in naturalist settings (Lutz et al., 2007). Studies have highlighted the utility and possible benefits of studying the practices of the most effective therapists (e.g., American Psychological Association, 2006; Brown, Lambert, Jones & Minami , 2005; Okiishi, Lambert, Neilsen &

Ogles, 2003; Okiishi et al 2006,). However, studies so far using MLM have shown that therapist variables such as type and amount of training, theoretical orientation and gender are not predictive of patient outcome (Okiishi et al., 2006).

The study of the most effective therapists may provide useful insights into their characteristics, and what makes them more effective, which could have implications for training and recruitment. However, focusing on effective therapists can detract from acknowledging that the average group of therapists in the present study were themselves effective, with a patient recovery rate of 60%, and that, in terms of any service delivery model, these therapists comprise the bulk of professional resources.

In contrast to both the effective and average therapists, it is those who consistently produce below average outcomes (19 in the current study) after adjusting for case-mix and caseload that should be a cause of professional concern. Only around 9 in 20 of their patients recovered despite completing treatment, while for the above average therapists the figure was 16 in 20. That is, the probability of recovery was almost twice as likely with the most effective therapists than with the least effective therapists. In addition, the deterioration rate for the least effective therapists was around 3 times that of other therapists. When the 19 least effective therapists and their patients were removed, we found an improvement in overall patient recovery rate of about 3.0%. In our dataset, we calculated that an additional 265 patients would have recovered had they been seen by therapists with average recovery rates. If all practicing therapists and their patients were considered, and considered over time, then this would equate to many thousands of additional patients who could benefit from therapy (Baldwin & Imel, in press)

In the current study, in common with routine data collection generally, there was minimal information held on therapists. This militated against our being able to investigate what it was about some therapists that made them more effective than others. In order to carry forward

this area of research, there is a pressing need for more complete information on the practitioners in routine practice samples.

In our study, practitioners were counselors working in a range of primary care mental health settings and utilizing a range of treatment types to varying degrees. Adherence to a treatment protocol, a desideratum in trials but also a component in treatment guidelines for routine practice as espoused by the UK National Institute for Health and Clinical Excellence (NICE), may reduce therapist variation. However, a single level study found that adherence to protocol was not predictive of patient outcome (Webb, DeRubeis & Barber, 2010) and it would be informative to study therapist effects in services with greater adherence to a treatment protocol.

The methods used in this study (i.e., MLM and the use of residuals to assess the relative effectiveness of therapists) have been taken largely from education research. They arose from the development of 'performance indicators' designed to make quantitative comparisons between schools and were in answer to cruder methods, such as the simple ranking of schools outcomes (Goldstein & Speigelhalter, 1996). At the present time, 'performance indicators' are being developed and introduced in health care services, including psychological therapies, and it will be important that the appropriate methods are used to make comparisons both between services and between practitioners. We found a considerable overlap of the ranges of recovery rates between the three groups of therapists and some therapists we identified as average had recovery rates lower than some therapists identified as below average. This was due to our methods and adjustments for case-mix and caseload but it is an indication of the perils of using simplistic methods, such as comparisons based solely on therapist outcomes. If such methods were used, some less effective therapists may not be identified and a number of average therapists may be deemed to be under-performing.

The limitations of the present study are those that can be leveled against studies within the paradigm of practice-based evidence and have been well documented and addressed (for a detailed summary and discussion, see Barkham, Stiles, Lambert, & Mellor-Clark et al., 2010b; Stiles et al., 2006, 2008b). Crucial is the issue of the representativeness of included data (Brown et al., 2005). In order to control for any bias due to the failure to collect measures from patients, only those therapists with a pre- and post-therapy measure return rates of over 90% of their treated patients were included in our sample. Including only those patients who completed their planned treatment may have inflated the overall effectiveness figures reported here and it will be important to consider how therapist variability is affected by the inclusion of patients who dropout of treatment. The study by Lutz et al. (2007) suggests the therapist effect may be slightly larger. Also, results here are only generalizable to therapists who have treated more than 30 patients and therapist effects may be larger if trainees and less experienced therapists are included in a sample.

Implications for Clinical Practice and Research

In terms of implications for clinical practice, our findings of greater therapist variation in the outcomes of more severe patients, and the effect of higher risk therapist caseloads on outcomes, may indicate support for the careful allocation of patients to therapists, as suggested elsewhere (e.g., Brown et al., 2005; Okiishi et al., 2003, 2006). There is also a responsibility on service managers to understand and then act appropriately in light of data that shows a therapist to consistently yield poor outcomes for their patients. Both approaches require service monitoring at a therapist level, monitoring patient allocation, and managing therapist caseloads. Furthermore, services need to adopt appropriate and responsive methods for assessing the relative effectiveness of therapists, identifying those therapists falling below the average range, and providing the necessary additional and ongoing professional training. In terms of protecting patient safety, the quality of treatment delivered, and the considerable

investment in training of practitioners, it is imperative that supervisors and service managers take collective responsibility for ensuring that appropriate action is taken where there is consistent evidence of outcomes that are appreciably below average. Equally, understanding what aspects of practice make some therapists particularly effective needs to be understood and fed back into principles of good practice.

In relation to research approaches, methods such as MLM, may seem unfamiliar and complex but they are increasingly being adopted as a means of understanding what is a complex intervention, namely psychological therapy, and efforts are being made to make these methods more accessible to practitioners and others (see Adelson & Owen, 2011). Vital to these methodologies is a large sample size and routine data are now being collected more widely in psychological services. By collecting clinically useful data, it should be possible to use the data systems and appropriate statistical methods to monitor therapist outcomes regularly and provide feedback to therapists and services. The benefits and problems of this development are described elsewhere (Goldstein and Spiegelhalter, 1996, Baldwin & Imel, in press), but Goldstein and Spiegelhalter (1996) emphasize that the use of monitoring and feedback to improve service outcomes should be approached sensitively and be a collaborative rather than confrontational process (Goldstein and Spiegelhalter, 1996).

In conclusion, we have shown that reports of therapist effects of around 8.0% are robust and after controlling for case-mix, the effect was still significant, at 6.6%. Accordingly, we conclude that most of the variation in patient outcome due to therapists is attributable to other untested variables. In addition, our results indicate a larger therapist effect as patient non-risk severity increases and a greater therapist risk caseload to be associated with poorer patient outcomes. However, even after controlling for these variables we found a considerable difference in effectiveness between therapists. This study illustrates that the reporting of simple aggregated outcomes for services and practitioners is limiting and can be misleading,

masking important factors for effective service delivery. It adds to the growing body of research, using large routine datasets and sophisticated methodologies such as MLM, that is moving beyond establishing the existence and size of therapist effects in practice to investigating the reasons for the variability, its impact on patient outcomes, and the implications for therapist training and service provision. Future research should test the model on other large datasets and consider further the relationships between patient severity, risk, therapist caseload, other therapist variables, and patient outcome.

References

Adelson, J. L., & Owen, J. (in press). Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. *Psychotherapy: Theory, Research, Practice, Training.* doi: 10.1037/a0023990

American Psychological Association Task Force on Evidence-Based Practice (2006). Report of the 2005 Presidential Task Force on Evidence-Based Practice in Psychology. *American Psychologist, 61,* 271-285. doi: 10.1037/0003-066X.61.4.271

Baldwin, S. A., & Imel, Z. E. (in press). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change*. 6th edition. Wiley and Sons.

Barkham, M., Gilbert, N., Connell, J., Marshall, C. & Twigg, E. (2005). Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings. *British Journal of Psychiatry, 186,* 239-246. doi:10.1192/bjp.186.3.239

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C.… & McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Towards practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*, 184-196. doi:10.1037/0022-006X.69.2.184

Barkham, M., Mellor-Clark. J., Connell, J., & Cahill J. (2006). A CORE approach to practice-based evidence: A brief history of the origins and applications of the CORE-OM and CORE System. *Counselling & Psychotherapy Research, 6,* 3-15. doi:10.1080/14733140600581218

Barkham, M., Mellor-Clark, J., Connell, J., Evans, R., Evans, C., & Margison, F. (2010a). The CORE measures & CORE system: Measuring, monitoring, and managing quality evaluation in the psychological therapies. In M. Barkham, G. E. Hardy, & J. Mellor-

Clark (Eds.), *Developing and delivering practice-based evidence: A guide for the psychological therapies.* (pp. 175-219). Chichester: Wiley.

Barkham, M., Mullin, T., Leach, C., Stiles, W. B., & Lucock, M. (2007). Stability of the CORE-OM and BDI-I: Psychometric properties and implications for routine practice. *Psychology & Psychotherapy: Theory, Research & Practice, 80,* 269-278. doi:10.1348/147608306X148048

Barkham, M., Stiles, W. B., Lambert, M. J., & Mellor-Clark, J. (2010b). Building a rigorous and relevant knowledge-base for the psychological therapies. In M. Barkham, G.E. Hardy, & J. Mellor-Clark (Eds.), *Developing and delivering practice-based evidence: A guide for the psychological therapies* (pp. 21-61). Chichester: Wiley.

Borkovec, T. D., Echemendia, R. J., Ragusea, S. A., & Ruitz, M. (2001). The Pennsylvania practice research network and future possibilities for clinically meaningul and scientifically rigorous psychotherapy effectiveness research. *Clinical Psychology: Science and Practice, 8*, 155-167.

Bridgewater, B., Grayson, A. D., Jackson, M., Brooks, N., Grotte, G. J., Keenan, D. J. M.… & Jones, M. (2003.) Surgeon specific mortality in adult cardiac surgery: comparison between crude and risk stratified data. *BMJ, 327,* 13-17. doi:10.1136/bmj.327.7405.13

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

Browne, W. J. (2009). *MCMC estimation in MLwiN Version 2.13.* Centre for Multilevel Modelling, University of Bristol.

Cahill, J., Barkham, M., & Stiles, W. B. (2010). Systematic review of practice-based research on psychological therapies in routine clinic settings. *British Journal of Clinical Psychology, 49,* 421-454. doi:10.1348/014466509X470789

Clark, D. M., Ehlers, A., Hackmann, A., McManus, F., Fennell, M., Grey, N.… & Wild, J. (2006). Cognitive therapy versus exposure and applied relaxation in social phobia: A randomised controlled trial. *Journal of Consulting and Clinical Psychology, 74*, 568-578. doi:10.1037/0022-006X.74.3.568

Connell, J., Barkham, M., Stiles, W. B., Twigg, E., Singleton, N., Evans, O., & Miles, J. N. V. (2007). Distribution of CORE-OM scores in a general population, clinical cut-off points, and comparison with the CIS-R. *British Journal of Psychiatry, 190*, 69-74. doi:10.1192/bjp.bp.105.017657

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapy. *Journal of Consulting and Clinical Psychology, 54*, 20-26. doi:10.1037/0022-006X.59.1.20

Department of Health, Mental Health Programme (2008). *Improving Access to Psychological Therapies Implementation Plan: National guidelines for regional delivery*. Department of Health. Crown Copyright 2008.

Elkin, I., Falconnier, L., Martinovitch, Z., & Mahoney, C. (2006). Therapist effects in the NIMH Treatment of Depression Collaborative Research Program. *Psychotherapy Research, 16*, 144-160. doi:10.1080/10503300500268540

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F.,… & Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: general effectiveness of treatments. *Archives of General Psychiatry, 46,* 971–892.

Evans, C., Connell, J., Barkham, M., Margison, F., Mellor-Clark, J., McGrath, G. & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, 180, 51-60. doi:10.1192/bjp.180.1.51

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change (*4th ed.) New York: Wiley.

Goldstein, H. & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 158,* 175-177

Goldstein, H. & Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance-with discussion. *Journal of the Royal Statistical Society*. *159*, 385-443.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 23*, 723-744. doi:10.1177/014920639802400504

Hollon, S. D., DeRebeis, R. J., Evans, M. D., Wiener, M. J., Garvey, M. J., Grove, W. M., & Tuason, V. B. (1992). Cognitive therapy and pharmacotherapy for depression: Singly and in combination. *Archives of General Psychiatry, 49*, 774-781.

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. doi:10.1037/0022-006X.69.5.747

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19. doi:10.1037/0022-006X.59.1.12

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. doi:10.1080/10503300500264911

Kolehmainen, N., MacLennan, G., Francis, J. J., & Duncan, E. A S. (2010). Clinicians' caseload management behaviours as explanatory factors in patients' length of time on caseloads: a predictive multilevel study in paediatric community occupational therapy. *BMC Health Services Research, 10*: 249. doi:10.1186/1472-6963-10-249

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. dx.doi.org/10.1080/10503307.2011.563249

Larrison, C. R., & Schoppelrey, S. L. (2011). Therapist effects on the disparities experienced by minorities receiving services for mental illness. *Research on Social Work Practice, 21,* 727-736. doi: 10.1177/1049731511410989

Lewis, M., Morley, S., van der Windt, D. A. W. M., Hay, E., Jellema, P., Dziedzic, K., & Main, C. J. (2010). Measuring practitioner/therapist effects in randomised trials of low back pain and neck pain interventions in primary care settings. *European Journal of Pain, 14,* 1033-1039. doi:10.1016/j.ejpain.2010.04.002

Luborsky, L., Crits-Christoph, P., Woody, G. E., Piper, W. E., Imber, S., & Pilkonis, P. A. (1986). Do therapists vary much in their success? Findings from four outcome studies. *American Journal of Orthopsychiatry, 51*, 501-512.

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. doi:10.1111/j.1468-2850.1997.tb00099.x

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 Counseling Psychology, 54,* 32-39. doi:10.1037/0022-0167.54.1.32

Maas, C. J. M. & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58,* 127-137. doi:10.1046/j.0039-0402.2003.00252.x

Martindale C. (1978). The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology, 46*, 1526-1530. doi:10.1037/0022-006X.46.6.1526

McArdle, C. S. (2000). ABC of colorectal cancer. Primary treatment – Does the surgeon matter? *BMJ, 321,*1121-1123. doi:10.1136/bmj.321.7269.1121

McArdle, C. S., & Hole, D. (1991). Impact of variability among surgeons on postoperative morbidity and mortality and ultimate survival. *BMJ, 302,* 1501-1505. doi:10.1136/bmj.302.6791.1501

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. doi:10.1111/0022-4537.00103

National Health Service Performance Indicators 2002. http://www.performance.doh.gov.uk/nhsperformanceindicators/2002/index.html

Okiishi, J. C., Lambert, M. J., Eggett, D., Nielson, S. L., Vermeersch, D. A., & Dayton, D. D. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their patients' psychotherapy outcome. *Journal of Clinical Psychology, 62,* 1157-1172. doi:10.1002/jclp.20272

Okiishi J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy, 10*, 361-373. doi:10.1002/cpp.383

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2009). *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). *A User's Guide to MLwiN, v2.10*. Centre for Multilevel Modelling, University of Bristol.

Richards, D.A., & Suckling R. (2009). Improving access to psychological therapies: Phase IV prospective cohort study. *British Journal of Clinical Psychology*. *48*, 377–396. doi:10.1348/014466509X405178

Saxon, D., Ricketts, T., & Heywood, J. (2010). Who drops-out? Do measures of risk to self and to others predict unplanned endings in primary care counselling? *Counselling and Psychotherapy Research, 10,* 13-21. doi:10.1080/14733140902914604

Shapiro, D. A., & Firth, J. A. (1987). Prescriptive vs. Exploratory psychotherapy: Outcomes of the Sheffield Psychotherapy Project. *British Journal of Psychiatry, 151*, 790-799.

Shapiro, D. A., Firth-Cozens, J., & Stiles, W. B. (1989). The question of therapists' differential effectiveness. A Sheffield Psychotherapy Project addendum. *The British Journal of Psychiatry, 154*, 383-385.

Shaw, B. F., Elkin, I., Yamaguchi, J., Olmsted, M., Vallis, T. M., Dobson, K. S.… Watkins, J. T. (1999) Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology, 67,* 837-846. doi:10.1037/0022-006X.67.6.837

Snijders, T., & Bosker, R. (2004). *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* London: Sage Publications Ltd.

Soldz, S. (2006). Models and meanings: Therapist effects and the stories we tell. *Psychotherapy Research, 16*, 173-177. doi:10.1080/10503300500264937

Stiles, W. B., Barkham, M., Connell, J., & Mellor-Clark, J. (2008a). Responsive regulation of treatment duration in routine practice in United Kingdom primary care settings. *Journal of Consulting and Clinical Psychology, 76,* 298-305. doi:10.1037/0022-006X.76.2.298

Stiles, W. B., Barkham, M., Mellor-Clark, J., & Connell, J. (2008b). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary care routine practice: Replication with a larger sample. *Psychological Medicine, 38,* 677-688. doi:10.1017/S0033291707001511

Stiles, W. B., Barkham, M., Twigg, E., Mellor-Clark, J., & Cooper, M. (2006). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies as practiced in UK National Health Service settings. *Psychological Medicine, 36,* 555-566. doi:10.1017/S0033291706007136

Trepka, C., Rees, A., Shapiro, D. A., Hardy, G. E., & Barkham, M. (2004). Therapist competence and outcome of cognitive therapy for depression. *Cognitive Therapy and Research, 28*, 143-157.

Vocisano, C., Arnow, B., Blalock, J. A., Vivian, D., Manber, R., Rush, A. J., ...Thase, M. E. (2004). Therapist variables that predict symptom change in psychotherapy with chronically depressed outpatients. *Psychotherapy, 41*, 255-265.

Wampold, B. E., & Bolt, D. M. (2006). Therapist effects: Clever ways to make them (and everything else) disappear. *Psychotherapy Research, 16*, 184-187. doi:10.1080/10503300500265181

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. doi:10.1037/0022-006X.73.5.914

Wampold, B. E., & Serlin, R. C., (2000). The consequences of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5,* 425-433. doi: 10.1037//1082-989X.5.4.425

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*, 200-211. doi: 10.1037/a0018912

Wilson, G. T., Wilfley, D. E., Agras, W. S., & Bryson, S. W. (2011). Allegiance bias and therapist effects: Results of a randomized controlled trial of binge easting disorder. *Clinical Psychology: Science and Practice, 18*, 119-125. doi:10.1111/j.1468-2850.2011.01243.x

Acknowledgements

Support for this work was provided by a development grant from Sheffield Health and Social Care NHS Foundation Trust. We thank the following colleagues for their helpful advice: William Browne (University of Bristol), Mike Campbell (University of Sheffield), Louis G Castonguay (Penn State University), Glenys Parry (University of Sheffield), William B Stiles (Miami University), and Bruce E Wampold (University of Wisconsin-Madison). We also thank John Mellor-Clark (CORE Information Management Systems Ltd) for facilitating the collection of the data set and the reviewers for their helpful contributions and comments on earlier drafts

Footnotes

[1]Full data on model estimates and diagnostics are available from the first author

Appendix A

MCMC model

$$\text{LNoutcome}_{ij} = \beta_{0j} + \beta_{1j}(\text{Ln\_NR\_pre-gm})\text{^}1_{ij} + 0.122(0.020)(\text{Ln\_NR\_pre-gm})\text{^}2_{ij} + 0.042(0.007)(\text{Ln\_R\_pre-gm})_{ij} + 0.057(0.015)(\text{TRisk\_Pre-gm})_j + e_{ij}$$

$$\beta_{0j} = 2.016(0.017) + u_{0j}$$

$$\beta_{1j} = 0.786(0.024) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.026(0.004) & \\ 0.010(0.002) & 0.005(0.002) \end{bmatrix}$$

$$e_{ij} \sim \text{N}(0, \sigma_e^2) \quad \sigma_e^2 = 0.366(0.005)$$

Deviance(MCMC) = 19764.443 (10786 of 10786 cases in use)

Note: All variables are centered around their grand means (gm). LNoutcome, Ln_NR_pre and Ln_R_pre are log transformed patient outcome scores and non-risk and risk scores at intake. TRisk_Pre is a therapist level variable for aggregated patient risk

Table 1: Patient and therapist level intake and outcome scores on CORE-OM (non-risk and risk items)

	Intake		Outcome	
	Mean (SD)	Range	Mean (SD)	Range
Patient level				
CORE-OM	17.5 (6.0)	0 – 37.9	8.2 (5.9)	0 – 35.3
Non-risk	20.5 (6.7)	0 – 39.3	9.8 (6.9)	0 – 38.2
Risk	3.5 (5.1)	0 – 36.7	1.0 (2.6)	0 – 32.0
Therapist level				
CORE-OM	17.6 (1.2)	15.0 – 20.4	8.6 (1.8)	3.9 – 13.4
Non-risk	20.6 (1.3)	17.8 – 23.3	10.2 (2.1)	4.6 – 15.8
Risk	3.6 (1.1)	1.3 – 6.8	1.1 (0.6)	0.1 – 2.8

Table 2: Number and percentages of therapists and patients in each group and the group recovery rate

	Group		
	Below Average N (%)	Average N (%)	Above Average N(%)
Therapists	19 (16.0)	79 (66.4)	21 (17.7)
Patients	1947 (18.1)	5951 (55.2)	2888 (26.8)
Patients scoring above clinical level at intake	1704 (87.5)	5328 (89.5)	2641 (91.4)
Patients Recovered (Recovery rate[a])	786 (46.1)	3155 (59.2)	2019 (76.5)

[a] The percentage recovery rate is based on patients above clinical cut-off at intake

Table 3: Pre and post therapy CORE scores for therapists in the 3 groups

	Below Average	Average	Above Average	F value df 2,116	p value
CORE-OM					
Pre-therapy	17.3 (1.1)	17.6 (1.3)	17.8 (0.9)	.921	.401
Post-therapy	10.4 (1.5)	8.8 (1.4)	6.4 (1.2)	44.07	<.001
Non-Risk					
Pre-therapy	20.2 (1.2)	20.6 (1.4)	20.9 (1.1)	1.26	.287
Post-therapy	12.4 (1.7)	10.4 (1.6)	7.7 (1.4)	45.71	<001
Risk					
Pre-therapy	3.5 (1.0)	3.7 (1.1)	3.4 (1.1)	1.09	.341
Post-therapy	1.4 (0.6)	1.2 (0.6)	0.6 (0.4)	13.63	<.001

Table 4: Therapist recovery rates (mean percentage, SD and range) for each group,

	Group		
	Below Average	Average	Above Average
Therapists N	19	79	21
Mean %(SD)	43.3 (10.2)	58.0 (10.1)	75.6 (9.5)
Range (%)	23.5 – 58.6	29.2– 79.6	62.0 – 95.6

Figure 1: Variance Partition Coefficients (VPC) for Intake CORE-OM non-risk scores, with a histogram of the frequency of scores

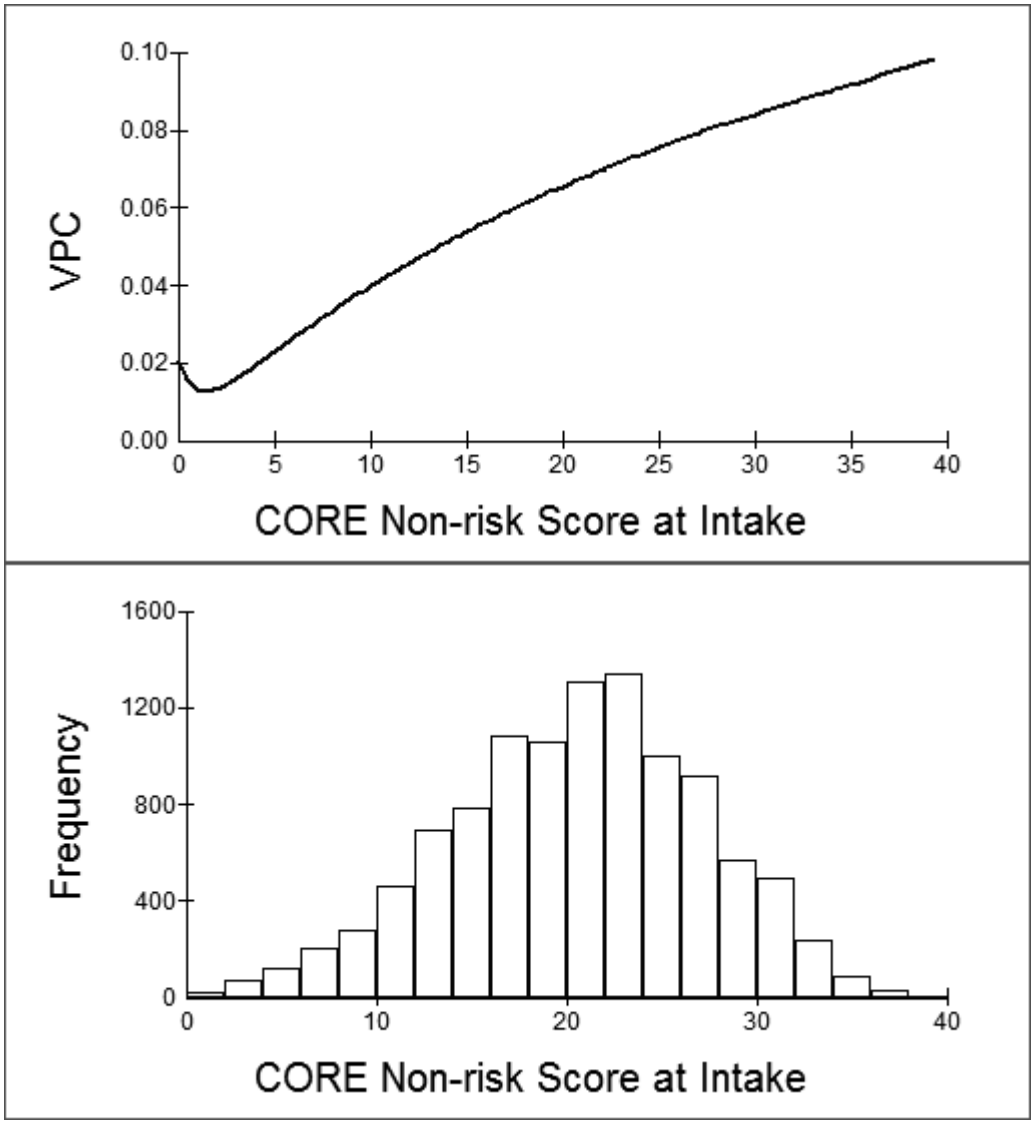

Figure 2: Intercept residuals for therapists, ranked, with 95% confidence intervals

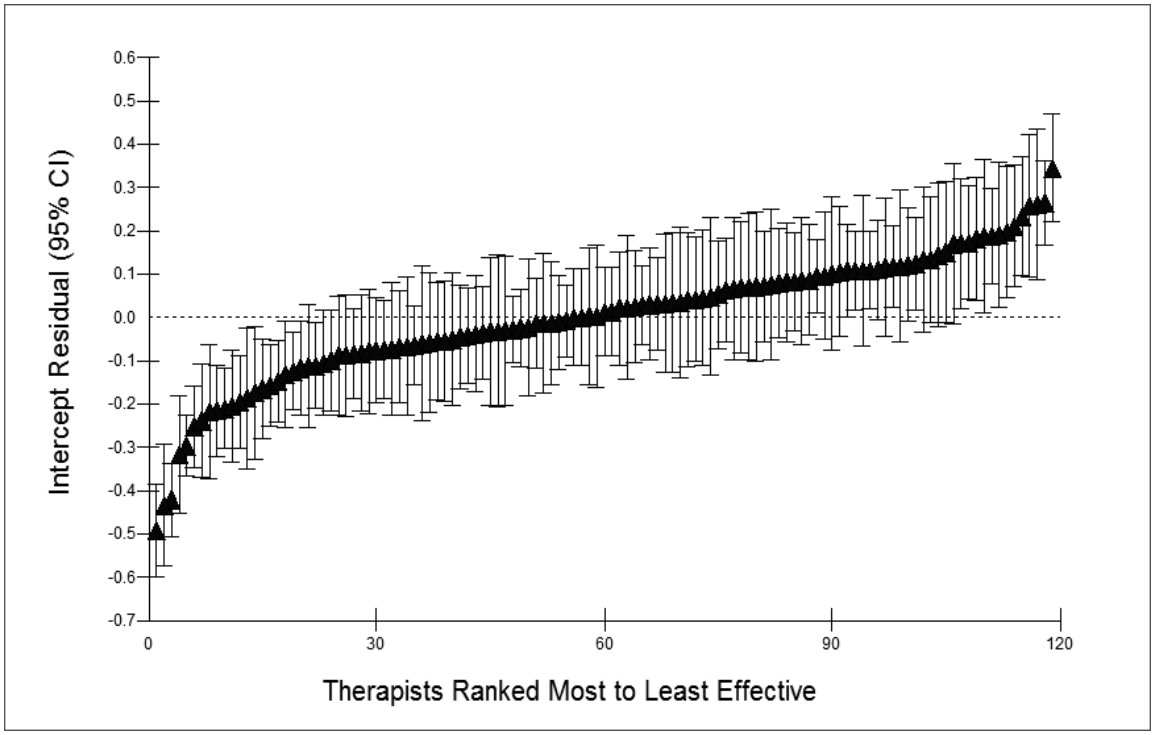

Figure 3: Patient outcome predictions for levels of patient risk and non-risk, and therapist risk caseload