Optum

Achievements in Clinical Excellence (ACE) Measuring Effectiveness

In 2009, Optum established the *Campaign for Excellence* (CFE) to measure clinical quality outcomes in the provider network. CFE measured clinical effectiveness using the Severity Adjusted Effect Size (SAES) derived from Wellness Assessments. In November 2014, CFE was replaced with ***Achievements in Clinical Excellence (ACE).*** ACE broadens the scope of measurement to include both effectiveness *and* efficiency data so that Optum can recognize and reward providers' strengths and elevated performance. As was done with the CFE program, the effectiveness measure for ACE is the Severity Adjusted Effect Size.

Research has supported the importance of the clinician in treatment outcomes. Analyses have revealed that the percentage of variance in treatment outcomes attributed to therapists exceeds the variance due to the type of treatment (Luborsky et al., 1986; Crits-Christoph et al, 1991; Crits-Christoph & Mintz, 1991; Wampold & Serlin, 2000; Wampold, 2001; Kim, Wampold, & Bolt, 2006). In addition, almost three decades of randomized clinical trials and meta-analyses of these trials have supported that psychological therapies appear to have similar treatment effects (Smith & Glass, 1977; Smith & Glass, 1980; Shapiro & Shapiro, 1982; Robinson, Berman, & Neimeyer, 1990; Wampold, Mondin, Moody, & Ahn, 1997; Ahn & Wampold, 2001; Wampold, 2001; Luborsky et al., 2002; Lambert & Ogles, 2004). These findings suggest that the contributions made by individual therapists exceed the effects of the type of treatment.

The research that supports the importance of clinician impact on outcomes also demonstrates that not all clinicians achieve comparable outcomes (Luborsky et al., 1986, Okiishi, Lambert, et al., 2003, Brown et al., 2005). Therefore, it is important to have a process that allows continual measurement of clinician effectiveness and recognizes those with superior outcomes.

The clinical effectiveness data for ACE is generated through the submission of Wellness Assessments, the Optum program for promoting and managing outcomes-informed outpatient treatment. Clinicians are asked to administer the Wellness Assessment (WA) at the first session and once more between the third and fifth sessions, providing an important measure of early change in treatment. Clinicians are welcome to submit additional WAs should they wish. Optum mails the WA directly to the patient four months after the baseline WA (first one received) in order to obtain a measure of longer-term change.

The core scale on the WA measures Global Distress (GD), a construct encompassing symptoms, functioning, and perceptions of self-efficacy. The adult WA uses a 15-item GD scale including some items based on the Symptom Checklist 90 (SCL-90). Adult GD scores range from 0 to 45, with higher scores indicating greater levels of distress. The youth GD scale is the 14-item Child and Adolescent Measurement Scale developed by Ann Doucette, Ph.D., of George Washington University. Youth GD scores range from 0 to 28. As with the adult GD scale, higher scores indicate greater levels of distress. Psychometric evaluations of the instruments were conducted, including two rounds of extensive Item Response Theory (IRT) analyses. Reliability of the adult scale is α = .90 while the youth scale is α = .80. For a full description of the psychometric properties of the WA, please refer to the papers posted on the Wellness Assessment section of Provider Express.

United Behavioral Health and United Behavioral Health of New York, I.P.A., Inc. operating under the brand Optum
U.S. Behavioral Health Plan, California doing business as OptumHealth Behavioral Solutions of California

Measuring Outcomes

Simple measures of change observed between the first and last GD scores allow for interpretation of clinical outcomes within a treatment episode. By incorporating the psychometric properties of the WA (e.g., clinical cut-off thresholds and the standard error of measurement), the change in GD reported by patients can be categorized as being clinically and/or statistically meaningful. In recent years, there have been a number of publications that addressed methodology for benchmarking outcomes in behavioral health care, many of which used data collected by Optum or its legacy companies (Brown et al., 2005; Azocar et al., 2007, Minami et al., 2007, Minami et al., 2008).

A different methodology is required when evaluating clinical outcomes at the level of the individual clinician. For example, within a clinician's caseload, variability in patient severity is not fully accounted for when changes that his or her patients report is simply being aggregated. Sophisticated statistical methods are instead required to accommodate this complexity. Moreover, in order to be credible, the methodology should be transparent, incorporate input from external statisticians and subject matter experts, and integrate benchmarks. After reviewing various methodologies with external consultants and the Behavioral Specialty Advisory Council, we elected to use Severity Adjusted Effect Size (SAES) to measure clinical outcomes at the clinician level because it meets these requirements.

Episode Severity Adjusted Effect Size (SAES)

Effect size is a standardized measure of change commonly used in the social sciences to describe the effectiveness of treatments. It is the raw change score between measurements within an episode of care divided by the standard deviation of the instrument. SAES takes the concept of effect size one step further by incorporating statistical adjustments to account for patient characteristics (e.g., clinical severity). Optum calculates the SAES for each treatment episode with a minimum of two complete WAs and where the patient has a baseline GD score in the clinical range[1].

In order to understand SAES, it is important to distinguish the various ways in which episode-level change in GD can be evaluated:

- **Actual Change** is the difference in GD scores between the baseline and last WA within a patient's treatment episode.
- **Predicted Change** is derived from a general linear regression model (using SAS PROC GLM) that predicts the expected change in GD within each treatment episode, given the patient's characteristics. The regression model is based on a number of variables that have been consistently shown in past analyses to be significantly related to change in GD scores. While baseline GD score has most often been found to be the strongest predictor (i.e., the greater the distress reported at baseline, the greater the likelihood of significant improvement), the regression model also includes other patient variables that account for severity such as age, gender, length of time between assessments, workplace impairment, and health status. The model is dynamic in that it is re-run each year, and thus is being fit to the growing, normative Optum dataset each time. Different models are used to predict change in adults and youth.

[1] For the purpose of this document, a treatment episode begins with the receipt of WA for a patient by a specific provider and concludes with the last WA received. A WA received 180 days after the last WA will initiate a new treatment episode The clinical range is defined by a GD score greater than 6 for youth and greater than 11 for adults.

- **Residualized Change score** is the difference between the predicted change and the actual change. Residualized change scores indicate how the actual change for each patient compares to that of other patients in the population, after adjusting for the factors in the regression model. If the actual change is greater than the predicted change, the patient improved more than was expected given his or her characteristics. Similarly, if the actual change is less than the predicted change, the patient's clinical status did not improve as much as expected.

Using the change scores described above, three versions of effect sizes are used in the SAES methodology:

1. ***Population Effect Size*** represents the amount of change we would expect for a typical patient in treatment. It is derived by taking the mean of the observed effect size (i.e., the actual change score divided by the standard deviation of the baseline GD scale[2]) and typically averages *d* = .80. The population effect size is based on an Optum national dataset that currently includes over 100,000 GD scores and will continue to grow over time.
2. ***Residulized Effect Size*** differs from the observed effect size in that it is the residualized change score (not the actual change score) divided by the standard deviation of the baseline GD score.
3. ***Severity Adjusted Effect Size (SAES)*** is then calculated by adding the residualized effect size for the treatment episode to the population effect size. In essence, the SAES re-standardizes the residualized effect size score so that improvement can be interpreted with respect to the amount of change expected for a typical patient after accounting for severity.

The following example of a fictional patient named Mark illustrates how the scores are derived. Mark is a 25-year-old engineer who reported a moderate GD score of 21 at his first session. By the 5th session, his GD score had dropped to 10, for an ***actual change*** score of -11 points. The regression model, however, had only predicted a decrease of 4 points (***predicted change***). As a result, Mark's ***residualized change*** score was 7, indicating he reported more improvement than had been predicted. With these change scores, effect size statistics can be calculated. The ***residualized effect size*** was 0.9 (i.e., the residualized change score of 7 divided by the standard deviation of the adult GD scale of 7.7). The ***SAES*** for Mark's treatment was 1.7 (i.e., the residualized effect size of 0.9 plus the observed effect size for the population). The SAES of 1.7 was higher than the population effect size, confirming that Mark's treatment was very effective.

Determining Clinician SAES

In the same way that SAES can be used to measure the effectiveness of a single treatment episode, it can also be the basis for deriving a measure of effectiveness for the clinician or group practices. To derive clinician and group practice SAES, Optum uses hierarchical linear modeling (HLM) with random effects (Minami, et al. 2011). HLM provides a more optimal methodology for fitting nested data, as is the case when patients' SAES are nested within clinicians or group practices. Additionally, random effects modeling is used instead of fixed effects because we assume that the patient sample

[2] The standard deviation differs between the youth (3.8) and adult (7.7) GD scales.

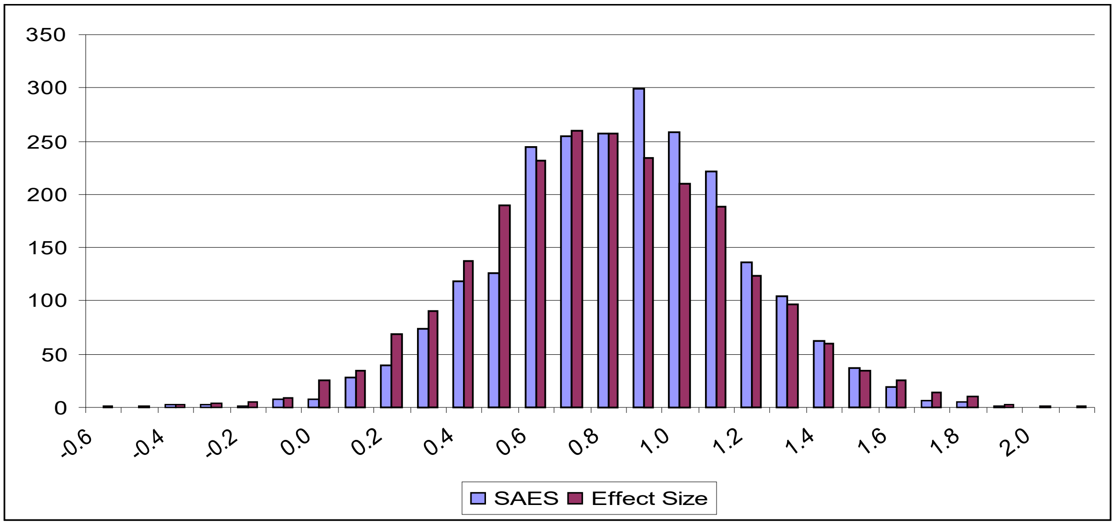
for which we have measurements is not necessarily representative of a clinician's or group's entire caseload. This HLM methodology uses the episode SAES data to generate a fixed intercept, random residuals, and standard errors. The fixed intercept is the mean SAES for all clinicians or groups eligible for evaluation, while the random residuals and standard errors are unique to each clinician. To determine each clinician's or group's SAES, the random residual for that clinician or group is added to the fixed intercept.

A comparison of clinician SAES to the effect size in its standard form (i.e., the observed effect size, which does not incorporate severity adjustment or HLM) shows a similar distribution, although the SAES methodology produces slightly higher effect sizes.

Table 1. Distribution of Clinician SAES and Observed Effect Size

	Mean	STD	25th Percentile	50th Percentile (Median)	75th Percentile
Observed Effect Size	0.80	0.36	0.56	0.78	1.0
SAES	0.83	0.33	0.62	0.84	1.1

Figure 1. Distribution of Clinician SAES and Observed Effect Size



In order for a clinician or group practice to have an SAES computed, he or she must have a minimum of 10 treatment episodes in the past 24 months for which the baseline GD scores were at or above the clinical cut-off (i.e., a clinical sample). Clinicians with more than 30 treatment episodes in the past 24 months with baseline scores in the clinical range are evaluated on the most recent 30 episodes. Group practices with more than 66 treatment episodes in the past 24 months with baseline scores in the clinical range are evaluated on the most recent 66 episodes. For clinicians and groups who see fewer Optum patients, extending the measurement period for up to 24 months maximizes their opportunity to meet the minimal sample size. However, for those clinicians and group practices with a more than sufficient sample, limiting the measurement to the most recent treatment episodes allows optimal opportunity to show more recent change over time measurements, as historical cases are not included in the measurement.

The SAES computation is limited to clinical episodes for two primary reasons. First, the GD scale, like many outcomes instruments, has a floor effect that prevents it from adequately measuring change for patients at the lower end of the severity spectrum. Since we do not have an adequate measure of change for these patients, we cannot apply the SAES methodology. The second reason for excluding these patients is to create a more homogeneous sample of patients for all clinicians and group practices whose effectiveness will be evaluated. An analysis of the effectiveness of treatment by practitioners using a dataset from PacifiCare Behavioral Health (an affiliate of Optum) similarly eliminated the non-clinical patients from the sample and demonstrated comparable effect sizes to those found in a meta-analysis of psychotherapy studies on depression (Minami et al., 2008).

Understanding SAES Designations of Effectiveness

As with any measurement, there is error associated with estimating clinical effectiveness. Because SAES is often based on small sample sizes it is subject to greater variability. Therefore, Optum also uses HLM to construct a confidence interval around a clinician's and group's SAES to take into account error or random variability. The confidence interval defines the range between which the "true" mean SAES may fall. Constructing a confidence interval requires one to make a decision about an acceptable degree of confidence or "confidence level." Within our methodology, we have chosen the 90% confidence level. Therefore, we can say, "with 90% confidence, we believe that the true mean SAES for clinician A falls between X and Y." X and Y in this context represent the lower and upper confidence limits respectively, and when taken together they represent the confidence interval.

The SAES confidence interval is the basis for determining clinical effectiveness. Determination of clinical effectiveness is based on (1) the number of treatment episodes with SAES and (2) comparison of a clinician's or group's SAES lower and upper confidence limits and an effect size threshold of 0.50. Using Cohen's conventional guidelines (Cohen, 1988), as is typically reported in the social sciences, $d = 0.50$ equates to a moderate effect. For further explanation of our approach, see Minami et al., 2011. SAES is used to identify clinicians and groups whose patients, on average, consistently report positive outcomes;

- Clinicians or groups with an SAES *lower* confidence limit greater than or equal to .50 are designated as '**Meets Effectiveness Criteria**'. More than 80% of the clinicians for whom SAES could be measured typically fall into this designation.
- Clinicians or groups with an SAES *lower* confidence limit below 0.50 and an *upper* confidence limit above 0.50 are given the designation of '**Insufficient Data to Determine Effectiveness Designation**.' This designation reflects the wide variability in a clinician's or group's patient's SAES. With additional patients or measurements, the reliability of measurement of average patient outcomes will increase, in turn also increasing the likelihood that clinicians and groups can be deemed effective.
- Clinicians or groups with an SAES *upper* confidence limit below 0.50 are given a designation of '**Does Not Meet Effectiveness Criteria**.'
- Fewer than 10 clinical cases will result in a designation of '**Insufficient Cases to Assess Effectiveness**.'

Table 2. Effectiveness Designations

Designation	Metric
Meets Effectiveness Criteria	LCL >= 0.50
Insufficient Data to Determine Effectiveness Designation	LCL < 0.50 and UCL >= 0.50
Does Not Meet Effectiveness Criteria	UCL < 0.50
Insufficient Cases to Assess Effectiveness	< 10 clinical cases

SAES is available only for ACE participating clinicians and groups with 10 or more patients who completed a minimum of two WAs and who began treatment in the clinical range. Clinicians and groups can increase their opportunity to be identified as meeting the *Effectiveness* designation by increasing the rate at which they submit WAs on their patients. It is especially important to submit both the baseline WA at the first session (or second session if necessary) and a minimum of one additional WA to measure change.

Optum gratefully acknowledges the contributions of Drs. Jeb Brown, Takuya Minami, Bruce Wampold, Warren Lambert, and members of the Behavioral Specialty Advisory Council Metrics Workgroup in the development and review of the SAES methodology.

References

Ahn H, Wampold BE. (2001). Where oh where are the specific ingredients? A meta- analysis of component studies in counseling and psychotherapy. J Counsel Psychol, 48, 251-7.

Azocar, F., Cuffel, B., McCulloch, J., McCabe, J., et al (2007) Monitoring patient improvement and treatment outcomes in managed behavioral health, Journal for Healthcare Quality, 29 (2), 4 -13.

Brown, G.S., Jones, E., Lambert, M.J., & Minami, T. (2005) Identifying highly effective psychotherapists in a managed care environment. American Journal of Managed Care. 2(8), 513-520.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crits-Christoph, P., Baranackie, K., Carrilm, K., Luborsky, L., et al (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. Psychotherapy Research, 1, 81-91.

Crits-Christoph P, Mintz J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. Journal of Consulting and Clinical Psychology 59 (1),20-26.

Kim, D., 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.

Lambert, M. J., & Ogles, B. M. (2004). The Efficacy and Effectiveness of Psychotherapy. In M. Lambert (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change (5th ed., pp. 139-193). New York, NY: John Wiley and Sons.

Luborsky L, Rosenthal R, Diguer L, Andrusyna, T., et al (2002). The dodo bird verdict is alive and well—mostly. Clinical Psychology Science Practice, 9, 2-12.

Luborsky L, Crits-Christoph P, McLellan A, Woody, G., et al (1986). Do therapists vary much in their success? Findings from four outcome studies. American Journal of Orthopsychiatry , 56, 501-12.

Minami, T., Brown, G. S., McCulloch, J., & Bolstrom, B. J. (2011). Benchmarking therapists: Furthering the benchmarking method in its application to clinical practice, Quality & Quantity. Advance online publication. doi:10.1007/s11135- 011-9548-4

Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (2007). Benchmarks for psychotherapy efficacy in adult major depression, Journal of Consulting and Clinical Psychology, 75, 232-243.

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G. et al (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment. Journal of Consulting and Clinical Psychology, 76, 116-124.

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

Robinson, L.A., Berman, J.S., & Neimeyer, R. A. (1990). Psychotherapy for treatment of depression: A comprehensive review of controlled outcome research. Psychological Bulletin, 108, 30-49.

Shapiro, D.A., & Shapiro D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin*, 92, 581-604.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcomes studies. *American Psychologist*, 32, 752-760.

Smith, M. L., & Glass, G. V. (1980). *The benefits of psychotherapy*. Baltimore: The John Hopkins University Press.

Wampold, B.E., (2001) *The Great Psychotherapy Debate: Models, Methods and Findings.* Mahwah, J.J.: Lawrence Erlbaum Associates, Inc.

Wampold, B. E., Mondin, G.W., Moody, M., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "All must have prizes." *Psychological Bulletin*, 122, 203-15.

Wampold, B,E., & Serlin, R.C. (2000). The consequences of ignoring a nested factor on measures of effect size in analysis of variance designs. *Psychological Methods*, 4, 425-33.