

Pushing the Quality Envelope: A New Outcomes Management System

Draft: Do not copy or disseminate without the authors' permission.

Word Count: Approx. 7500

Type of Article: Special Article

To appear in *Psychiatric Services*, a journal of the American Psychiatric Association.

The article is based on the authors' experience designing and implementing outcomes management systems for large managed care organizations. Topics addressed include design of instruments, use of cost effective technology, development of computerized decision support tools and methods for case mix adjustment. Case mix adjustment models are based upon a data repository of several-thousand treatment cases with multiple measurement points across the episode of treatment. Data from controlled and field studies are described suggesting that the outcomes management methods described in this article can result in significantly improved clinical outcomes and more rational allocation of care resources.

G. S. (Jeb) Brown, Ph.D.  
Center for Clinical Informatics, LLC  
1821 Meadowmoor Drive  
Salt Lake City, UT 84117  
[jebbrown@clinical-informatics.com](mailto:jebbrown@clinical-informatics.com)  
Voice: 801 541-9720  
Fax: 801 278-2329

Gary M. Burlingame, Ph.D.  
Brigham Young University

Michael J. Lambert, Ph.D.  
Brigham Young University

Edward Jones, Ph.D.  
PacifiCare Behavioral Health

Jerome Vaccaro, M.D.  
PacifiCare Behavioral Health

**Pushing the Quality Envelope: A New Outcomes Management System**

## **Introduction**

The provision of behavioral healthcare services has undergone radical change over the past decade due to market and political forces creating pressure to contain spending on healthcare. This forces healthcare providers to confront the reality of finite resources and the necessity to make difficult decisions about the allocation of those resources. The shift towards increased accountability for the cost of services has evolved in the past few years to a growing focus on the quality and outcomes of the services. Unfortunately, the implementation of clinical information systems for measuring and improving quality has lagged far behind the success in containing financial costs.

Consequently, the ongoing debate regarding the impact of cost containment efforts on clinical outcomes, while often emotionally charged, has remained anecdotal and data poor. In recent years there has emerged something of a consensus on the desirability of incorporating the measurement of outcomes into quality improvement efforts. A few managed care companies have funded large-scale efforts to build outcomes management systems (1-3).

It is necessary to make a clear distinction between outcomes *measurement* and *management*. Outcomes measurement is the process of assessing the clinical outcome of treatment through use of standardized measures of clinical severity. Since outcome is a measure of change, at least two data points are necessary, one at the start of treatment and another at some later point in time, presumably the conclusion of treatment or at some follow-up point. Ideally, measurement of change is achieved through repeated measures at regular intervals so that it is possible to estimate not only the magnitude of change, but also the rate of change.

On the other hand, outcomes management is an effort to improve the *effectiveness* of treatment services throughout a system of care by the evaluation of outcome data. The key performance indicator for an outcomes management program is its ability to make a difference over time (i.e. to measurably improve outcomes). While reliable and valid outcomes measurement is an essential element of an outcomes management program, to be effective they must go well beyond simple storage and tabulation of data.

The outcomes evaluation process must be systemic. That is, the outcomes management program should be integrated into how care is delivered and managed for all patients rather than existing in isolation as a research study focusing on a selected sample. However, translating this wish into reality presents several technical, practical and scientific challenges, including the need for: (a) reliable, valid, and easy to use measures of outcomes, (b) economical and user friendly technology to capture data (e.g. scanners, PC based software, etc...), (c) a large normative sample of patients with multiple measurement points in treatment in order to evaluate measurement tools and create norms for change profiles, (d) empirically validated statistical models for *case mix adjustment*, (e) clinical reports and other decision support tools designed to foster improvement in clinical outcomes and allocation of treatment resources, and (f) clinician acceptance of and participation in efforts to systematically improve outcomes.

This article describes the authors' attempts to solve these problems for large managed behavioral health organizations. The authors draw on lessons learned over a several year period designing and implementing an outcomes management system for Brigham Young University Clinic, Human Affairs International, Inc. (now a part of

Magellan Health Services, Inc.), and most recently, for PacifiCare Behavioral Health, Inc (a subsidiary of PacifiCare Health Services, Inc.).

PacifiCare Behavioral Health (PBH) is a managed behavioral healthcare company with over 3,000,000 (commercial and public sector) covered lives in 9 western states. The methods described in this article are currently being implemented with PBH's largest commercially insured population and with one of its public sector populations, with near term plans for full implementation across its entire system of care.

### **Outcomes measurement method**

Measurement instruments with known validity and reliability are essential. However, data sets that are comprehensive and meet high standards for scientific rigor can become excessively burdensome to the line staff and consumers when employed in real world service-delivery settings. With this in mind, the authors have attempted to keep the time requirements for data collection at PBH to five minutes or less for both the clinician and consumer. Collecting data at specified intervals rather than at every session further reduces the labor demand. Clinical outcomes are assessed from both patients' and clinicians' perspectives.

Analyses of large data sets of commercially insured outpatients indicate there are systematic differences in outcomes obtained from patient versus clinician report. Clinician assessment tends to underestimate improvement for consumers reporting rapid improvement (4). Conversely, analyses of change scores using patient self-report and clinician GAF (Global Assessment of Functioning Scale) ratings suggest that providers may significantly underestimate deterioration and risk for premature termination (3).

The authors therefore advocate a system in which the patient's rating of improvement in symptoms and quality of life is the "gold standard" for assessing the system's performance. Patient satisfaction, relapse rates, and other variables are important, but the global change score reflects the overall reduction in patient distress and has the advantage of broad applicability across multiple diagnoses and settings.

Another argument in favor of a patient-centered system is cost. Clinician rating scales are time consuming to complete and may require training (and retraining) of the clinical staff in order to maintain adequate reliability. [See (5) for a more complete discussion of this and other measurement issues.] This problem is complicated when the results of the evaluation are to be used for performance monitoring. Clinician concerns about the use of outcome tools can introduce hidden sources of bias that are difficult to detect statistically. While patients may also have idiosyncratic ways of understanding and rating items, given a large enough sample this source of error is randomly distributed across providers and therefore is much less likely to contaminate results.

In keeping with the principle of maintaining the cost and effort of data collection as low as possible, Lambert and Burlingame developed two thirty-item self report questionnaires (adult and child/adolescent versions) for the PBH outcomes management program. Over a period of many years they have accumulated a data repository of behavioral health related items from several different instruments, including the widely utilized Outcome Questionnaire-45 (OQ-45) and Youth Outcome Questionnaire (YOQ). [See (6-9) for a more complete description of the item content, reliability and validity of these questionnaires.] The data included repeated measures in treatment of thousands of adults and children treated at hundreds of different sites across the country. Most of the

data was collected under the auspices of several different managed care companies as part of ongoing research agreements, and was redacted on any patient identifiers.

The investigational instruments developed for the PBH program are named the Life Status Questionnaire (LSQ) and the Youth Life Status Questionnaire (YLSQ). Items were selected for these shorter instruments based on their tendency to improve during treatment while remaining relatively stable in a sample of matched non-treatment controls. This approach to item selection created instruments with presumptively sound psychometric properties despite the fact that they had not been previously administered in this 30-item format. Subsequent experience with the investigational instruments has confirmed this presumption.

Since early 1999 the investigational instruments have been employed in PBH's outcomes management program. Nineteen private sector group practices and 5 public sector clinics are providing ongoing outcomes data. In addition PBH's highest volume solo providers are also participating in the data collection. PBH named its outcomes management program the ALERT system, an acronym for Algorithms for Effective Reporting and Treatment (ALERT). The ALERT system connects the patient, the provider, and PBH in an information loop that provides timely reports on critical risk factors and changing levels of patient distress. Aggregate level reports summarize clinical outcomes for entire systems of care and for specific provider groups.

Before addressing specific methods employed by the system, a brief discussion of the enabling information technology is helpful. The entire enterprise is dependent on the ability to *rapidly* and *cheaply* capture, analyze, and report on the data in a variety of formats for target audiences. The cost of data capture must be minimal, and the tools to

program the complex logic for data analysis must be powerful yet extremely flexible. The technology must allow rapid development and deployment of various reports and decision support tools.

The ALERT system is based on an approach to data capture and management that retains maximum flexibility and timely reporting while minimizing cost. The ALERT system starts with data being captured via bubble sheet paper forms that are faxed to a central location data capture, using the Teleform™<sup>1</sup>. There are several software products available for this purpose. The use of paper as the primary data capture interface with the consumer and clinician has the advantage of familiarity and low cost of implementation. Even if the data set is subsequently altered (as will be the case in a system that is learning), the only cost is in printing and distributing new forms. Once the raw data are captured, SAS™<sup>2</sup> is used to manage the data and construct a clinical information system to provide clinical decision support algorithms and reporting. SAS is utilized in conjunction with Microsoft Office products such as Excel and Access for end-user viewing of reports and relevant clinical data. SAS can also be interfaced to an organization's legacy systems.

Using these off the shelf products, the cost of developing the information technology infrastructure necessary for outcomes management is kept to a minimum while retaining flexibility to modify the data set or logic as needed. Over time an

---

<sup>1</sup> Teleform™ is a product of Cardiff Software, Inc. that permits the user to design forms and then perform optical character and optical mark recognition from fax or scanner produced images of the completed forms.

<sup>2</sup> SAS™ is a versatile and comprehensive program for performing data analysis, managing data, creating data warehouses, and other data processing operations. It is the product of the SAS Institute in Cary, N.C. and is widely used throughout government and industry when an "industrial strength" tool is needed. While not known for its user friendliness, the SAS scripting language permits rapid development of SAS code capable of performing virtually any data management and decision support chore.

organization may develop custom software applications fully integrated into the primary operational databases, but the cost of this development is postponed until the organization has had a chance to test and refine the data set and accompanying decision support logic and performance indicators.

The performance indicators and decision support tools are the most critical elements of an outcomes management program. As stated previously, the key performance indicator is the change score on the LSQ and YLSQ from one session to the next. However, knowing the change score for a patient tells us little unless we know what improvement was reasonable to expect. Outcome results are impossible to interpret if one does not have a valid method for statistically accounting for variations in the severity and difficulty of a case, commonly called “case mix adjustment.” A case mix model utilizes data collected at intake to predict the change score at the end of treatment. Arguably one of the most powerful methods for managing outcomes is to use a case mix model that includes *trajectory of change* predictions (10). Such a method depends on repeated measures at regular intervals in treatment so that progress of each individual case can be monitored against these norms. Obviously, development of valid case mix adjustment and trajectory of change models requires a large normative sample with repeated measures.

The PBH outcomes management program utilizes a repeated measures design, with the frequency of data collection being greater during the initial phase of treatment. PBH utilizes assessment at the first, third, and fifth sessions before dropping to a frequency based on risk and complexity of the case. This repeated measure design enables the ability to track trajectory of improvement as part of the clinical management

of the case. The trajectory of improvement during the first few sessions tends to be highly predictive of the eventual outcome of the case (3).

The data repository provided the means to model the expected trajectory of recovery for the most common diagnoses found in outpatient samples. A sample of over 3200 adults and 800 children/adolescents was used to calculate the expected change. The cases selected were drawn from over 15,000 cases in the data repository. They all contained data on the primary diagnosis, intake score and at least one other assessment point in treatment and were drawn from treatment populations thought to be similar to the commercially insured population served.

The data repository also contains test protocols from community volunteers not currently receiving any mental health treatment. This sample was used to estimate a cutoff score on the instruments that constitutes a boundary between normal and clinically significant levels of life distress (11-13). The formula is:

$$\text{Cut-off} = (SD_1 * M_2 + SD_2 * M_1) / (SD_1 + SD_2)$$

For example, in the case of the OQ-45, the sample of community non-patients (n=1353) has a mean of 45 with 19 as a standard deviation (14). The mean intake score for the clinical sample drawn for the PBH project is 82 with a standard deviation of 24. Using the formula, the cut-off score between the clinical sample used for the PBH norms and the community sample is 61, with scores above this point being more likely to become a clinical sample. The data repository enables the design of a system that would track the improvement of each individual patient against that of similar patients in the normative sample as well as means to determine the point at which a patient was within

the normal level of life distress. As we shall see, this feature is critical to the success of the outcomes management program.

### *Case mix adjustment*

The authors' analyses of multiple outpatient samples indicates that the single best predictor of the change score for any given treatment episode is the score on the measurement instrument at the beginning of treatment. Other variables such as diagnosis, chronicity, and treatment population alter the relationship to some degree, yet every sample analyzed during the course of this project produced the same finding. The change score was found to have an essentially linear relationship to the intake score, with higher levels of distress at intake predicting a higher change score and a steeper trajectory of recovery. The relationship between severity at intake and change can be easily communicated visually by providing a plot of the regression line, with the intake score on the *x-axis* and the improvement on the *y-axis*. For consistency sake and ease of interpretation, the change score will be expressed as effect size in the following examples drawn from the data repository.

Effect size is calculated by dividing the raw score change by the standard deviation of the measure. The use of effect size to express the change score is useful in that it conveys the magnitude of change in a way that allows pooling of data from different instruments, an important consideration when aggregating data for an entire system of care. However, a word of caution regarding interpretation of effect size is in order. Care must be taken when comparing mean effect sizes from different populations.

Large, heterogeneous outpatient samples taken from the field will tend to average significantly lower effect sizes than commonly reported in published outcome research.

This fact does not necessarily derive from a reality that academic research studies get better results. Rather, it is do to the fact that the range of intake scores in a research sample is often restricted in some way. For example, a study on the treatment of major depression would naturally contain subjects screened to meet the criteria of depression. Consequently, the mean intake score would be higher than the score for a broad sample of outpatients. As the following graphs will illustrate, higher intake scores tend to show more change. Thus a restricted sample of very distressed patients would also certainly average more change than a more heterogeneous sample. Furthermore, since the homogeneous sample would not contain cases at the lower end of the severity distribution, the standard deviation for the sample would also be truncated.

The importance of sample variability is that the standard deviation of the sample is used as the denominator in calculating effect size. Thus, greater variability (S.D.) will result in smaller effect sizes even if absolute change remains the constant. In the relatively homogeneous sample of patients selected for the hypothetical depression study, the result is a higher numerator (change score) and smaller denominator (standard deviation), resulting in a larger effect size in clinical trials than would be present in a typical sample of outpatients.

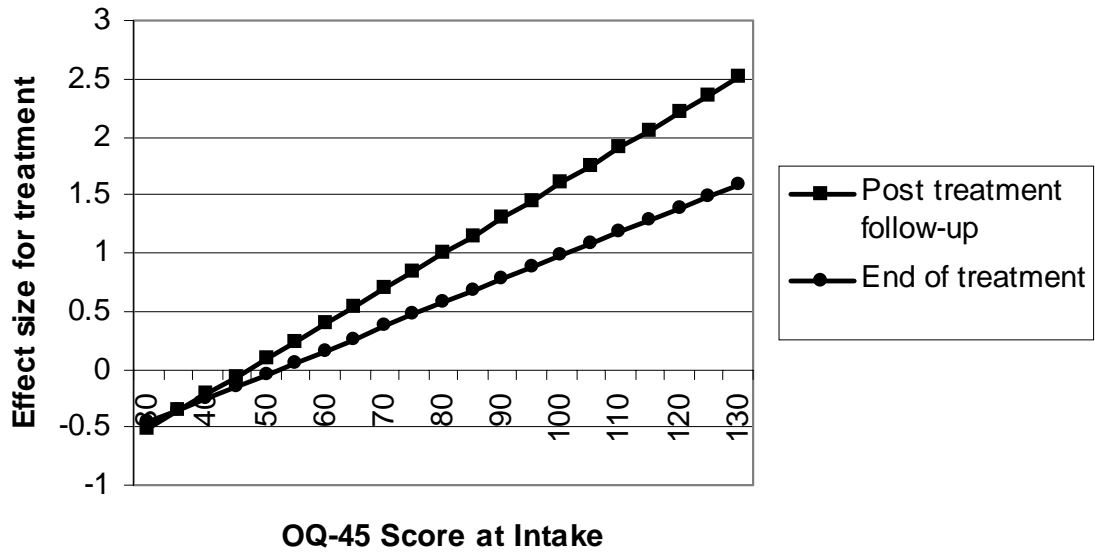
Graph 1 displays outcomes for a subset of the sample (n=219) with repeated OQ-45 administrations during treatment and a single follow-up measure at an average of 10 months after the last recorded treatment session. The intake score is plotted on the  $x$  axis while the effect size from intake to end of treatment and post treatment follow-up is

graphed on the y axis. As can quickly be seen, the higher the intake score, the greater the change during treatment and during the post treatment follow-up period.

*Insert Graph 1 about here.*

Graph 1

Change as a Function of Intake Scores



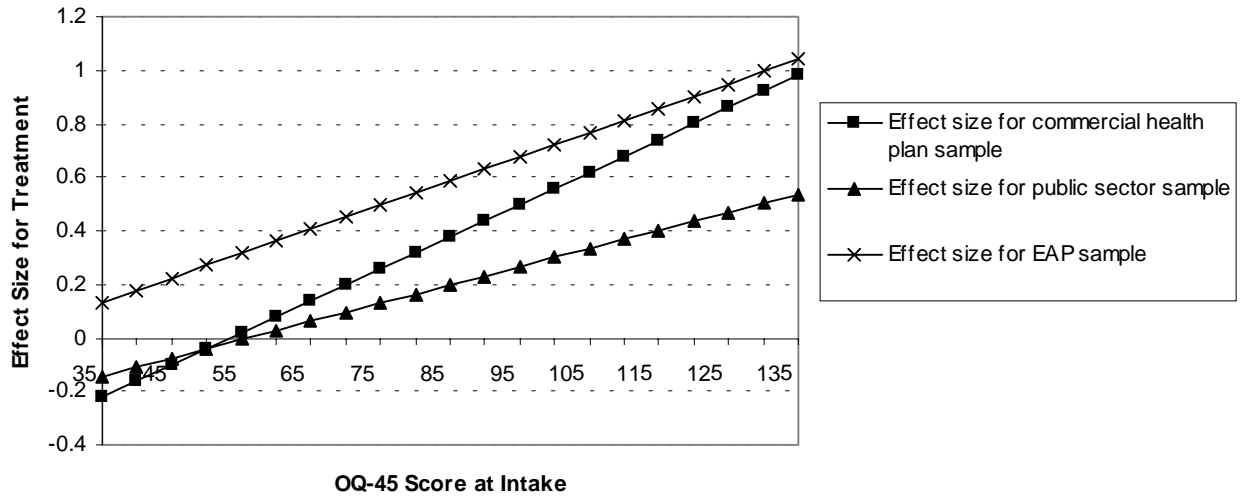
The percentage of variance of the change score accounted for by the intake score appears to range from 10-20% of the variance depending on the sample, the instrument and the length of time to the final assessment. The intake score accounted for 15% of the variance at the end of treatment and 20% at the follow-up assessment.

Graph 2 presents the regression line for three different adult treatment populations: (a) a national sample of commercially insured patients, (b) a statewide public sector sample, and (c) a national sample of clients seen through various Employee Assistance Programs. As can be seen here, the use of plotted regression lines conveys information about the results for these different populations that would have been lost by simply reporting mean change scores. The different samples show marked differences in the slope and and/or intercept line. The EAP sample shows greater effect size at the low end of intake severity compared to the other two. The public sector sample appears to receive a similar benefit to the commercially insured sample at the lower end of severity, but fares worse at higher levels of severity.

*Insert Graph 2 about here.*

Graph 2

Comparison of Results for Different Treatment Populations



We would not argue that this case mix model accounts for all relevant factors. The best process for improving its predictive ability is to identify reasonable variables such as socioeconomic level or chronicity of problems to investigate, and then collect data on these variables for analysis and modeling. Ideally, as data accumulate for a system of care, the model is tested and refined on a continuous basis. The case mix model permits an estimate of the expected improvement for any given patient at the start of treatment. Predicted change remains constant throughout an episode of care, serving as a benchmark to measure treatment progress and outcome.

For example, the ALERT system utilizes a Change Index indicator. The Change Index is simply a residualized change score calculated by subtracting the predicted change from actual change for each case. Positive values indicate above average results. When one knows the trajectory of change during the first few sessions, it becomes feasible to evaluate the individual patient response to treatment early on and make any necessary adjustments to the treatment plan in a timely manner. Furthermore, calculation of an expected trajectory of change opens up the possibility of addressing the question of which patients require what amount of treatment.

#### *Predicting and tracking change*

A number of researchers have investigated dose response curves for psychotherapy (15,16,4). The research supports the premise that most patient change occurs during the early stages of treatment, with diminishing benefit per session as the length of treatment increases. However, it is risky to extrapolate from expected dose response curves established in research settings where the length of treatment is

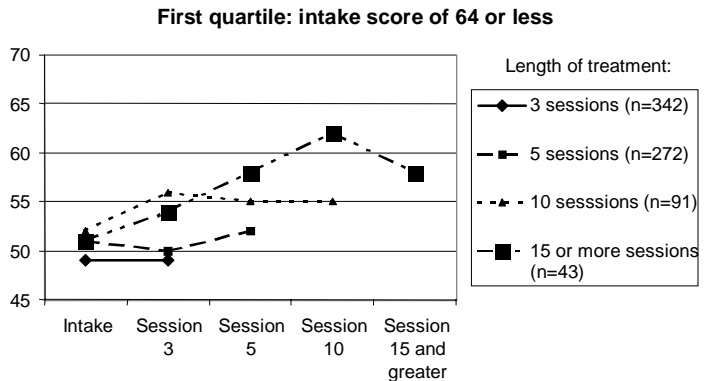
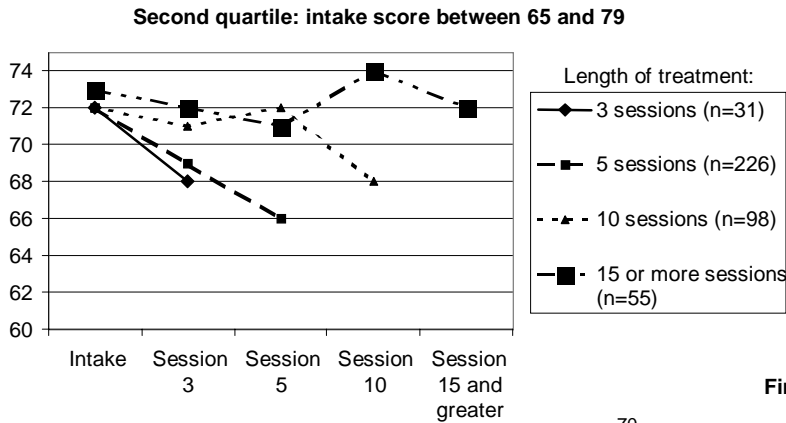
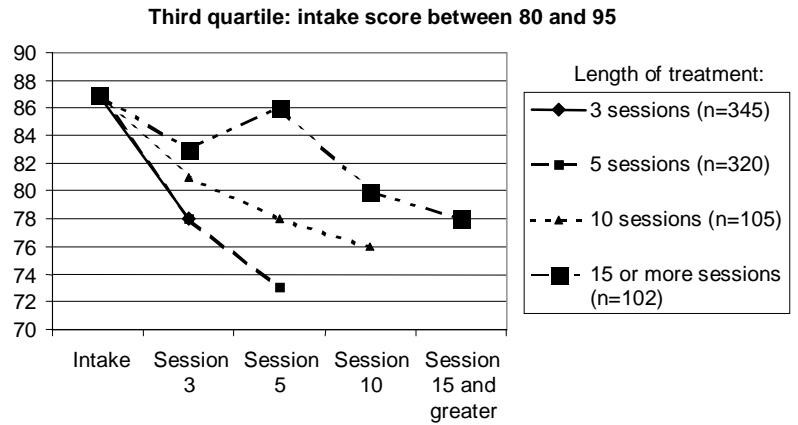
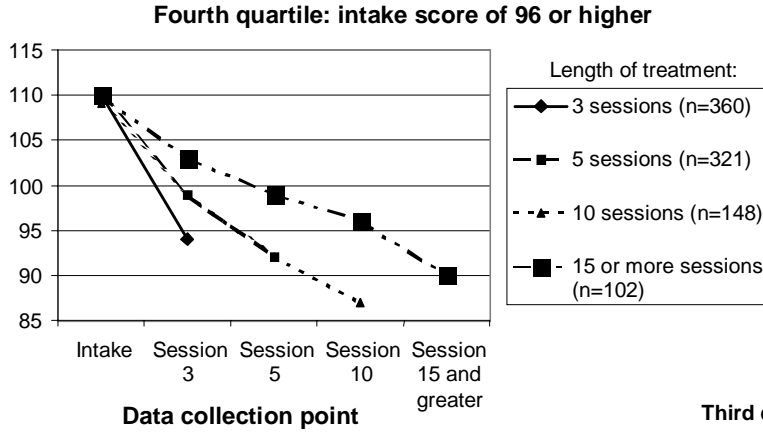
prolonged or held constant for study purposes. In real world settings, whether fee-for-service or managed care environments, the length of stay in treatment is determined by many factors and one would anticipate different patterns of dose response. Conventional wisdom would argue that longer lengths of treatment result in superior outcomes. Much of the criticism of managed care plans is that they arbitrarily limit the length of treatment. However, the reality is somewhat more complex than this, and surprisingly encouraging for managed care. Data from the commercially insured managed care populations in the data repository fail to show a significant correlation between length of treatment and outcome. In fact, short lengths of treatment are common, even for some of the most distressed patients, and the results often indicate rapid improvement rather than poor outcome.

The following series of graphs on adult outcomes help to illustrate this point. Data were collected at the first, third and fifth sessions and at every fifth session thereafter. The graphs show the trajectory of change for patients receiving three, five, ten, and fifteen or more sessions. The sample is divided into quartiles based on the intake score to more readily demonstrate trajectory of change as a function of severity at intake. Length of treatment is determined by the last session at which data were collected. Many of the cases are assumed to have had additional sessions, but they did not continue in treatment to the next data collection point. Only cases with at least two data points are included in this analysis.

**Insert Graphs 3 – 6 about here**

## Graphs 3 - 6

### Length of Treatment and Score Change After Intake



The graph for the fourth quartile, the most severely distressed patients at intake, provides some support for longer lengths of treatment. Patients for whom data was collected at the 10<sup>th</sup> session or later show a few points of improvement more than those for which the last data point was the third or fifth session. However, the rate of improvement for patients with the shorter length of treatment is so rapid that it could be argued that they had no need to stay in treatment to the next data collection point. On the other hand, patients in the other three quartiles had a very different outcome: lengths of treatment of ten sessions and greater are associated with worse outcomes, and for those in the mildest range at intake, actual deterioration. It would appear that these patients and therapists continue to meet *because* the patient is not improving or is doing worse, not because the treatment is helping.

The average length of treatment in all of the quartiles, even the most severe, was less than seven sessions. Likewise, over 75% of the treatment episodes were completed before the 10<sup>th</sup> session in all four quartiles. It is apparent that shorter lengths of treatment are associated with a higher rate of change. Cases in the first quartile are an exception to the above generalizations. However, these cases are more characteristic of a non-treatment sample than a clinical population. In fact, the mean intake score in the first quartile is 50, only 5 points higher than the mean of the non-treatment sample cited above. This group shows no change with shorter lengths of treatment and slight deterioration for cases averaging ten sessions and more. One must wonder if this represents a “good as it gets” phenomenon.

At the other end of the spectrum, treatment length for the most severe cases bears closer analysis. It is highly unlikely that the cases in the fourth quartile with lengths of

treatment below five sessions (42%) are a result of managed care limitations or intentional termination by the clinicians. Despite the rapid improvement seen by the third session, the severity of distress remains relatively high compared to the entire clinical sample. At termination these cases are on average above the 50th percentile of the larger clinical sample from which they were drawn. The managed care companies had access to the OQ-45 scores, as did the clinicians, and there are clinical and business reasons for maintaining these patients in treatment. They are still demonstrating a level of distress that warrants further clinical intervention, and without such intervention this group might arguably pose the greatest risk for deterioration and need for more costly, higher levels of care.

The self reported theoretical orientation of the provider seems to have little impact on the length of treatment. Treatment orientations as diverse as Psychodynamic, Cognitive Behavioral and Brief Solution Focused therapies appear to result in essentially equivalent outcomes and length of treatment when applied in the real world of commercial managed care (17). The most tenable hypothesis is that the patients themselves are the primary determinant of length of treatment, and that the decision to terminate treatment is based on the rate of improvement. The faster the improvement, the sooner they are terminated.

Seen in this light, the regression equations used for case mix adjustment are actually an estimate of how much improvement is necessary for the average patient of a given severity to decide that treatment has been adequate. Since the patients appear to be determining the length of treatment, it more accurate to say that the length of treatment is a function of the speed of recovery rather than that the outcome is a function of the length

of treatment. From the perspective of outcomes management and quality improvement, this finding suggests that the focus should be on ensuring that patients achieve a given level of outcome rather than a certain length of treatment. There are obvious implications from these findings regarding the optimal allocation of treatment resources. Outpatient treatment resources allocated to the most distressed patients realize greater benefits per dollar invested than resources allocated to the healthiest patients.

It is also evident that if a system of care wishes to improve outcomes, the greatest opportunity lies with those most in need. While the evidence shows in aggregate that patients tend to remain in treatment until reaching a certain outcome, the variability of this phenomenon at the individual patient level is quite large. The average change for the entire sample is nine points (effect size=.39). However, the standard deviation of the change score is over 19 points.

What this means is that even after controlling for severity using regression techniques the predicted change for any given patient is still only a gross estimate. Despite this limitation, the use of predicted change as a benchmark against which to compare actual change remains a useful tool to evaluate results. If nothing else, it quantifies the importance of case mix and provides a useful reference point for evaluating outcomes in aggregate.

The wide variability in outcomes is further evident when one looks at the 10-20% of cases with the worst outcomes. These patients are substantially worse off than patients with average results in that they show either no improvement, or even substantial worsening in symptoms. Just as patients showing rapid improvement tend to terminate early, patients who fair poorly early in treatment will tend to terminate before

experiencing any substantial benefit. The challenge for an outcomes management program is to target these at-risk cases as early as possible in hopes of averting a premature termination. The fact that early change is predictive of final outcome permits the use of statistical models to target at-risk cases as soon as there are at least two data points in treatment.

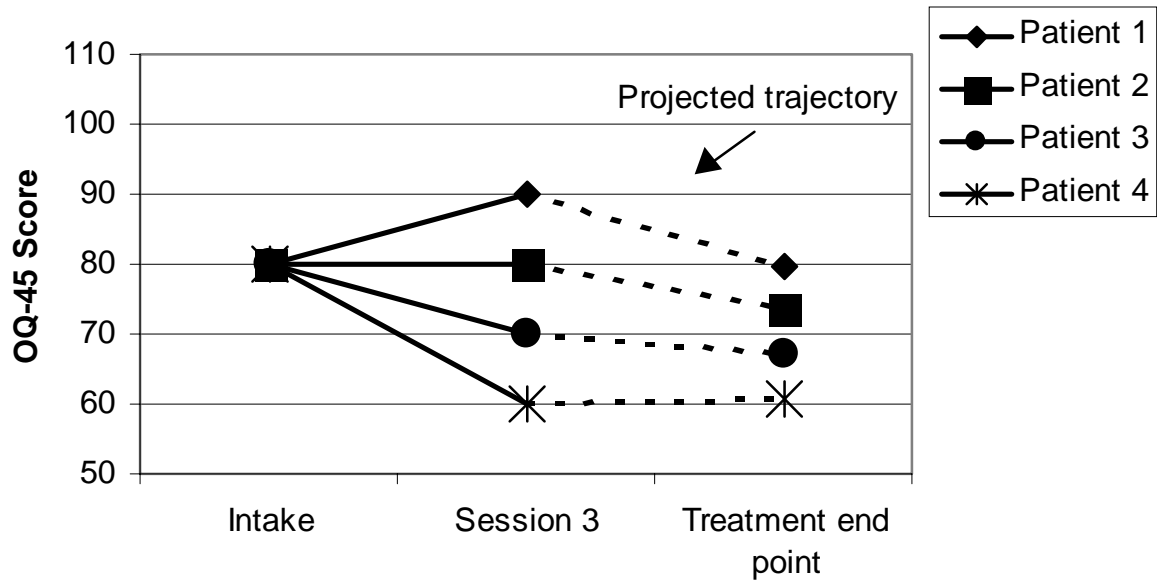
This is accomplished by using regression models employing the intake score and the change score at a given session to predict the final outcome in treatment. The model assumes continued treatment to at least the next data collection point, but does not attempt to predict outcome as a function of  $x$  number of additional sessions. This is in keeping with the finding that the length of treatment is mostly a function of speed of improvement. Once data for the first and third session are obtained, the multiple regression formula accounts for approximately 40% of the variance in final outcomes. Of course the predictive formulas incorporating change at later sessions are different depending on the number of sessions already delivered.

Graph 7 presents four hypothetical cases beginning treatment with the same diagnosis and an intake score on the OQ-45. The expected improvement (utilizing the case mix adjustment model) is 9 points. At the third session they exhibit markedly different responses to treatment. The lines extending from the third to final sessions represent projected change for each patient if they continue treatment beyond the third session.

**Insert Graph 7 about here**

Graph 7

Predicting Change After Session 3



Patient 1 has deteriorated substantially by the third session. Even if the patient continues in treatment, prognosis for substantial improvement from the baseline score is poor. However, if the patient stops at this point, the outcome is much worse, 24 points worse than expected. So despite the overall poor treatment prognosis, the patient still stands to gain significantly by remaining in treatment. Unfortunately, in the data repository over 60% of the patients with deterioration at session three did not remain in treatment till the fifth session.

Patient 2 shows no change by session three, but the probability is that if treatment continues the patient will still achieve close to nine points improvement. Patient 3 has achieved the nine points improvement by the third session, and is likely to gain only modest additional benefit by remaining in treatment. Patient 4 has shown substantial benefit, is and is unlikely to benefit significantly from additional treatment.

Fortunately, in practice there are far more patients following a trajectory similar to Patients 3 and 4 than to Patient 1. In principle this means that for a large system of care the cost of focusing resources to keep Patient 1 in treatment can be offset by a modest shift in resources away from the large number of patients doing very well and unlikely to benefit from further treatment.

A simple analysis of the adult cases based on test scores at the third session. As cited previously, the mean intake score on the OQ-45 in the sample used to develop the PBH norms was 82, with a standard deviation of 24. Twenty percent (20%) of these cases had intake scores at or below 61, the cutoff between the normal and clinical samples. At the other end of the severity spectrum, 23% had scores of 100 or greater.

By the third session, 29% cases still in treatment had scores below 62. The mean of these cases in the normal range was 46, only one point higher than the mean of the community sample. In contrast, at the third session the percentage of cases with scores in the severe range above 100 had dropped to 12%. The point of this analysis is to track what happens to these two groups of patients after the third session. Of the cases with scores in the normal range at the third session, 39% continued to at least the fifth session and averaged 3.8 additional sessions. For those with scores of 100 or more at the third session, 54% continued in treatment, meaning that almost half (46%) of the most severely distressed patients at the third session did not continue in treatment as far as the fifth session. Given the severity of distress, these cases are clearly premature terminators and treatment failures.

Looking at the pattern of change is also telling. Those patients with scores below 62 at the third session as a group had experienced substantial improvement at that point, averaging almost 12 points change in the three sessions (.5 effect size). However, for the cases that continued on in treatment there was no further measurable benefit on average. In fact their OQ-45 scores rose by an average of 3 points from the third session to the end of treatment, similar to the pattern illustrated by Patient 4 in Graph 8. Looking at these cases it seems clear that there is room for reduction in the average length of treatment without sacrificing outcomes. In sharp contrast, the 12% of cases with scores of at least 100 at the third session had deteriorated an average of over 1 point at the third session. This is cause for alarm, considering the elevation of the scores. However, the encouraging news is that of those that did continue, the average improvement between

the third session and the end of treatment (mean additional sessions = 5.7) was 15 points (.63 effect size).

From a quality improvement standpoint, the focus should be on keeping a higher percentage of these high-risk cases in treatment for a sufficient duration to realize benefit. This data supports the assertion that the cost of this effort can be more than offset by reduction in the length of treatment for patients experiencing sub-clinical levels of distress.

#### *Performance feedback and decision support tools*

This section will address the challenge of providing clinicians and clinical managers with outcomes based information to assist in treatment planning and monitoring. Effective decision support tools should result in improved outcomes and much more efficient allocation of resources.

The ALERT system produces clinical outcomes reports on a daily, weekly and monthly basis. The system is also able to track and identify high risk cases based on a number of clinical variables such as diagnoses, clinician and patient reports of suicidal ideation and substance abuse, and treatment history. On a daily basis ALERT scans in data from the most recent encounters, evaluates risk indicators and calculates trajectory of change. The system utilizes algorithms as a clinical aid to Case Managers and generates an individual case report on those patients targeted as high-risk. These algorithm reports are produced on a daily basis, utilizing both patient self-report data (LSQ/YLSQ) and provider-reported data. The algorithm reports are intended to be decision support tools for sorting those cases requiring no intervention from those

requiring active case management. The algorithm report for high-risk cases is faxed to the provider and provides a starting point for a dialogue to determine how best to serve the patient. The focus of the discussion is first on how best to keep the at-risk patient engaged in treatment and second, what changes, if any, are warranted in the treatment plan.

The algorithms evaluate nine variables: age, diagnosis, LSQ/YLSQ scores, trajectory of change projections, critical items from the LSQ/YLSQ regarding substance abuse and suicidal ideation, and clinician assessments of substance abuse and risk for suicide. These algorithms contain over 42,000 separate decision rules encompassing all possible combinations of the variables and their values. As noted earlier, the computerized algorithms are coded using the SAS scripting language. This permits easy modification of the clinical variables and logic as data accumulates and the system “learns.”

Case mix adjustment is achieved by indexing the actual change score against the baseline statistical projection of change. The baseline projected change is calculated from data collected at intake using the aforementioned case mix model. This variable remains constant for that patient throughout the episode of care. The outcomes are indexed by subtracting the baseline projected change from the actual change score to create a residualized change score. Positive values indicate more improvement than expected.

While the system tracks outcomes at the level of the individual patient, it also provides regular reports of aggregated results across multiple patients for use by treating clinicians, clinical managers and administrators. The system provides two sets of outcomes reports. One is for closed cases, referred to as the Aggregate Outcome Report.

The other is the Change Index Report, which projects outcomes for active cases. Used together these reports are powerful tools for managing and monitoring results. These reports are provided on a monthly basis to contracted group practices seeing a high volume of PBH cases. Daily and weekly reports provide similar information on active high-risk cases only.

The Aggregate Outcomes Report (AOR) provides outcomes on closed cases using the Change Index, or average of residualized change scores. The AOR separates results for adults and children/adolescents, and further breaks out results by severity level (four quartiles as determined by intake scores). It provides information on the number of cases seen, the number and percentage of cases with at least two data points in treatment, and the average number of sessions for those cases.

The report provides three key pieces of information needed to evaluate outcomes: (a) the expected outcome based on the average of the baseline predicted change scores using the case mix model, (b) the actual outcome as measured from the first session to the end of treatment, and (c) the Change Index, which is the average of the residualized change scores calculated by subtracting the baseline predicted change from the actual change. Results are expressed as effect size for ease of interpretation and to permit pooling of results from the LSQ and YLSQ.

Figure 1 depicts a sample report for a group practice where the effect size across 288 cases exceeded the case mix adjusted predicted outcome by .10 effect size units. The report indicates that this is an above average result utilizing an alpha level of  $p < .1$ . A ninety (90) percent confidence level is utilized, rather than the more conservative ninety-five (95), because this report is intended to provide general feedback for ongoing quality

improvement purposes and the implications of an increased Type I error rate are seen as less serious.

**Insert Figure 1 about here.**

Figure 1

## PacifiCare Behavioral Health Aggregate Outcomes Report

Provider ID: AAAA

Date of report: 9/20/99

Cases included in this report began treatment between 3 months and 15 months prior to the date of the report.

Age Group Severity at intake	Total Cases	> 1 data point		Change (effect size)		Change Index (actual-expected)
		Number cases	Sessions/ Case	actual	expected	
<i>Normal range</i>	55	18	4.00	0.02	-0.03	0.05
<i>Mildly distressed</i>	60	45	5.50	0.21	0.20	0.01
<i>Moderately distressed</i>	49	28	6.55	0.65	0.49	0.17
<i>Severely distressed</i>	61	40	7.10	0.90	0.75	0.16
<b>Combined Adult</b>	<b>225</b>	<b>131</b>	<b>6.01</b>	<b>0.49</b>	<b>0.40</b>	<b>0.09</b>

Children & Adolescents						
<i>Normal range</i>	15	6	3.40	-0.06	-0.05	-0.01
<i>Mildly distressed</i>	13	6	5.00	0.40	0.23	0.17
<i>Moderately distressed</i>	18	9	7.20	0.55	0.33	0.22
<i>Severely distressed</i>	17	10	7.30	0.80	0.66	0.14
<b>Combined Child/Adolescent</b>	<b>63</b>	<b>31</b>	<b>6.07</b>	<b>0.48</b>	<b>0.34</b>	<b>0.14</b>

### Aggregate Results for All Age Groups

Total number of cases: 288  
 Number of cases with > one data point: 162  
 % of cases with > one data point: 56%  
 Sessions Per Case: 6.02

Change		Change Index
actual	expected	(actual-expected)
<b>0.49</b>	<b>0.39</b>	<b>0.10</b>

Above average

While the AOR provides information on closed cases, the Change Index Report (CIR) is designed to provide similar information on open cases while there is still an time to alter the treatment plan. The CIR presents the Current Change Index, the residualized change score at the most recent session. It goes further in also providing a Predicted Change Index based on the likely outcome if the patient remains in treatment. This is calculated by using the regression formulas incorporating the intake score and change score (see Graph 8) to estimate the change score at the end of treatment. The new projected change score is then indexed by subtracting the baseline expected change from this value. The CIR is designed so that the provider can quickly review patients' progress and focus efforts on those patients that are at the greatest risk. The report utilizes raw score change rather than effect size because it is specific to one instrument. Feedback from consumers of the reports supports the use of raw scores in this context.

Cases are sorted so that those with a projected change index score furthest below the expected change are at the top of the list. Cases with Projected Change Index Scores below zero at the 75% confidence level are highlighted in *italics*. These tend to be the cases with both highest risk for premature termination and greatest likelihood to benefit from further treatment. A weighted average of the current and projected change index scores is used to derive an aggregated estimate of the most likely outcome for the entire patient cohort. This weighted average makes the assumption that 40% of cases will improve with continued treatment, based on analysis of the data repository. Note that these are scores from a 30 item instrument rather than the 45 items used elsewhere in this article, and so average 67% of the OQ-45 scores. The clinical cut-off score in the LSQ is 41. **Insert Figure 2 about here**

Figure 2

## Change Index Report

Provider ID: Example

Date of Report: 09/21/99

Summary Statistics			
Average Baseline Expected Change	3.9	Average Current Change Index (assumes all cases stop now)	-1.0
Average Current Change Score	2.9	Average Projected Change Index (assumes all cases continue)	1.0
Average Projected Change Score (assumes all cases continue in treatment)	4.9	Change Index - most likely outcome (assumes 40% of cases continue)	-0.2

Average results

***Bold italics*** indicates case has >75% probability of below average outcome.

Name	Intake Date	Baseline LSQ Score	Baseline Expected Change	Most Recent Session Date	Most Recent Session Number	Most Recent Score	Current Change Score	Current Change Index	Projected Change Index <i>(assumes continued treatment)</i>
Patient A	<b><i>4/23/99</i></b>	<b><i>60</i></b>	<b><i>8.12</i></b>	<b><i>7/16/99</i></b>	<b><i>4</i></b>	<b><i>99</i></b>	<b><i>-39</i></b>	<b><i>-47.1</i></b>	<b><i>-20.7</i></b>
	5/7/99	55	7.34	8/6/99	3	55	0	-7.28	-1.68
	4/29/99	16	-6.52	6/3/99	3	21	-5	1.46	-1.04
	8/19/99	61	9.11	9/1/99	3	58	3	-6.11	-0.38
	4/10/99	22	-0.55	5/14/99	2	17	5	5.55	0.58
	5/13/99	38	1.12	7/13/99	5	35	3	1.88	0.99
	8/20/99	16	-10	9/14/99	9	18	-2	8.02	1.8
	4/14/99	90	19.35	5/12/99	3	78	12	-7.53	2.19
	7/29/99	49	4.94	8/26/99	5	38	11	5.66	4.19
	7/14/99	65	5.64	8/5/99	3	61	4	-1.64	4.71
Patient B	7/17/99	33	-0.62	8/7/99	2	21	12	12.62	7.67
Patient C	5/25/99	58	8.39	7/12/99	5	27	31	22.61	13.61

This example is based on an actual sample of cases from a single provider. While the overall results are averages, it illustrates several of the points made previously in association with Graph 7. Patient A shows the pattern illustrated by Patient 1 on Graph 7. There is significant deterioration and the outcome is likely to be relatively poor even with continued treatment. However continued treatment is projected to result in 27 points of improvement from the current session forward if the patient remains in treatment. The most critical intervention based on this report is to take steps to keep this patient engaged in treatment. In a well-organized clinical setting, this case might be targeted for clinical staffing and special monitoring.

Like Patient 4 in Graph 7, Patients B and C have achieved significant improvement. Both have scores well within the non-clinical range and are therefore unlikely to achieve measurable benefit from additional sessions. Clinical judgement must be the final determinant in any single case. However, from a population-based perspective, this type of decision support tool has the potential to direct available resources to those most in need and most likely to benefit.

#### *Field and experimental evidence on decision support*

The impact of feedback on patient improvement and allocation of care was tested in a recent experimental study conducted at the Comprehensive Counseling Center at Brigham Young University (14). This study involving 609 patients and 31 clinicians utilized the OQ-45 at every session. Treatment continued until termination was deemed appropriate by the clinician and/or patient. Half of the cases were randomly assigned to the experimental condition in which the clinician had the benefit of feedback on

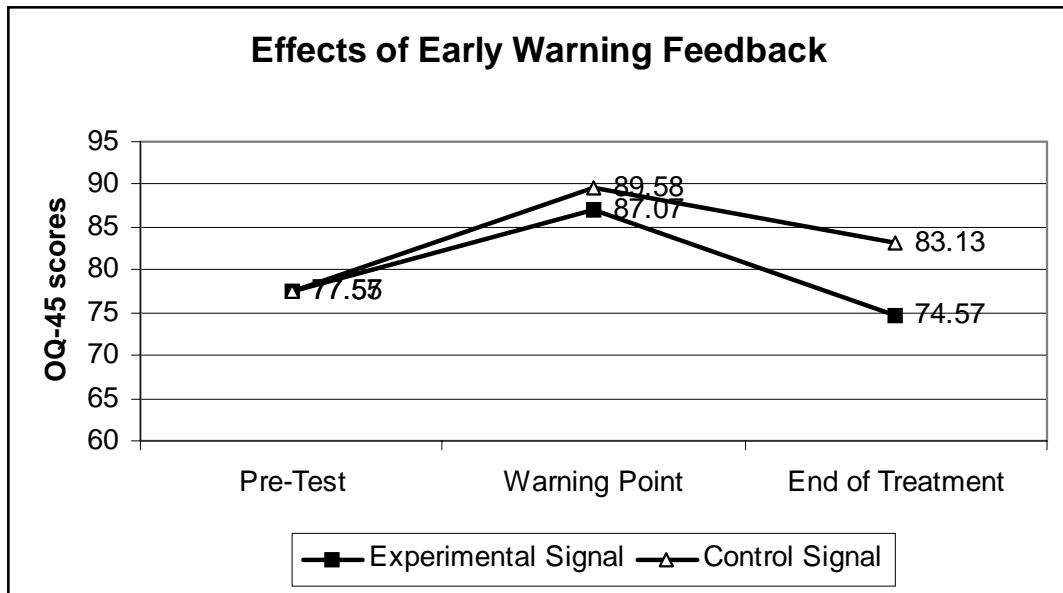
trajectory of change and severity range. The feedback was determined by a set of algorithms that were a function of the number of sessions completed, current level of distress, and assumed likelihood that the patient would fail to recover. In the control condition, the patient completed the questionnaire but the results were withheld from the clinicians.

Space does not permit a complete summary of this study, but for purposes of the present discussion, it is instructive to look at the effects of feedback for the most at-risk patients in the study. Using the algorithm logic, 35 of cases in the experimental condition (11%) and 31 (10%) from the control condition were identified as “signal cases”, those most at risk for premature termination and poor outcome (similar to Patient 1 in Graph 8). The feedback message included a warning that the patient was improving less than expected, along with suggestions to review the treatment plan and to guard against premature termination.

In the experimental condition, the feedback resulted in the signal patients receiving almost twice as many sessions of treatment than the control signal cases ( $p < .001$ ). More importantly, the signal cases in the feedback condition showed significantly more improvement post warning than the control group ( $p < .05$ ). Graph 8 displays the trajectory of change and outcomes for the two signal groups.

**Insert Graph 8 about here.**

Graph 8



At the same time the signal cases in the experimental group were receiving additional services, those non-signal cases that were proceeding well in treatment ended up averaging fewer sessions in the experimental group than in the control group ( $p < .05$ ). As our previous discussion would lead us to expect, this occurred without any degradation in outcomes for these cases. In fact, the non-signal cases in the experimental group averaged more change than the control group, but this difference was non-significant. The additional sessions that are provided to the signal cases was more than offset by the modest reduction in sessions to the non-signal cases, so that the experimental group utilizes overall 4% fewer sessions than the control group.

This study provides support to the premise that it is possible to focus resources and improve results for the most at risk cases without increasing the overall cost of care. Interestingly, one of the questions asked in the post research interviews with the clinicians was “how did you choose to use the feedback you received?”. Of the 29 clinicians interviewed, none reported that they increased or decreased the number of sessions given to a client as a result of the feedback. When the results of the study were presented to the staff, there was considerable surprise expressed over the attendance data.

Results from the field are likewise encouraging. The data repository contains a large number of cases ( $n=4825$ ) treated by large group practices that had participated in previous outcomes management initiatives and had the benefit of some form of decision support and feedback similar to the reports described in this article.. These were compared to cases 1412 cases treated by clinicians that did not receive this kind of feedback. (Note: this sample of larger than the sample used to develop the PBH norms because cases were included even if they were missing therapist generated data such as

DSM-IV diagnoses.) In this analysis, cases treated at one of the sites receiving feedback averaged over 25% more improvement than cases seen by practitioners at sites without this feedback (.29 versus .37 effect size,  $p < .001$ ). Of course any number of other factors could contribute to this result, but it does offer a tantalizing hint of what may be possible with outcomes management techniques.

ALERT was first implemented by PBH for its commercially insured population in February 1999. All of the decision support tools were provided to the nineteen large group practices seeing a high volume of patients. Individual practitioners not associated with one of the groups are contacted by a care manager if the algorithms determine that a cases was at risk, but other wise these practitioners do not receive reports like the Change Index or Aggregate Outcomes Reports. This project provides a mechanism to further explore outcomes in natural settings and to investigate the impact of outcomes management methods on the delivery of care in varied clinical environments.

### *Summary and conclusions*

The preliminary field results of these outcomes management methods are encouraging. However, the early findings must be treated with some caution. Future work will focus on refining the case mix model and identifying process variables such as treatment and case management methods that are associated with superior results. Psychotherapy research has given us more than a quarter century of valuable information on how to assess change associated with behavioral health treatments. While the science of outcomes measurement might be judged to be relatively mature, the implementation of

outcomes management programs like ALERT within large systems of care is in its infancy. Much work needs to be done to validate and refine the methods.

This article is intended to encourage behavioral healthcare organizations to pursue outcomes management programs and contribute to the growing body of knowledge about what works. Organizations like NCQA and JCAHO should drive this process forward since it is now possible to insist that behavioral health delivery systems demonstrate the clinical outcomes associated with their services. Behavioral health practitioners should welcome this development since the evidence suggests clinical outcomes in the field are generally positive and monitoring outcomes during treatment can contribute to even better outcomes.

### *References*

1. Bartlett J, Cohen J: Building an accountable, improvable delivery system. *Administration and Policy in Mental Health* 21(1):51-58, 1993
2. Brown GS, Fraser JB, Bendoraitis, TM: Transforming the future - the coming impact of CIS. *Behavioral Health Management* 14(5):8-12, 1995
3. Brown GS, Lambert MJ: Tracking patient progress: decision making for cases who are not benefiting from therapy. Paper presented at the 29<sup>th</sup> Annual Meeting of the Society for Psychotherapy Research at Snowbird, Utah, 1998
4. Howard KI, Kopta SM, Krause MS, et al: The dose effect relationship in psychotherapy. *American Psychologist* 41:59-164,1986

5. Lambert MJ, Hill CE: Assessing psychotherapy outcomes and processes. Handbook of Psychotherapy and Behavior Change 4<sup>th</sup> ed. Edited by Bergin AE, Garfield SL, New York, Wiley, 1994
6. Lambert MJ, Hansen NB, Umphress V, et al: Administration and scoring manual for the Outcome Questionnaire (OQ-45.2). Wilmington, DL, American Professional Credentialing Services, 1996
7. Wells MG, Burlingame GM, Lambert MJ, et al: Conceptualization and measurement of patient change during psychotherapy: development of the Outcome Questionnaire and Youth Outcome Questionnaire. *Psychotherapy* 33(2):275-283, 1996
8. Lambert MJ, Finch AE: The Outcome Questionnaire. The Use of Psychological Testing for Treatment Planning and Outcomes Assessment 2<sup>nd</sup> ed. Edited by Maruish, ME Mahway, NJ, Lawrence Erlbaum, 1999
9. Wells MG, Burlingame GM, Lambert MJ: Youth Outcome Questionnaire (Y-OQ). In *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment* 2nd ed. Edited by Maruish ME, Mahway, NJ, Lawrence Erlbaum, 1999
10. Lyons LS, Howard KH, O'Mahoney MT, et al: *The Measurement and Management of Clinical Outcomes in Mental Health*, New York, John Wiley & Sons, 1997
11. Jacobson NS, Follette WC, Revenstorf, D: Toward a standard definition of clinically significant change. *Behavior Therapy* 17:308-311, 1986

12. Jacobson NS, Truax P: Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology* 59:12-19, 1991
13. Lambert MJ, Brown GS: Data-based management for tracking outcome in private practice. *Clinical Psychology: Science and Practice* 14(2):172-178, 1996
14. Lambert MJ, Whipple JL, Smart DW, et al: The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? To appear in *Psychotherapy Research*, in press
15. Kopta SM, Howard KI, Lowry JL, et al: The psychotherapy dosage model and clinical significance: Estimating how much is enough for psychological symptoms. Paper presented at the Society for Psychotherapy Research, Berkeley, CA, 1992
16. Howard I, Lueger J, Martinovich Z, et al: The cost-effectiveness of psychotherapy: Dose-response and phase models. In *Cost-effectiveness of Psychotherapy: A Guide for Practitioners, Researchers, and Policymakers*. Edited by Miller NE, Magruder KM, New York, Oxford University Press, 1999
17. Brown GS, Dreis S, Nace D: What really makes a difference in psychotherapy outcomes? And why does managed care want to know? In *The Heart and Soul of Change*. Edited by Miller S, Hubble M. Washington DC, American Psychological Association, 1999