Outcomes Measurement 2.0:
Emerging technologies for managing treatment outcomes in behavioral healthcare
(2009)

Authors:

1. Warren Lambert, PhD; Center for Evaluation and Program Improvement at Vanderbilt University
2. Takuya Minami, PhD; Dept. of Counseling and Applied Educational Psychology at Northeastern University
3. Eric Hamilton, MS; Vice President for Clinical Informatics at Value Options
4. Joyce McCulloch, MS; Vice President for Health Informatics at United Behavioral Health
5. John Peters, PhD; Director of Outcomes Management at Kaiser Permanente – Northern California region
6. Marni Selway; Director of Project Management at APS Healthcare
7. Carson Graves, MHA, MA; Assistant Director for Provider Services and Wellness at Regence BC/BS
8. Joanne Cameron, PhD; Center for Clinical Informatics
9. Jeb Brown, PhD; Center for Clinical Informatics

Table of Contents

Summary:

The article summarizes recent advances in outcomes measurement and statistical methods used to identify highly effective behavioral healthcare practitioners. The authors argue for actively identifying and rewarding such clinicians, including the potential for preferential reimbursement based on treatment outcomes.

1. Introduction

This paper represents a collaboration of individuals from a wide range of public and private organizations and academic institutions. These individuals share a commitment to improving treatment outcomes in behavioral healthcare through the routine use of patient-completed outcome questionnaires concurrently with treatment. The use of routine outcome assessment, combined with ongoing feedback to the treating clinicians, is often referred to as “outcomes informed care.” The measurement technology involved in the broader adoption of outcomes informed care is collectively referred to as "Measurement 2.0," indicating accumulated innovations that are a substantial departure from 20th century practices.

The application of psychological measurement to clinical practice in the late 20th century was shaped by the training of the vast majority of psychologists in practice at that time. Naturally, those practitioners had received their training in measurement practices in prior decades from faculty trained even earlier going back to the post World War II blossoming of clinical psychology and psychotherapy. Measurement practices, and just as importantly, practitioners’ beliefs about how measurement can be applied to their practices, has been very slow to catch-up with the substantial theoretical and technological advances in measurement over the past 30 years.

The 20th century model of applied psychological measurement, which we call “Measurement 1.0,” was characterized by proprietary symptom rating scales, which were sold in identical forms to all users. Large-scale users, such as behavioral healthcare organizations, had to buy copyrighted commercial test forms, often at a cost as high as $10 per administration, or pay for expensive remote scoring services. Test authors and publishers placed idiosyncratic limitations on how the test could be used, sometimes charging different prices for academic and commercial users. By the end of the 20th century, there were so many proprietary mental health symptom measures that it was hard to believe any single one was the best.

In the 21st century we are beginning to see a new model take hold, which we refer to as “Measurement 2.0.” This model of applied psychological measurement recognizes the power of advances in psychometric methods and the widespread availability of statistical software to overcome many of the limitations of Measurement 1.0. Rigorous psychometric investigations can now be conducted inexpensively on any questionnaire to determine whether or not all items work in the way they were intended. When organizations need to revise items or create new ones, the psychometric properties of the new items can be evaluated in less than a day once the data are collected. Thus, with appropriate psychometric knowledge, proprietary measures are no longer the only source of valid measurement.

Outcomes informed care in the 21st century, powered by Measurement 2.0 technology and principles, consists of the following components:

(1) patient self-report (or caregiver report) items and questionnaires,

(2) data collection technology,
(3) data standardization and analytical methods, and
(4) timely feedback system.

As a general workflow, the clinician asks the patient in therapy to complete a brief self-report questionnaire with common and specific psychological symptoms and/or patient-therapist treatment alliance indicators. The self-report forms completed by the patients are transmitted (typically by fax) to a central server that processes the information into a central database. Then, based on the overall collected data, the patients' responses are standardized so that they can be compared within and between patients. The standardized outcomes of the patients are then communicated back to the clinicians via a web-based toolkit, allowing clinicians to keep track of their clients' progress as well as reevaluate their current course of treatment with each of their patients. Depending on infrastructure, organizations can also implement algorithms that inform clinicians of specific patients identified as being at risk based on significant levels of distress or impairment as indicated by the collected data.

2. The Building Blocks of Change

a. Psychometrics

In the past, outcomes informed care initiatives in behavioral health were necessarily based on normed, copyrighted questionnaires originating in academic research. However, from the standpoint of large organizations equipped with psychometric sophistication, relying on copyrighted measures is advantageous only if the organization lacks access to large samples of normative data. However, large organizations can generate data at rates far surpassing any research study. Therefore, for large organizations, the size, quality, and relevance of the normative data quickly exceed the normative data available for research-based instruments.

Copyrighted questionnaires have additional disadvantages in large-scale outcomes informed care initiatives. Aside from the obvious issue of cost, the end user (e.g., clients, therapists, organizations) has no control over the items. Test publishers are generally reluctant to permit modifications of their questionnaires, including rewording items for special populations, reducing the number of items, and adding new items. Such prohibitions severely limit the end user's capacity to optimize the questionnaires to their caseload.

These limitations of copyrighted questionnaires led some larger organizations to devise their own outcome measures. With the aid of in-house and consulting psychometricians, organizations can create items and questionnaires that are as valid and reliable as research-based measures yet are tailored to their measurement needs. The Wellness Assessment (Adult and Youth versions) employed by Optum Health Behavioral Solutions (OHBS) is an example of such a questionnaire. The Wellness Assessment was developed collaboratively with external partners, tested within national samples of OHBS membership, and evolved in response to multiple psychometric evaluations and changing business needs. In addition, other organizations represented by the authors of this article continue to develop items and questionnaires based on analyses incorporating classical test theory, factor analysis, and IRT.

Applying a battery of psychometric tools to new items and questionnaires is nothing new to researchers, and can also work in applied measurement settings. The specific combination of rigorous psychometric analyses applied by the authors is sometimes referred to as the “psychometric torture test.”   This set of analyses was established at Vanderbilt University and used to develop the Peabody, which is a set of 10 free measures for providing feedback to clinicians about the progress of child clients.[1] The purpose of the Peabody battery was to make item evaluation objective and public, so there is an explicit evidence-based rationale when items are retained, revised, or dropped. In the “torture test,” many figures of merit are measured for every item and for the full scale in order to weed out or revise items that do not work.

The “torture test” is ecumenical. It views the three main approaches to psychometrics (classical test theory, factor analysis, and IRT) as useful tools, rather than as competing religions. A number of tools are based on the classical test theory, such as Cronbach’s alpha [2,3], item-total correlations, and the Spearman-Brown [4,5]approximation. Factor analysis may be exploratory and/or confirmatory, and it provides useful indicators such as scree plots, factor loadings, and, most importantly, model-fit statistics [6] IRT has two broad streams. The one-parameter Rasch models offer many tools for practical test construction. [7,8] The other is the multiple-parameter IRT , which offers a great depth of features for model-based measurement research. [9,10]

The torture test provides the test developer with exhaustive metrics for every test item and for the test as a whole. In addition, the torture test helps the novice test developer understand the statistical characteristics of items that measure well. In test development, all statistics are initially interpreted by the psychometrician as (statistically) good or bad, with the understanding that no single statistic is a definitive representation of an item’s value. Others involved in the test development who are not psychometricians are able to make evidence-based decisions on item revision and retention based on explanations provided by psychometricians, and make reasonable decisions about the tradeoffs, such as between test length and reliability. Readers interested in the technical details will find examples in the free manual for the Peabody Battery [1], or in two articles by one of the co-authors (Lambert), of which the Psychological Assessment article contains the most technical detail. [1,11,12]

b. Therapist Variability in Client Outcomes

The idea that clinicians differ in clinical competence and effectiveness is a hypothesis that has been entertained since perhaps the inception of psychotherapy. [13,14] However, 20th century results were obscure until the statistical method of hierarchical linear modeling (HLM) became widely available to psychologists. [15] This approach has various names, such as random, multilevel modeling, and mixed modeling.[16-18] By the late 1990's dependable software was available, e.g. from SAS, HLM,,, SPSS,, and S+ or R.[18-21] HLM is ideal for behavioral health data because there are multiple patients who are seen by the same clinician (i.e., nesting). HLM supports estimates of variance both at the client and therapist levels. A third level, such as the clinic, may simultaneously be modeled as well.

Before user-friendly software was available for HLM, clinical trials and treatment evaluation studies ignored the problem of nested data. As such, they likely overestimated the effect of the type of treatment because they failed to correctly attribute the variance due to therapists. [22-24]Recent analyses have since revealed that the percentage of variance in treatment outcomes attributed to therapists far exceed the variance due to the type of treatment.[22,25-28] In addition, almost three decades of randomized clinical trials and meta-analyses of these trials have consistently supported the "dodo bird effect," that all bona fide psychological therapies appear to have similar treatment effects for majority of psychological disorders [29-36] Further, such effects have been found in studies investigating the effects of antidepressants.[28,37] In particular, Wampold and Brown found that for patients who received medication and psychotherapy concurrently, the effect of medication was impacted by the clinician's competence in providing psychotherapy.[37] These findings suggest that the contributions made by individual therapists far exceed the effects of the different types of treatment.

Our collective Measurement 2.0 data warehouse containing data from multiple organizations shows that therapists do indeed vary in their clinical outcomes. In addition, these differences appear to be stable over time. When patients are referred to clinicians who have above average outcomes in the past, these patients are more likely to report better than average outcomes. In addition, the data has shown that clinicians who support and actively participate in the outcomes informed care initiatives, on average, report above-average results compared to those who participate sparsely. Therefore, it is in the best interest of effective clinicians to participate in outcomes informed care initiatives because participation would likely provide favorable evidence of their clinical effectiveness. On the other hand, for clinicians who are getting below-average results, efforts to improve their outcomes by use of the feedback tool may also lead to favorable results. Likewise, it is in the interest of employers and payers of mental health services to identify these effective clinicians in order increase patient access to these highly effective clinicians. Conversely, it is also in the interest of payers to identify clinicians whose clinical competence is less than optimal so as to implement strategies to improve their clinical competence. Most importantly, outcomes informed care initiatives are in the patients' best interest. Outcomes informed care can help connect clients with the most effective therapists and they can improve outcomes by the effects of outcome monitoring itself. After all, what is good for clients should be the common ground for therapists, clinics, and payers.

c. Benchmarking

To say whether particular outcomes are "good," we ask the question, "Compared to what?" That there is variability among clinicians in treatment effectiveness need not be a problem if even the least effective therapists are still effective enough based on some meaningful criterion (i.e., benchmark). Conversely, variability among clinicians is less of an issue if those who are considered relatively highly effective are still below the benchmark. Therefore, the question of criterion becomes central. Evaluations that compare data obtained from outcomes informed care initiatives and other real-world clinical settings against established criteria are called benchmarking studies. Over the past decade, a growing number of research articles have addressed methods of benchmarking behavioral health treatment. [38-44]

Traditionally, benchmarks established from research have been considered more valid than real-world outcome data. A recent study compared data from outcomes informed care initiatives against benchmarks derived from clinical trials and found that behavioral health services provided in a managed care context have similar treatment effects compared to well-controlled clinical trials. [45] The benchmark used for this study was based on published clinical trials of adult depression.[46]

As previously mentioned, data accumulated from a number of large-scale outcomes informed care initiatives using Measurement 2.0 principles has now far exceeded the amount of data obtained from any single clinical trial. These huge real-world samples will likely provide the best benchmarks for effectiveness research on therapy in real-world conditions in the near future, as opposed to efficacy trials using carefully-selected special samples.

An additional reason for using samples from large-scale outcomes informed care is that, unlike randomized clinical trials, real-world clinical data include patients that would be selected out of clinical efficacy trials (e.g., low initial severity, co-morbid conditions). Clinical trials require homogeneity in their patients, but clinics in the real world sees anyone who comes through their door. Therefore, the benchmarks currently established from randomized efficacy trials are not easily generalized to patients in these real-world settings. In the following section, we present a solution to the problem of benchmarking real world against results from clinical trials through the use of a "Severity Adjusted Effect Size" statistic.[47]

3. From Research to Practice: Putting Together a New Model

a. Patient Self-Report Items and Questionnaires

Although there have always been academic criticisms against patient self-reports, clinician ratings might be even more problematic. There has been consistent evidence pointing to better treatment effects when measured by clinicians, even when the evaluating clinician is not the one providing treatment. [30,46,48] Clinicians may consider patients doing better than what the patients perceive. Second, there has been a historical shift in behavioral health with regards to who has the "authority" to determine outcome. In the past, clinicians were vested with this authority in their role as experts. But in recent years, patients' own sense of distress has become a crucial indicator of improvement. Even in primary and specialty medical care, there is growing acknowledgement that patient self-reported information is critical to a valid assessment of outcome. In fact, a major component of the NIH Roadmap to improving health care research is called the PROMIS initiative: Patient Reported Outcomes Management Information System.[49] This approach recognizes that patient symptoms, distress, and functioning are in some ways more important than clinical indicators such as x-rays and lab tests in assessing clinical effectiveness. Therefore, from many converging perspectives, it is unreasonable, if not inappropriate, to dismiss patients' self-reported clinical symptoms as invalid.

The authors also believe that it is untenable to base effectiveness evaluation on the clinician's report. Clinicians, overwhelmed with administrative responsibilities as it is, are often unwilling to increase their responsibilities. [50] Also, treating clinicians may be so motivated to find "progress" that they lose the patient's perspective. Finally, cost would prohibit a system that requires independent clinician assessments. For these reasons we believe that effectiveness measurement based on clinician assessments would not be practically feasible.

However, this is not to say that clinicians' inputs are unnecessary in outcomes informed care initiatives. Feedback from communities of clinicians has been thoroughly incorporated throughout the development of Measurement 2.0, including professional organizations such as the American Psychiatric Association and the American Psychological Association. In particular, Measurement 2.0 has moved away from assessment of progress based on a single questionnaire, regardless of how global these items may be. In Measurement 2.0, numerous items have been developed based on input from communities of clinicians so as to better reflect the issues that their patients face. Incorporating a thorough psychometric evaluation of how each item behaves relative to other items, standardization has been possible at the item level so that meaningful comparisons can still be done even if two forms do not have identical items.

b. Technology

It is a common misunderstanding that "newer is better" for all systems. The same could be said about outcomes informed care initiatives that often aim to be too technically "advanced." Based on over two decades of outcomes informed care implementation, it is now firmly established that Measurement 2.0 cannot yet be widely implemented if data collection is done purely online. The main reasons are all related to cost. First, as compared to traditional paper-and-pencil methods, the cost of electronic self-reports becomes prohibitive if the clinicians have to collect data from more than one patient simultaneously, which would require multiple computers or other expensive hardware. Secondly, computers require much more space than paper. Third, whereas paper-and-pencil questionnaires require no instructions, computer-administered assessments require instructions given by a staff member. Fourth, it is more costly to protect patients' confidentiality when they are filling out the questionnaires electronically than when using paper. The technology involved with large-scale outcomes informed care initiatives must be low-cost and available anywhere—traditional paper-and-pencil questionnaires that are optimized for optical character and mark recognition software. These forms can easily be faxed to a central data server that scans the forms and electronically scores the patients' responses.   Computer assisted assessment, on the other hand, has several potential advantages and is being implemented in some large institutional settings. As computer technology becomes more ubiquitous and clinics increasingly adopt electronic medical records, computer-based assessment may become more commonplace.

c. Severity Adjusted Effect Size: Data Standardization and Analytical Methods

The various organizations represented by the authors of this article have collectively agreed on a common metric for benchmarking provider: the Severity Adjusted Effect Size (SAES).

SAES benchmarks are derived through three steps: (a) calculation of observed effect size, (b) calculation of predicted effect size based on case mix variables, and (c) adjusting the observed effect size based on the difference between the observed effect size and the predicted effect size (i.e., residual). This approach is the widely used "residual gain score" of Cronbach and Furby .[51]

The observed pre-post effect size is translated into a Cohen's *d* effect size calculated by dividing the difference between the intake score and the score at last self-report with the standard deviation of the intake score.[52.53] Although this statistic provides the absolute magnitude of the treatment effect, in outcomes informed care initiatives with patient heterogeneity, differences in case mix prevent the use of the raw observed effect to compare across patients. Therefore, in the second step, the effect of the case mix is statistically adjusted using multiple regression. Through the use of regression, the impact of each of the case mix variables is estimated, and a prediction is made with regards to the expected effect size given a particular case mix. As a result, in the third step, SAES is calculated by adding the difference between the expected effect size and the observed effect size (i.e., residual) to the observed mean effect size of the outcomes informed care population. In this way, the effect sizes are standardized across patients allowing for meaningful comparisons across patients, clinicians, and organizations even if there are differences in case mix.

Among numerous case mix variables, the intake severity of the patient's distress has invariably been the strongest predictor of change. Specifically, as greater distress has consistently led to greater observed effect size, SAES adjusts the observed effect size based on the patient's initial severity.   This makes sense because patients with little pathology, the "worried well," have little room for improvement.

In a related issue, data from outcomes informed care initiatives have consistently revealed that a significant percentage (25~35%) of patients enter treatment with intake scores indicating low levels of distress no different from individuals in the community who have never sought behavioral health services. This group of patients, although typically satisfied with the services they receive, on average show no improvement on the outcome questionnaires. In fact, many trend upwards in the course of treatment (i.e., become “worse”). Currently, we exclude patients who initially score at this low level of distress from the analysis because (a) including this group would make any comparisons with clinical trials benchmarks invalid and (b) the general clinical symptom measure as it is currently developed do not appear to be appropriate in measuring treatment progress for this population.

d. Estimating Clinicians’ Treatment Effectiveness

Once SAES is calculated for each patient/case, data are then aggregated by clinicians to estimate their treatment effectiveness. In doing so, rather than taking the simple average SAES of the clinicians’ caseload, clinicians’ effectiveness is estimated using a random effects model HLM taking into consideration the nested nature of the data (i.e., clinicians seeing multiple patients). [15,22,47] As this modeling considers the estimate of the clinician’s average effectiveness as a random factor rather than fixed, the modeling results in providing a “benefit of the doubt” effect for clinicians whose average outcomes are below the overall average across clinicians. In other words, as the modeling considers clinicians’ caseload as a sample from a larger pool of possible cases, their estimated effectiveness is adjusted toward the overall average across clinicians. Although this method also leads to adjusting the above-average clinicians’ effectiveness toward the average (e.g., lower estimated effect size rather than a simple average of their caseload), this does not become a practical issue because (a) the relative standing among clinicians based on effectiveness remains practically unchanged among clinicians with a large enough number of cases (e.g., n = 15, r = .99 between simple average and adjusted estimate of effectiveness) and (b) the higher the clinician’s caseload, the closer their estimated effectiveness is to the simple average of their caseload.

Statistical estimates contain error. As such, a confidence interval for the clinician’s estimated treatment effectiveness is calculated based on the standard error of the estimate. Of interest in evaluating the clinician’s estimated effectiveness is the lower-bound confidence limit. In the context of treatment effectiveness, the “true” treatment effectiveness is considered to be higher than the lower-bound confidence limit value with a certain percentage of confidence (e.g., 95% confident that the “true” treatment effect is larger than d = 0.3). For a clinician to be designated as “effective,” the lower-bound confidence limit of the clinician’s estimated treatment effectiveness needs to exceed a certain criterion. This numerical criterion is currently set at d = 0.5, which is a magnitude that is approximately the halfway mark between the benchmark for adult depression treatment in clinical trials (d = 0.8) and the natural symptom reduction without treatment (d = 0.15).[46] For example, if the clinician’s estimated effectiveness is d = 0.7 and the lower-bound confidence limit is d = 0.6 (which is larger than d = 0.5), the clinician is designated as “effective.” The method also implies that the clinician need not have an estimated treatment effectiveness that is at or greater than the benchmark derived from clinical trials or the observed average in the data to be designated as “effective.”

A crucial point is that a clinician's lack of designation as effective does not mean that the clinician is ineffective. As the lower-bound confidence limit is partly a function of the clinician's caseload, the lower-bound confidence limit may not exceed $d = 0.5$ simply because the clinician has only a handful of cases in the database. For example, it is possible that a clinician with two cases in the database with an estimated effectiveness of $d = 1.0$ has a lower-bound confidence limit that is lower than $d = 0.5$ because of the small number of cases. On the other hand, a clinician with an estimated effectiveness of $d = 0.6$ may be designated as effective because the clinician has many cases in the database, resulting in a lower-bound confidence limit at or above $d = 0.5$. Therefore, as the cutoff criterion is well below the average treatment effectiveness observed in a managed care setting ($d = 0.83$), the majority of the clinicians with a sizeable caseload in the database will likely be designated as effective.[42] In addition, as the designation that a clinician is "effective" is not based on a comparison against other clinicians in the database but is based on a credible benchmark, it is possible that all clinicians in the database are designated as "effective."

e. Clinician Feedback System

Arguably the most crucial aspect of outcomes informed care is the feedback given to the clinicians with regard to their patients, as without this feedback, it is unlikely that the patients will benefit. In Measurement 2.0, a web-based application (i.e., Clinician's Toolkit) provides clinicians access to their data to view both aggregated results and individual session-by-session patient-level scores. Clinicians differ in what they like to see, and the Clinician's Toolkit allows for customization. These interactive aspects of the Toolkit help to engage clinicians, encouraging them to monitor their own outcomes so they can become what we call "outcomes informed clinicians."

4. Getting Clinicians On Board

The psychometric and technological challenges addressed in this article are small problems compared to the difficulty of engaging clinicians in the process. Not surprisingly, the message "I'm from the managed care company and I'm here to measure you" has not been met with enthusiasm by clinicians.

Lack of interest in outcomes informed care initiatives takes many forms. Most common are the complaints about the additional workload for clinicians and staff. There is no getting around the problem that even the simplest questionnaires will require work for someone, and the frequent initial reaction from providers is to view measurement efforts as yet another administrative burden of serving the managed care organization, but have no benefit to patients or clinicians.

Clinicians' participation rates vary even when there are clear incentives, such as an increase in referrals and/or higher reimbursement rates. The reasons take many forms, such as concerns about measurement methods and case mix adjustment. Providers often express the concern that because their cases are qualitatively different and/or more difficult than those of other providers, they might be unfairly penalized by measurement. Many providers express anxiety regarding measuring their own outcomes and making results available to potential patients. A close look at actual outcome data may provide some clues regarding the source of anxiety. The fact is that outcomes are highly variable, even for the most effective clinicians. Some patients get significantly worse, some experience rapid improvement, and most are somewhere in between. The simple fact of such high variability makes it cognitively impossible to "guesstimate" what their average outcome is. In addition, for any given case, clinicians likely have a number of different hypotheses as to why treatment was or was not effective for their clients. Therefore, the anxiety is likely rooted in the fact that clinicians, except for those who are actively participating in outcomes informed care initiatives, have difficulty in objectively assessing what their actual outcomes are or how they might compare to their peers. It is natural to be anxious under these circumstances.

At the same time, there is a subset of clinicians who do appear willing to participate without too much encouragement. While these clinicians are currently a minority, our data suggests that these clinicians tend to have exceptionally good outcomes. Therefore, it is clearly to the advantage of potential patients, employers, and payers that these clinicians be recognized and every effort made to steer referrals to them. Likewise, strong evidence of effectiveness could be used to justify increased reimbursement, regardless of level of training.

Therefore in summary it appears in the best interest of both those organizing and financing a system of behavioral healthcare and the patients being served to support those clinicians who participate in outcomes informed care. Transparency in how data is analyzed and reported, combined with forthrightness about use of the data may eventually reduce clinicians’ anxiety. Once clinicians see that their results are actually very good and they are recognized and rewarded for their efforts, enthusiastic participation may eventually develop.

Similarly, there is no reason for any organization to harm a provider's reputation or to their ability to earn a living. However, this unwillingness to say something negative about any provider should not prevent organizations from saying something positive about a clinician who has strong evidence of effectiveness. Clearly, patients are better served if managed care organizations make effort to steer referrals towards clinicians with a track record of high-level effectiveness with a variety of patients.

5. Future Directions

We have described the evolution of outcomes informed care from the perspective of advances in psychometrics and treatment effectiveness research and the benefits of collaboration across multiple organizations. As we have shown, the measurement methodology and information technology needed to measure treatment outcomes systematically and cost effectively now exists.

Many questions remain, but the rapid accumulation of data provides the necessary conditions for further learning. Unanswered questions include: What is the relationship between clinicians’ use of feedback systems such as the Clinician’s Toolkit and treatment outcomes? How do the most effective clinicians use outcomes feedback? Which clinicians are most likely to benefit from feedback? Answers to these and similar questions will form the basis of the “best practices” of the future.

While we believe we have a solid foundation for measuring treatment outcomes, the ethos of Measurement 2.0 encourages continuous experimentation. Exploring new variables to improve case mix models and to provide clinicians with useful clinical information, combined with the open sharing of technological information, will result in best measurement practices being propagated across organizations.

In our view, it is crucial to support clinicians who implement outcomes informed care. The various organizations represented by the authorship of this paper are reaching out to these clinicians to find ways to facilitate the collection of data with as little burden to the clinicians as possible. By centralizing many of the processes for data capture and warehousing, the workload on the clinicians is reduced to the effort to give the questionnaire to the patient and later fax the form to a tool-free number.

Making it easy to collect data is not enough. We also need to support clinicians in the use of data to improve outcomes by providing ongoing training and consultation to increase providers' sophistication and comfort with the methodology of outcomes informed care.

We need to recognize and reward highly effective clinicians, and consider the use of preferential payment for demonstrated effectiveness. Cost-benefit analyses can be performed to see if preferential reimbursement rates are justified. Unfortunately, in the near term the practicalities of provider contracting, not to mention variations in state laws and regulations, make pay-for-performance strategies difficult to implement. However, this should not keep us from planning for a future when many more clinicians are ready to use outcome data for their benefit, which ultimately benefits their patients.

References

1. Bickman L, Breda C, Dew SE, et al. Peabody Treatment Progress Battery Manual Electronic version. Retrieved 09/01/2006, from http://peabody.vanderbilt.edu/ptpb/
2. Cronbach LJ, Shavelson RJ. My Current thoughts on Coefficient Alpha and Successor Procedures. Educational and Psychological Measurement. 2004; 64(3):391-418.
3. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297-334.
4. Brown W. Some experimental results in the correlation of mental abilities. British Journal of Psychology. 1910;3:296-322.
5. Spearman C. Correlation calculated from faulty data. British Journal of Psychology. 1910;3:171-195.
6. Yu CY. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes [unpublished doctoral dissertation]. Los Angeles (CA):UCLA; 2002.
7. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Mahwah (NJ): L. Erlbaum; 2001.
8. Rasch G. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Expanded ed. Chicago: University of Chicago Press; 1980.
9. Embretson SE, Hershberger SL. (Eds.). The new rules of measurement: what every psychologist and educator should know: Mahwah (NJ): Lawrence Erlbaum Associates; 1999.
10. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah (NJ): L. Erlbaum Associates; 2000.
11. Greco LA, Lambert W, Baer RA. Psychological inflexibility in childhood and adolescence: development and evaluation of the Avoidance and Fusion Questionnaire for Youth. Psychological Assessment. 2008; 20(2):93-102.
12. Walker LS, Beck JE, Garber J, Lambert W. Children's Somatization Inventory: Psychometric Properties of the Revised Form CSI-24. J Pediatric Psychology;2008.
13. Luborsky L.The personality of the psychotherapist. Menninger Quarterly.1952; 6:1-6.
14. Rosenzweig S. Some implicit common factors in diverse methods of psychotherapy. American Journal of Orthopsychatry. 1936;*6*:412-415.
15. Raudenbush SW, Bryk AS. Hierarchical linear models: applications and data analysis methods. 2nd ed. Newbury Park (CA): Sage Publications Inc; 2002.
16. Hedeker D, Gibbons RD. Longitudinal data analysis. Longitudinal data analysis. Hoboken (NJ): Wiley-Interscience; 2006.
17. Snijders T, Bosker R. Multilevel Analysis: an introduction to basic and advanced multilevel modeling. 2000 ed. Thousand Oaks (CA): Sage Publications; 1999.
18. Pinheiro JC, Bates DM. Mixed-effects models in S and S-PLUS. New York: Springer; 2000.

19. Littell RC, Milliken GA, StroupWW, Wolfinger RD, Schabenger O. SAS System for Mixed Models. 2nd ed. Cary (NC): SAS Institute; 2006.
20. Bickel R. Multilevel analysis for applied research: it's just regression! New York: Guilford Press; 2007.
21. Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modeling Using Stata. 2nd ed. College Station (TX): StataCorp; 2005.
22. Wampold BE, Serlin RC. The consequences of ignoring a nested factor on measures of effect size in analysis of variance designs. Psychological Methods. 2000;4:425-33.
23. Wampold BE, The Great Psychotherapy Debate: Models, Methods and Findings. Mahwah, JJ: Lawrence Erlbaum Associates Inc; 2001.
24. McKay, K. M., Imel, Z. E., & Wampold, B. E. (2006). Psychiatrist effects in the psychopharmacological treatment of depression. Journal of Affective Disorders, 92, 287-290.
25. Luborsky L, Crits-Christoph P, McLellan T, et al. Do therapists vary much in their success? Findings from four outcome studies. American Journal of Orthopsychiatry. 2006;56:501-12.
26. Crits-Christoph P, Baranackie K, Carrilm K, Luborsky L, McLellan T, Woody G. Meta-analysis of therapist effects in psychotherapy outcome studies. Psychotherapy Research. 1991;1:81-91.
27. Crits-Christoph, P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. J Consulting and Clinical Psychology. 1991;59:20-6.
28. Kim D, Wampold BE, Bolt DM. Therapist effects in psychotherapy: a random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. Psychotherapy Research. 2006;16:161-172.
29. Smith ML, Glass GV. Meta-analysis of psychotherapy outcomes studies. American Psychologist. 1977;32:752-760.
30. Smith ML, Glass GV. The benefits of psychotherapy. Baltimore: The John Hopkins University Press; 1980.
31. Shapiro DA, Shapiro D. Meta-analysis of comparative therapy outcome studies: A replication and refinement. Psychological Bulletin. 1982;92:581-604.
32. Robinson LA, Berman JS, Neimeyer RA. Psychotherapy for treatment of depression: a comprehensive review of controlled outcome research. Psychological Bulletin. 1990;108:30-49.
33. Wampold BE, Mondin GW, Moody M, Ahn H. A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, “All must have prizes.” Psychological Bulletin. 1997;122;203-15.
34. Ahn H, Wampold BE. Where oh where are the specific ingredients? a meta-analysis of component studies in counseling and psychotherapy. Journal of Counseling Psychology. 2001;48:251-7.
35. Luborsky L, Rosenthal R, Diguer L, et al. The dodo bird verdict is alive and well—mostly. Clinical Psychology Science Practice. 2002;9:2-12.

36. Lambert MJ, Ogles BM. The Efficacy and Effectiveness of Psychotherapy. In M. Lambert (Ed.), Bergin and Garfield’s Handbook of Psychotherapy and Behavior Change. 5th ed. New York (NY): John Wiley and Sons. 2004;139-193
37. Wampold BE, Brown GS. Estimating therapist variability: A naturalistic study of outcomes in private practice. Journal of Consulting and Clinical Psychology. 2005;75: 914-923.
38. Brown, G.S., Burlingame, G.M., Lambert, M.J., Jones, E. & Vacarro, J. (2001) Pushing the quality envelope: A new outcomes management system. Psychiatric Services, 52 (7), 925-934.
39. Brown GS, Jones E, Lambert M.J, Minami T. Evaluating the effectiveness of psychotherapist in a managed care environment. American Journal of Managed Care. 2005;2(8):513-520.
40. Matumoto K, Jones E, Brown J. Using clinical informatics to improve outcomes: a new approach to managing behavioral healthcare. Journal of Information Technology in Health Care. 2003; 1:135-150.
41. Merrill KA, Tolbert VE, Wade WA. Effectiveness of cognitive therapy for depression in a community mental health center: a benchmarking study. Journal of Consulting and Clinical Psychology. 2003;71:404-409.
42. Minami T, Wampold BE, Serlin RC, Hamilton EG, Brown GS, Kircher JC. Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment. Journal of Consulting and Clinical Psychology. 2008;76:116-124.
43. Wade WA, Treat TA, Stuart GL. Transporting an empirically supported treatment for panic disorder to a service clinic setting: a benchmarking strategy. Journal of Consulting and Clinical Psychology. 1998;66:231-239.
44. Weersing VR, Weisz JR. Community clinic treatment of depressed youth: benchmarking usual care against CBT clinical trials. Journal of Consulting and Clinical Psychology. 2002;70:299-310.
45. Minami T, Serlin RC, Wampold BE, Kircher JC, Brown GS. Using clinical trials to benchmark effects produced in clinical practice. Quality & Quantity. 2008;*42:*513-525.
46. Minami T, Wampold BE, Serlin RC, Kircher JC, Brown GS. Benchmarks for psychotherapy efficacy in adult major depression, Journal of Consulting and Clinical Psychology. 2007;75:232-243.
47. Minami T, Brown GS. Benchmarking therapists: furthering the benchmarking method in its application to clinical practice. Manuscript in preparation. Expected publication date: 2010.
48. Lambert, MJ, Hatch, DR, Kingston, MD, Edwards, BC. Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: A meta-analytic comparison. Journal of Consulting & Clinical Psychology. 1986; 54:54-59.
49. Ader D. Developing the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care. 2007; 45(5): S1-S2.
50. Rupert, PA, Baird, KA. Managed care and the independent practice of psychology. Professional Psychology: Research and Practice. 2004; 35:185-193.
51. Cronbach, L. J., Furby, L. How should we measure "change" - or should we? Psychological Bulletin. 1970; 74:68-80.

52. Becker BJ. Synthesizing standardized mean-change measures. British Journal of Mathematical and Statistical Psychology. 1988;41:257-278.
53. Morris SB. Distribution of the standardized mean change effect size for meta-analysis on repeated measures. British Journal of Mathematical and Statistical Psychology. 2000;53:17-29.