KNOWLEDGE ASSESSMENT: A MISSING LINK BETWEEN KNOWLEDGE DEVELOPMENT AND APPLICATION

H. Stephen Leff, Ph.D., Virginia Mulkern, Ph.D., Robert E. Drake, M.D., Ph.D., I. Elaine Allen, Ph.D., Clifton M. Chow, Ed.M.

Corresponding Author:

H. Stephen Leff, Ph.D.

Human Services Research Institute

2336 Massachusetts Avenue

Cambridge, MA 02140

Phone: 617-844-2507

E-mail: Sleff@hsri.org

H. Stephen Leff is Director of The Evaluation Center@HSRI and Assistant Professor in the Harvard Medical School Department of Psychiatry at The Cambridge Hospital. Virginia Mulkern is Associate Director of The Evaluation Center@HSRI. Robert E. Drake is Professor of Psychiatry at Dartmouth Medical School and Director of the New Hampshire-Dartmouth Psychiatric Research Center. I. Elaine Allen is Visiting Associate Professor of Statistics at Babson College. Clifton M. Chow is Program Manager for the Knowledge Assessment and Application Program in The Evaluation Center@HSRI.

Acknowledgements: Work on this paper was supported by a cooperative agreement from the Center for Mental Health Services, Substance Abuse and Mental Health Services Agency: SM53128-01-1. The paper profited from comments by colleagues in The Cambridge Hospital Center for Mental Health Services Research and Training, The Evidence-based Practices Project and Crystal Blyler, Betsy McDonel-Herr, and Neal Brown at the Center for Mental Health Services, Substance Abuse and Mental Health Services Agency.

KNOWLEDGE ASSESSMENT: A MISSING LINK BETWEEN KNOWLEDGE DEVELOPMENT AND APPLICATION

Knowledge Assessment:A missing link between knowledge development and application

Abstract

Health care systems throughout the U.S. are adopting evidence-based practices. This paper addresses issues involved in evaluating evidence-based practices for psychosocial rehabilitation (PSR) interventions. Because PSR interventions are complex, they address components of the intervention contexts that are generally deemed outside the bounds of traditional clinical interventions. To build a solid evidence base for PSR interventions, the authors consider a structured approach such as that suggested by the Federal Food and Drug Administration (FDA) and other organizations. The criteria for building and determining the evidentiary base for pharmaceuticals as developed by the FDA have been subjected to the widest review. Evaluating the criteria as stated by the FDA, the authors address a subset of those deemed most relevant to building and assessing the evidence base of PSR interventions. A set of questions are raised to evaluate how the FDA and other criteria can be applied to studies on PSR interventions to determine the evidence base for them and thus, make maximum use of the existing research literature, a process the authors term as *knowledge assessment*.

INTRODUCTION

Health care systems throughout the U.S. are adopting evidence-based practices because it is believed that such interventions should result in the safest, most potent, and efficient services([1]). This same logic applies to the psychosocial rehabilitation (PSR) interventions included in community support programs for persons with severe mental illness([2]). PSR interventions are complex ones aimed at enhancing role performance and quality of life as well as changing affects and cognitions. They address elements of the intervention context such as administrative policies, provider training, and consumer interactions with employers and landlords that would be considered outside the bounds of traditional clinical interventions. This paper addresses two major issues. First, what can we learn from more structured approaches to research staging so that the evidence base for PSR interventions can be built on a more firm foundation? Second, how can we make maximum use of the existing research literature on PSR interventions to better understand the evidence base of current PSR practices?

Overview of Approaches

Several noteworthy efforts have set explicit criteria for conducting studies to establish the evidence base for an intervention, for using completed studies to determine the degree to which an intervention is evidence-based, and for weighing the evidence about an intervention to decide whether it should be recommended for adoption. These include: The United States Food and Drug Administration Center for Drug Evaluation and Research (widely known as the FDA) Guidelines for the Format and Content of the Clinical and Statistical Sections of an Application

([3]), The International Conference on Harmonisation (ICH) Guidance on Statistical Principles for Clinical Trials([4]), The Criteria for Empirically Validated Treatments developed by a task force of Division 12 of the American Psychological Association (APA) ([5]), a report commissioned by Division 12 of the American Psychological Association published as A Guide to Treatments That Work([6]), and What Works for Whom, a review of psychotherapy efficacy prepared for the National Health Service in the United Kingdom by Roth and Fonagy([7]).

The FDA and ICH efforts address criteria for building and determining the evidentiary base for drugs and biologics. However, they provide the most detailed and exhaustive descriptions of criteria for assessing the degree to which interventions in general are evidence-based. Additionally, they have been in existence for the longest period and have been subjected to the widest review. For these reasons, this paper focuses on how the FDA and ICH criteria might be adapted for determining the evidentiary basis for PSR interventions. However, we also discuss other criteria when they add to or differ from the FDA and ICH criteria in instructive ways and draw upon related articles([8,9]).

This paper deals with a subset of the FDA and ICH criteria that seem most relevant to the task of building and assessing the evidence base of psychosocial interventions generally and PSR interventions specifically. The paper's emphasis is on how the FDA and ICH criteria can be applied to completed studies to retrospectively determine the evidence base for interventions. However, it will be apparent that the criteria can also be used to assess multi-site data and prospectively to plan and conduct research agendas and studies.

The FDA Approach

The FDA approach to determining the evidence base for an intervention assumes an ordered process, involving multiple studies, during which specific criteria are examined to assess the evidence resulting from the studies. We refer to this when applied to a body of completed studies as the *knowledge assessment process*. The process combines the logic of classic experimental science and a concern for the public's safety.

The FDA process requires three phases of testing and a final phase of post- marketing studies and surveillance([10]). The underlying logic of this phase model is that ideally each phase is completed before the next is begun, and that a phase is considered completed only when multiple studies have met specific criteria.

Phase 1 studies involve testing an intervention for feasibility and safety. These preliminary tests are usually conducted with normal or healthier volunteers and build the rationale for later studies. They address whether the intervention under study, or focal intervention, is feasible, safe, and tolerated by consumers and also are used to work out research and evaluation design issues. In some cases, these studies may explore possible intervention impacts. Phase 1 studies usually involve from 10-100 persons([10]).

Phase 1 studies of a psychosocial rehabilitation intervention, by analogy, focus on insuring readiness for efficacy trials. The tasks involve defining the intervention (typically with a manual and fidelity measure), making sure that the intervention fits into the relevant mental health settings and is acceptable to clients, and conducting small pilot tests (typically pre-post, or

mirror image, studies) to show that some clients are helped by the intervention and what the magnitude of the effect is likely to be.

Phase 2 studies address the efficacy of a specified (focal) intervention versus no intervention, placebo or active control under rigorously controlled conditions. The standard for efficacy studies is multiple randomized clinical trials. Phase 2 studies also include bioequivalence testing in which a focal drug is compared with competing medications to investigate not only whether they have different impacts, but also whether this difference is clinically meaningful. Phase 2 studies usually involve up to several hundred persons([10]). These highly controlled studies have also been referred to as efficacy studies. Emphasis has been placed on the importance of evidence from randomized clinical trials in mental health services research. However, in most cases, the equally important issue of the nature of comparison groups is overlooked.

The FDA and the ICH state that, scientifically, impact is most convincingly established by demonstrating superiority to placebo or by showing superiority to an *active* [our italics] control treatment([4]). According to the ICH, a suitable active control should be a widely used therapy whose superiority to placebo or placebo-equivalent in the relevant indication has been clearly established and quantified in well-designed and well-documented trials, and that can be reliably expected to exhibit similar superiority in the contemplated active control trial. From this perspective, the comparison to an intervention that is not an "active control" as defined above, whether or not randomization is employed, is not an acceptable source of evidence of impact. Such comparisons are ambiguous at best and misleading at worst. If a comparison intervention

performs worse than placebo, a focal intervention that outperforms it may perform only equal to placebo or worse than placebo. In the former case, if the focal intervention is more expensive or risky than placebo, its worth is questionable. In the latter case, in which the focal intervention may only be the lesser of two evils, the placebo is preferable.

The issue of placebo has been discussed in the psychotherapy research literature([11-13])in a manner that is useful for assessing PSR interventions. The psychotherapy literature conceptualizes placebo impacts as non-specific effects due to shared relational mechanisms present in placebo and active treatment situations that engender hope and positive self-regard in service recipients that result in positive clinical change. These mechanisms include the treator's positive regard for the recipient, acceptance of the recipient, and belief that the recipient can improve. From this perspective, non-placebo interventions must have ingredients in addition to these shared relational mechanisms to qualify as unique interventions. This provides a model for efficacy studies of PSR interventions.

Placebo comparisons in psychotherapy research have been questioned for reasons that might also apply to PSR interventions. It has been argued that behavioral interventions cannot really be discriminated from shared relational mechanisms, that as long as behavioral interventions have positive impacts it does not matter whether they work better than only shared relational mechanisms, that unlike drugs in comparison to pill placebo, behavioral interventions are not likely to have "noxious side effects" absent in shared relational mechanisms, and that the differential costs of behavioral interventions and shared relational mechanisms are not of interest([11,14]). If these arguments were once persuasive, they are far less so today. As mental

health interventions have become more standardized, as consumers have become more concerned about achieving recovery and avoiding negative treatment impacts, and as payers have become more concerned about the efficiency and costs of services, the arguments for placebo controls have become more compelling.

For psychosocial rehabilitation interventions, efficacy studies are typically conducted in university or research settings with clients who have severe mental illness but who also meet specific criteria designed to maximize the potential impact of the intervention. For example, a family intervention might be tested with clients who have a diagnosis of schizophrenia, who have recently been hospitalized, who are willing to take medications, and who live with families that are willing to participate regularly in a family intervention and a research project. The intervention might be provided by experienced Ph.D.-level clinicians, and researchers might carefully measure the outcomes of family interactions, client symptoms, and hospitalization. Clients and families in the control condition might receive a discussion group or an educational intervention without the focal family intervention. As this example illustrates, the efficacy study is designed to maximize the impacts of the intervention, but transfer and generalizability to routine settings, clients, clinicians, and other conditions of effectiveness are not addressed.

Phase 3 studies emphasize effectiveness under everyday conditions of practice, by including routine, or typical clients, clinicians, and programs in controlled trials. In addition, these studies often address a wider set of outcomes, some of which may be identified by clients and their families rather than professionals, and cost-effectiveness in relation to mental health system expenditures or sometimes total societal costs. In other words, the focus is on

generalizability, on whether or not the interventions are helpful in the real world of routine mental health programs, clients, and settings. The FDA and ICH do not make the distinction between efficacy and effectiveness; however, both recognize the need for effectiveness information representative of the possible influences of populations, geography, the practices of the particular provider(s), and clinics and so on([4]). Effectiveness studies can involve either comparison to placebo or other interventions.

Services investigators have increasingly emphasized the efficacy-effectiveness distinction because their studies have shown that the impacts of interventions implemented under efficacy conditions can differ from those of the same interventions implemented under effectiveness conditions([15]). These findings suggest it may be the special features of the efficacy conditions as well as the intervention services that are responsible for intervention impacts([9]). The analysis of why efficacy and effectiveness impacts differ can suggest that interventions include as specific manualized components aspects of interventions thought of originally as context.

The transfer from efficacy testing to effectiveness studies is challenging for psychosocial rehabilitation interventions. New psychosocial interventions may not be applicable to a large proportion of clients, may not be reimbursed in routine practice, may not be compatible with the training and preferred approaches of most clinicians, or may not address the outcomes in which clients are most interested. All of these concerns have been raised regarding assertive community treatment([16]) and psychoeducational family interventions([17]). For these reasons, we believe psychosocial rehabilitation interventions should be developed, refined, and tested directly in routine mental health programs using manuals and other materials that specify how they can

be implemented with mental health center staff, minimal client selection (other than self-selection), and outcome goals that have been consistently identified by clients as well as professionals as the standard([18]).

Phase 4. Post approval studies encompass post-marketing studies and surveillance after an intervention has FDA approval. Post approval studies are performed to determine the incidence of adverse reactions, investigate long-term impacts, study impacts in populations not previously studied, and to compare interventions with new products. Post- marketing surveillance data are collected on a voluntary basis from health professionals and manufacturers([10]).

The analogs for psychosocial rehabilitation interventions include studies that examine broader uses for proven interventions and mental health services research([19]). Given the complexities of organization and financing issues, researchers would like to know if interventions can have an impact at the system level and, given the limited budgets of most mental health systems, if they are cost-effective relative to alternative care.

THE KNOWLEDGE ASSESSMENT PROCESS

PSR research, like many other areas of mental health research, is "anarchic"([20])and does not follow the phases or criteria indicated by the FDA model. Studies may not have been randomized clinical trials, designs may have had other flaws, the number of replications addressing important questions may be small, some important issues, such as side effects, may not have been investigated, and inconsistent study findings are rarely reconciled([21]). The

remainder of this paper poses a set of questions to guide the way we assess the currently available evidence for PSR practices and plan new studies.

Knowledge Assessment and Meta-Analysis

The FDA and ICH emphasize that knowledge assessments should include all relevant studies([3,10]). The knowledge assessment process draws upon meta-analysis to integrate all of the evidence available, taking into account study methods and results. Meta-analysis, like the FDA model, has tended to assume intervention research and evaluation that follow the logic of orderly, classic experimental research. Our approach to knowledge assessment stretches meta-analysis in two ways. First, it envisions relying on data from non-experimental and quasi-experimental studies because so few randomized clinical studies of PSR interventions are available. Second, it envisions the possibility of reorganizing the knowledge available and comparing groups (intervention arms) with other groups that were not included in the same studies to synthetically create the comparisons called for by the FDA process, once again, because of the paucity of studies. This latter approach allows us to use the information from well-designed pre-post studies ([22,23])and it allows us to combine data from intervention arms representing similar intervention types, but in different studies, that cannot be combined in study level analyses because comparison interventions differed. Synthetic analyses of PSR interventions are problematic because they are complex and must take into account organizational and community variables, but such analyses are plausible when their findings are consistent within study comparisons([24]).

Of course, the problem in knowledge assessment and meta-analysis is that persons served, intervention characteristics, study contexts, and study methods will vary from study to study and among intervention arms making it difficult to tease out the effects of interventions. In knowledge assessment, as in meta-analysis, we attempt to adjust for this by coding, and statistically controlling for differences in subjects, the characteristics of interventions and intervention contexts, and study designs([25]). Given the information that is routinely available from studies, this presents a challenge. The extent to which it currently will be possible to implement knowledge assessments that satisfactorily code for threats to internal validity for given interventions and dependent variables can only be resolved empirically([26]).

Knowledge Assessment Questions

Knowledge assessment rates study results based on study and intervention arm characteristics. Interventions supported by results based on studies and intervention arms that have desirable characteristics are considered more evidence-based. Below, we discuss fifteen knowledge assessment questions addressing key study characteristics suggested primarily by the FDA and the ICH. To the degree studies with these characteristics showed interventions had positive outcomes, the interventions would be considered evidence-supported. Each of these questions implies the need for coding categories and some method for weighting study findings according to study characteristics. Specifying these is beyond the scope of this paper.

1 Are there sufficient numbers of studies with the required data, within each phase, to support a knowledge assessment?

As described above, the FDA holds that interventions should be tested by ordered programs of multiple studies. A knowledge assessment should begin by organizing intervention studies or study components into groupings suggested by the FDA and the ICH. Although PSR interventions differ from drugs in involving more contextual factors, the logic of study phases is the same.These groupings would reflect the following breakdown of studies:

Phase 1: Studies assessing safety and feasibility. Feasibility includes studying whether the intervention is adequately and clearly specified.

Phase 2: Studies assessing the impact of a focal intervention versus no intervention, placebo or active control under rigorously controlled conditions

Studies assessing the impact of a focal intervention versus alternative interventions under rigorously controlled conditions

Phase 3: Studies assessing the impact of a focal intervention versus no intervention, placebo or active control under conditions of routine practice

Studies assessing the impact of a focal intervention versus alternative interventions under conditions of routine practice, including mental health services research using primarily observational methods and administrative data([19]).

The first step after identifying reports is to determine which impact variables have data that can be integrated to assess performance in one or more of the FDA-like groupings. Useful data consist of data describing or data for computing pre- and post-intervention proportions or

means and measures of variation for each separate study group and data about commonly used demographic and clinical characteristics for each separate study group.

Addressing the questions posed by the FDA-like phases requires multiple studies or study components within each phase([3]). The FDA and the American Psychological Association Division 12 materials also specify that studies should be carried out by independent investigators([3,5]). They note that this is consistent with the general scientific demand for replicability. However, there is no clear-cut answer to the question of how many studies or intervention arms with what numbers of service recipients are necessary within FDA-like groupings. The more studies or intervention arms available within groupings and the larger the samples, the greater the statistical power([27, 28]). A systematic knowledge assessment process also reduces the risk of the "file drawer problem" or the failure to include studies with negative results([29]). Multiple studies or study arms can also increase external validity since they are more likely to represent a diversity of investigators, service recipients, and contexts and may permit more subgroup analyses. Nevertheless, more guidance is needed on this issue. At one extreme, the Task Force standard requires only two studies by independent investigators showing positive impacts to label an intervention as an evidence-supported treatment. It makes no recommendations about sample sizes or the handling of conflicting results. At the other extreme, Bangart-Drowns([26])considers fewer than 10 studies prohibitively small for a meta-analysis and Takkouche et al.([30]) consider 20 studies to be only a small number.

2 Are interventions manualized and were manuals used in the existing studies?

The FDA requires that chemistry, pertinent properties, manufacturing and control should be described for any drugs studied([3])and that evidence be related to the dosage and dose interval recommended([3]).

For PSR interventions, the analog to this criterion is that interventions should be described in an operationally defined manner with sufficient specificity to make discrimination from other interventions, replication, and fidelity measurement possible. The most common way of specifying a PSR intervention in this way is to provide manuals([31]) although it may be that training protocols may also provide some specificity. The existence of a manual is one of the American Psychological Association Division 12 criteria for an intervention to be labeled an empirically supported psychological intervention([5]). Knowledge assessments should code whether manuals were employed for focal and comparison interventions.

3 Did studies measure fidelity to determine the degree to which the intervention was implemented as intended?

The FDA and the ICH state that drugs under investigation and placebos should be studied to show that they are manufactured to chemical standards, and administered in planned doses and at planned intervals([3]). The equivalent to this for PSR interventions would be fidelity measurement([32]). Fidelity refers to the degree to which interventions were implemented as specified by experts, in manuals, or in other materials. Knowledge assessments should code whether studies employed fidelity measures for focal and comparison interventions and the degree of fidelity achieved.

4 Did studies present data on collateral interventions?

Both the FDA and the ICH discuss the need for studies to present information on collateral interventions delivered in addition to focal interventions. Collateral interventions are common for persons receiving PSR interventions and can occur in both efficacy and effectiveness studies. Their impacts should be carefully considered and controlled in both types of studies. Knowledge assessments should include what collateral interventions were present along with focal interventions and present analyses of the impacts of the focal intervention adjusted for the presence of collateral interventions.

5 Were clinical and rehabilitation, as opposed to surrogate, outcomes assessed?

Studies sometimes address what the ICH refers to as surrogate variables for clinical and rehabilitation outcomes([4]). The ICH notes that there have been many instances when treatments showing a highly positive effect on a proposed surrogate have ultimately been shown detrimental to clinical outcomes([4])and vice-versa. The ICH recommends that when surrogate variables are used, their relationship to the clinical variables of concern should be known or studied. Reduced hospitalization, a measure often used in PSR studies, would be an example of a surrogate variable, if it were assumed that this measure was a proxy measure of symptom reduction. Knowledge assessments should code whether impacts are clinical, rehabilitation or surrogate variables and for the latter whether relationships to clinical and rehabilitation outcomes have been demonstrated. Findings for clinical, rehabilitation and surrogate variables should be combined only if a relationship has been shown. Changes in service use and cost, while

legitimate targets for PSR interventions, should not be confused with clinical and rehabilitation outcomes.

6 Did studies include adverse events as key impact measures?

The FDA and the ICH repeatedly stress the necessity of monitoring in all study phases for negative impacts/adverse events that may affect contraindications, warnings, precautions, and adverse reactions([3]). It may not, indeed often will not, be possible to decide whether a particular serious event is intervention-induced, but such events should be noted for future review and consideration as an intervention is disseminated([3]). If PSR interventions can have positive impacts, it is logical that they might also have negative impacts([33]). Knowledge assessments should code whether PSR studies monitor and explicitly report negative impacts/adverse events and summarize evidence for these events.

7 Did studies define changes in key impact variables that have clinical, rehabilitation, or policy significance and report them in such a way that they can be aggregated?

Not all statistically significant differences have clinical or policy significance. As various technologies make larger studies possible, the probability of over-powered studies, i.e., ones finding statistically significant, but otherwise trivial differences, becomes larger([34]). The FDA and the ICH recognize this([4]). Several procedures could be employed to define clinical significance: Choosing variables that are widely accepted to be important outcomes (e.g. mortality); using statistical methods, including bioequivalence, for operationally defining clinical meaningfulness([35, 34]); and, measuring the utilities (e.g. willingness to pay) associated with

outcomes([36]). PSR studies have not established conventions for clinical or policy significance. Although bioequivalence methods have not been used in psychosocial and PSR research and evaluation, recently there have been calls for their use([14,34,37]). Knowledge assessments should code whether impacts have been assessed for clinical or policy significance and focus on impacts that have such significance.

8 Did studies measure long-term impacts?

The FDA([3])notes that drugs for chronic use are not usually studied for the full intended period of use, but are generally studied for periods of 6 months to one year. After the study periods, the FDA apparently relies on post-market surveillance. The FDA approach seems inadequate for PSR interventions that take longer than 6 months to one year to have positive impacts. Knowledge assessments should synthesize impacts for different follow-up periods and short-term and long-term impacts should be discussed separately.

9 Were key impact variables identified in a priori hypotheses?

The importance of theory-driven evaluation and testing a priori hypotheses both within and across studies has been widely discussed in the scientific literature. The ICH states that studies without a priori hypotheses do not constitute formal proof of efficacy([4])and that unplanned analyses should be used sparingly([4]). Extant PSR studies may not distinguish between planned and unplanned analyses. Knowledge assessments should code whether findings followed from planned or unplanned analyses and should themselves distinguish between planned and ad hoc analyses.

10 Did studies use reliable and validated instruments to measure primary impacts?

The ICH recommends that studies use reliable and validated instruments and measures on which experience has been gained in earlier studies or the published literature([4]). A recent study found systematic differences for syntheses of instruments with and without known reliability and validity([38]). If measures are newly developed, or if only portions of previously tested instruments are used, reliability and validity should be tested and reported. A number of instruments well suited to PSR studies have been developed([39,40]), but PSR studies vary in instruments and measures used. Knowledge assessments should code whether measures meet these criteria and focus on ones that do.

11 Did studies provide data on recipient characteristics by intervention arm?

When combining different studies or intervention arms from different studies, information about recipient differences should be used to examine and statistically control for group differences in recipient characteristics([41]). This is particularly important in PSR studies conducted at different sites that can involve sociodemographically and clinically diverse consumers. Knowledge assessments should code and statistically control for differences in the demographic and clinical characteristics of persons from different studies.

12 Did studies employ blinded designs?

Questions about blinds are routinely asked in scales for rating study quality. We focus on this question in this paper because, unlike randomization, little attention is being given to employing blind designs in PSR studies. In most cases it is impossible to keep providers and

consumers ignorant of a PSR intervention. However, it may be possible to keep interviewers and/or data analysts blind until data are collected and analyzed. Knowledge assessments should code whether impacts were associated with blinded designs.

13 Did studies adhere to their protocols?

Changes in designs are made for a number of reasons, including "pipeline" problems and breakdowns in randomization procedures. Skinner([42])has chronicled the types of "accidents" and problems that change studies and notes that these things tend not to be chronicled in study reports. The FDA and the ICH take these incidents seriously and require that they be carefully evaluated, sometimes by independent groups (Data Safety Monitoring Boards), and the implications explicitly discussed. PSR studies encounter various circumstances, such as pipeline problems, requiring protocol changes. Knowledge assessments should code whether studies had protocol violations. Violations should be rated in terms of the possibility that they biased study results.

14 Did studies provide impact data for all persons assigned to interventions?

The intent-to-treat principle implies that the primary study results include all randomized persons or persons otherwise assigned to conditions([4]). The use of all persons assigned to conditions may provide estimates of intervention effects that are more likely to mirror those observed in subsequent practice([4]). Even if data are analyzed for a reduced subset of persons, an additional intent to treat analysis should use all persons assigned([3]). PSR studies often focus on populations that are difficult to retain in services and studies, raising issues related to drop-outs

and missing data. Knowledge assessments should code whether complete or incomplete samples were used in analyses and any differences between results for intent-to-treat and per protocol samples should be discussed and interpretated([4]). Knowledge assessments also should record reasons for all post-assignment discontinuations and describe and evaluate strategies to compensate for missing data([3,4]).

15 Did studies focus on limited numbers of impacts to avoid the inflation of type 1 error (Alpha Inflation)?

Studies that have multiple, related impact measures run the risk of producing significant findings by chance([4,43,26]). Studies can avoid this problem by: focusing on a small number of key impacts chosen because of their grounding in theory or psychometric properties, focusing only on the most representative impacts of variables judged to be in the "same family" based on their content or correlation, combining variables into indicators, and adjusting the alpha level for the number of related tests conducted([43]). PSR studies understandably focus on multiple outcomes, given the complexity of the interventions. Nevertheless, often some of these are in the same family, increasing the probability of chance findings unless an a priori strategy is used to reduce the number of analyses conducted. Knowledge assessments should both record whether studies used these approaches and consider whether they would strengthen the knowledge assessment.

16 Did studies provide information on intervention costs?

Although costs are not used to determine the evidence base for interventions under FDA, ICH, or Task Force criteria, cost is considered in deciding whether to recommend or adopt

interventions. Given resource constraints and payment mechanisms like managed care, PSR studies must address cost issues. Therefore, knowledge assessment findings should be supplemented with cost information.

SUMMING UP A KNOWLEDGE ASSESSMENT: THE KNOWLEDGE ASSESSMENT PROFILE

The culmination of a knowledge assessment is a knowledge assessment profile, summarizing the answers to the above questions across studies or intervention arms for specific dependent variables and time periods. Ultimately, it should be possible to summarize knowledge assessments by giving interventions numeric or letter grades. Several approaches to doing this using from three([44])to 10 categories([45-47])have been suggested for medical and for mental health interventions, although there is no grading system specifically for PSR services.

A recent systematic review of employment programs([38]) enables us to anticipate what knowledge assessments of PSR programs might show. This review found 11 randomized clinical trials comparing employment programs for persons with severe mental illness with community alternatives. No information is provided on whether interventions were manualized. None of the eleven studies was classified as comparing an employment intervention to a placebo-like control, as defined above. One was classified as comparing supported employment to standard community care, 5 as comparing prevocational training to standard community care, and 5 as comparing supported employment to prevocational care. Competitive employment rates organized by follow-up periods ranging from 4 to 24 months showed supported employment

outperformed prevocational training and standard community care in most periods. Addressing the knowledge assessment questions cited above using the information in this review suggests the need for Phase 2 studies comparing manualized supported employment to placebo to investigate whether supported employment outperforms non-specific interventions. The review contains limited information on key clinical, rehabilitation, and policy related impacts such as job tenure, program costs, and adverse events, suggesting the need for studies of these variables or reviews that have a wider focus. The review also suggests the need for Phase 3 studies involving larger numbers of more diverse consumers (Ns for intervention arms ranged from 12 to 76), providers, and settings. The need for additional evidence, suggests that it might be useful to assess quasi-experimental studies excluded from the review. Finally, for decision makers desiring to implement explicitly vocational programs before additional studies are available, the review suggests supported employment appears more promising than prevocational training.

DISCUSSION AND CONCLUSIONS

Determining the evidence base for PSR interventions leaves us with a number of methodological issues including: how to think about placebo controls for PSR services, determining the necessary number of studies or intervention arms for a knowledge assessment, translating the complex features of PSR interventions and contexts into measurable variables, moving beyond efficacy to effectiveness studies, and encapsulating knowledge assessment information. It also poses policy issues. There is no federal agency, like the FDA, with a legislative mandate and equivalent funding to decide if PSR interventions are ready for

widespread adoption. Nor are there organizations, like drug companies, with financial incentives and their own funds to implement knowledge assessments. The Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, the Agency for Health Care Research and Quality (which currently supports evidence-based practice centers), and the National Institute of Mental Health are organizations that might assume this responsibility since they are independent, have some staying-power, have all engaged in similar kinds of activities, and have at least some resources they could devote to knowledge assessments. Academic and voluntary organizations that might conduct knowledge assessments for PSR interventions include: the evidence-based practice centers created by the Agency for Health Care Research and Quality, the various Centers funded by the Substance Abuse and Mental Health Services Agency and the National Institute of Health, the Campbell and the Cochrane Collaborations (voluntary organizations of social scientists committed to meta-analysis located in the United States and England respectively), and the Centre for Evidence Based Medicine at Oxford University.

One next step in developing knowledge assessment is for implementers and users of knowledge syntheses to critique and improve the framework presented here, considering, in particular, the evidence base for the knowledge assessment criteria([38]). A second is for agencies and organizations that fund evaluations to use the principles of knowledge assessment to shape requests for proposals, review applications, monitor study implementation, and review study reports. A third is to translate the principles of knowledge assessment into guidance on how studies are reported in journals([48]) and to consumers. The FDA works with drug manufacturers to develop simple explanations for using medications correctly to facilitate consumer choice and

participation in treatment([10]). A fourth step is for synthesists to conduct knowledge assessments for PSR interventions to document and evaluate the current evidence base for PSR interventions. A fifth will be to use knowledge assessments to identify and design new studies of PSR interventions.

A rigorous knowledge assessment of PSR interventions may show that there are few, if any, for which there is a substantial evidence base. If this turns out to be the case, it may be necessary to accept the use of less proven interventions until more evidence exists. As Levin and Roberts([49])note "if many studies ranging over a long period of time are needed to build a good scientific account of some social phenomenon, the social practitioner will probably not have time to wait." However, this should not be interpreted as lessening the requirements for evidence, since this might result in the requisite studies never being conducted.

References

1. Millenson ML: Demanding Medical Excellence: Doctors and Accountability in the Information Age. Chicago: University of Chicago Press, 1997
2. Rotter J: On the Evaluation of Methods of Intervening in Other People's Lives. The Clinical Psychologist 24: 1-2, 1971
3. Food and Drug Administration/Center for Drug Evaluation Research: Guideline for the Format and Content of the Clinical and Statistical Sections of an Application, 1988
4. International Conference on Harmonisation: Guidance on Statistical Principles for Clinical Trials. Federal Register 63: 49583-49598, 1998
5. Chamblesss D, Ollendick TH: Empirically Supported Psychological Interventions: Controversies and Evidence. Annual Review of Psychology. 2001
6. Nathan P, Gorman JM: A Guide to Treatment that Work. New York: Oxford University Press, 1998. Psychiatric Services 50: 1365-1366, 1999
7. Roth A, Fonagy P: What Works for Whom?: A Critical Review of Psychotherapy Research. New York: The Guilford Press, 1996
8. Burns BJ: A Call for a Mental Health Services Research Agenda for Youth with Serious Emotional Disturbance. Mental Health Services Research 1: 5-20, 1999
9. Weisz J, Jensen PS: Efficacy and Effectiveness of Child and Adolescent Psychotherapy and Pharmacotherapy. Mental Health Services Research 1: 125-157, 1999

10. Food and Drug Administration/Center for Drug Evaluation and Research: The CDER Handbook. Washington, D.C.:U.S. Department of Health and Human Services, 1998

11. Parloff M: Placebo Controls in Psychotherapy Research: A sine qua non or a placebo for research problems? Journal of Consulting and Clinical Psychology 54: 79-87, 1986

12. Healy D: The Anti-Depressant Era. Cambridge, MA: Harvard University Press, 1999

13. Shapiro AK, Morris LA: Placebo Effects in Medical and Psychological Therapies. In: Garfield SL, Bergin Ae, eds. Handbook of Psychotherapy and Behavior Change: An Empirical Analysis Second Edition ed. New York: John Wiley and Sons, 1978

14. Chambless D, Hollon SD: Defining Empirically Supported Therapies. Journal of Consulting and Clinical Psychology 66: 7-18, 1998

15. Clarke GN: Improving the Transition from Basic Efficacy Research to Effectiveness Studies: Methodological Issues and Procedures. Journal of Consulting and Clinical Psychology 63: 718-725, 1995

16. Phillips S, Burns B, Edgar E, Mueser K, et al: Assertive Community Treatment: Moving an evidence-based intervention into standard practice. Psychiatric Services, in press

17. Dixon L, McFarlane W, Lefley H, et al: The State of evidence-based practices for services to family members of people with psychiatric disabilities. Psychiatric Services, in press

18. Bond G, Becker D, Drake R, et al: Implementing Supported Employment as an evidence-based practice. Psychiatric Services 52:313-322, 2001

19. Donenberg GR, Lyons JS, Howard KI: Clinical Trials Versus Mental health Services

Research: Contributions and Connections. Journal of Clinical Psychology 55 : 1135-1146, 1999

20. Hunt M: How Science Takes Stock: The Story of Meta-Analysis. New York: Russell Sage Foundation, 1997

21. Linden W, Wen FK: Therapy Outcome Research, Health Care Policy, and the Continuing Lack of Accumulated Knowledge. Professional Psychology: Research and Practice 21: 482-488, 1990
22. Lipsey MW, Wilson DB: Practical Meta-Analysis. Thousand Oaks, CA: Sage Publications, 2001
23. Becker B: Synthesizing Standardized Mean-Change Measures. British Journal of Mathematical and Statistical Psychology 41: 257-278, 1988
24. Banks S: Personal Communication, 2001
25. Lohr K, Carey TS: Assessing "Best Evidence": Issues in Grading the Quality of Studies for Systematic Reviews. Journal of Quality Improvement 25, 1999
26. Bangert-Drowns RL: Review of Developments in Meta-Analytic Method. Psychological Bulletin 99: 388-399, 1986
27. Cooper H, Hedges L: The Handbook of Research Synthesis. New York: Russell Sage Foundation, 1994
28. Cappelleri J, Ioannidis J, Schmid C, et al: Large Trials vs. Meta-Analysis of Small Trials. Journal of the American Medical Association 276:1332-1338, 1996

29. Rosnow R, Rosenthal R: Focused Tests of Significance and Effect Size Estimation in Counseling Psychology. Journal of Counseling Psychology 35:202-208, 1988

30. Takkouche B, Cadarso-Suarez C, Spiegelman D: An Evaluation of Old and New Tests of Heterogeneity in Epidemiologic Meta-Analysis. American Journal of Epidemiology 150: 206-215, 1999

31. Torrey WC, Drake RE, Dixon L, et al: Implementing Evidence-Based Practices for Persons with Severe Mental Illnesses. Psychiatric Services 52: 45-50, 2001

32. Bond G, Williams J, Evans L, et al: Psychiatric Rehabilitation Fidelity Toolkit. Cambridge, MA:The Evaluation Center at HSRI, 2000

33. Beutler L: David and Goliath: When Empirical and Clinical Standards of Practice Meet. American Psychologist 55: 997-1007, 2000

34. Rogers J, Howard K, Vessey J: Using Significance Tests to Evaluate Equivalence Between Experimental Groups. Psychological Bulletin 113: 553-565, 1993

35. Jacobson N, Truax P: Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research. Journal of Consulting and Clinical Psychology 59: 12-19, 1991

36. Tsevat J, Weeks J, Guadagnoli E, et al: Using Health-Related Quality-of-Life Information: Clinical Encounters, Clinical Trials, and Health Policy. Journal of General Internal Medicine 9: 576-582, 1994

37. Stegner BL, Bostrom AG, Greenfield TK: Equivalence Testing for use in Psychosocial and

Services Research: An Introduction with Examples. Evaluation and Program Planning 19: 193-198, 1996

38. Crowther RE, Marshall M, Bond GR, et al: Helping People with Severe Mental Illness to Obtain Work: Systematic Review. British Medical Journal 322:204-208, 2001
39. Lehman A, Postrado L, Rachuba L: Convergent Validation of Quality of Life Assessments for Persons with Severe Mental Illnesses. Quality of Life Research 2:327-333, 1993
40. Tessler R, Fisher G,Gamache G: A Role Strain Approach to the Measurement of Family Burden: the Properties and Utilities of a New Scale. Paper presented at Rutgers University, Institute for Health, Health Policy, and Aging Research, 1988
41. Lyons JS, Howard KI: Main Effects Analysis in Clinical Research: Statistical Guidelines for Disaggregating Treatment Groups. Journal of Consulting and Clinical Psychology 59: 745-748, 1991
42. Skinner B: A Case History in Scientific Method. American Psychologist 11:221-233, 1956
43. Dar R, Serlin R, Omer H: Misuse of Statistical Tests in Three Decades of Psychotherapy Research. Journal of Consulting and Clinical Psychology 62:75-82, 1994
44. Agency for Health Care Policy and Research Depression Guideline Panel: Depression in Primary Care: Volume 1. Detection and Diagnosis. Washington, DC:US Department of Health and Human Service, 1993
45. Centre for Evidence-based Medicine: Levels of Evidence. Oxford University, UK, 2000
46. Battista R, Fletcher S: Making Recommendations on Preventive Practices: Methodological Issues. In: Battista R, Lawrence R, eds. Implementing Preventive Services:

Supplement to the American Journal of Preventive Medicine 4(4). New York: Oxford University Press: 53-67, 1988

47. Spitzer W: Report of the Task Force on the Periodic Health Examination. Canadian Medical Association Journal 121:1193-1254, 1979

48. Begg C, Cho M, Eastwood S, et al: Improving the Quality of Reporting of Randomized Controlled Trials. Journal of the American Medical Association 276:647-649, 1996

49. Levin G, Roberts EB: The Dynamics of Human Service Delivery. Cambridge, MA: Ballinger Publishing, 1976