Treatment Effectiveness Score as an Outcome Measure in Clinical Trials

Walter Ling, Steven Shoptaw, Donald Wesson, Richard A. Rawson, Margaret Compton, and C. James Klett

A variety of measures are used for evaluating patients' responses to substance abuse treatments. These range from physical measures (such as samples of urine, breath, hair, or blood), self-reports of drug use (such as the Addiction Severity Index (ASI) or the Time Line Follow-Back), self-reports of psychological or physiological functioning (such as symptom checklists or craving or mood ratings), and collateral reports. Physical indices of recent drug use, such as urine toxicology screens, are preferable to self-report or collateral reports for evaluating patients' responses to drug abuse treatments because of their objectivity. In order to optimize the likelihood of both detecting individual episodes of problem drug use and correctly inferring drug abstinence based on urine toxicology results, guidelines have been suggested for collection procedures and timing for collection of urine specimens (Blaine et al. 1994; Cone and Dickerson 1992; Jain 1992). However, the difficult task of aggregating urine toxicology results remains, whether when interpreting the response of a single patient to a specific treatment or when evaluating a treatment's effectiveness based on a group of patients' responses in a clinical trial. Difficulties in aggregating urine toxicology results include, but certainly would not be limited to, such problems as the frequency and sensitivity of toxicology screens, early termination of some patients from treatment (or, conversely, the continued participation of some patients who respond poorly to treatment), and problems of analyzing a data matrix that contains a large number of missing datapoints. This chapter reviews the objective indices of treatment response that have traditionally been used and suggests three composite methods for evaluating these data: the Treatment Effectiveness Score (TES), the Joint Probability score (JP), and the Clinical Stabilization Score (CSS).

TRADITIONAL OBJECTIVE METHODS FOR MEASURING TREATMENT EFFECTIVENESS

Many of the traditional objective measures of treatment effectiveness have been characterized as imperfect indices (Ling et al. 1976) due to

their inability to accurately and completely describe the various aspects of treatment response. For example, retention in treatment is one commonly used method for measuring clinical response. However, reliance on retention as a sole indicator of treatment efficacy can be misleading if concomitant use of illicit drugs is not taken into account. A patient who shows little to no alteration in drug use cannot be considered an unqualified therapeutic success regardless of how long he or she remains in treatment.

In addition to retention, clinical reports typically include some type of urine toxicology results to indicate treatment efficacy. Several approaches have been used for interpreting drug use or drug abstinence based on analyzed urine samples. These have included single-point urine test results (such as urine samples collected at posttreatment followup), percent of urine samples during the trial that are negative for drug metabolite, and percent of patients able to achieve a specific criterion (such as a varying number of consecutive weeks of samples negative for drug metabolite).

Most commonly used to document long-term followup status, single-point urine samples can detect recent drug use, but cannot indicate patterns of drug use throughout the followup interval. Further, patients who provide urine samples at followup are usually those who can be located, which further increases the threat to the internal validity of this treatment response indicator. Still, many trials of substance abuse treatments will include long-term followup urine results as a primary indicator of treatment efficacy.

Clinical trials of substance abuse treatments depend on urine toxicology results gathered during treatment to evaluate efficacy. Researchers have debated the merits of using qualitative versus quantitative values for interpreting urine toxicology results (Cone and Dickerson 1992). However, these data are most commonly reported as percent of samples negative for the metabolite of the problem drug. Simple percent-negative indices can provide some indication of patients' overall response, but do not characterize accurately those patients who terminate early despite all samples being negative for metabolite or those who remain in the trial, yet continue to use the problem drug. One alternative to a simple percent-negative index is the achievement of a specific criterion based on achieving some number of consecutive weeks of negative urinalysis (Carroll et al. 1991; Higgins et al. 1991) or percent of patients with continued abstinence. Criterion-linked indices can suffer from the problem of setting cutoff levels. That is, liberal cutoff levels can inflate the

actual clinical utility of a specific treatment, while conservative cutoffs can underrepresent treatment efficacy. There are at least two other problems associated with this approach. One is the loss of ability to discriminate among patients with various drug use patterns, e.g., using an 8-week criterion in a 16-week trial would count each of the following patients as a success: (1) a patient who consistently gives drug-free urines for 16 weeks; (2) a patient who uses drugs at the beginning of the trial but then "cleans up" and gives drug-free urines for the last 8 weeks; and (3) a patient who initially is fully compliant, gives consistently drug-free urines for 8 weeks, and then relapses to drug use or drops out of the program altogether. A second problem is that the use of such a criterion yields a noncontinuous (i.e., categorical) dependent variable that is not optimal from a statistical point of view.

All traditional approaches to interpreting urine toxicology results are vulnerable to the effects of missing data. At a minimum, missing datapoints are a nonrandom influence on the data matrix. Further, missing data likely indicate treatment inadequacy—patients typically do not attend a clinic regularly when treatment is ineffective. Missing data heavily influence single-point urine results since the reason for the missing data cannot be accurately represented in subsequent analyses. It is often unknown whether missing data are due to patients' resolution of their drug problem, to patients' continuing drug use, or to patients' refusal to participate. By contrast, percent-negative urinalysis methods can overrepresent patients' responses when patients discontinue treatment early but provide all samples negative for drug during the trial. Least affected by missing data are estimates of achievement of specific criteria, since patients with missing data usually fail to meet the specific criteria.

Traditional methods for interpreting objective clinical indices also commonly focus on one indicator to the exclusion of others. Synthesis of information that describes patients' treatment response (as measured by urine toxicology), treatment compliance (as measured by retention), and treatment toleration (as measured by lack of severe side effects/toxicity) allows for a more complete evaluation of various aspects of the efficacy of a given treatment. Dissatisfaction with the limitations of traditional methods for interpreting objective measures led the authors to experiment with new ways to compile and interpret these data to address specific concerns that have been encountered while conducting clinical trials.

NEW METHODS OF INTERPRETING OBJECTIVE MEASURES OF CLINICAL RESPONSE

The Treatment Effectiveness Score

One important concern for any trial is objective evidence of treatment efficacy. As an alternative to the methods reviewed earlier for interpreting retention and urine data, the authors developed the TES, which is a different approach to interpreting retention and urine toxicology results with conceptual advantages. Using the TES, "clean" urines rather than "dirty" urines can be counted. This simple shift emphasizes patient success rather than failure, but avoids the explicit imputation of a missing specimen as "dirty." That is, a patient either provides a "clean" urine as scheduled, or does not. "Clean" urines are counted for the full scheduled tenure of each patient in the trial.

For example, in a study of 17 weeks' duration requiring three urine samples each week, there would be 51 scheduled urine specimens. If each "clean" urine earns a point, a metric is established with a range of 0 to 51. The most successful therapeutic outcome is represented by a patient who attends the clinic reliably, completes the full duration of the trial, gives urine specimens as requested, and whose urine samples are consistently clean. Such a patient would obtain a score of 51. Patients may achieve scores of less than 51 in two ways: either by providing one or more urines positive for the drug of abuse being tested or by providing fewer than 51 specimens due to missed clinic visits or leaving the trial early. The TES provides a measure of relative standing in comparison to other patients in the trial. In the above example, each patient has the opportunity to earn 51 points by complying with the therapeutic expectations.

Conceived in this way, sample attrition is not a concern. Every patient who is randomized has a score and is included in the analysis. There are no dropouts in the usual sense and there is no assumption of whether or not patients who are no longer actively participating have returned to illicit drug use. Within a single clinical trial, there is no need to convert the score to a percentage because all patients have the same denominator, although doing so facilitates comparison across studies of different duration or different scheduling of urine collection. It is important to understand that the TES is not a pure measure of illicit drug use, nor is it a measure of retention, although it is heavily influenced by both. It is also influenced by other clinically important parameters such as adherence to clinic policy, drug craving,

and withdrawal symptoms, to the extent that these affect retention and drug use. Thus, the TES is intended as a composite score that reflects multiple aspects of therapeutic success.

The authors have applied the TES to data gathered as part of two large pharmacotherapy trials. In an opiate pharmacotherapy trial, the TES was compared with the more commonly used percent of urine samples negative for opiate metabolite (Ling et al., in press). Results of comparisons of patients' responses to different opiate medication treatment conditions using the two indices showed similar patterns when using the two measures. This similarity of results indicates that the TES can provide a valid alternative to percent-negative urine samples yet also captures retention. The advantage of the TES over the percent-negative urine samples is that this measure provides a clear indication of treatment response: averaged TES scores represent the expected value of negative urine samples for similar patients who receive an identical treatment.

The TES has also been applied to urine toxicology data generated from a cocaine pharmacotherapy trial and has been found to correlate significantly with traditional objective and subjective measures of treatment outcome (Ling et al. 1995). Specifically, the data showed the TES to exhibit significant positive associations with the percent of patients who achieved criteria of 3 and 8 consecutive weeks of urine samples negative for cocaine metabolite, with the average number of weeks of retention, and with the average number of counseling sessions attended by patients. Significant negative associations were found between the TES and the ASI drug scale and the Profile of Mood States (POMS) depression score (McNair et al. 1992).

These findings provide strong evidence for the validity of using the TES as an outcome indicator of clinical response. Application of the TES to data from these two large pharmacotherapy trials has indicated that the TES is a conceptually encompassing and succinct indicator of outcome. Implicit in its measurement, the TES provides an indication of another important factor: patients' acceptance of the treatment. Patients can reject treatment for a variety of reasons that range from resumption of drug use, to being incarcerated, to resolution of the drug problem. Assumptions of "automatic positive" for missing data when using traditional methods for interpreting urine toxicology results are avoided with the TES. By not imputing the cause of missing data, the TES simply interprets missing data as an indirect measure of patients' acceptance of the treatment.

While the TES is an improvement over unidimensional scores, it is not perfect. There will always be a few patients who are unable to complete the trial for reasons beyond their control and for reasons that have nothing to do with the treatment. However, the authors' current position is that early termination is either drug related or it is random. A relatively few random events could distort trial results, but this risk seems preferable to layers of assumptions that might have the same effect. It is obvious that information about drug use during the trial is lost whenever the vector of test results is collapsed into a single score. Patients with the same score can have different drug use profiles with quite different therapeutic or prognostic implications. The authors are interested in this and intend to explore other approaches.

The Joint Probability Score

Another limitation of traditional outcome measures involves the lack of a conceptually linked method for understanding the clinical relevance of the trial. A method for estimating a given patient's probability of successful outcome at a specific point in time would be useful to both clinicians and researchers. Most reports of clinical trials customarily present a retention curve and an illicit drug use curve as a means for summarizing objective treatment response data. Using these to estimate patients' responses can result in biased appraisals, since both of these indicators are vulnerable to nonrandom influences. Although broad statements about the value of a particular treatment can be inferred for a group of patients, such retention and drug use aggregate estimates cannot provide accurate information about the probability of treatment success over time.

One method for compiling retention and illicit drug use data that approaches the purpose of estimating patients' response is to plot the number of samples negative for illicit drugs during a given week divided by the number of scheduled urine tests for that week times the number of patients still active in the trial (Ling et al., in press). This technique intends to correct for patients who terminate participation early, though plots of such data are likely to demonstrate a gradual upward trend, which a casual reader might interpret as clinical improvement in patients. Such an association is likely to be spurious, since in most clinical trials the number of patients who terminate early increases over time, and attrition in this group is likely not due to a random process. Rather, dropouts are more likely to be those patients who have more severe levels of drug dependence and/or who

show poor response to the treatment than those who resolve their dependency or who have external forces that preclude continued participation. Thus, plots that illustrate the performance of the residual sample will likely show an upward trend since those remaining in the trial are those who tolerate the treatment and who may show positive treatment responses.

A correction to this problem is to multiply each point of the plot described above by the probability of retention to that point. In essence, the plot is converted to a JP curve. For example, in a trial requiring one sample per week, the point at week X would be p_1 (i.e., the number of patients still in the study at week X divided by number of patients who started treatment) times p_2 (i.e., the number of urines negative for illicit drugs at week X divided by the number of patients still in the study at week X). Since the numerator of p_1 and the denominator of p_2 cancel out, the curve can be constructed simply by dividing the number of negative urines obtained each week by the number of patients who started the study. This curve will tend to take a downward path unless the loss of patients over time is fully compensated for by better performance of the residual sample. As presented, the JP is a conservative measure of treatment efficacy in a clinical trial. Upward drift over time can be attributed to the effectiveness of the treatment program rather than to influences on the data of differential dropout of treatment nonresponders.

Validated using data in a large opiate pharmacotherapy trial (Ling et al., in press), the JP has yet to be applied to data from clinical trials of cocaine or other drug abusers. However, the logic underlying the JP index argues for its use in trials using these other drug-dependent patients. Knowing the retention rates, the number of samples negative for illicit drugs over the weeks of a trial, and the original number of patients, researchers and clinicians can easily calculate accurate probabilities that their patients will produce a negative urine sample at a given point when using a specific type of treatment. Plotting the JP index produces a curve that can also be useful in comparing outcomes from different studies of the same medication.

The Clinical Stabilization Score

The need for a composite index of treatment response, retention, and acceptance has been identified by the authors when conducting dose-ranging studies of new medications for substance abusers. In such trials, information that describes the safety and efficacy of a particular medication at a particular dose level is crucial, yet often

incomplete. Measurements of good therapeutic response to a medication should a priori indicate the elements that demonstrate that response. The CSS is an index developed by the authors to address this point.

The CSS is based on a set of criteria devised to study therapeutic responses to variable doses of medications in the treatment of drug dependency. As the name implies, the CSS is used to indicate that a specific dose of a specific medication has stabilized the patient's drug dependence problem. The criteria that comprise the CSS are based on a logic that incorporates clinically important elements of the patient's response to medication: reduction of illicit drug use, continued treatment compliance, lack of adverse symptoms, and absence of drug toxicity. CSS criteria are framed in a 2-week time period. The window of observation moves forward in real time as the patient remains in the trial. The clinical assessment consists of three elements:

1. *Urine toxicology.* Monitored urine samples are collected at a set rate over the course of the clinical trial. Samples are collected on Mondays, Wednesdays, and Fridays, with no substitutions allowed. Urine samples are immediately analyzed (within 24 hours) for the presence of metabolite of the problem drug. The sample must be free of this drug for the patient to earn a CSS point.

2. *Clinic attendance.* To earn a CSS point, patients must attend the clinic as scheduled on a Monday, Wednesday, or Friday. Patients who receive a CSS point comply with treatment. Conversely, patients unable to comply with treatment likely will not attend clinic and, hence, cannot earn a CSS point.

3. *Adverse signs and symptoms.* At each occasion for submitting urine samples, the patient must report that he or she is free of moderate to severe medication-related or withdrawal-related symptoms and adverse medical events to earn a CSS point. For a drug abuse medication to be clinically useful, it cannot induce symptoms or effects that produce moderate to high levels of discomfort in patients. Patients who report moderate to high levels of adverse signs and symptoms cannot earn a CSS point.

Using these criteria, the authors provided the opportunity for patients to earn CSS points three times per week, which corresponds with each occasion for providing a monitored urine specimen. Patients must achieve all three CSS criteria (come to the clinic, provide a drug-free

urine specimen, and be free of moderate to severe symptoms) to achieve one CSS point. Using the scheduled visits in the authors' research, patients can earn a possible six points over any 2-week period. Patients who earn five or six out of the six possible CSS points in a 2-week period, and who earn one of those points on the most recent assessment occasion, are considered to have stabilized on a therapeutic dose of medication. Study designs that use different numbers of assessment points per week will have correspondingly different ranges of possible CSS points. However, the rolling 2-week period for evaluating CSS scores should be retained.

The CSS is not conceptualized as an outcome evaluation tool for comparison among patients. Rather, it is a measure of how well a given dose of study medication is helping a particular patient reduce his or her problem drug use, without causing untoward symptoms and adverse effects. In a clinical pharmacotherapy trial, the CSS can be used in dose runup phases of studies or in studies that have variable medication levels to monitor patient safety and to trigger study medication dose changes. Unless a satisfactory CSS is achieved (e.g., a CSS of five or six out of the six possible points), the dose of study medication is increased by one increment at each weekly review. If the occurrence of adverse symptoms reduces the CSS, the medication is not increased or may be decreased by one increment. If a satisfactory CSS is achieved, the dose remains unchanged.

It is conceivable that some patients could show positive response to a study medication such that good therapeutic response can be maintained with less frequent clinic attendance than the three times per week required by the authors' studies. Further, good therapeutic response may be affected by a medication, though some patients may find it inconvenient to attend the clinic on scheduled days. The CSS would be unable to discriminate between such instances and poor response to medication. Another problem is that the CSS suffers from all indices that use a cutoff for classifying response outcomes. For some patients, four of six scheduled urine samples being negative for illicit drugs over a 2-week period could be classified as a treatment "success." At this point, the authors are planning to evaluate the sensitivity and specificity of various cutoff levels using the above criteria for the CSS. Finally, the CSS was conceived as an index to address needs specific for a certain type of pharmacotherapy trial and has been used by the authors for this purpose.

CONCLUSIONS

It is agreed that objective methods for assessing patients' responses to clinical trials offer the best indication of treatment efficacy. However, the authors maintain that traditional methods of interpreting such data are imperfect. Development of alternative methods for interpreting objective data should be driven by researchers' needs to understand various aspects of treatment response during the trial. The three indices suggested in this chapter are intended to provide empirically derived integration of retention and urine toxicology measures to indicate treatment outcome (TES), probable treatment response (JP), and good therapeutic response (CSS). Although these indices are still in the development and evaluation phase, they offer clear advantages to traditional methods for assessing patients' responses in clinical trials.

REFERENCES

Blaine, J.D.; Ling, W.; Kosten, T.R.; O'Brien, C.P.; and Chiarello, R.J. Establishing the efficacy and safety of medications for the treatment of drug dependence and abuse: Methodologic issues. In: Prien, R.F., and Robinson, D.S., eds. *Clinical Evaluation of Psychotropic Drugs: Principles and Guidelines*. New York: Raven Press, 1994.
pp. 593-623.

Carroll, K.M.; Rounsaville, B.J.; and Gawin, F.H. A comparative trial of psychotherapies for ambulatory cocaine abusers: Relapse prevention and interpersonal psychotherapy. *Am J Drug Alcohol Abuse* 17(3):229-247, 1991.

Cone, E.J., and Dickerson, S.L. Efficacy of urinalysis in monitoring heroin and cocaine abuse patterns: Implications in clinical trials for treatment of drug dependence. In Jain, R.B., ed. *Statistical Issues in Clinical Trials for Treatment of Opiate Dependence.* National Institute on Drug Abuse Research Monograph 128. DHHS Pub. No. (ADM)92-1947. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1992. pp. 46-58.

Higgins, S.T.; Delaney, D.D.; Budney, A.J.; Bickel, W.K.; Hughes, J.F.; Foerg, F.; and Fenwick, J.W. A behavioral approach to achieving initial cocaine abstinence. *Am J Psychiatry* 9:1218-1224, 1991.

Jain, R.B. Analysis of clinical trials for treatment of opiate dependence: What are the possibilities? In: Jain, R.B., ed. *Statistical Issues in Clinical Trials for Treatment of Opiate*

Dependence. National Institute on Drug Abuse Research Monograph 128. DHHS Pub. No. (ADM)92-1947. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1992. pp. 116-134.

Ling, W.; Charuvastra, V.C.; Kaim, S.C.; and Klett, C.J. Methadyl acetate and methadone as maintenance treatment for heroin addicts. *Arch Gen Psychiatry* 33:709-720, 1976.

Ling, W.; Shoptaw, S.; Rawson, R.A.; and Klett, C.J. A composite score for evaluating treatment response in cocaine pharmacotherapy trials (TES). In: Harris, L.S., ed. *Problems of Drug Dependence, 1994: Proceedings of the 56th Annual Scientific Meeting of the College on Problems of Drug Dependence, Inc.* Vol. II. National Institute on Drug Abuse Research Monograph 153. NIH Pub. No. (ADM)95-3883. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1995. p. 495.

Ling, W.; Wesson, D.R.; Charuvastra, C.; and Klett, C.J. A controlled trial comparing buprenorphine and methadone maintenance. In press.

McNair, D.; Lorr, M.; and Droppleman, L.F. *POMS Manual: Profile of Mood States.* San Diego, CA: EdITS/Educational & Industrial Testing Service, 1992.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Los Angeles Addiction Treatment Research Center; Matrix Center, Inc.; Friends Medical Science Research Center, Inc.; National Institute on Drug Abuse, Medications Development Division; and the West Los Angeles Veterans Administration Medical Center.

AUTHORS

Walter Ling, M.D.
Associate Chief of Psychiatry for Substance Abuse
West Los Angeles Veterans Administration
Medical Center
Research Services (B151S)
Building 257, Room 7
11301 Wilshire Boulevard
Los Angeles, CA 90073

Steven Shoptaw, Ph.D.
Research Director
Los Angeles Addiction Treatment
Research Center (LAATRC)
Suite 340
10350 Santa Monica Boulevard
Los Angeles, CA 90025

Donald Wesson, M.D.
Medical Director and Scientific Director
Merritt-Peralta Institute Treatment Services
435 Hawthorne Avenue
Oakland, CA 94609

Richard A. Rawson, Ph.D.
Executive Director
Matrix Center and
Matrix Institute on Addictions
Suite 330
10350 Santa Monica Boulevard
Los Angeles, CA 90025

Margaret Compton, Ph.D.
Researcher
Los Angeles Addiction Treatment
Research Center (LAATRC)
Suite 340
10350 Santa Monica Boulevard
Los Angeles, CA 90025

C. James Klett, Ph.D.
Co-Director
Los Angeles Addiction
Treatment Research Center (LAATRC)
Suite 340
10350 Santa Monica Boulevard
Los Angeles, CA 90025

Click here to go to page 221