Journal of Consulting and Clinical Psychology
1992, Vol. 60, No. 2, 225-239

0022-006X/92/$3.00

METHODOLOGICAL DEVELOPMENTS

Evaluating the Effects of Drugs on Behavior and Quality of Life: An Alternative Strategy for Clinical Trials

Curtis A. Bagne
Department of Psychiatry
Wayne State University and Lafayette Clinic

Ronald F. Lewis
Lafayette Clinic and Wayne State University

Conventional clinical trials involve tests of hypotheses with statistics computed from values of dependent variables alone. An alternative is to test hypotheses with statistics computed from benefit/harm scores that measure *longitudinal associations* between dose and values of the dependent variables. The proposed standardized measure of benefit/harm quantifies the strength of evidence that a patient did either better or worse while on treatment. A benefit/harm score, particularly when obtained from a randomized, *N*-of-1 trial, indicates a beneficial or harmful treatment effect for the individual. Benefit/harm scores from samples of patients are evaluated with standard statistical tests, with or without group comparisons, to make inferences about populations. The proposed alternative strategy can yield *within-patient* indicators of treatment effect that are more reliable, valid, comprehensive, and detailed. This, in turn, could help make many population-based clinical trials more informative, cost-effective, and clinically useful for participants.

This article introduces an alternative strategy for conducting clinical trials. The proposed alternative strategy is particularly advantageous for evaluating the way drugs, administered in doses that vary over time, affect dependent variables that can fluctuate in level. Examples of dependent variables suitable for the alternative strategy include many measures of physiological, psychological, and social functioning and quality of life. In contrast, the conventional strategy is more appropriate for evaluating the effects of all-or-none treatments on generally nonrecurrent, all-or-none outcomes such as death, remission, or cure, which often are quantified with group rates and proportions.

The primary distinction between the conventional strategy and the proposed alternative strategy for clinical trials can be identified by answering the question: "What type of measure is tested to make inferences about the effects of therapy on populations?" Investigators conducting conventional clinical trials test hypotheses with statistics computed from values of dependent variables alone. These dependent variables are measures of signs, symptoms, and behavior. The proposed alternative is a two-phase analytic strategy that can be applied when both the dose of treatment and values of the dependent variables can fluctuate over time for individuals.

Phase 1, the measurement phase of analysis with the alternative strategy, is to score within-patient *longitudinal associations* between dose and values of the dependent variables. The proposed standardized measure of longitudinal association that characterizes the alternative strategy quantifies the *strength of evidence that a patient did either better or worse while on treatment.* Values of the measure will be called *benefit/harm scores* or simply *benefit/harm.* The set of procedures and options for computing benefit/harm scores will be called *benefit/harm scoring.* A benefit/harm score, particularly when computed from data obtained under experimental conditions that isolate the effects of treatment from the effects of other variables, is *interpreted as a within-patient indicator of treatment effect.*

Phase 2, the statistical phase of analysis, tests hypotheses using sample statistics computed from benefit/harm scores to make inferences about populations. Benefit/harm scores in the following demonstrations will not be used to test hypotheses about individual subjects.

Benefit/harm scoring can facilitate the use of several techniques to help make within-patient indicators of treatment effect more reliable, valid, comprehensive, and detailed. These techniques include use of information from assessments repeated on two or more occasions, use of more than one dependent variable, and use of randomized *N*-of-1 clinical trials. The combination of benefit/harm scoring with these techniques can, in turn, help make many population-based clinical trials more informative, cost-effective, and clinically useful for participants.

Demonstration 1: A Parallel-Groups, Multiple *N*-of-1 Clinical Trial

Demonstration 1 consists of a coordinated set of *N*-of-1 clinical trials for each of two groups obtained by randomizing a

We acknowledge Paul Sherwood, who wrote the original benefit/harm scoring program. The Information Technology Institute at Wayne State University supported development of new scoring software. We thank Dr. David Gurevitch of Inpatient Geriatrics at Lafayette Clinic for his advice and encouragement.

Correspondence concerning this article should be addressed to Curtis A. Bagne, who is now at Parke-Davis Pharmaceutical Research Division, 2800 Plymouth Road, Ann Arbor, Michigan 48105.

sample of patients. The results will be analyzed to make inferences about benefit/harm for the population. Demonstration 1 illustrates a parallel-groups, multiple *N*-of-1 clinical trial.

Benefit/harm scoring can be applied both to multiple *N*-of-1 designs and to intensive studies with more conventional pre–post and crossover designs. The multiple *N*-of-1 design was selected for Demonstration 1 because it is uniquely suitable for providing within-patient indicators of treatment effect. However, selection of an unconventional design apparently precludes direct comparisons of, for example, statistical power achieved with conventional and alternative strategies because there appears to be no standard procedure for analyzing Demonstration 1 data. Demonstration 1 uses hypothetical data.

Study Questions

Assume there was a need to evaluate and compare the effectiveness of two antipsychotic drugs, such as oral thioridazine and oral haloperidol, for managing the behavior of patients with chronic schizophrenic symptoms. There were two primary questions:

1. Do the two drug therapies differ from each other in terms of benefit/harm with respect to a set of weighted dependent variables that assess patient behavior and quality of life?
2. Are the individual drug therapies either better or worse than placebo in terms of benefit/harm with respect to the set of all dependent variables?

In addition, do the drugs have different profiles of benefit/harm across the dependent variables? Questions such as these can be answered with a parallel-groups, multiple *N*-of-1 clinical trial in which the parallel-groups component would compare the two drug therapies and the multiple *N*-of-1 component would evaluate the effects of dose (drug vs. placebo) for each drug.

Study Design and Intervention

Eight patients who comprised a representative sample were randomly assigned to either Drug Group 1 (thioridazine) or Drug Group 2 (haloperidol)—4 patients to each group. In addition, each patient participated in a randomized, *N*-of-1 trial. Each *N*-of-1 trial consisted of two pairs of treatment periods as described by Guyatt et al. (1986, 1988). The patients received drug during one period of each pair and placebo during the other period. This yields four possible dose patterns—0101, 0110, 1001, and 1010—where placebo and drug are represented by 0 and 1 respectively. The four patients in each group were randomized to one of these dose patterns. Thus each patient was randomized twice, once to drug group and once to dose pattern. Assessments were repeated on two occasions during each period. Table 1 shows the dose pattern and the dependent variable data for each patient in Group 1. Assume each patient's dose was optimal as determined by previous experience. Larger samples of patients would be recommended for actual studies, especially if the results were to be used for identifying subgroups of responders.

Selection and Assessment of Dependent Variables

Benefit/harm scoring allows the measurement of apparent overall benefit/harm across many dependent variables. In this example, one statistical test will be used to evaluate overall benefit/harm across seven dependent variables.

The seven dependent variables in Demonstration 1 include several types of scales. The Brief Psychiatric Rating Scale (BPRS; Overall & Gorham, 1962) represents ordinal rating scales of the type often used in psychiatry. Clinical Global Impression (CGI; Guy, 1976) is a 7-point subjective rating of the overall severity of a patient's condition. Extrapyramidal symptoms (EPS) were rated as absent, mild, moderate, or severe. The Trailmaking Test–Part B (Trails B; Boll, 1981) is a neuropsychological task that was timed in seconds. Sedation was measured on a 100-mm visual analog scale labeled *drowsy* (0) and *alert* (100). Both dry mouth and drooling were rated as being *absent* (0) or *present* (1). Higher values on all dependent variables except sedation indicate that the patient was in worse condition. Assume that all assessments were conducted double blind with respect to both drug group assignment and dose pattern.

Each patient was scheduled to have repeated assessments on eight occasions. Assume that the assessment interval was long enough for drug effects to have become evident after beginning drug therapy and for drug effects to have ceased after changes to placebo. No baseline assessments were used for scoring benefit/harm. Some assessments were not completed. Benefit/harm scores were computed from the results of all completed assessments.

Computation of Benefit/Harm Scores

The next three sections describe computation of summary benefit/harm scores such as those that appear in the right-hand column of Table 1. For simplicity, the basic procedure will be described for the case in which both drug and a symptom were either present or absent on each assessment occasion. Series that can have only one of two values (0 or 1) for each occasion will be called *dichotomous series.* Also for simplicity, the scoring procedure demonstrated here uses scoring options that assume stability of disorder.

The benefit/harm score for dry mouth, Patient 1, Table 1, is −1.53. This score was obtained as follows: Each of the eight occasions was assigned to one cell of a 2 × 2 table associating the presence or absence of drug and dry mouth. The cells of the 2 × 2 table and the number of occasions assigned to each cell are identified by *a*, *b*, *c*, and *d*. For example, the upper left hand cell is labeled *a*, and *a* equals the number of occasions where both drug and dry mouth were present ($a = 2$). The total number of occasions, $a + b + c + d$, is represented by lower case *n* to distinguish it from *N*, a common abbreviation for the number of subjects. The data, cell assignments, and 2 × 2 table are shown here.

Drug	0	0	1	1	0	0	1	1
Dry mouth	0	0	1	0	0	0	0	1
Cell assignment	d	d	a	c	d	d	c	a

		Drug: Present 1	Drug: Absent 0
Dry Mouth	Present, 1	1, 1 $a = 2$	0, 1 $b = 0$
	Absent, 0	1, 0 $c = 2$	0, 0 $d = 4$

The *magnitude* or absolute value of the raw or unstandardized benefit/harm score, $|B_{raw}|$, is

$$|B_{raw}| = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} = 2.67.$$

Although this quantity, $|B_{raw}|$, is the sum of the squared differences between the observed and expected cell frequencies, the value is not referred to the chi-squared distribution to assess a probability or significance level. Notice that the value of $|B_{raw}|$

Table 1
Demonstration 1: Raw Data and Summary Benefit/Harm Scores for Group 1 of a Hypothetical Study Comparing Two Drugs Against Each Other and Each Against Placebo in Schizophrenic Patients, by Pair, Period, and Assessment Occasion

	Pair 1				Pair 2				
	Period 1		Period 2		Period 1		Period 2		
Variable	1	2	3	4	5	6	7	8	Summary benefit/harm score
Patient 1									
Drug	0	0	1	1	0	0	1	1	
BPRS	34	47	37	29	51	42	30	23	2.55
CGI	2	3	3	2	4	2	2	1	1.00
EPS	0	0	1	0	1	0	0	0	0
Trails B	73	106	96	85	94	87	100	78	0
Sedation	95	65	70	83	91	78	83	68	−1.53
Dry mouth	0	0	1	0	0	0	0	1	−1.53
Drooling	0	0	0	0	0	0	0	0	0
Patient 2									
Drug	1	1	0	0	0	0	1	1	
BPRS	40	34	43	54	48	49	37	37	4.18
CGI	4	3	4	5	4	4	2	3	2.55
EPS	1	2	1	1	1	1	1	1	−1.00
Trails B	232	121	164	108	219	183	251	170	−1.53
Sedation	34	78	65	54	19	34	21	50	1.00
Dry mouth	1	1	1	1	1	1	1	1	0
Drooling	0	0	0	0	0	0	0	0	0
Patient 3									
Drug	0	0	1	1	1	1	0	0	
BPRS	37	27	32	25	33	22	40	29	1.53
CGI	3	2	1	1	3	1	3	2	2.55
EPS	2	2	2	1	1	1	1	1	0.28
Trails B	155	148	138	150	99	162	108	176	1.00
Sedation	73	56	56	67	84	45	93	84	−1.05
Dry mouth	1	0	1	1	0	0	0	0	−0.28
Drooling	0	0	0	0	0	1	1	0	0
Patient 4									
Drug	1	1	0	0	1	1	0	0	
BPRS	39	32	—	43	31	35	40	41	3.61
CGI	3	2	—	3	2	2	3	3	2.08
EPS	0	1	—	0	1	0	1	1	0.14
Trails B	161	163	—	110	130	155	140	144	−2.08
Sedation	12	30	—	50	25	27	37	44	−3.61
Dry mouth	1	1	—	0	0	0	0	0	−1.15
Drooling	0	0	—	0	0	1	0	0	−0.87

Note. BPRS = Brief Psychiatric Rating Scale; CGI = Clinical Global Impression; EPS = extrapyramidal symptoms; Trails B = Trailmaking Test–Part B. Dash = information not available.

depends on both the strength of association and the number of occasions.

The *sign* of B_{raw} is determined by computing the expected value of a, $E(a)$,

$$E(a) = \frac{(a + b)(a + c)}{n} = 1.00,$$

and following a two-part rule so that positive values describe evidence for benefit and negative values describe evidence for harm.

1. If higher values on the dependent variable indicate higher severity or worse function, then

$$\text{If } O(a) < E(a),\ B_{raw} = |B_{raw}|$$

$$\text{If } O(a) > E(a),\ B_{raw} = -|B_{raw}|.$$

2. If higher values on the dependent variable indicate lower severity or better function, then

$$\text{If } O(a) < E(a),\ B_{raw} = -|B_{raw}|$$

$$\text{If } O(a) > E(a),\ B_{raw} = |B_{raw}|.$$

The B_{raw} for these data is −2.67, which describes evidence that the patient did worse—had more dry mouth—while on therapy. Positive scores describe evidence for benefit or improvement while on therapy. Notice that $B_{raw} = 0$ when $O(a) = E(a)$.

Standardization of benefit/harm scores makes it more appropriate for them to be compared, summarized, and averaged. Each standardized score is one score from a distribution of scores that has a mean of 0, and the standard deviation is 1 unless 0 is the only potential score. This distribution is described by a set of *potential* scores and their corresponding probabilities. The potential scores consist of all benefit/harm scores that are possible given the marginal frequencies of the observed 2 × 2 table. The probability corresponding to each potential score, given the observed marginal frequencies, is based on the assumption of random association.

Table 2 illustrates the standardization procedure for the aforementioned 2 × 2 table. The marginal frequencies of this observed 2 × 2 table are $a + b = 2$, $c + d = 6$, $a + c = 4$, and $b + d = 4$. Column 1 in Table 2 shows the three 2 × 2 tables that are possible given these marginals. Column 2 lists B_{raw} for each of these 2 × 2 tables. The probability of getting each 2 × 2 table and the corresponding value of B_{raw} under the assumption of random association was computed with the same formula that is used in the Fisher exact probability test:

$$P(B_{raw}) = \frac{(a + b)!(c + d)!(a + c)!(b + d)!}{n!a!b!c!d!}.$$

These probabilities, which are shown in column 3, cannot be used to determine the significance level of the benefit/harm score for an individual patient when there is serial dependency in the dichotomous series. However, the more modest claim that benefit/harm scores help quantify the strength of evidence that a patient did either better or worse while on treatment appears to be justified. The extent to which the measurement of benefit/harm can serve important functions is largely a matter for empirical study.

Columns 2 and 3 of Table 2 describe a discrete probability distribution. The mean or expected value of B_{raw}, $E(B_{raw})$, for this distribution is

$$E(B_{raw}) = \Sigma(B_{raw})P(B_{raw}) = 0.$$

The variance, σ^2, of the distribution of B_{raw} in Table 2 is

$$\sigma^2 = \Sigma[B_{raw} - E(B_{raw})]^2 P(B_{raw}) = 3.0476.$$

The standardized benefit/harm score, B, that corresponds to each B_{raw} is computed

$$B = \frac{B_{raw} - E(B_{raw})}{\sigma}.$$

The distribution of potential scores, given the observed marginal frequencies and the assumption of random association, is described by columns 3 and 4 of Table 2. The value of B that corresponds to the observed 2 × 2 table is −1.53.

The entire standardization procedure was repeated for every 2 × 2 table that was used to score benefit/harm for the demonstrations in this article. Appendix A, which provides two lines of evidence confirming that the mean or expected value of B, $E(B)$, equals zero when the longitudinal or within-patient association between dose and values of the dependent variable is random, may help provide insight about how the scoring system works. Appendix B provides information about how n and $(a + c)/n$ can affect the shapes of distributions of potential scores and thus affect the comparability of benefit/harm scores from different 2 × 2 tables. No attempt was made to normalize the distributions of potential B_{raw} values before standardization.

Benefit/Harm Score Arrays From Dimensional Dependent Variables

Unlike the data for dry mouth and drooling, all other dependent variables in Demonstration 1 could have had more than two values. The longitudinal association between the dichotomous dose series for each patient and each series of values for a dimensional dependent variable was described by a one-di-

Table 2
Derivation of a Distribution of Potential Standardized Benefit/Harm Scores for a 2 × 2 Table With Fixed Marginal Frequencies

2 × 2 table			B_{raw}	$P(B_{raw})$	B
0	2	2	2.67	.2143	1.53
4	2	6			
4	4	8			
1	1	2	0	.5714	0
3	3	6			
4	4	8			
2	0	2	−2.67	.2143	−1.53[a]
2	4	6			
4	4	8			
Total				1.0000	

[a] Observed summary benefit/harm score for Patient 1, dry mouth.

mensional array of benefit/harm scores and summarized by the most extreme score in the array. A series of values for a dimensional variable will be called a *dimensional series.*

A dimensional series can be transformed into a *set* of dichotomous series. This transformation is not the same as converting ordinal data into categorical data because a series of values with L levels can be transformed, without loss of information, into a set of dichotomous series with $L - 1$ members. All data for Demonstration 1 were scored at an ordinal level. Additional series could be used to transform interval data, including decimal values from continuous scales, in a way that would achieve any desired degree of dimensional resolution without loss of information.

The BPRS data for Patient 2, Table 1, were transformed as follows. In this transformation, 1 indicates a BPRS score greater than or equal to a specified value, and 0 indicates lesser values. Each dichotomous series in this set is labeled with the lowest BPRS score that yielded a 1 in the series and corresponds to the use of a different cutoff score for defining the presence of a health state.

Dimensional BPRS score series with seven levels

40 34 43 54 48 49 37 37

Set of dichotomous BPRS score series with six members

Label	Dichotomous Series							
≥37	1	0	1	1	1	1	1	1
≥40	1	0	1	1	1	1	0	0
≥43	0	0	1	1	1	1	0	0
≥48	0	0	0	1	1	1	0	0
≥49	0	0	0	1	0	1	0	0
=54	0	0	0	1	0	0	0	0

After this transformation, benefit/harm was scored as shown above. Each of the six dichotomous BPRS score series was cross-classified with the drug dose series for Patient 2 as shown in Table 3. This yielded six observed 2×2 tables with different marginal frequencies. The value of B is shown for each 2×2 table.

The one dimensional array of benefit/harm scores, shown in the right-hand column of Table 3, provides a detailed description of the longitudinal association between drug dose and BPRS score for Patient 2. The single dimension of this array corresponds to levels of BPRS score. These levels are identified by the labels for members of the set of dichotomous series that are shown above.

Summarizing Benefit/Harm Score Arrays

Arrays of standardized benefit/harm scores can be summarized to help make generalizations and inferences. The recommended procedure for summarizing the array corresponding to one dependent variable for one patient is to select the most extreme value. Thus, the array in the right-hand column of Table 3 is summarized by 4.18. Extreme values of equal magnitude but opposite sign are summarized by 0. The right-hand column of Table 1 uses summary scores to profile benefit/harm across the seven dependent variables for each patient in Group 1.

One reason for summarizing arrays by extreme values is that the location of the summary benefit/harm score in the array identifies the conditions, as defined by the levels of array's dimensions, that provided the most benefit/harm. The most evidence for an association between treatment and health states in Table 3 ($B = 4.18$) was obtained when BPRS was ≥43: All BPRS scores for Patient 2 were ≥43 on placebo and <43 on drug. It would not seem meaningful to summarize an array by the mean of scores in the array when the levels of a dimension are identified by values on an integrated scale—values greater than or equal to specified values.

Selection of extreme values within arrays for individuals in the sample apparently does not increase the probability of Type I error. The Monte Carlo simulations in Appendix A show how positive and negative summary benefit/harm scores from different individuals in a sample tend to offset each other when the within-patient associations are random. For this reason, both of the individual drug groups in Demonstration 1 will be evaluated with a one-sample t test in an attempt to reject the null hypothesis that mean benefit/harm equals 0.

Computing Overall Benefit/Harm Scores

Overall benefit/harm scores can be used to evaluate hypotheses about benefit/harm across many dependent variables with one statistical test. Investigators who compute overall benefit/harm scores have the option of differentially weighting benefit/harm scores for the individual dependent variables in accord with clinical significance and patient preferences. Methodologies described by Froberg and Kane (1990) for measuring health state preferences could be applied to determine these weights. Figure 1 includes the weights for Demonstration 1. The summary benefit/harm score for each dependent variable was multiplied by the assigned weight. The sum of the resulting products was divided by the sum of the weights to obtain the overall benefit/harm score for each patient. Publication of unweighted benefit/harm scores would assist readers who want to reanalyze the results of studies using different weights.

Results for Demonstration 1: Testing Benefit/Harm Scores

Figure 1 summarizes results for Demonstration 1 and helps answer the previously stated study questions. The difference between the mean overall benefit/harm scores for Drug Groups 1 (2.97) and 2 (2.48) did not approach statistical significance, $t(6) = -0.46$. This univariate test was based on the results of assessments repeated on from six to eight occasions (because of missing data) of seven differentially weighted dependent variables for each of 8 patients. This demonstrates how overall benefit/harm can be scored over time (two or more occasions) and across any number of dependent variables for each patient in a sample and how the resulting scores can be used to make inferences about benefit/harm in the population.

The mean of the overall benefit/harm scores was significantly greater than 0 for both drug groups: Group 1, $t(3) = 4.39$, $p = .0220$; Group 2, $t(3) = 5.96$, $p = .0094$. In other words, the data provide evidence that the longitudinal association between drug dose and health status in the population was not random for either drug. Such nonrandom associations can be attributed to drug effect when study designs effectively isolate the effects of drug from the effects of other variables. These one-sample t tests demonstrate how the results of several ran-

Table 3
Demonstration 1: Procedure for Computing Benefit/Harm Scores for the BPRS Score Data, Patient 2

Dichotomous series and cell assignments									2 × 2 table	Benefit/harm score array
Drug	1	1	0	0	0	0	1	1		
BPRS ≥ 37	1	0	1	1	1	1	1	1	3 4	1.00
Cell assignment	a	c	b	b	b	b	a	a	1 0	
Drug	1	1	0	0	0	0	1	1		
BPRS ≥ 40	1	0	1	1	1	1	0	0	1 4	2.55
Cell assignment	a	c	b	b	b	b	c	c	3 0	
Drug	1	1	0	0	0	0	1	1		
BPRS ≥ 43	0	0	1	1	1	1	0	0	0 4	4.18
Cell assignment	c	c	b	b	b	b	c	c	4 0	
Drug	1	1	0	0	0	0	1	1		
BPRS ≥ 48	0	0	0	1	1	1	0	0	0 3	2.55
Cell assignment	c	c	d	b	b	b	c	c	4 1	
Drug	1	1	0	0	0	0	1	1		
BPRS ≥ 49	0	0	0	1	0	1	0	0	0 2	1.53
Cell assignment	c	c	d	b	d	b	c	c	4 2	
Drug	1	1	0	0	0	0	1	1		
BPRS ≥ 54	0	0	0	1	0	0	0	0	0 1	1.00
Cell assignment	c	c	d	b	d	d	c	c	4 3	

Note. BPRS = Brief Psychiatric Rating Scale.

domized, placebo-controlled, *N*-of-1 clinical trials—one for each patient in a sample—can be used to make inferences about the population.

Figure 1 includes the mean benefit/harm profiles for both drug groups. The bar corresponding to each dependent variable for Group 1 is the mean of four unweighted summary benefit/harm scores—one for each patient in Table 1. Profiles for the two groups appear to differ. Independent sample *t* test results, shown in Figure 1, suggest that there may be more EPS and drooling with haloperidol and more dry mouth with thioridazine.

Demonstration 2

We have demonstrated how benefit/harm can be scored when drug is considered to be either *present* (1) or *absent* (placebo, 0) on each occasion. However, individual patients often receive more than two doses when the amount of drug needs to be optimized or increased and decreased gradually. In addition, clinicians may change doses or discontinue a drug to help determine if, for example, an adverse state is being caused by drug. Demonstration 2 extends benefit/harm scoring to situations that involve more than two doses per patient. Benefit/harm

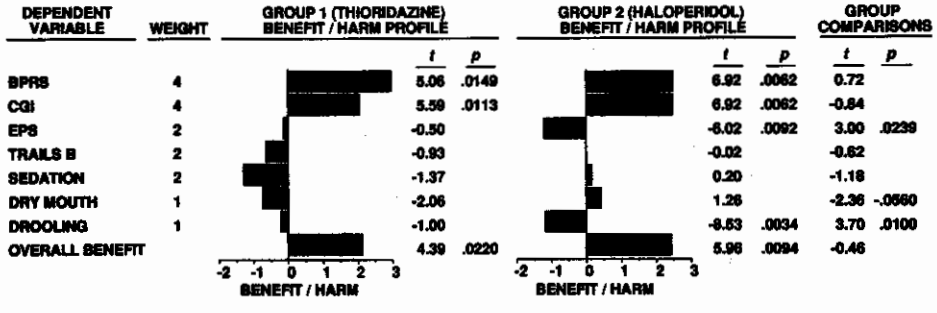

Figure 1. Demonstration 1: Average benefit/harm profiles for the two drug groups, weights used for computing overall benefit/harm, and results of the one- and two-sample *t* tests. (BPRS = Brief Psychiatric Rating Scale; CGI = Clinical Global Impression; EPS = extrapyramidal symptoms; Trails B = Trailmaking Test–Part B.

will be evaluated as a function of dose in a manner that permits different doses to be considered as different levels of a single treatment rather than as different treatments.

Responses to changes in dose are seldom immediate. For example, the antidepressant effects of tricyclic drugs may be delayed for 2 or more weeks. Demonstration 2 also presents one procedure for evaluating delay of apparent response to therapy. Limiting factors in the resolution of temporal analyses are the length and the regularity of the assessment intervals.

The hypothetical data for Demonstration 2 are included in Table 4 and consist of dimensional series for dose and two dependent variables for one patient. The dependent variables, assessed on 15 equally spaced occasions, are the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960) and a 4-point rating scale for assessing a side effect such as dry mouth. Higher values on both scales indicate that the patient was in worse condition. Assume that actual dose was masked by placebo dosage forms and that ratings were made double-blind. The observation that a patient did better or worse on therapy may not mean that the apparent effect is due to therapy—especially when the dose pattern was not obtained by randomization.

Scoring Benefit/Harm When There Are More Than Two Doses

The procedure for scoring benefit/harm for a pair of dimensional variables will be demonstrated with the dose and dry mouth series from Table 4. Both dimensional series were transformed into sets of dichotomous series as done for BPRS scores in Demonstration 1. Table 4 includes this transformation for Demonstration 2.

Each dichotomous dose series was cross-classified with each dichotomous dry mouth series to yield a two-dimensional array of observed 2 × 2 tables. The 2 × 2 tables have different marginal frequencies. The value of B was computed for each

Table 4
Demonstration 2: Raw Data (Dimensional Series) and Transformed Data (Dichotomous Series) Resulting From 15 Equally Spaced Assessments of Dose and Two Dependent Variables for One Patient

Variable levels	Assessment occasion														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Dimensional dose series														
	0	0	50	75	100	125	125	125	125	100	75	50	0	0	0
	Set of four dichotomous dose series														
≥50	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0
≥75	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0
≥100	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0
=125	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	Dimensional HRSD score series														
	25	23	25	20	21	20	14	8	10	6	10	16	23	18	23
	Set of nine dichotomous HRSD score series														
≥8	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
≥10	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1
≥14	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1
≥16	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1
≥18	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1
≥20	1	1	1	1	1	1	0	0	0	0	0	0	1	0	1
≥21	1	1	1	0	1	0	0	0	9	0	0	0	1	0	1
≥23	1	1	1	0	0	0	0	0	0	0	0	0	1	0	1
=25	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	Dimensional series for ratings of dry mouth														
	0	0	1	3	2	3	2	1	2	3	2	1	2	0	2
	Set of three dichotomous series for ratings of dry mouth														
≥1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1
≥2	0	0	0	1	1	1	1	0	1	1	1	0	1	0	1
=3	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0

Note. HRSD = Hamilton Rating Scale for Depression.

2×2 table to yield the following array of benefit/harm scores, which is summarized by the extreme value −4.334.

	Dose			
Dry mouth level	≥50	≥75	≥100	=125
≥1	−4.334	−2.496	−1.515	−0.839
≥2	−0.687	−2.994	−1.271	−0.303
=3	−1.144	−1.946	−0.603	−0.002

Evaluating Benefit/Harm as a Function of Delay of Apparent Response

Delay of apparent response is an indicator of how long it takes for the benefit/harm of therapy to become evident after an increase in dose. The procedure for evaluating benefit/harm as a function of delay of response to therapy is shown in Table 5 using the dichotomous series for dose = 125 and HRSD ≥14 from Table 4. Delay was evaluated by shifting the health state series for HRSD score to the left relative to the dichotomous dose series in increments of 0, 1, 2, 3, and 4 assessment intervals. This particular procedure does not distinguish delay of response after increases in dose from persistence of response after decreases in dose. Each occasion with paired dose and HRSD score values was assigned to the appropriate cell of the corresponding 2×2 table. The value of B was computed for each observed 2×2 table. The most extreme benefit score (7.31) in Table 5 was obtained with a delay of two assessment intervals. This procedure for evaluating delay of response was repeated for each combination of a dichotomous dose series with a dichotomous dependent variable series from Table 4.

Scoring Benefit/Harm as a Function of Both Dose and Delay of Response

The procedures for evaluating benefit/harm as a function of dose and delay of response were applied concurrently to yield a three-dimensional array of standardized benefit/harm scores for both dependent variables. The dimensions correspond to dose, dependent variable level, and delay of apparent response.

Figure 2 shows benefit/harm with respect to HRSD score as a joint function of dose and delay of response for Table 4 data. Each bar in Figure 2 is the summary benefit score across nine levels of HRSD score. The summary benefit score across all three dimensions is 7.31 and occurred when dose = 125, delay of response = 2, and HRSD ≥14.

The top part of Figure 3 shows benefit/harm as a function of

Table 5
Demonstration 2: Benefit/Harm Scoring Procedure for Evaluating Delay of Apparent Response to Therapy

Dichotomous series and cell assignments	2×2 table	Benefit/harm score
Delay of response = 0		
Dose = 125 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0		
HRSD ≥ 14 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1	2 9	0.81
b b b b b a a c c d d b b b b	2 2	
Delay of response = 1		
0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0		
1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1	1 9	3.32
b b b b b a c c c d b b b b	3 1	
Delay of response = 2		
0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0		
1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1	0 9	7.31
b b b b b c c c c b b b b	4 0	
Delay of response = 3		
0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0		
1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1	1 7	2.59
b b b b d c c c a b b b	3 1	
Delay of response = 4		
0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0		
1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1	2 5	0.26
b b b d d c c a a b b	2 2	

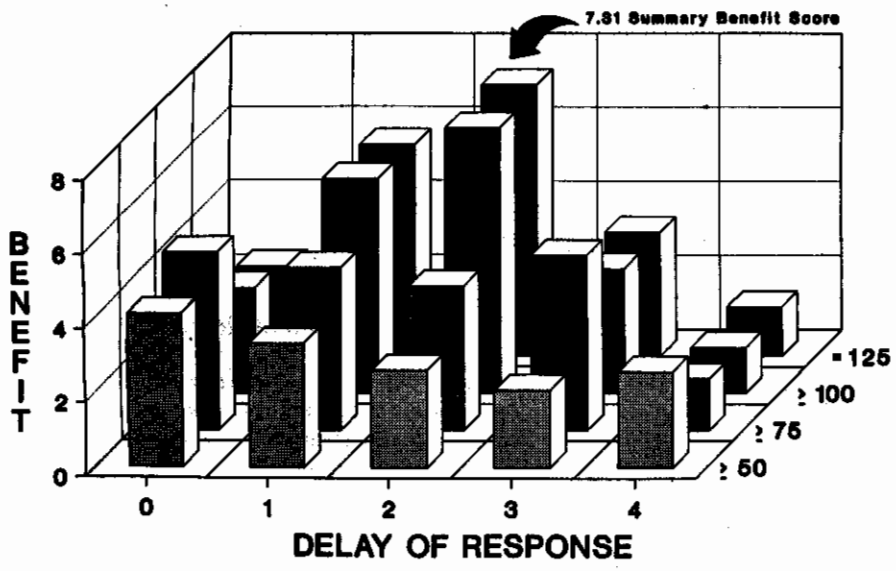

Figure 2. Demonstration 2: Benefit with respect to HRSD score as a function of dose and delay of response. (HRSD = Hamilton Rating Scale for Depression.)

dose for both HRSD score and ratings of dry mouth. Levels of the integrated dose scale are labeled in the same way as the dichotomous dose series are labeled in Table 4. Each point on a curve is the summary benefit/harm score (extreme value) across levels of the dependent variable and delay of response values 0 through 4. The dose–benefit curve for HRSD score is the silhouette of Figure 2 viewed across delay of response. This curve shows that the highest doses yielded the most extreme benefit scores. The lowest nonzero dose (≥50) yielded the most extreme harm score with respect to ratings of dry mouth, a nonmonotonic relationship. Notice that each point in Figure 3 is based on data from most of the 15 repeated assessments. Any number of dose–benefit/harm curves for different dependent variables could be differentially weighted and averaged. The most extreme positive value on the overall dose–benefit/harm curve possibly could be used to help identify the minimum dose that provided the most benefit.

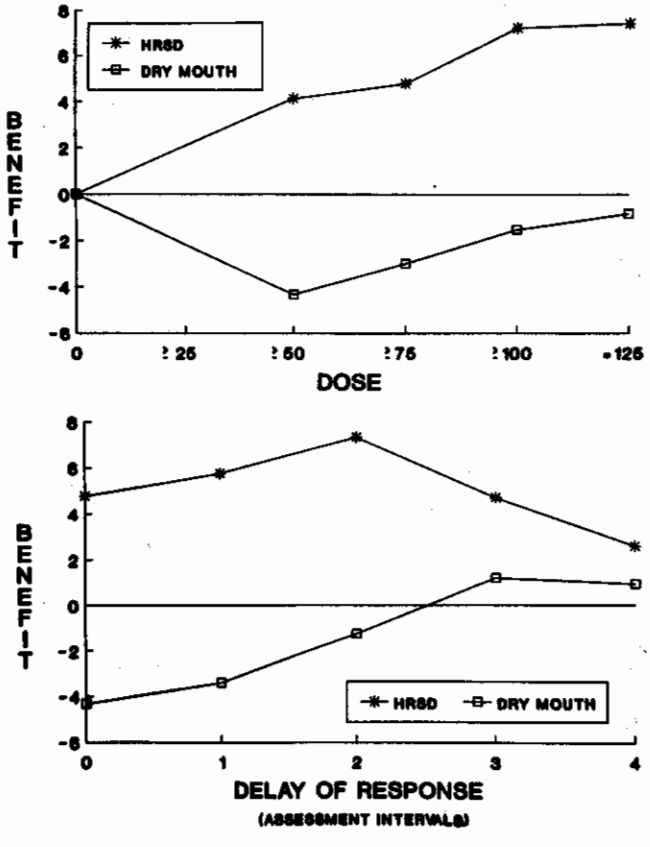

Figure 3. Demonstration 2: Benefit/harm as a function of dose (top) and delay of apparent response (bottom) for HRSD score and ratings of dry mouth. (HRSD = Hamilton Rating Scale for Depression.)

The bottom part of Figure 3 shows benefit/harm as a function of delay of apparent response for both HRSD score and ratings of dry mouth. Each point on both curves is the summary benefit/harm score across dose and levels of the dependent variable. Evidence for benefit with respect to HRSD score was greatest after two assessment intervals. Delay of response with respect to ratings of dry mouth appeared to be no more than the length of one assessment interval.

Demonstration 2 provided a detailed, quantitative description of benefit/harm for one patient as evaluated with particular benefit/harm scoring options. The same type of description could be obtained from each patient in a sample using a randomized, parallel-groups design. Such descriptions could be analyzed statistically to test a hypothesis about treatment effect in the population. Similarly, the average dose–benefit/harm curve across dependent variables for each patient in a sample could, in turn, be averaged to help make dose recommendations for the patient population.

Benefit/Harm Scoring Options

This article has demonstrated only a few available benefit/harm scoring options. There are, for example, many ways to

transform a dimensional series into a set of dichotomous series. The general guideline for developing optional transformations is to state explicit rules for determining whether treatment states, defined in terms of dose, and whether health states, defined in terms of dependent variables, are considered to be either *present* (1) or *absent* (0) on the occasion of each assessment. Each variable in the criteria used to define the presence of either treatment or health states would correspond to a dimension of the resulting benefit/harm score arrays. Demonstration 2 yielded a three-dimensional array—dose, delay of apparent response, and dependent variable level—for both dependent variables. Additional dimensions could be used, for example, to define episodes of treatment or health states. Another option would be to form sets of dichotomous series from the residuals of a regression line that relates a dependent variable to assessment number or time. This could be used to help distinguish treatment effects from longer term changes due to practice, fatigue, spontaneous recovery, or progression of disease. Still another option, which could be used when change in level of the independent variable is more important than absolute level, would be to form dichotomous series from differences between successive values. Investigators should select or develop options to meet the needs of their studies. All scoring procedures and options should be specified in research protocols. The objective would be to use options that yield the most extreme benefit/harm scores and maximize predictive power.

Benefit/harm scoring also can be applied to study associations between recurrent events rather than states. Events could be defined in terms of transitions from 0 to 1 or from 1 to 0 in dichotomous treatment and health state series. Thus, for example, health events could be related to events such as the initiation or termination of specific treatment states.

Discussion

Benefit/Harm Scores Versus Some Other Measures of Association

Benefit/harm scores can be contrasted with measures of association that generally function as estimates of population parameters. Phi and correlation coefficients usually are obtained from cross-sectional data, and their magnitudes describe the *strength of associations* between variables across subjects. In contrast, benefit/harm scores are obtained from repeated measures data, and their magnitudes describe the *strength of evidence* for longitudinal associations between treatment and health states across occasions for individual subjects. The maximum magnitude of phi and correlation coefficients is 1. This value can be obtained with two subjects. In contrast, the magnitude of a benefit/harm score can increase without limit as the number of occasions, which can contribute to the strength of evidence for an association, increases. Unlike chi-squared, benefit/harm scores can have positive and negative values. Each benefit/harm score is one score from a distribution of potential scores that has a mean of 0 and a standard deviation of 1.

Measures of strength of association, which could be useful for studying the relationship between clinical and statistical significance, can be computed from benefit/harm scores. One measure, derived from the distribution of potential scores for an observed 2×2 table, consists of the ratio of B over the maximum value of B in the same positive or negative direction. A similar but more stringent measure can be obtained by forcing zeros into one or the other diagonal of the 2×2 table and changing the marginal frequencies for the dependent variable as may be necessary. Thus if $a = 2$, $b = 5$, $c = 6$, and $d = 1$, the value of the first measure would be $2.56/5.77 = 0.44$, and the value of the second would be $2.56/7.71 = 0.33$.

Reliable Within-Patient Indicators of Treatment Effect

Several one-group t tests in Demonstration 1 yielded statistically significant results with only 4 subjects. This suggests that benefit/harm scoring uses information from assessments repeated on more occasions to help make within-patient indicators of treatment effect more reliable and thus improve statistical power. This contrasts with the conventional strategy in which the primary hypothesis often is tested with assessments obtained at only one or two occasions (at endpoint, before and after treatment, or at the ends of two periods in a crossover design). The ability of benefit/harm scoring to improve statistical power through the use of information from more occasions has been supported by a simulation (unpublished) in which values of t from t tests for a given N were studied as functions of the size of a treatment effect signal, noise level (based on random normal deviates), and n. Improvements in statistical power, achieved without increasing N, can increase the cost-effectiveness of many clinical trials.

The reproducibility of benefit/harm scores, given the raw data, is not a problem because the scores are obtained by computation. This contrasts with subjective ratings of response to therapy that often are used as dependent variables in clinical trials.

The issue of "How many occasions?" that arises when there is need for reliable within-patient indicators of treatment effect is important for many of the same reasons that "How many subjects?" is important in population-based clinical trials that involve soft data (cf. Kraemer & Thiemann, 1987, 1989). The number of occasions becomes crucial when there is unavoidable unreliability in measuring dependent variables, when there are unavoidable fluctuations in dependent variable levels as a result of biological rhythms and homeostatic processes, and when dependent variables can be affected by uncontrollable extraneous variables.

Benefit/harm scoring is an alternative to the standard procedure of using the mean of several repeated assessments, obtained under the same treatment condition, to improve reliability (Fleiss, 1986). Compared with this standard, benefit/harm scoring allows use of the techniques described below to help make within-patient indicators valid, comprehensive, and detailed.

Valid Within-Patient Indicators of Treatment Effect

Valid within-patient indicators of treatment effect can make population-based trials more clinically useful for participants and facilitate the development of patient classifications that

predict response to therapy. Development of these classifications would be valuable for identifying more homogeneous groups of drug responders, particularly during early phases of clinical drug evaluations.

This section contrasts two strategies for conducting randomized clinical trials designed for making valid inferences about the effects of treatments on populations. Assume we are comparing one drug with placebo in terms of effects on a dimensional dependent variable that can fluctuate in level over time. In brief, the conventional strategy generally does not provide valid within-patient indicators of treatment effect and therefore requires group comparisons to isolate the effects of treatment. The alternative strategy can provide valid within-patient indicators of treatment effect so that group comparisons are optional.

To elaborate, the conventional strategy generally is implemented with endpoint, pre–post, and two-period crossover designs. Patients studied with an endpoint design are randomized to parallel groups that receive either drug or placebo but not both. Endpoint designs do not provide any data that can be used to compute values of a within-patient indicator of treatment effect. Hypotheses are tested by comparing groups.

Pre–post designs evaluate change from one time (pretreatment) to a later time (posttreatment). The conceptualization and measurement of change has engendered an extensive literature (cf. Francis, Fletcher, Stuebing, Davidson, & Thompson, 1991). However, there are several common conditions under which change, regardless of how it is measured, provides weak or misleading evidence about the effects of treatment on individual subjects. Unlike benefit/harm scores from randomized *N*-of-1 clinical trials, measures of change do not distinguish response to active treatment from placebo response. In addition, measures of change may not provide valid within-patient indicators of treatment effect when the levels of dependent variables are affected by time-dependent processes such as disease progression, spontaneous recovery, practice, or fatigue. For such reasons, investigators who use pre–post designs test hypotheses by comparing groups rather than using single-group tests of no change.

Unlike measures of change from most pre–post designs, a difference score from a two-period crossover design does contrast the level of the dependent variable while a patient was on drug with the level on placebo. As such, difference scores have an advantage over change scores as within-patient indicators of treatment effect. However, difference scores can be invalidated by the same time-dependent processes that can invalidate measures of change. Because of these processes, two-period crossover designs are recommended only for stable disorders, and group comparisons are used in an attempt to rule out order of treatment effects before evaluating treatment effects.

The alternative strategy, illustrated by Demonstration 1, makes group comparisons optional because *N*-of-1 trials can provide valid within-patient indicators of treatment effect. Of course, results from individual *N*-of-1 trials may not be valid for other patients, and the results of a population-based trial, regardless of study strategy, may not be valid for an individual. Nevertheless, it is valuable to use sample data for making inferences about populations. Benefit/harm scoring is an important part of the alternative strategy because it can quantify and describe evidence from a coordinated set of *N*-of-1 trials for a sample of patients in order to make inferences about the population. In addition, although not illustrated by the demonstrations, previously described scoring options can be used to help separate treatment effects from long-term changes that result from spontaneous recovery, disease progression, practice, and fatigue. Procedures such as those illustrated in Demonstration 2 can be used to study delay and persistence of response. Such options and procedures can help make the alternative strategy feasible when disorders are not stable, when treatment effects are not rapid, and when carryover effects are possible.

Benefit/harm scoring offers a highly flexible system for obtaining within-patient indicators of treatment effect. Much of this flexibility is achieved by using dose information to compute benefit/harm scores rather than as a factor to define groups for statistical analyses. The scores can be computed for any pattern of variable dose. Dose patterns resulting from within-patient randomization, which helps assure the validity of each score, would not have to be restricted to the ordering of doses within pairs of periods as shown in Table 1. Every patient could be randomized to a different dose pattern. Benefit/harm scoring makes it more practical to analyze studies with multiple crossovers (Demonstration 1) and multiple doses (Demonstration 2). The flexibility of benefit/harm scoring also would facilitate nonexperimental studies of how signs, symptoms, and behaviors may be associated over time with actual dose or with levels of drugs or endogenous substances in body fluids. Benefit/harm, resulting from routine clinical care, could be monitored and used to generate hypotheses for experimental test.

The history of psychology includes notable efforts to develop and successfully apply methods for the experimental study of individuals (Chassan, 1979; Kazdin, 1982). In contrast, experimental medicine has focused almost exclusively on population-based trials. The proposed alternative strategy—consisting of the combined use of *N*-of-1 trials, benefit/harm scoring, and standard statistical tests—would appear to help bridge the gap between clinical and statistical prediction (Meehl, 1954) and to help promote "a unity of clinical and research decision-making principles" (Kraemer et al., 1987, p. 1104).

Comprehensive Treatment Evaluations

Comprehensive treatment evaluations would address multiple dependent variables in one study. This is important because drugs generally affect many potential dependent variables, and many disorders are characterized by syndromes of signs and symptoms. Evaluations conducted with dependent variables that represent more of the clinically significant effects of treatment and more components of syndromes would be more informative guides to patient care.

It is important for clinical trials, regardless of strategy, to evaluate a primary hypothesis with one statistical test (Friedman, Furberg, & DeMets, 1985). The conventional strategy, which tests statistics computed from dependent variables alone, makes it difficult to obtain comprehensive treatment evaluations. For example, trials that focus on dependent variables that measure relatively discrete concepts such as depressed mood or systolic blood pressure tend to provide an inadequate basis for treatment decisions. Most often, many de-

pendent variables are used to evaluate safety and efficacy even though this usually entails multiple statistical tests with results that are weighted and combined subjectively.

The conventional use of composite dependent variables such as the Alzheimer's Disease Assessment Scale (Rosen, Mohs, & Davis, 1984) provides evaluations that are more comprehensive at the expense of making it difficult to profile apparent treatment effects across scale items and to study the way individuals vary in response. In contrast, Demonstration 1 used seven dependent variables to show how benefit/harm can be profiled across dependent variables or scale items for both individuals and groups. In addition, Demonstration 1 showed how benefit/harm can be evaluated across many dependent variables with one statistical test. Benefit/harm scoring can help make treatment evaluations more comprehensive because it is a common metric for evaluating evidence for treatment effect with respect to different dependent variables.

Detailed Treatment Evaluations

Demonstration 2 showed how the alternative strategy can provide detailed treatment evaluations. Figure 2, for example, shows how benefit/harm with respect to each dependent variable for an individual patient can be described by an array of benefit/harm scores. Dimensions of the detailed array summarized by Figure 2 correspond to levels of two independent variables, dose and delay of apparent response to therapy, as well as levels of the dependent variable, HRSD score. The set of arrays, consisting of one array for each combination of a dependent variable with a patient, can provide a very detailed quantitative description of the benefit/harm observed in a study.

Conclusion

Benefit/harm scoring, in combination with randomized *N*-of-1 clinical trials and standard statistical tests, has the potential to improve the way we evaluate and use drugs that affect behavior and quality of life. This article has addressed a broad constellation of methodological problems. Additional theoretical and empirical research is required to refine the alternative strategy and to help identify when and how it can be used to maximal advantage. Benefit/harm scoring can be applied to many research applications, in addition to clinical drug trials, that could be addressed more adequately with multivariate longitudinal data.

References

Boll, T. J. (1981). The Halstead-Reitan Neuropsychological Battery. In S. B. Filskov & T. J. Boll (Eds.), *Handbook of Clinical Neuropsychology* (pp. 577–607). New York: Wiley.

Chassan, J. B. (1979). *Research design in clinical psychology and psychiatry* (2nd ed.). New York: Irvington.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments.* New York: Wiley.

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Davidson, K. C., & Thompson, N. M. (1991). Analysis of change: Modeling individual growth. *Journal of Consulting and Clinical Psychology, 59,* 27–37.

Friedman, L. M., Furberg, C. D., & DeMets, D. L. (1985). *Fundamentals of clinical trials* (2nd ed.). Littleton, MA: PSG Publishing.

Froberg, D. G., & Kane, R. L. (1990). Methodology for measuring health-state preferences: 1. Measurement strategies. *Journal of Clinical Epidemiology, 42,* 342–354.

Guy, W. (Ed.). (1976). *ECDEU assessment manual for psychopharmacology* (rev.; DHEW Publication No. ADM 76-338). Rockville, MD: National Institute of Mental Health.

Guyatt, G., Sackett, D., Adachi, J., Roberts, R., Chong, J., Rosenbloom, D., & Keller, J. (1988). A clinician's guide for conducting randomized trials in individual patients. *Canadian Medical Association Journal, 139,* 497–503.

Guyatt, G., Sackett, D., Taylor, D. W., Chong, J., Roberts, R., & Pugsley, S. (1986). Determining optimal therapy—Randomized trials in individual patients. *New England Journal of Medicine, 314,* 889–892.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry, 23,* 56–62.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Kraemer, H. C., Pruyn, J. P., Gibbons, R. D., Greenhouse, J. B., Grochocinski, V. J., Waternaux, C., & Kupfer, D. J. (1987). Methodology in psychiatric research. *Archives of General Psychiatry, 44,* 1100–1106.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage.

Kraemer, H. C., & Thiemann, S. (1989). A strategy to use soft data effectively in randomized controlled clinical trials. *Journal of Consulting and Clinical Psychology, 57,* 148–154.

Meehl, P. E. (1954). *Clinical versus statistical prediction.* Minneapolis: University of Minnesota Press.

Overall, J. E., & Gorham, D. R. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports, 10,* 799–812.

Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A new rating scale for Alzheimer's disease. *American Journal of Psychiatry, 141,* 1356–1364.

Appendix A

Expected Value of Benefit/Harm Scores

The first line of evidence that $E(B) = 0$ when there is random within-patient association between treatment and health status is based on examination of the 133 different 2×2 tables that are possible when $n = 8$ (as in Demonstration 1) and none of the marginal frequencies equals 0. (If any marginal frequency of a 2×2 table equals zero, the table cannot provide any evidence for benefit/harm, and the benefit/harm score is zero.) Values of B_{raw} and B were computed for each of the 133 tables. As expected, each positive score was offset by a negative score with the same magnitude and the same probability of occurring by chance so that the means for all 133 values of B_{raw} and all 133 values of B were both 0.

The 133 different 2×2 tables for $n = 8$ constitute 49 distributions of potential scores—one discrete probability distribution for each combination of values from 1 through 7 for $a + c$ with values from 1 through 7 for $a + b$. Seven distributions, which include 23 different 2×2 tables, are possible when $a + c = b + d = 4$ in accord with the study design for Demonstration 1. Table A1 shows these seven distributions with their means and standard deviations. Distributions of potential scores are symmetric when $a + c = b + d$, as shown in Table A1, or when $a + b = c + d$. Symmetric distributions of potential benefit/harm scores, both raw and standardized, have a mean of 0.

The standard deviations of the distributions of potential B_{raw} values shown in Table A1 range from 1.143 to 1.912. Standardization adjusts the distributions of B_{raw} so that all the distributions of B have a standard deviation of 1.

Treatment states may be present on different proportions of occasions, $(a + c)/n$, because of experimental design, missing data, or as a result of observation under natural conditions. Whatever this proportion for a given value of n, each positive benefit/harm score would be offset by a negative score with the same magnitude and probability of occurrence when all possible 2×2 tables are scored and when there is no longitudinal association between treatment and health status.

We also conducted two Monte Carlo simulations to confirm that the $E(B)$, both for all scores in the arrays and the scores that summarize the arrays, is 0 when there is random longitudinal association between treatment and health status for each individual in a population. These simulations used $n = 8$ and $n = 40$, respectively. Both simulations had 130 subjects. For both simulations, treatment was either present or

Table A1
All Distributions of Raw and Standardized Benefit/Harm Scores Possible When $a + c = b + d = 4$ With Means and Standard Deviations for Each Distribution

Marginal frequencies			Quantity	a 0	1	2	3	4	Distribution statistics M	σ
a b		1	B_{raw} =	1.143	−1.143				0	1.143
c d		7	B =	1.000	−1.000				0	1.000
4	4		p =	.500	.500					
		2	B_{raw} =	2.667	0.000	−2.667			0	1.746
		6	B =	1.528	0.000	−1.528			0	1.000
4	4		p =	.214	.571	.214				
		3	B_{raw} =	4.800	0.533	−.533	−4.800		0	1.880
		5	B =	2.553	0.284	−0.284	−2.553		0	1.000
4	4		p =	.071	.429	.429	.071			
		4	B_{raw} =	8.000	2.000	0.000	−2.000	−8.000	0	1.912
		4	B =	4.183	1.046	0.000	−1.046	−4.183	0	1.000
4	4		p =	.014	.229	.514	.229	.014		
		5	B_{raw} =		4.800	0.533	−0.533	−4.800	0	1.880
		3	B =		2.553	0.284	−0.284	−2.553	0	1.000
4	4		p =		.071	.429	.429	.071		
		6	B_{raw} =			2.667	0.000	−2.667	0	1.746
		2	B =			1.528	0.000	−1.528	0	1.000
4	4		p =			.214	.571	.214		
		7	B_{raw} =				1.143	−1.143	0	1.143
		1	B =				1.000	−1.000	0	1.000
4	4		p =				.500	.500		

Table A2
Results of Two Monte Carlo Simulations With 130 Subjects Each

Variable	All scores in the 130 benefit/harm score arrays: B_{raw}, $n = 8$	B_{raw}, $n = 40$	B, $n = 8$	B, $n = 40$	Summary benefit/harm scores, $n = 8$	Summary benefit/harm scores, $n = 40$
Number of benefit/harm scores	616	1,146	616	1,146	130	130
Number of distinct values	12	110	12	107	12	64
Mean	−0.09	−0.05	−0.04	−0.03	−0.13	0.07
Standard deviation	1.62	1.44	0.98	0.85	1.54	1.77
Skewness	−0.15	−0.06	−0.08	−0.03	0.02	−0.03
Kurtosis	2.51	4.14	0.76	3.82	−0.89	−0.76
Minimum observed	−8.00	−7.03	−4.18	−4.01	−4.18	−4.01
Minimum possible	−8.00	−40.00	−4.18	−22.70	−4.18	−22.70
Maximum observed	4.80	7.62	2.55	4.35	2.55	4.35
Maximum possible	8.00	40.00	4.18	22.70	4.18	22.70

absent on the occasion of each assessment, and treatment was present on 50% of the occasions. Response variable data consisted of consecutive single digits from a random number table. Benefit/harm scores were computed and summarized as shown in Demonstration 1. All benefit/harm scores were rounded to the third decimal place.

Results of the two Monte Carlo simulations are shown in Table A2. Zero falls within the 80% confidence interval about the mean of the 130 summary benefit/harm scores for both simulations. Furthermore, the mean of the summary benefit/harm scores was computed after addition of the score for each of the 130 consecutive subjects. The sign of these "running means" changed 11 times when $n = 8$ and 10 times when $n = 40$. Both types of evaluation indicate that the expected value of summary benefit/harm scores for the population is 0 when the longitudinal association for each subject is random.

Appendix B

Comparability

Although each value of B is one score from a distribution of potential scores that has a mean of 0 and a standard deviation of 1, distributions of potential scores can vary in shape. Benefit/harm scores with the same values may not indicate the same strength of evidence for treatment effect when the distributions of potential scores differ substantially in shape. These differences in shape effect comparability. This section addresses two factors that affect the way distributions of potential benefit/harm scores are shaped. These factors are n and the proportion of occasions on which treatment is present, $(a + c)/n$.

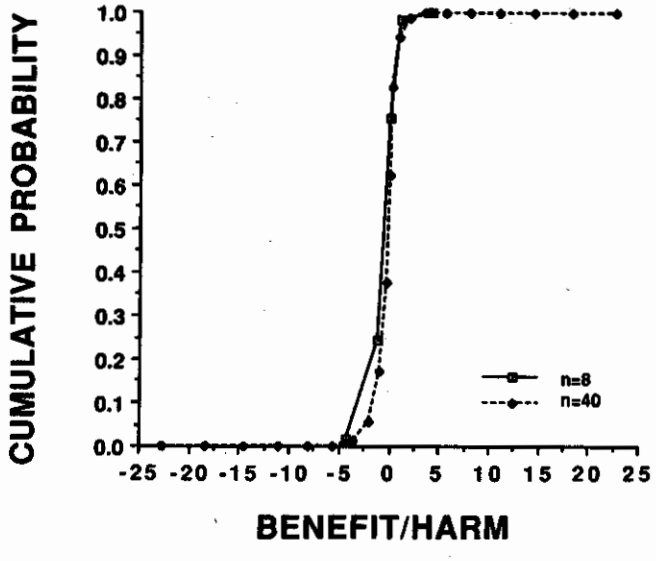

Figure B1. Cumulative probabilities for distributions of potential benefit/harm scores when all marginal frequencies are equal and when n equals either 8 or 40.

Figure B1 compares cumulative probabilities as a function of B for the distributions of potential scores that are theoretically possible when all marginal frequencies of the 2×2 tables are equal and when n equals either 8 or 40. The distribution for $n = 40$ has much longer and narrower tails. Note, however, the similarity of the observed distributions of summary benefit/harm scores that resulted from the Monte Carlo simulations with $n = 8$ and $n = 40$ as indicated by the descriptive statistics in Table A2.

With regard to the proportion of occasions on which treatment is present, distributions of potential benefit/harm scores are symmetric when $(a + c)/n = .5$ regardless of the proportion of occasions on which health states or events are present, $(a + b)/n$. This is illustrated in Table A1. However, distributions of potential scores are skewed if neither the column nor the row marginal frequencies are equal. The following is a highly skewed distribution of potential scores that can be obtained when $n = 8$.

a	b	1
c	d	7
1	7	8

	$a = 0$	$a = 1$	M	σ
B_{raw} =	0.163	−8.000	−0.857	2.700
B =	0.378	−2.646	0	1
p =	.875	.125		

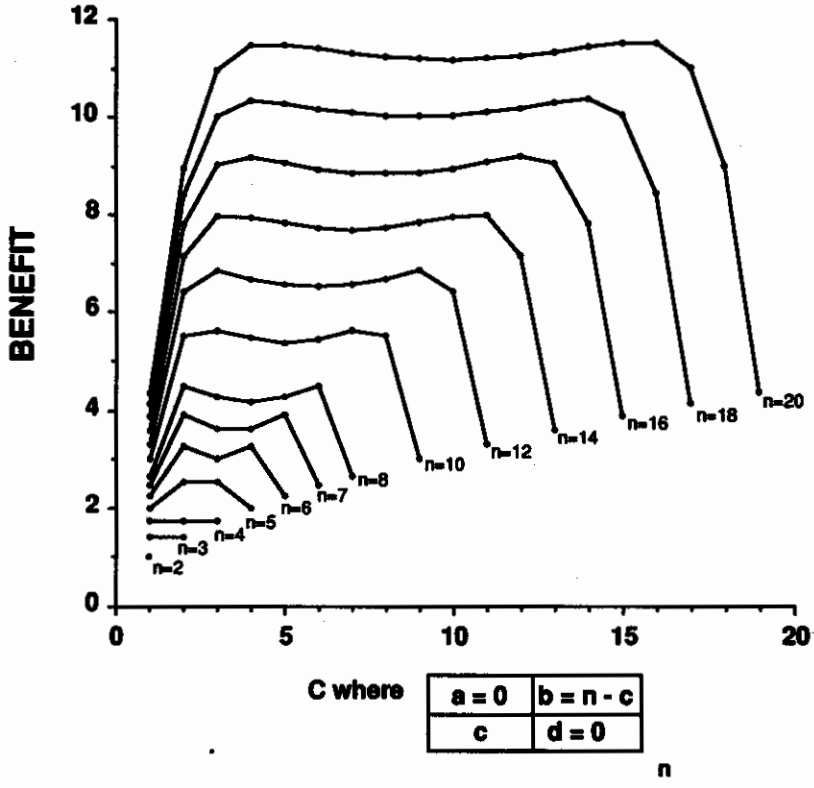

Figure B2. Benefit as a function of c for various values of n under the conditions shown.

Within limits, different values of $(a + c)/n$ can yield benefit/harm scores with similar magnitudes as shown in Figure B2. The comparability of benefit/harm scores from different 2×2 tables can be maintained with a rule of thumb based on Figure B2: Select study design and scoring options for which $(a + c)/n$ is in the range of approximately $.2n$ through $.8n$. For example, a study with $n = 10$ should have a minimum of two occasions on the lowest dose and two occasions on the highest dose. Acceptable values of $(a + c)/n$ can yield benefit/harm scores with magnitudes that are on the relatively flat portion of the curves in Figure B2.

Received March 1, 1990
Revision received June 23, 1991
Accepted August 26, 1991 ■