Drug and Alcohol Review (September 2008), 27, 524–528

informa
healthcare

How large must a treatment effect be before it matters to practitioners? An estimation method and demonstration

WILLIAM R. MILLER & JENNIFER KNAPP MANUEL

Center on Alcoholism, Substance Abuse and Addictions (CASAA), The University of New Mexico, Albuquerque, New Mexico, USA

Abstract

Introduction and Aims. *Treatment research is sometimes criticised as lacking in clinical relevance, and one potential source of this friction is a disconnection between statistical significance and what clinicians regard to be a meaningful difference in outcomes. This report demonstrates a novel methodology for estimating what substance abuse practitioners regard to be clinically important differences.* ***Design and Methods.*** *To illustrate the estimation method, we surveyed 50 substance abuse treatment providers participating in the National Institute on Drug Abuse (NIDA) Clinical Trials Network. Practitioners identified thresholds for clinically meaningful differences on nine common outcome variables, indicated the size of effect that would justify their learning a new treatment method and estimated current outcomes from their services.* ***Results.*** *Clinicians judged a difference between two treatments to be meaningful if outcomes were improved by about 10–12 points on the percentage of patients totally abstaining, arrested for driving while intoxicated, employed or having abnormal liver enzymes. A 5 percentage-point reduction in patient mortality was regarded as clinically significant. On continuous outcome measures (such as percentage of days abstinent or drinks per drinking day), practitioners judged an outcome to be significant when it doubled or halved the base rate. When a new treatment meets such criteria, practitioners were interested in learning it.* ***Discussion and Conclusions.*** *Effects that are statistically significant in clinical trials may be unimpressive to practitioners. Clinicians' judgements of meaningful differences can inform the powering of clinical trials.* [Miller WR, Manuel JK. How large must a treatment effect be before it matters to practitioners? An estimation method and demonstration. *Drug Alcohol Rev* 2008;27:524–528]

Key words: clinical significance, dissemination, effect size, evidence-based practice, treatment.

Introduction

If two treatments, when compared, show a statistically significant difference in outcomes, should clinicians care? The standard of significance testing at $p < 0.05$, long honoured in clinical research [1], has also been roundly criticised [2]. If a study is underpowered, as often happens with small sample sizes, a clinically meaningful difference may be obscured. In large samples, by contrast, relatively small and clinically meaningless effects may pass the standard of $p < 0.05$. This may be one reason for practitioner scepticism regarding the importance and relevance of treatment research to clinical practice.

Consider, for example, the outcome measure of percentage of days abstinent (PDA), a common endpoint in efficacy research (e.g. [3–6]). In the Combined Pharmacotherapies and Behavioural Interventions for Alcohol Dependence (COMBINE) study of treatments for alcohol dependence [5], baseline PDA was 25% (± 25); in other words, patients had been abstaining on 7–8 days and drinking on about 23 days per month on average before treatment. Using Cohen's [7] criteria, a small effect ($d = 0.20$) would increase abstinence from 7 to 9 days per month; with a medium effect ($d = 0.50$) the result would be 11 abstinent days; and a large ($d = 0.80$) effect would produce 15 abstinent days per month. Would clinicians regard it as a treatment success if alcohol-dependent patients drank on only 15 instead of 23 days per month? If not, even large statistical effects in a clinical trial may appear to practitioners to be unimpressive.

This conundrum has led to the reporting of supplemental 'clinical significance' measures, often individual

William R. Miller PhD, Center on Alcoholism, Substance Abuse and Addictions (CASAA), 2650 Yale SE, The University of New Mexico MSC11 6280, Albuquerque, New Mexico, USA 87131-0001, Jennifer Knapp Manuel MS, Center on Alcoholism, Substance Abuse and Addictions (CASAA), 2650 Yale SE, The University of New Mexico MSC11 6280, Albuquerque, New Mexico, USA 87131-0001. Correspondence to William R. Miller PhD, 1706 Notre Dame NE, Albuquerque, NM 87106-1010, USA. Tel: (505) 265 3318. Fax: (505) 925 2379. E-mail: wrmiller@unm.edu

Received 24 October 2007; accepted for publication 7 January 2008.

ISSN 0959-5236 print/ISSN 1465-3362 online/08/050524–05 © Australasian Professional Society on Alcohol and other Drugs
DOI: 10.1080/09595230801956165

patient outcomes reported in a way that is more meaningful to practitioners [7–9]. In Project MATCH [4] and the COMBINE study [5], in addition to the mean values for treatment groups that served as the primary outcome measures, the investigators also reported the outcome status for patients, classifying each individual as totally abstinent, drinking in a moderate and problem-free manner, improved or unimproved [10]. These indices, however, still involve subjectivity in judging significance: if two different treatments yield 12-month total abstinence rates of 21% versus 24%, is that a clinically meaningful difference?

In traditional power analysis, one first estimates an expected effect size (based ideally on previous studies or pilot data) and then computes the sample size needed to detect a statistically significant difference at $p < 0.05$. To detect a small effect, one needs a relatively large and commensurately expensive sample. Even a rather small effect may be clinically important if the treatment can be delivered to large populations at relatively low cost; e.g. taking one low-dose aspirin daily to reduce cardiovascular risk.

With higher-cost interventions, a different approach may be warranted. Behavioural and pharmacotherapies for substance use disorders vary widely in cost of delivery [10–13]. When allocating limited resources for service delivery, how large of a difference matters, warranting a change in practice? One approach would be to ask how large a difference clinicians regard to be meaningful. Studies might then be powered to detect an effect of this size or larger.

Methods

A questionnaire was developed to survey the subjective judgement of substance abuse practitioners regarding meaningful treatment effects (http://casaa.unm.edu/csq.html). For a range of dependent measures, we asked how much better outcomes a treatment would have to produce before the clinicians would (a) regard it as a clinically significant improvement, and (b) be interested in learning the new treatment. We also asked respondents to estimate what outcomes their patients were currently showing on these same measures.

Nine different treatment outcome measures often used in efficacy research were included in the survey. Respondents were asked to complete the first question (clinically meaningful difference) for all nine outcome measures before proceeding to the second question (interest in learning a new treatment), then finally the third question (estimated current outcome for your patients). The nine outcome indices included in this survey were: (a) abstinent cases – the percentage of treated cases who abstained totally from alcohol during the first 12 months following treatment; (b) drinking days – the average percentage of days following treatment on which people drank any alcohol; (c) drinks per drinking day – the amount that people drank on days if they did drink any alcohol; (d) driving while intoxicated (DWI) arrests – the percentage of people arrested for driving while intoxicated during the 5 years following treatment; (e) employment – the percentage of people working full-time in legal employment one year after treatment; (f) liver function – the percentage of people with abnormal elevations of liver enzymes (ALT, AST, GGT) related to drinking; (g) time to first drink – the average number of days after treatment before a person had one or more drinks; (h) all cause mortality – the percentage of people who died from any cause during the 5 years following treatment; and (i) treatment retention – the number of treatment sessions completed.

In order to pose these questions, it was necessary to specify a base-rate level of outcome against which improvement could be judged. For example, the size of a clinically meaningful improvement in the percentage of patients remaining totally abstinent may differ if the base rates against which it is compared were 7% or 67%. Given the length of the survey, it was feasible to ask about only one base rate for each of the nine outcome measures. For four of the measures (a, b, c and g) we were able to use actual average 12-month outcome rates observed across seven US multi-site trials of treatments for alcohol use disorders [3]. For the remaining five we estimated base rates for US public alcohol treatment populations.

The survey was formatted for administration online, and required approximately 15 minutes to complete. To pilot-test this measurement approach, we e-mailed the survey web address to potential participants in March and April of 2006. The population surveyed was all providers ($n = 133$) listed in the current national directory of the National Institute on Drug Abuse Clinical Trials Network (CTN, http://www.nida.nih.gov/CTN/index.htm), with the title of programme director, interventionist, counsellor, psychologist, nurse practitioner, study nurse, substance abuse counsellor, therapist, counselling supervisor, clinical counselling social worker or study physician. The study was reviewed by the University of New Mexico Institutional Review Board and judged to be exempt.

Results

Participants

Fifty providers responded to the survey, who reported a mean of 21 years of experience in treating substance use disorders. At educational level, 28% held doctoral degrees, 40% masters degrees and 32% reported no graduate degree. Almost half (46%) were licenced alcohol and drug counsellors, and 40% indicated that

they were themselves in recovery. More than half (58%) reported that they worked in out-patient substance abuse treatment settings, 27% in in-patient substance abuse treatment programmes, 11% in out-patient mental health and 4% in out-patient research settings. They reported treating an average of 22 substance abuse clients per week, and treating their clients for a mean of 21 sessions. Average typical session length varied widely, from less than 10 minutes (9%) to 10–30 (18%), 45–60 (40%), 60–90 (21%) and more than 90 minutes (12%). Asked about the most common primary drug problem that they treated, 61% named alcohol and 18% named heroin, with smaller percentages indicating cocaine (7%), methamphetamine (7%), cannabis (5%) and other opiates/analgesics (2%). Asked about their primary approach to treatment, 39% endorsed a cognitive-behavioural model and 33% an eclectic approach, with others identifying 12-Step (15%), humanistic (4%), behavioural (2%) and other approaches (7%). Most (85%) reported that they endorsed a disease model of alcoholism.

The respondents were asked how often they read scholarly psychology or addiction journals. Only one (2%) reported never reading journals, with the remainder saying they read journals once or twice a year (4%), three to six times per year (16%), seven to 11 times per year (7%), once a month (13%), two to three times a month (31%), once a week (13%) and more than once a week (13%). When the respondents were asked how interested they were in learning new treatment methods, 64% indicated that they were very interested.

Survey responses

The mean (SD) and median responses of participants are reported in Table 1, along with the base rates against which respondents judged clinically significant levels of improvement. We used one-way analyses of variance to compare clinicians' judgements of current outcomes, clinically meaningful improvement and an effect that would justify learning a new technique. There were significant differences on seven of the nine outcome measures. In all seven cases, *post-hoc* contrasts (Tukey's *t* test) showed no difference between practitioners' judgements of a clinically meaningful difference and a level of improvement that would be sufficient to interest them in learning a new treatment. That is, if a new treatment produced better outcomes that met their standards of a clinically meaningful improvement, practitioners said they would be interested in learning the new treatment method. For the remaining two measures (all-cause mortality and time to first drink), means for the three ratings did not differ significantly. The standard deviations on these measures were relatively large, indicating broad diversity in what practitioners regarded to be a significant improvement.

Clinicians' estimates of typical outcomes from current practices were generally close to (and may have been biased by) the base rates that we provided, although there was more variance from base rates that we estimated rather than drawing them from large practice samples. On the rate of 5-year DWI recividism, for example, we suggested a base rate of 35%, whereas

Table 1. *Practitioners' estimates of current outcomes, clinically significant improvement and threshold for learning a new treatment*

	Base rate	Estimated current average outcome			Clinically significant improvement			Threshold to learn a new treatment		
		Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
% Totally abstinent at 1 year	25	30.1	13.4	25	38.2	8.7	35	38.3	12.1	35
% Drinking days at 1 year	19	26.8	15.3	25	9.6	4.6	10	11.1	4.3	10
Drinks per drinking day at 1 year	7	6.6	3.2	6	2.8	1.5	3	2.9	1.8	3
% DWI arrests/ 5 years	35	5.0	8.3	2	22.7	7.4	25	23.8	8.0	25
% Full-time employed at 1 year	60	51.9	25.0	59	72.8	6.6	71	72.2	7.5	70
% Abnormal liver enzymes	19	23.6	20.8	15	11.5	3.7	11	12.0	4.2	12
Days to first drink	35	49.2	43.4	35	68.0	37.1	56	64.9	39.4	50
% Mortality in 5 years	12	6.5	5.6	5	7.0	2.9	7	7.2	3.0	7
Treatment sessions completed	7	20.7	15.3	15	10.0	1.1	10	10.1	1.2	10

DWI: driving while intoxicated; SD: standard deviation.

practitioners estimated only 5% for their own patients. This is a figure unlikely to be known by most clinicians, and would be influenced by the percentage of clientele who enter treatment as a result of a previous DWI offence. For 5-year mortality, we suggested a base rate of 12% and clinicians estimated 6.5% among their own patients. On treatment retention, we provided a base rate of seven sessions completed, whereas respondents reported completing a mean of 20.7 sessions. On these dimensions, they perceived current practice as already yielding outcomes far better than the base rates we specified.

Discussion

We tested a method for measuring what substance abuse practitioners regard to be meaningful treatment effects. Stated briefly, our clinician respondents opined that the difference between two treatments would be clinically significant if one treatment produced outcomes that were better by about 10 – 12 percentage points for the proportion of patients totally abstaining, arrested for DWI, employed or having abnormal liver enzymes. For patient mortality, the improvement margin judged to be clinically significant was smaller: a reduction of 5 percentage points in death rate. On three continuous outcome measures, practitioners judged a difference to be significant when it doubled or halved the base rate. Finally, our respondents indicated that a significant improvement in retention would be completion of 10 versus the base rate of seven of 12 sessions, a 43% improvement.

Clearly, our data cannot be taken as representative of a larger population of practitioners. Our sample was small, and being drawn from programmes in the Clinical Trials Network it consists of providers with more than average familiarity with treatment outcome research. Some of the base rates that we provided were arbitrary, and changing these anchor points in the survey could yield different judgements about significant levels of improvement. Our primary point is that providers' judgements about clinically meaningful differences are measurable, and can inform the selection of target effect sizes for clinical trials.

Our data suggest that statistically significant between-group differences observed in outcome trials of behavioural and pharmacotherapies may sometimes fall well short of effects that substance abuse practitioners regard to be meaningful. In the COMBINE study [5], for example, naltrexone significantly ($p < 0.02$) reduced return to heavy drinking from 71.4% to 68.2%. In our data, this is about one-quarter of the minimum difference that practitioners regarded to be clinically meaningful. In the same study, however, naltrexone and/or behavioural therapy increased good clinical outcomes by 10 percentage points on average, meeting our respondents' threshold for clinical significance. Also in COMBINE, the post-treatment standard deviation for percentage of days abstinent was 25, so that our respondents' estimate of a clinically meaningful difference (9 percentage points) on this variable would be a relatively modest effect ($d = 0.36$) that larger clinical trials should have adequate power to detect.

Other clinical trials have clearly produced treatment effects that meet these subjective standards of clinical significance. Among out-patients in Project MATCH, 12-Step facilitation therapy yielded total abstinence rates that were 10 percentage points higher throughout 3 years of follow-up, relative to those from cognitive – behavioural therapy [4,5,14]. The community reinforcement and family training (CRAFT) approach for working with family members engaged more than twice as many 'unmotivated' substance users in treatment, relative to Al-Anon facilitation or the Johnson Institute intervention [15]. Some studies observe treatment differences that meet practitioners' standards of clinical significance, but fail to reach statistical significance. For example, a quasi-experimental study found 18% mortality at 2 years among those receiving in-patient alcoholism treatment, compared with 32% among patients referred out while admissions were closed – a risk ratio that fell short of statistical significance [16].

Next steps

A clinically meaningful effect size can be calculated across a wide variety of outcome measures. Although most of the illustrations in this pilot test focused on mean differences, the same approach can be used to estimate clinically significant effects on other benchmarks such as the number needed to treat (NNT) or relative risk of mortality.

Size of effect – the relative advantage of a treatment method – is only one influence on dissemination and adoption of innovations. Diffusion is also promoted by factors such as perceived simplicity, cost and compatibility with current practices [17]. Nevertheless, efforts to encourage the adoption of 'evidence-based' treatments might emphasise those that most reliably produce improvements of a magnitude that providers regard to be clinically meaningful [18]. Our data suggest that when a treatment meets these subjective criteria, practitioners are more motivated to learn it.

The most direct application of our approach would be to power clinical trials to detect clinically meaningful differences in outcome, rather than the more usual approach of estimating the probable effect size and constructing a sample that would render this effect statistically significant. Undertaking power calculations on the basis of effects considered to be clinically significant could result in trials with smaller sample sizes that are therefore less costly to complete. Perhaps

a greater value, however, is that this approach is likely to generate useful discussion of the desired and intended outcomes of prevention and treatment interventions, rather than simply relying on mean differences on standard metrics. The priority given to various outcomes will differ across audiences. This study involved addiction treatment providers and focused primarily on reduction in substance use. Other social agents might place higher value on reduction in crime, lost work days or health care costs. The value of an intervention is, to this extent, in the eye of the beholder. 'Significance' has been given increasing emphasis as a review criterion for grant proposals, and this estimation method offers an empirical basis for evidence of significance, beyond the subjective arguments of investigators or reviewers. Starting with estimation of desired effects requires an explicit, up-front consideration of the intended benefits of the treatments being tested, and thereby of the clinical research itself.

Acknowledgements

This work was supported in part by grants U10-DA015833 from the National Institute on Drug Abuse and no. 049533 from the Robert Wood Johnson Foundation.

References

[1] Cowles M, Davis C. On the origins of the .05 level of statistical significance. Am Psychol 1982;37:553–8.
[2] Cohen J. The earth is round ($p < .05$). Am Psychol 1994;49:997–1003.
[3] Miller WR, Walters ST, Bennett ME. How effective is alcoholism treatment in the United States? J Stud Alcohol 2001;62:211–20.
[4] Project MATCH Research Group. Matching alcoholism treatments to client heterogeneity: Project MATCH post-treatment drinking outcomes. J Stud Alcohol 1997;58:7–29.
[5] Anton RF, O'Malley SS, Ciraulo DA, *et al*. Combined pharmacotherapies and behavioral interventions for alcohol dependence. The COMBINE study: A randomized controlled trial. JAMA 2006;295:2003–17.
[6] Carroll KM, Ball SA, Nich C, *et al*. Motivational interviewing to improve treatment engagement and outcome in individuals seeking treatment for substance abuse: A multisite effectiveness study. Drug Alcohol Depend 2006;81:301–12.
[7] Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
[8] Atkins DC, McGlinchey JB, Beauchaine TP. Assessing clinical significance: Does it matter which method we use? J Consult Clin Psychol 2005;73:982–9.
[9] Jacobson NS, Truax, P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59:12–19.
[10] Zweben A, Cisler R. Composite outcome measures in alcoholism treatment research: Problems and potentialities, Subst Use Misuse 1996;31:1783–805.
[11] Finney JW, Monahan SC. The cost-effectiveness of treatment for alcoholism: A second approximation. J Stud Alcohol 1996;57:229–43.
[12] Holder HD, Longabaugh R, Miller WR, Rubonis AV. The cost effectiveness of treatment for alcoholism: A first approximation. J Stud Alcohol 1991;52:517–40.
[13] Holder HD, Cisler RA, Longabaugh R, *et al*. Alcoholism treatment and medical care costs from Project MATCH. Addiction 2000;95:999–1013.
[14] Project MATCH Research Group. Matching alcoholism treatments to client heterogeneity: Project MATCH three-year drinking outcomes. Alcohol Clin Exp Res 1998;22:1300–11.
[15] Miller WR, Meyers RJ, Tonigan JS. Engaging the unmotivated in treatment for alcohol problems: A comparison of three strategies for intervention through family members. J Consult Clin Psychol 1999;67:688–97.
[16] Willenbring ML, Olson DH, Bielinski J. Integrated outpatient treatment for medically ill alcoholic men: Results from a quasi-experimental study. J Stud Alcohol 1995;56:337–43.
[17] Rogers EM. Diffusion of innovations, 5th edn. New York: Free Press, 2003.
[18] Miller WR, Sorensen JL, Selzer J, Brigham G. Disseminating evidence-based practices in substance abuse treatment: A review with suggestions. J Subst Abuse Treat 2006;31:25–39.

Copyright of Drug & Alcohol Review is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.