Addiction

FOR DEBATE

doi:10.1111/j.1360-0443.2011.03473.x

Primary outcome indices in illicit drug dependence treatment research: systematic approach to selection and measurement of drug use end-points in clinical trials

Dennis M. Donovan[1], George E. Bigelow[2], Gregory S. Brigham[3,13], Kathleen M. Carroll[4], Allan J. Cohen[5], John G. Gardin[6], John A. Hamilton[7], Marilyn A. Huestis[8], John R. Hughes[9], Robert Lindblad[10], G. Alan Marlatt[11]*, Kenzie L. Preston[8], Jeffrey A. Selzer[12], Eugene C. Somoza[13], Paul G. Wakim[14] & Elizabeth A. Wells[15]

Alcohol and Drug Abuse Institute and Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA,[1] Behavioral Pharmacology Research Unit, Johns Hopkins University School of Medicine, Baltimore, MD, USA,[2] Maryhaven, Columbus, OH, USA,[3] Division of Addiction, Department of Psychiatry, Yale University School of Medicine, West Haven, CT, USA,[4] Bay Area Addiction Research and Treatment, Sherman Oaks, CA, USA,[5] ADAPT, Inc., Roseburg, OR, USA,[6] Regional Network of Programs, Inc., Shelton, CT, USA,[7] National Institute on Drug Abuse, Intramural Research Program, Baltimore, MD, USA,[8] Department of Psychiatry, University of Vermont, Burlington, VT, USA,[9] The EMMES Corporation, Rockville, MD,[10] Addictive Behaviors Research Center, Department of Psychology, University of Washington, Seattle, WA, USA,[11] Committee for Physician Health, Albany, NY, USA,[12] Department of Psychiatry and Behavioral Neuroscience University of Cincinnati, Cincinnati, OH, USA,[13] National Institute on Drug Abuse, Center for the Clinical Trials Network, Bethesda, MD, USA[14]and School of Social Work, University of Washington, Seattle, WA, USA[15]

ABSTRACT

Aims Clinical trials test the safety and efficacy of behavioral and pharmacological interventions in drug-dependent individuals. However, there is no consensus about the most appropriate outcome(s) to consider in determining treatment efficacy or on the most appropriate methods for assessing selected outcome(s). We summarize the discussion and recommendations of treatment and research experts, convened by the US National Institute on Drug Abuse, to select appropriate primary outcomes for drug dependence treatment clinical trials, and in particular the feasibility of selecting a common outcome to be included in all or most trials. **Methods** A brief history of outcomes employed in prior drug dependence treatment research, incorporating perspectives from tobacco and alcohol research, is included. The relative merits and limitations of focusing on drug-taking behavior, as measured by self-report and qualitative or quantitative biological markers, are evaluated. **Results** Drug-taking behavior, measured ideally by a combination of self-report and biological indicators, is seen as the most appropriate proximal primary outcome in drug dependence treatment clinical trials. **Conclusions** We conclude that the most appropriate outcome will vary as a function of salient variables inherent in the clinical trial, such as the type of intervention, its target, treatment goals (e.g. abstinence or reduction of use) and the perspective being taken (e.g. researcher, clinical program, patient, society). It is recommended that a decision process, based on such trial variables, be developed to guide the selection of primary and secondary outcomes as well as the methods to assess them.

Keywords Clinical trials, drug dependence, end-points, primary outcome, self-report, toxicology, treatment research.

Correspondence to: Dennis M. Donovan, Alcohol and Drug Abuse Institute and Department of Psychiatry and Behavioral Sciences, University of Washington, 1107 NE 45th Street, Suite 120, Seattle, WA 98105-4631, USA. E-mail: ddonovan@u.washington.edu

Submitted 15 December 2010; initial review completed 15 March 2011; final version accepted 13 April 2011

INTRODUCTION

Despite the development of evidence-based behavioral and pharmacological interventions, the use of and dependence upon illicit drugs and prescription medications continue to be major public health concerns. Clinical trials contributed the empirical basis to support these interventions [1]; however, there is a lack of consensus

*Deceased, 14 March 2011.

among clinical researchers and practitioners about what outcomes are most important and how best to define the efficacy of drug dependence treatment. Further, given the interdependence between outcomes and methods of assessment, a related question also still lacking consensus is what measures or methods best assess those outcomes. Drug use is measured variably, including self-report, collateral report and qualitative or quantitative assessment of a variety of biological markers, with different measures potentially yielding different results, thus making cross-study comparisons difficult [2,3]. This is the case not only in the United States but also internationally [4,5]. While drug-taking behavior is the cardinal feature of drug dependence and the primary target for interventions, it is not the only outcome of interest, with other important biopsychosocial dimensions to consider [6–11]. This dilemma is long-standing, as exemplified in the seminal paper by Dole & Nyswander [12] on methadone as a treatment for opiate dependence. They pointed to methadone's positive benefits of relief from 'drug hunger' and blockade of heroin's euphoria, yielding an overall reduction in opiate use, but also noted improved functioning in educational, employment and familial arenas. This is consistent with the three key outcome domains measured in virtually every published evaluation of addiction treatment since the 1960s, namely substance use, employment/self-support and criminal activity [8].

An important factor contributing to this lack of consensus is the differing perspectives of the investigator, practitioner, drug user or policymaker. Researchers may be interested primarily in the impact of interventions on drug use, regardless of how measured. Clinicians and program administrators are typically interested in a broader array of outcomes, defined as clinically meaningful change [13]. From the client's perspective, the experience of reduced symptom severity and improved quality of life are important [14,15]. Policymakers, third-party payers and society more generally expect treatment to yield improvements in multiple areas of psychosocial function, increased public safety and decreased financial burden [8].

A second factor in determining drug treatment outcomes is whether an intervention focuses on a single or multiple drugs, reflecting the purported specificity versus generalizability of treatment effects [16]. The stated intervention goal with respect to substance use, namely abstinence or reduced use, harm and/or consequences, also has an impact on how a treatment is appraised [17,18]. The European College of Neuropsychopharmacology's consensus panel on efficacy of interventions [4] distinguishes between trials that focus on either full recovery ('cure') versus drug use stabilization and harm minimization ('care') as the main treatment goal. Type of intervention, behavioral or pharmacological, may also influence the potential therapeutic target and the associated outcome [19]. This latter factor also may influence outcome assessment timing. Some interventions, directed at the immediate effects on drug use, have a relatively short time-frame, while others focusing on improved psychosocial dimensions of recovery usually have a much longer perspective. Differences in outcome assessment timing result in different periods of relapse risk and resumed use [3], suggesting different trial efficacies.

Continued research focusing on determining appropriate outcomes and the standardization of measurement across drug dependence treatment clinical trials has been recommended [2]. This includes further examination of the properties and benefits of alternative and potentially complementary substance use/drug-taking, medical and psychosocial outcomes; different approaches to their measurement; and methods of evaluating the timing and sequencing of assessments consistent with the natural course of recovery and relapse processes. Several previous expert panels, both in the United States [9,20,21] and Europe [4], were unable to reach consensus on these issues.

This paper summarizes the recommendations of a panel of substance abuse treatment and research experts (see the list of participants in Appendix I), convened by the US National Institute on Drug Abuse (NIDA) in December 2009 to select an appropriate primary outcome measure for drug dependence treatment clinical trials, and in particular to consider the feasibility of selecting a common outcome measure to be included in all or most drug dependence treatment trials. First we provide a brief overview of perspectives concerning outcomes used in treatment trials with tobacco and alcohol; we next summarize the state of the science regarding toxicology and self-report measures of drug use; and finally we summarize the discussion and recommendations from the panel's deliberation.

PERSPECTIVES ON PRIMARY OUTCOMES FROM TRIALS WITH LEGAL DRUGS: TOBACCO AND ALCOHOL

Outcome data for interventions targeting the two most prevalent legal substances of abuse, tobacco and alcohol, were considered.

Tobacco experience

Work-groups from the Society for Research on Nicotine and Tobacco and others [22–24] have published relevant recommendations on measuring craving and withdrawal [25], self-report criteria for abstinence in clinical trials [26] and biochemical verification of abstinence [27].

Nicotine dependence differs from other drug dependencies in that it does not cause urgent psychosocial problems, yet even low levels of use are harmful [28]. Therefore, neither temporary abstinence nor drug use reduction is an acceptable goal; the focus is on long-term abstinence as the primary treatment outcome. Most trials enter current smokers, establish quit dates and tie outcome assessments to that date. Three processes are measured typically: abstinence initiation within 24 hours, onset of a smoking lapse and transition from lapse to relapse (return to regular smoking), as interventions have different effects on each of these outcomes [29]. The major goal of abstinence is measured typically in one of four ways [26,30]. Continuous abstinence (CA) is abstinence that begins on the quit date and continues till the last assessment. Prolonged abstinence (PA) is continuous abstinence with a short (usually 2 weeks) grace period after the quit date. Point prevalence abstinence (PPA) is a short abstinence (usually 7 days) period immediately preceding an assessment time-point. Repeated point prevalence abstinence (RPPA) is PPA at repeated follow-ups. The major advantages of CA, PA and RPPA are requiring longer abstinence and better prediction of the ultimate goal—life-long abstinence or health benefits. PA, and especially CA, are more closely tied to and influenced by treatment. Furthermore, PPA and PA count the common occurrence of some initial lapses with later complete abstinence as successes, and they can capture this effect even if a treatment has a delayed effect (e.g. a blocking agent). Finally, only PPA can be verified biochemically [27]. These four measures are correlated highly, with one no more valid than the other [31,32].

Falsification of abstinence self-reports is substantial in trials with face-to-face treatments, but minimal in less intensive treatments. Biochemical verification provides additional assurance that the participant's self-reports are accurate [27,30]. Self-reports of amount of smoking (e.g. number of cigarettes smoked) collected via the timeline follow-back (TLFB) have poor validity [33], due in part to digit bias [34]. Observer verification also appears to add little [35].

Alcohol experience

Unlike the situation described for smoking cessation, alcohol consumption within certain limits is not viewed as harmful, and in some cases moderate alcohol intake may be health-enhancing. The National Institute on Alcohol Abuse and Alcoholism (NIAAA) published guidelines and limits/thresholds that specify low-, moderate- and high-risk drinking based on number of standard drinks consumed per day and per week by men and women [36,37]. While guidelines such as these and those of the British National Health Service [38] and the World Health Organization [39] are helpful clinically, their utility as outcome measures is complicated by differing sizes of beverage containers, varying alcohol concentrations and individuals' difficulties in judging the size of their drinks accurately [40–43]. This has led to a strong recommendation that a common method for reporting alcohol consumption be adopted internationally [43].

The degree to which an individual's drinking is considered harmful, based on exceeding thresholds and associated negative consequences, and whether criteria for an alcohol use disorder are met, influence intervention goals. These may range from reduced consumption and harm-reduction for high-risk drinkers or alcohol abuse (e.g. college binge-drinking) to total abstinence for alcohol-dependent individuals [17,18].

Given this range of interventions and goals, there is a lack of consensus about best outcome(s) for clinical trials [44]. Besides abstinence, still important in alcoholism treatment trials, there is an increased array of outcomes, including continuous measures such as percentage of days abstinent, drinks per drinking day, number of heavy drinking days, presence of alcohol-related problems and dependence symptoms and biological markers of drinking or heavy drinking to verify self-reports, or integrated with self-reports for a combined drinking index [45–47].

Some trials employed empirically based composite indicators that integrate aspects of both alcohol consumption and alcohol-related problems, allowing classification of outcomes as abstinent, moderate drinking without problems, heavy drinking or problems or heavy drinking and problems [48]. Such an approach accommodates both abstinence-oriented and harm-reduction interventions. Outcome selection may vary based on the intervention's mechanism of action [49]. A NIAAA expert panel selected days of heavy drinking (at least four or five drinks per drinking day for women and men, respectively) as the 'optimal' outcome for clinical trials, as assessed by drinking estimation methods such as the TLFB procedure [50,51]. More recently, the US Food and Drug Administration (FDA) promoted percentage of subjects with no heavy drinking days as an end-point in pharmacotherapy trials for alcohol dependence; individuals who are abstinent or who engage in low-risk drinking are considered successful treatment responders [45].

While there is a pronounced difference in the relative weight given to biological indicators and self-report of use in tobacco and alcohol research, substance-taking behavior is the primary outcome measure for tobacco and alcohol treatment trials. Researchers acknowledge the importance of psychosocial and physical consequences of smoking and drinking and endorse potential outcomes in these domains, but the primary and more proximal treatment focus is to reduce or eliminate substance use [45]. Issues concerning methods providing the most valid indicator of use and endorsement of substance-taking behavior as the defining outcome in clinical trials apply equally well to drug dependence treatment.

BIOLOGICAL MONITORING OF ILLICIT DRUG USE AS A PRIMARY OUTCOME

Drug exposure detection has traditionally utilized urine testing; alternative matrices including oral fluid, sweat and hair are now available. Advantages of urine testing include adequate specimen volume, on-site and laboratory-based test availability, relatively high drug concentrations, proven accuracy and reliability, easy automation, low-cost, well-established quality control programs and a large scientific database for interpretation. Disadvantages include the need for private collection facilities and same-sex collectors for observed collections, short detection windows and ease of test adulteration (addition of chemicals, or even high fluid consumption, reduce sensitivity). There are multiple advantages for alternative matrix testing, as each offers unique information about the participants' drug use. Specimen collection may be less invasive and less subject to adulteration, may improve drug stability, may provide opportunities for multiple samples and may lower the risk of disease during handling and analysis. Furthermore, alternative matrices offer a choice in drug detection windows, frequently measure parent drug concentrations that may correlate more effectively with drug effects and permit easier shipment and storage. If parent drug and metabolites are present, test interpretation is frequently improved. There also are disadvantages associated with each matrix, including cost, turnaround time, adsorption of drug to collection devices, extensive specimen preparation and limited controlled drug administration data to aid in the interpretation of test results.

Urine tests reflect drug use ranging from 12 hours for alcohol [52] to typically 2–4 days for other drugs of abuse [53–56]. An important exception is cannabis; chronic daily smoking may result in positive tests for an extended time, creating a situation where it is difficult to differentiate new drug use from residual drug excretion [57]. If urine cannabinoid concentrations are normalized to urine creatinine concentrations, noting the time between urine specimen collections, predictive models can be employed to estimate new cannabis use [57,58]. A recently published model differentiates for the first time new cannabis exposure from residual drug excretion in chronic, daily cannabis smokers, taking into consideration cannabinoid concentrations at the time of treatment admission and variable times between urine collections [59]. Models have also been developed to identify new cocaine exposure [60]. Urine monitoring usually requires a minimum of twice- or thrice-weekly collection to identify new drug use adequately. Alcohol's detection window in urine may be extended to 5–7 days by quantifying the non-oxidative metabolites ethyl glucuronide and ethyl sulfate [52], although caution in test interpretation is encouraged due to positive results with unintended alcohol exposure [61].

Drug testing is available for three additional matrices, oral fluid (or saliva), sweat and hair, although additional research is needed to document the utility of each of these as an outcome in clinical trials. Each of these has advantages and disadvantages relative to urine testing and to each other. The choice of which biological measure to employ should be determined primarily by the desired time-frame of detection.

Oral fluid testing is increasing rapidly due to interest in drug treatment, work-place drug testing and drugged driving programs. Presence of drug in oral fluid documents drug exposure, particularly recent drug use, and also may correlate more effectively with blood concentrations. A short detection time may be useful if recent drug use is the focus of monitoring, but could require more frequent testing to ensure sustained abstinence. On-site and laboratory-based screening and confirmation tests for oral fluid are becoming more available and assay costs are reasonable due to competition; however, on-site tests to date have not achieved adequate sensitivity and specificity. Unfortunately, many manufacturers have simply tried to modify urine assays for the new matrix. Drug metabolites are found primarily in urine, while oral fluid generally contains both parent drug and metabolites. Drug concentrations may also be much lower in oral fluid than in urine. In many cases, assays have not been modified adequately to target different analyte profiles and lower cut-offs.

On-site oral fluid tests are available for many drug classes and are being developed and improved for many other drugs. The advantages of these tests are that oral fluid is collected non-invasively and under direct observation without the need for specialized facilities. It is preferable to utilize collection devices without chemicals to stimulate saliva flow, because stimulation increases oral fluid pH, changing the distribution of drug between plasma and saliva. Stimulation also increases the amount of fluid, diluting drug concentrations and reducing sensitivity. Many drugs, including cocaine and sympathomimetic amines, reduce saliva production, making specimen collection more difficult. Testing for Δ9-tetrahydrocannabinol (THC) can also be problematic due to passive contamination and contamination of the oral mucosa during smoking, not through diffusion of drug from the blood, and because the highly lipophilic THC may be tightly adsorbed to the specimen collection device, making it difficult to elute and greatly reducing test sensitivity. Oral fluid monitoring requires a minimum of twice- or thrice-weekly collection to adequately identify new drug use. Additional research is needed to exclude the possibility of passive contamination of oral fluid from smoked and/or oral drugs.

Sweat testing is a more recent and less investigated method of monitoring drug use. The primary advantage is that the sweat patch is worn for 1 week, and accumulates drug excreted throughout the week. Drugs in sweat can also reflect use as much as 24–48 hours prior to patch application. Both parent drug and metabolites are excreted. Thrice-weekly urine and weekly sweat specimens from participants in a methadone maintenance treatment program were tested for opiates and cocaine [55,62]. Weekly sweat testing was equivalent or better than thrice-weekly urine tests in identifying cocaine and opiate drug use. Another advantage of sweat testing is that it decreases the opportunity for adulteration. Each sweat patch has a unique identification number and will not adhere to the skin if removed; punctures are also readily visible. Limitations to sweat testing include intra- and inter-subject variability in sweat production, low analyte concentrations, occasional skin sensitivity, failure of the patch to adhere to the skin and possible contamination of the patch from the environment and during specimen handling. Some portion of the drug also may be reabsorbed into the skin, degraded on the patch and/or escape through the semi-permeable membrane. Sweat patch testing identifies drug use but is not quantitative. As is the case for oral fluid, this new technology has few scientific data from controlled drug administration studies to guide the interpretation of test results. There does not appear to be a dose–concentration relationship and the question of residual excretion of drug in heavy chronic users long after last use has not been resolved fully [63]. In addition, only a single laboratory in the United States is offering routine testing of the sweat patch and the cost has been estimated to be 1.7 times more than for urine testing for cocaine [64].

Hair is another alternative matrix that may be helpful in evaluating drug use. The primary advantage of hair testing is the long window of drug detection, although that is dependent upon hair length. For example, several studies found that hair analysis identified more cocaine use than urine tests [65]. Three centimeters of hair can be collected every 3 months to detect drug use efficiently over this time-frame. Although the cost of hair testing is much higher than urine testing, the number of specimens required, visits for hair collection and staff time to collect specimens is greatly reduced. Other advantages are the stability of drug analytes in hair at room temperature and the resultant ease of storage, handling and shipping of specimens. If a repeat specimen is required, a new specimen that reflects the original time of sampling is easy to collect. An individual cannot abstain from drug use for a short period of time prior to hair collection and avoid detection, as can be the case with urine, oral fluid and sweat. Further, adulteration of hair by bleaching, dyeing or straightening is easily apparent.

A limitation of hair testing is the differential incorporation of basic drugs into hair according to its melanin content [66,67]. Darker hair contains more melanin than lighter-colored hair and will most probably contain greater concentrations of basic drugs, such as cocaine or methamphetamine, if exposed to the same amount of drug. This complicates hair test interpretation. In the future, we may find that normalization of basic drug concentrations to melanin concentrations in hair will reduce this apparent discrepancy [68]. Neutral and acidic drugs, i.e. THC, do not appear to bind preferentially to melanin and may have less variable disposition into hair of different colors and with different melanin content. Hair testing is a sensitive technology to detect basic drugs, such as cocaine, but lacks sensitivity to detect cannabis use compared to urine tests. Ethyl glucuronide, unlike alcohol, is incorporated into hair, suggesting that hair could be a valuable matrix for alcohol testing [69].

A second major limitation of hair testing is potential contamination by drugs in the environment [70,71]. Whether it is possible to differentiate contamination from actual drug use remains a controversial subject. Furthermore, highly sensitive tandem mass spectrometry methods are frequently required. Other limitations of hair testing are that recent drug use over the past 10 days may not be detected, there may be a high refusal rate for hair sampling and frequently too little specimen is collected.

SELF-REPORT OF ILLICIT DRUG USE AS A PRIMARY OUTCOME

Monitoring drug use via self-report can be flexible and provide a range of data and outcome measures that are sensitive to changes in patterns or intensity of substance use. Self-reports are not invasive and can be collected remotely (i.e. via interviews, over the telephone, through direct entry on questionnaires or computer screens or over the internet) and in a wide variety of formats. Finally, self-reports may be collected retrospectively over comparatively long periods of time and minimize missing data (as participants who miss assessment visits can provide data later covering the missing time-periods).

The accuracy of self-reported substance use in clinical trials remains highly controversial, with some studies pointing to impressive reliability, validity, sensitivity and consistency with other indicators [72–77] and others suggesting poor accuracy and substantial under-reporting with respect to biological measures

[64,78–81]. A key consideration is that the reliability and validity of substance users' self-reports are not fixed properties of the reports themselves or the data collection instruments; rather, these vary with sample and the method and context of collection.

Self-reports can be accurate given appropriate context and investigators' use of methods to enhance accuracy. Factors and procedures to enhance accuracy of self-report among drug users in clinical settings and in clinical trials include assurance of confidentiality and absence of adverse consequences, use of appropriate recall cues and 'bogus pipeline' techniques (e.g. participants are convinced that their self-reported values will be verified by biomarkers or other types of measures), clarifying how data are used, collection from multiple sources (including biological indicators and collateral information) and standardized, consistent and clear instructions and procedures [82,83]. Factors reducing accuracy include significant consequences (positive or negative) of reporting substance use, lack of confidentiality or collection of self-report data with individuals who are cognitively impaired, have significant psychiatric comorbidity or are under the influence of drugs or alcohol, and collection in a non-clinical or non-research context [84]. Extensive reliability and validity data have been collected on the TLFB technique, a calendar-based interview collecting day-by-day use information over the past 60 days [50,72,85]. Variants of the TLFB approach with substantial psychometric support include the Form 90 [86–90] and the Substance Use Calendar [91] that incorporates a strategy for comparing self-report to urine results.

INTEGRATING SELF-REPORT AND TOXICOLOGY

Self-report methods and toxicology methods for assessing drug-taking behavior each have both strengths and weaknesses. Neither is a direct measure, but each is an indicator of drug-taking behavior. The strengths and weaknesses of each approach are summarized in Table 1. Although monitoring of drug use through urinalysis and other biological indicators is an important strategy of assessing recent drug use and monitoring treatment response in clinical trials, evaluation of treatment outcome is complex, and there are multiple considerations that often lead investigators to use and rely primarily on self-reports in treatment outcome research. These include greater flexibility and range of data and outcome measures than can be obtained via biological indicators which, as described earlier, tend to be limited to detection of relatively recent substance use, and provide limited information regarding anything other than point-prevalence abstinence unless they are collected frequently. Qualitative biological measures are usually insensitive to issues such as changes in patterns or intensity of substance use, as well as detection of significant reductions in drug use over time. Secondly, biological indicators carry with them costs of collection and assays, while collection of self-report data is comparatively less expensive [82]. Collection of biological samples is generally invasive and requires in-person contact with the participant, unlike self-report. Finally, as noted above, self-reports may be collected retrospectively over comparatively long periods of time and minimize missing data, whereas appropriate 'windows' for biological data collection are short and fixed.

One important factor in determining the extent of agreement between self-report and biological markers is the level of drug use. Agreement is greatest at the extremes of drug use frequency—either high or low levels of use. In addition, the measures will have better agreement when the time-frame covered by the questions is the same as the biological window for ascertainment. Other factors influencing agreement are the drug and the

Table 1 Summary of strengths and limitations of self-report and toxicology approaches to assessing drug-taking behaviour.

Method of assessment	*Strengths*	*Limitations*
Self-report	• Convenient, inexpensive • Usually good validity • Can provide temporal and quantitative detail • Missing data potentially retrievable at a subsequent visit	• Uncertain validity • Risk of willful or accidental distortion
Toxicology[a]	• Objective data • Adequate specimen volume • On-site and laboratory-based test availability	• Inconvenience, expense • Poor quantity/frequency resolution • Utility depends on frequency of sample collection • Poor sensitivity to reduced but continued use • Missing data permanently lost

[a] Each potential substrate for toxicology testing (urine, saliva, hair, breath, sweat, cuticles, etc.) has its own method-specific strengths and weaknesses, which vary depending upon the drug being assessed.

purpose and timing of biological measurements, for example at treatment intake, when motivation to report accurately is thought to be higher than at follow-up [92]. Disagreement between self-report and biological measures can occur when the specimen is collected outside the window of drug detection, when there is poor recall or deliberate misreporting or when the analytical method is not sensitive enough to detect drug use. There is a strong consensus that multiple assessments, including self-report and biological testing, yield the most accurate drug use information.

The panel proposed an outcome measure that combines results from both self-report and biological testing; a positive self-report or toxicology test indicates drug use during an assessment period. While such a measure can be used to determine a dichotomous abstinence/non-abstinence outcome, it also can accommodate reduction in the number or percent of days of use. Also in this context, it should be noted that periods of brief abstinence detected by toxicology testing or self-report are acknowledged as beneficial and are consistent with 'improvement' measures, but not by measures requiring periods of sustained abstinence. Strategies that support accurate self-report and combine self-reports with biological indicators have a number of advantages. Although the field has not yet achieved consensus on an ideal strategy for combining self-report and biological data [75], efforts towards that goal are being made, most notably for cocaine studies. The current algorithm used for this purpose in cocaine clinical trials was developed by the NIDA Division of Pharmacotherapy and Medical Consequences of Substance Abuse in conjunction with the College on Problems of Drug Dependence as a result of a consensus meeting held in 1999 [21]. It combines self-report based on the TLFB, quantitative urine benzoylecgonine levels and an estimate of the concordance between the two to determine the cocaine-use status of each study day. The resulting primary outcome variable, which has been used in a large percentage of cocaine clinical trials funded by this division of NIDA over the past 10 years, is called 'the weekly fraction of cocaine non-use days' [93]. As a more recent example, in 2010 NIDA's National Drug Abuse Treatment Clinical Trials Network (CTN) Treatment Effect and Assessment Measures (TEAM) Task Force (http://ctndisseminationlibrary.org/display/522.htm) recommended to use as primary outcome in CTN trials the number of days of drug use during the last 30 days of the active treatment phase, based on self-report corroborated by qualitative urine drug screening tests. Typically, a positive toxicology result overrules self-reported abstinence for the period covered by the toxicology procedure. Toxicology testing is also sometimes the only outcome, ignoring self-report. However, combining the objectivity of toxicology with the reduced data loss, the wider assessment window and the continuous data associated with self-reports provides a composite measure that has advantages over either of its components [93].

Such an approach works best when both the self-report and toxicology are collected at frequent intervals—for example, self-reports for each day with toxicology two or three times per week. For commonly abused drugs such as cocaine and short-acting opioids for which toxicology is sensitive for only 2–3 days after use, this provides a degree of temporal precision for each indicator that makes it relatively easy to combine them into a single classification. For example, each day can be categorized as positive or negative for drug use based on self-report, with a positive toxicology resulting in each of the past 2 or 3 days being categorized as positive for drug use regardless of the self-report. It is more complicated when toxicology testing is infrequent or when toxicology may remain drug-positive for a long duration after drug use as for cannabis, and there are no established guidelines for integrating self-report and toxicology data in such cases. Nevertheless, the panel recommended that clinical trials of drug dependence treatment should routinely collect both self-report and toxicology data at whatever greatest frequency is practical for the study design, patient participation logistics and available budget. While the requirement of such frequent assessments may provide more accurate measures of substance use, it may also restrict the individuals who are willing and able to enroll and remain in a trial having a rigorous data collection schedule, may result in a measurement effect that reduces the effect of the intervention relative to the comparison group [94] and may have an impact on external validity and generalizability of findings. However, such a limitation may be necessary in an efficacy trial to maximize internal validity and in a randomized trial the effects of repeated assessment should be constant across groups. Subsequent effectiveness trials, with less frequent assessments and greater external validity, would be necessary to determine treatment effectiveness and generalizability.

It is also important that self-report data be collected independently of toxicology data—i.e. that patients provide their self-reports before being informed of toxicology results. It is possible that additional information may be gained by repeating the self-report assessment after the toxicology data are available and presented to the patient, but the value of doing this is uncertain.

NO SINGLE CLINICAL OUTCOME METRIC

The panel reached consensus that the primary outcome measure should be an indicator of drug-taking behavior, and that there is no single clinical metric that is appropriate for inclusion in most drug dependence treatment trials. The panel noted that a wide variety of indicators are available, that the most appropriate one might vary by study methods and goals, that a decision process for determining an appropriate primary outcome would be useful and that there is probably substantial correlation between all the commonly used indicators, although this latter supposition is one on which data are needed for verification.

The panel noted that the array of outcome measures reported in contemporary studies remains as diverse as in reviews by Wells and colleagues [2] from more than 20 years ago. The most common self-report methods are questions about number of days used during the past 30 days from the Addiction Severity Index or the Maudsley Addiction Profile, and the TLFB procedure's categorization of each past day as a day of use or not.

Outcome measures used commonly in these studies include: percentage of days used, percentage of days abstinent, number of abstinent visits (a composite of abstinence and retention), percentage of positive (or negative) toxicology tests, longest duration of continuous abstinence and percentage of patients achieving abstinence of 'x' duration (often 2 or 3 weeks), among others. Studies vary in whether missing toxicology samples are treated as missing (unknown) or imputed to be drug-positive. Imputation of missing samples as drug-positive is quite common practice, but can lead to quite implausible conclusions when used indiscriminately; unfortunately, it is not clear where the boundary lies between appropriate and inappropriate imputation.

FACTORS TO CONSIDER IN A DECISION PROCESS FOR DETERMINING APPROPRIATE OUTCOME MEASURES

A number of questions to be considered as aids in the process of selecting an appropriate primary outcome variable for a specific study were proposed. These could be seen as key elements in a decision process for outcome selection:

- What specific type of drug of abuse is being studied?
- What is known about the pattern of use and temporal course of this particular drug?
- What toxicological options are available for assessing drug-taking?
- What are the strengths, weaknesses and precision of toxicology methods?
- Will patients enter the trial abstinent or as ongoing active users?
- Does the intervention's mechanism suggest appropriate outcome measures?
- How long are the treatment effects and outcomes expected to last—only during active treatment delivery or beyond? If beyond, what is the length of expected benefit beyond treatment to demonstrate durability?
- To what audience is this trial directed?
- Is reduced use of the target drug an acceptable benefit/outcome, or only abstinence?
- Are alcohol and other drugs being assessed to determine what impact reduction in the target drug has on other substance use, either contributing to a decrease, increase, or substitution effect?
- Are indirect/surrogate indices such as desire/intention/craving appropriate?
- Must the trial's outcome measure satisfy the FDA criterion of 'success'?

Figure 1 presents a graphic depiction of this process, indicating that the specific primary outcome for a given trial may be best determined in consideration of factors such as those listed above and, as such, the specific parameters of drug use behavior to be assessed, the specific biological matrix to be employed and the time-frames for assessments will vary across trials. Table 2 presents a number of examples of this trial-to-trial variability in primary outcomes as a function of trial-specific factors in both behavioral and pharmacological interventions, as well as providing an overview of how the factors presented above were operationalized. While these are five representative protocols conducted in the NIDA Clinical Trials Network [95–99], the principles involved in the outcome selection process are applicable to clinical trials conducted in a number of countries. As can be seen, while all five protocols include drug-taking behavior as the primary or a co-primary outcome, there is considerable variation across trials. This reflects the difficulty in recommending a single common outcome for clinical trials.

THE FDA STANDARD: A 'SUCCESS' METRIC

The panel's consensus that the cardinal feature of drug use disorders is the behavior of drug-taking, and that an indicator of this behavior should be the primary outcome variable in drug dependence treatment clinical trials, is in contrast to the view of the US FDA. The FDA suggests that drug-taking behavior is only a surrogate indicator for risks of health or behavioral problems, and that a clinical metric reflecting sufficient behavior change beyond drug use is needed to reasonably conclude a probable benefit in health and behavior domains. In particular, the metric must define clinical 'success' so that results and effect sizes can be expressed as the percentage of patients achieving success. Success must be clinically meaningful. The panel's consensus was that interventions that reduce drug use by half are clinically meaningful [13] and that a greater acceptance of the value of interventions

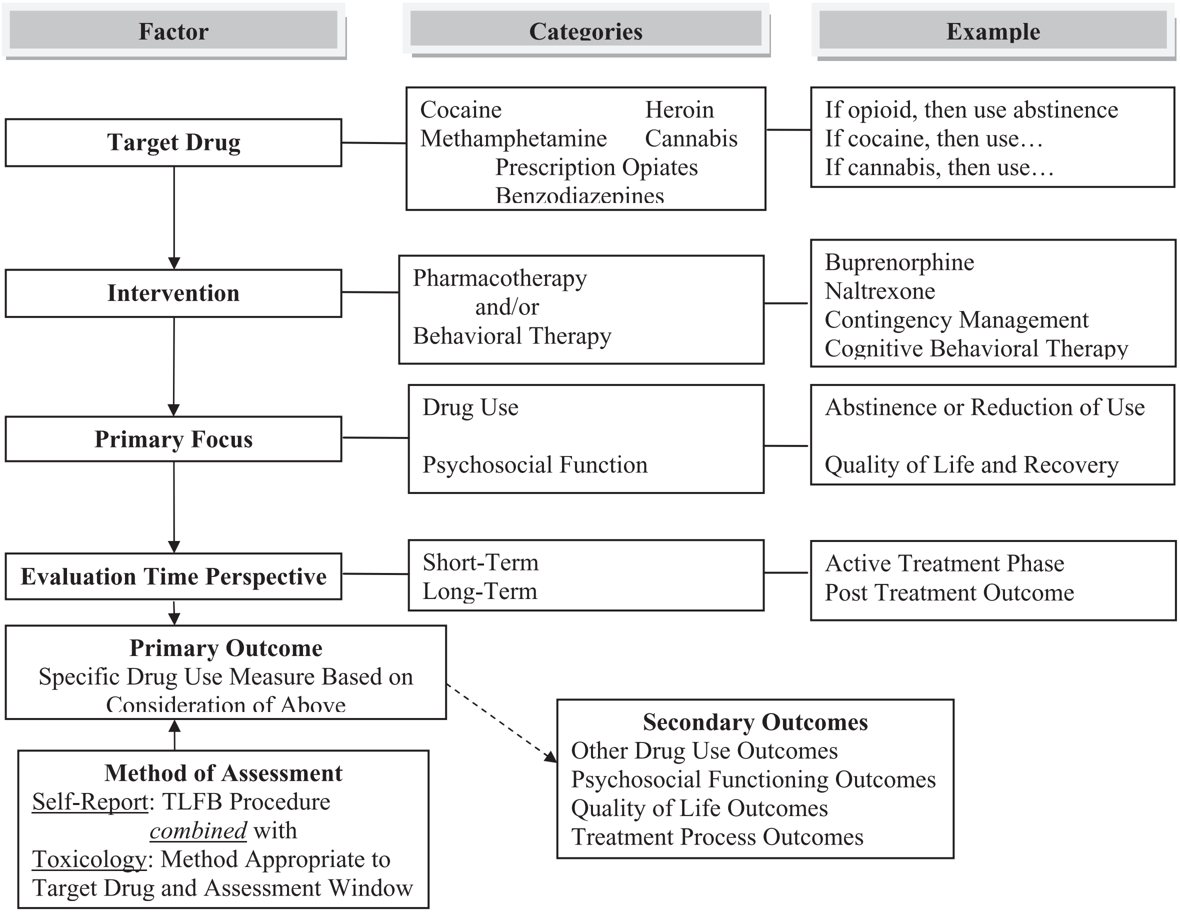

Figure 1 Heuristic model of factors to be considered in selecting outcome measures in drug dependence treatment clinical trials. TLFB: timeline follow-back

producing limited benefits may be appropriate, especially in a field where available treatment options are often quite limited.

The appropriate standard for considering an intervention to be of clinical benefit may depend on the treatment(s) already available and on the stage of therapy development for a particular substance abuse problem. Early in therapeutics development, clinical trials may ask the modest question 'does this intervention have any benefit?', whereas later, it becomes more relevant to address the question of 'does this intervention have benefit of clinically meaningful magnitude?'. While we focus here on drug-taking as proximal behavior and primary outcome, a companion paper addresses measures appropriate to broader domains associated with drug-taking [100].

ALTERNATIVES TO SEEKING A SINGLE CLINICAL OUTCOME METRIC

One impetus for selecting a common outcome for clinical trials is facilitating between-study comparisons, although within-study comparisons are often far more valid and informative than between-study comparisons. As separate studies always differ, it may well be that efforts to compare studies on a common clinical metric may be more misleading than informative.

One standard and accepted method for comparing and combining data across studies is that of meta-analysis, in which the magnitudes of within-study effects are compared or combined across different studies, commonly on a standardized metric of the statistical 'effect size'. This standardized effect size approach may be appropriate for comparison of substance abuse treatment trials that employed different clinical metrics for assessing their outcomes.

The panel concluded that producing appreciable change in drug-taking behavior itself can be sufficient to be considered clinically meaningful, without requiring benefit on more global or distant health or behavior outcomes. At the same time, the panel recognized and endorsed the value of assessing such health and behavior outcomes—especially functional behavioral and psychosocial outcomes as they often precede medical health consequences. These other domains of potential behavioral assessment are discussed in a companion paper [100].

Table 2 Examples from clinical trials of factors involved in primary outcome decisions.

	Examples of diversity of primary outcome variables as a function of trial characteristics				
Factors to be considered	*Opiate detox with buprenorphine/naloxone*	*Motivational enhancement*	*Motivational incentives*	*ADHD and adolescent drug use*	*Web-based cognitive behavioral therapy*
Target drug	Opiates	Any substance use disorder with use within the past 28 days	Stimulants—cocaine, amphetamine or methamphetamine	Multiple drugs of abuse	Primary drug of abuse
Type of intervention	Pharmacotherapy: 13-day buprenorphine/naloxone taper in in-patient setting	Behavioral: 3 sessions of MET at beginning of out-patient treatment	Behavioral: motivational incentives—provision of tangible incentives to clients for drug-free urines	Pharmacotherapy: 16 weeks of methylphenidate in adolescents with ADHD and substance use disorders	Behavioral: web-based version of the Community Reinforcement Approach plus abstinence incentives
Primary outcome focus	Response to taper/detox treatment	Treatment engagement and retention; substance use	Stimulant use reduction	ADHD symptom severity; substance use	Substance use; treatment retention
Operationalized primary outcome variable	Positive treatment response: participant completes 13-day detoxification and last urine specimen is negative for opiates	Weeks in treatment at CTP through 3-month follow up; Days of substance use, validated with urine and breath specimen results	Percentage of stimulant-free urines submitted; longest duration of abstinence from primary target drugs (cocaine, amphetamine, methamphetamine and alcohol)	Number of DSM-IV ADHD symptoms endorsed; number of days of alcohol and drug use during the past 28-day assessment periods	Abstinence from 10 tested drugs and heavy drinking during the 12 weeks of active treatment; number of days patients are in treatment until last face-to-face contact
Rationale for outcome measure	Evaluated detoxification, which is primarily concerned with safe and effective medical management of withdrawal and transfer to ongoing care	Evaluated a brief intervention designed to enhance engagement in treatment, which was expected to improve drug use outcomes	The primary focus of the incentives intervention was establishing and prolonging stimulant drug abstinence	Evaluated the impact of effective pharmacotherapy treatment for ADHD on substance use	The focus of the incentives intervention was on establishing and prolonging drug abstinence
Method of assessment	Urine specimen collected on day 13 or 14 is negative for opiates	Self-report drug and alcohol use based on Substance Use Calendar, as validated by urinalysis and breathe specimens	Full screen urinalysis and alcohol breathalyzer; self-reported drug use based on Addiction Severity Index	DSM-IV ADHD symptom checklist; self-report of use based on time-line follow-back, urine toxicology screens	Self-report of drug use and heavy drinking days based on time-line follow-back and urine toxicology screens
Secondary outcomes	Withdrawal symptoms, treatment retention, patient satisfaction	Cost–benefit, patient satisfaction, patient/staff attitudes	Treatment retention, counseling utilization	Psychosocial function, coping skills acquisition, HIV risk behaviors	Treatment retention, HIV risk behaviors, psychosocial functioning
Follow-up time perspective	Short-term, primarily interested in status at end of 13-day active detoxification period	Intermediate-term evaluated the durability of MET relative to standard treatment through a 3-month follow-up	Intermediate- to longer-term Evaluated outcomes throughout active treatment but also out to 6-month follow-up	Short-term Primary focus on 16-week active treatment phase and 1-month follow-up	Short- to longer-term Test for difference in substance use primary outcome based on last 4 weeks of active treatment but also will follow out to 6 months

ADHD: attention deficit hyperactive disorder; CTP: community treatment program; DSM-IV: *Diagnostic and Statistical Manual version IV*; HIV: human immunodeficiency virus; MET: motivational enhancement therapy.

CONCLUSIONS

- An indicator of drug-taking behavior is the appropriate primary outcome variable in most clinical trials of drug abuse treatments.
- An outcome measure that combines information from both self-report and objective toxicology testing is often preferable to either alone.
- There is no single outcome measure recommended as the standard index for incorporation into most clinical trials.
- Selection of a specific primary outcome measure for any specific study will depend upon the study and its goals.
- Use of alcohol and drugs other than the drug targeted specifically by the intervention should be assessed and included as potential secondary outcomes.
- For many trials it will be valuable to include secondary outcomes related to behavioral functioning and/or quality of life.
- Research is needed regarding correlations of drug-taking indicators with one another and with behavioral functioning and/or quality of life in both the short and long term.
- Research is needed regarding the effect of implementing and incorporating research-quality outcome measures into clinical treatment settings.
- Research regarding relationships of clinical trial data to functional behavioral outcomes may guide refinements in selecting primary outcome measures in future clinical trials.

Declarations of interest

None.

References

1. Wells E. A., Saxon A. J., Calsyn D. A., Jackson T. R., Donovan D. M. Study results from the Clinical Trials Network's first 10 years: where do they lead? *J Subst Abuse Treat* 2010; **38**: S14–30.
2. Wells E. A., Hawkins J. D., Catalano R. F. Jr. Choosing drug use measures for treatment outcome studies. I. The influence of measurement approach on treatment results. *Int J Addict* 1988; **23**: 851–73.
3. Wells E. A., Hawkins J. D., Catalano R. F. Jr. Choosing drug use measures for treatment outcome studies. II. Timing baseline and follow-up measurement. *Int J Addict* 1988; **23**: 875–85.
4. Van den Brink W., Montgomery S. A., Van Ree J. M., van Zwieten-Boot B. J. On behalf of the Consensus Committee. ECNP Consensus Meeting March 2003: guidelines for the investigation of efficacy in substance use disorders. *Eur Neuropsychopharmacol* 2006; **16**: 224–30.
5. Marsden J., Nizzoli U., Corbelli C., Margaron H., Torres M., Prada De Castro I. *et al.* New European instruments for treatment outcome research: reliability of the maudsley addiction profile and treatment perceptions questionnaire in Italy, Spain and Portugal. *Eur Addict Res* 2000; **6**: 115–22.
6. Donovan D. M., Marlatt G. A. *Assessment of Addictive Behaviors*. New York: Guilford Press; 1988.
7. Donovan D. M., Marlatt G. A., editors. *Assessment of Addictive Behaviors*, 2nd edn. New York: Guilford Press; 2005.
8. McLellan A. T., Chalk M., Bartlett J. Outcomes, performance, and quality: what's the difference? *J Subst Abuse Treat* 2007; **32**: 331–40.
9. Substance Abuse and Mental Health Services Administration. A report required by Congress on performance partnerships: a discussion of SAMHSA's efforts to increase accountability based on performance in its Block Grant Programs by instituting National Outcome Measures. Substance Abuse and Mental Health Services Administration, editor. Rockville, MD, 2004.
10. Darke S., Hall W., Wodak A., Heather N., Ward J. Development and validation of a multi-dimensional instrument for assessing outcome of treatment among opiate users: the Opiate Treatment Index. *Br J Addict* 1992; **87**: 733–42.
11. Marsden J., Gossop M., Stewart D., Best D., Farrell M., Lehmann P. The Maudsley Addiction Profile (MAP): a brief instrument for assessing treatment outcome. *Addiction* 1998; **93**: 1857–68.
12. Dole V. P., Nyswander M. A medical treatment for diacetylmorphine (heroin) addiction: a clinical trial with methadone hydrochloride. *JAMA* 1965; **193**: 646–50.
13. Miller W. R., Manuel J. K. How large must a treatment effect be before it matters to practitioners? An estimation method and demonstration. *Drug Alcohol Rev* 2008; **27**: 524–8.
14. Laudet A. B., Becker J. B., White W. L. Don't wanna go through that madness no more: quality of life satisfaction as predictor of sustained remission from illicit drug misuse. *Subst Use Misuse* 2009; **44**: 227–52.
15. Miller P. G., Miller W. R. What should we be aiming for in the treatment of addiction? *Addiction* 2009; **104**: 685–6.
16. Rounsaville B. J., Petry N. M., Carroll K. M. Single versus multiple drug focus in substance abuse clinical trials research. *Drug Alcohol Depend* 2003; **70**: 117–25.
17. Kellogg S. H. On 'Gradualism' and the building of the harm reduction–abstinence continuum. *J Subst Abuse Treat* 2003; **25**: 241–7.
18. Marlatt G. A., Witkiewitz K. Update on harm-reduction policy and intervention research. *Annu Rev Clin Psychol* 2010; **27**: 591–606.
19. Miller W. R., LoCastro J. S., Longabaugh R., O'Malley S., Zweben A. When worlds collide: blending the divergent traditions of pharmacotherapy and psychotherapy outcome research. *J Stud Alcohol* 2005; (suppl): 17–23.
20. Tai B. Workshop summary: Outcome measures and success criteria. In: Tai B., Chiang N., Bridge P., editors. *Medication development for the treatment of cocaine dependence: Issues in clinical efficacy trials*. NIDA Research Monograph 175; 2008. Rockville, MD: National Institutes of Health, National Institute on Drug Abuse, Office of Science Policy and Communications. p. 303–11. Available from http://archives.drugabuse.gov/pdf/monographs/monograph175/303-311_AppendixI.pdf (accessed 21 June 2011; archived by Webcite at http://www.webcitation.org/60GfMQWOz)

21. Vocci F., de Wit H. Consensus statement on evaluation of outcome of pharmacotherapy for substance abuse/dependence: report from a NIDA/CPDD meeting. In: National Institute on Drug Abuse Medications Development Division. Bethesda, MD, 1999. Available from: http://www.cpdd.vcu.edu/Media/FactSheets/nida_cpdd_report.pdf (accessed 17 June 2011; archived by Webcite at http://www.webcitation.org/5zXE1Axw7)
22. Aveyard P., Wang D., Connock M., Fry-Smith A., Barton P., Moore D.Assessing the outcomes of prolonged cessation-induction and aid-to-cessation trials: floating prolonged abstinence. *Nicotine Tob Res* 2009; **11**: 475–80.
23. Hughes J. R. Measurements of the effects of abstinence from tobacco. A qualitative reiview. *Psychol Addict Behav* 2007; **21**: 127–37.
24. West R., Hajek P., Stead L., Stapleton J. Outcome criteria in smoking cessation trials: proposal for a common standard. *Addiction* 2005; **100**: 299–303.
25. Shiffman S., West R. J., Gilbert D. G. Recommendation for the assessment of tobacco craving and withdrawal in smoking cessation trials. *Nicotine Tob Res* 2004; **6**: 599–614.
26. Hughes J. R., Keely J., Niaura R., Ossip-Klein D., Richmond R., Swan G. Measures of abstinence from tobacco in clinical trials: issues and recommendations. *Nicotine Tob Res* 2003; **5**: 13–25.
27. Society for Research on Nicotine and Tobacco Subcommittee on Biochemical Verification. Biochemical verification of tobacco use and cessation. *Nicotine Tob Res* 2002; **4**: 149–59.
28. Hughes J. R. Should criteria for drug dependence vary across drugs? *Addiction* 2006; **101**: S134–41.
29. Shiffman S., Scharf D., Shadel W., Gwaltney C., Dang Q., Paton S. *et al.* Analyzing milestones in smoking cessation: illustration in a nicotine patch trial in adult smokers. *J Consult Clin Psychol* 2006; **74**: 276–85.
30. Velicer W. F., Prochaska J. O., Rossi J. S., Snow M. G. Assessing outcome in smoking cessation studies. *Psychol Bull* 1992; **111**: 23–41.
31. Hughes J., Carpenter M., Naud S. Do point prevalence and prolonged abstinence measures produce similar results in smoking cessation studies? A systematic review. *Nicotine Tob Res* 2010; **12**: 756–62.
32. Velicer W. F., Prochaska J. O. A comparison of four smoking cessation outcome measures. *Addict Behav* 2004; **29**: 51–60.
33. Shiffman S. How many cigarettes did you smoke? Assessing cigarette consumption by global report, time-line follow-back, and ecological momentary assessment. *Health Psychol* 2009; **28**: 519–26.
34. Klesges R. C., Debon M., Ray J. W. Are self-reports of smoking rate biased? Evidence from the second national health and nutrition examination survey. *J Clin Epidemiol* 1995; **48**: 1225–33.
35. Hughes J. R. Observer reports of smoking status: a replication. *J Subst Abuse* 1993; **4**: 403–6.
36. National Institute on Alcohol Abuse and Alcoholism. Helping patients who drink too much: a clinician's guide. In: National Institutes of Health. Updated 2005 edition. Bethesda, MD: US Government Printing Office; 2007. Available from: http://pubs.niaaa.nih.gov/publications/practitioner/cliniciansguide2005/guide.pdf (accessed 17 June 2011; archived by Webcite at http://www.webcitation.org/5zXDEqfCm)
37. National Institute on Alcohol Abuse and Alcoholism. Rethinking drinking: alcohol and your health. In: National Institutes of Health. Bethesda, MD: US Government Printing Office; 2010. Available from http://pubs.niaaa.nih.gov/publications/rethinkingdrinking/rethinking_drinking.pdf (accessed 17 June 2011; archived by Webcite at http://www.webcitation.org/5zXD0T5cC)
38. Institute of Alcohol Studies (IAS). IAS fact sheet: what is problem drinking? In: Institute of Alcohol Studies, editor. St Ives, Cambridgeshire, UK, 2002.
39. Babor T. F., Higgins-Biddle J. C., Saunders J. B., Monteiro M. G. *AUDIT: The Alcohol Use Disorders Identification Test. Guidelines for Use in Primary Care*, 2nd edn. Geneva: World Health Organization; 2001.
40. Devos-Comby L., Lange J. E. 'My drink is larger than yours'? A literature review of self-defined drink sizes and standard drinks. *Curr Drug Abuse Rev* 2008; **1**: 162–76.
41. Kaskutas L. A., Graves K. An alternative to standard drinks as a measure of alcohol consumption. *J Subst Abuse* 2000; **12**: 67–78.
42. Kerr W. C., Greenfield T. K., Tujague J., Brown S. E. A drink is a drink? Variation in the amount of alcohol contained in beer, wine and spirits drinks in a US methodological sample. *Alcohol Clin Exp Res* 2005; **29**: 2015–21.
43. Miller W. R., Heather N., Hall W. Calculating standard drink units: international comparisons. *Br J Addict* 1991; **86**: 43–7.
44. Allen J. P. Measuring outcome in interventions for alcohol dependence and problem drinking: executive summary of a conference sponsored by the National Institute on Alcohol Abuse and Alcoholism. *Alcohol Clin Exp Res* 2003; **27**: 1657–60.
45. Falk D., Wang X. Q., Liu L., Fertig J., Mattson M., Ryan M. *et al.* Percentage of subjects with no heavy drinking days: evaluation as an efficacy endpoint for alcohol clinical trials. *Alcohol Clin Exp Res* 2010; **34**: 2022–34.
46. Finney J. W., Moyer A., Swearingen C. E. Outcome variables and their assessment in alcohol treatment studies: 1968–1998. *Alcohol Clin Exp Res* 2003; **27**: 1671–9.
47. Litten R. Z., Bradley A. M., Moss H. B. Alcohol biomarkers in applied settings: recent advances and future research opportunities. *Alcohol Clin Exp Res* 2010; **34**: 955–67.
48. Zweben A., Cisler R. A. Clinical and methodological utility of a composite outcome measure for alcohol treatment research. *Alcohol Clin Exp Res* 2003; **27**: 1680–5.
49. Anton R. F., Randall C. L. Measurement and choice of drinking outcome variables in the COMBINE study. *J Stud Alcohol* 2005; (suppl): 104–9.
50. Sobell L. C., Sobell M. B. Timeline followback: a technique for assessing self-reported alcohol consumption. In: Litten R. Z., Allen J. P., editors. *Measuring Alcohol Consumption: Psychosocial and Biological Methods*. Totowa, NJ: Humana Press; 1992, p. 41–72.
51. Sobell L. C., Sobell M. B., Connors G. J., Agrawal S. Assessing drinking outcomes in alcohol treatment efficacy studies: selecting a yardstick of success. *Alcohol Clin Exp Res* 2003; **27**: 1661–6.
52. Helander A., Bottcher M., Fehr C., Dahmen N., Beck O. Detection time of urinary ethyl glucuronide and ethyl sulfate in heavy drinkers during alcohol detoxification. *Alcohol Alcohol* 2009; **44**: 55–61.
53. Goodwin R. S., Darwin W. D., Chiang C. N., Shih M., Li S.-H., Huestis M. A. Urinary elimination of 11-Nor-9-Carboxy-D9-tetrahydrocannabinol in cannabis users

during continuously monitored abstinence. *J Anal Toxicol* 2008; **32**: 562–6.

54. Huestis M. A., Cone E. J. Methamphetamine disposition in oal fluid, plasma, and urine. *Ann NY Acad Sci* 2007; **1098**: 104–21.
55. Huestis M. A., Cone E. J., Wong C. J., Umbricht A., Preston K. L. Monitoring opiate use in substance abuse treatment patients with sweat and urine drug testing. *J Anal Toxicol* 2000; **24**: 509–21.
56. Huestis M. A., Mitchell J. M., Cone E. J. Detection times of marijuana metabolites in urine by immunoassay and GC-MS. *J Anal Toxicol* 1995; **19**: 443–9.
57. Huestis M. A., Cone E. J. Differentiating new marijuana use from residual drug excretion in occasional marijuana users. *J Anal Toxicol* 1998; **22**: 445–54.
58. Smith M. L., Barnes A. J., Huestis M. A. Identifying new cannabis use with urine creatinine-normalized THCCOOH concentrations and time intervals between specimen collections. *J Anal Toxicol* 2009; **33**: 185–9.
59. Schwilke E. W., Gullberg R. G., Darwin W. D., Chiang N., Cadet J. L., Gorelick D. A. *et al.* Differentiating new cannabis use from residual urinary cannabinoid excretion in chronic, daily cannabis users. *Addiction* 2010; **106**: 499–506.
60. Preston K. L., Silverman K., Schuster C. R., Cone E. J. Assessment of cocaine use with quantitative urinalysis and estimation of new uses. *Addiction* 1997; **92**: 717–27.
61. Center for Substance Abuse Treatment. The role of biomarkers in the treatment of alcohol use disorders. *Subst Abuse Treat Advisory* 2006; **5**: 1–7.
62. Preston K. L., Huestis M. A., Wong C. J., Umbricht A., Goldberger B. A., Cone E. J. Monitoring cocaine use in substance abuse treatment patients by sweat and urine testing. *J Anal Toxicol* 1999; **23**: 313–22.
63. Huestis M. A., Scheidweiler K. B., Saito T., Fortner N., Abraham T., Gustafson R. A. *et al.* Excretion of Delta9-tetrahydrocannabinol in sweat. *Forensic Sci Int* 2008; **174**: 173–7.
64. Winhusen T. M., Somoza E. C., Singal B., Kim S., Horn P. S., Rotrosen J. Measuring outcome in cocaine clinical trials: a comparison of sweat patches with urine toxicology and participant self-report. *Addiction* 2003; **98**: 317–24.
65. Dolan K., Rouen D., Kimber J. An overview of the use of urine, hair, sweat and saliva to detect drug use. *Drug Alcohol Rev* 2004; **23**: 213–7.
66. Rollins D. E., Wilkins D. G., Krueger G. G., Augsburger M. P., Mizuno A., O'Neal C. *et al.* The effect of hair color on the incorporation of codeine into human hair. *J Anal Toxicol* 2003; **27**: 545–51.
67. Fauci A. S., Braunwald E., Isselbacher K. J., Wilson J. D., Martin J. B., Kasper D. L. *et al.*, editors. *Harrison's Principles of Internal Medicine*, 14th edn. New York: McGraw-Hill, Health Professions Division; 1998.
68. Scheidweiler K. B., Huestis M. A. Simultaneous quantification of opiates, cocaine, and metabolites in hair by LC-APCI-MS/MS. *Anal Chem* 2004; **76**: 4358–63.
69. DuPont R., Goldberger B., Gold M. Clinical and legal considerations in drug testing. In: Ries R., Fiellin D., Miller S., Saitz R., editors. *Principles of Addiction Medicine*, 4th edn. Philadelphia, PA: Lippincott, Williams, and Wilkins; 2009, p. 1495–8.
70. Blank D. L., Kidwell D. A. External contamination of hair by cocaine: an issue in forensic interpretation. *Forensic Sci Int* 1993; **63**: 145–56.
71. Wang W. L., Cone E. J. Testing human hair for drugs of abuse. IV. Environmental cocaine contamination and washing effects. *Forensic Sci Int* 1995; **70**: 39–51.
72. Fals-Stewart W., O'Farrell T. J., Freitas T. T., McFarlin S. K., Rutigliano P. The timeline followback reports of psychoactive substance use by drug-abusing patients: psychometric properties. *J Consult Clin Psychol* 2000; **68**: 134–44.
73. Hersh D., Mulgrew C. L., Van Kirk J., Kranzler H. R. The validity of self-reported cocaine use in two groups of cocaine abusers. *J Consult Clin Psychol* 1999; **67**: 37–42.
74. Napper L. E., Fisher D. G., Johnson M. E., Wood M. M. The validity and reliability of drug users' self reports of amphetamine use among primarily heroin and cocaine users. *Addict Behav* 2010; **35**: 350–4.
75. Schuler M. S., Lechner W. V., Carter R. E., Malcolm R. Temporal and gender trends in concordance of urine drug screens and self-reported use in cocaine treatment studies. *J Addict Med* 2009; **3**: 211–7.
76. Sherman M. F., Bigelow G. E. Validity of patients' self-reported drug use as a function of treatment status. *Drug Alcohol Depend* 1992; **30**: 1–11.
77. Zanis D. A., McLellan A. T., Randall M. Can you trust patient self-reports of drug use during treatment? *Drug Alcohol Depend* 1994; **35**: 127–32.
78. Ehrman R. N., Robbins S. J. Reliability and validity of 6-month timeline reports of cocaine and heroin use in a methadone population. *J Consult Clin Psychol* 1994; **62**: 843–50.
79. Johnson M. B., Voas R. A., Miller B. A., Holder H. D. Predicting drug use at electronic music dance events: self-reports and biological measurement. *Eval Rev* 2009; **33**: 211–25.
80. Magura S. Validating self-reports of illegal drug use to evaluate National Drug Control Policy: a reanalysis and critique. *Eval Program Plann* 2010; **33**: 234–7.
81. Williams R. J., Nowatzki N. Validity of adolescent self-report of substance use. *Subst Use Misuse* 2005; **40**: 299–311.
82. Babor T. F., Steinberg K., Anton R., Del Boca F. Talk is cheap: measuring drinking outcomes in clinical trials. *J Stud Alcohol* 2000; **61**: 55–63.
83. Del Boca F. K., Noll J. A. Truth or consequences: the validity of self-report data in health services research on addictions. *Addiction* 2000; **95**: S347–60.
84. Langenbucher J., Merrill J. The validity of self-reported cost events by substance abusers: limits, liabilities, and future directions. *Eval Rev* 2001; **25**: 184–210.
85. Carey K. B. Reliability and validity of the time-line follow-back interview among psychiatric outpatients: a preliminary report. *Psychol Addict Behav* 1997; **11**: 26–33.
86. Meyer T. D., Hautzinger M. Assessing current alcohol use with the Form 90 in a student sample. *Psychol Psychother* 2009; **82**: 233–45.
87. Miller W. R., Del Boca F. K. Measurement of drinking behavior using the Form 90 family of instruments. *J Stud Alcohol* 1994; (suppl): 112–8.
88. Rice C. Retest reliability of self-reported daily drinking: Form 90. *J Stud Alcohol Drugs* 2007; **68**: 615–8.
89. Scheurich A., Muller M. J., Anghelescu I., Lorch B., Dreher M., Hautzinger M. *et al.* Reliability and validity of the Form 90 interview. *Eur Addict Res* 2005; **11**: 50–6.
90. Tonigan J. S., Miller W. R., Brown J. M. Reliability of Form 90: an instrument for assessing alcohol treatment outcome. *J Stud Alcohol* 1997; **58**: 358–64.

91. Carroll K. M., Fenton L. R., Ball S. A., Nich C., Frankforter T. L., Shi J. *et al.* Efficacy of disulfiram and cognitive-behavioral therapy in cocaine-dependent outpatients: a randomized placebo controlled trial. *Arch Gen Psychiatry* 2004; **64**: 264–72.
92. Fendrich M., Johnson T. P., Wislar J. S., Hubbell A., Spiehler V. The utility of drug testing in epidemiological research: results from a general population survey. *Addiction* 2004; **99**: 197–208.
93. Somoza E., Somoza P., Lewis D., Li S.-H., Winhusen T., Chiang N. *et al.* The SRPK1 outcome measure for cocaine-dependence trials combines self-report with urine toxicology to determine weekly fractions of cocaine use days. *Drug Alcohol Depend* 2008; **93**: 132–40.
94. Clifford P. R., Maisto S. A., Davis C. M. Alcohol treatment research assessment exposure subject reactivity effects: part I. Alcohol use and related consequences. *J Stud Alcohol Drugs* 2007; **68**: 519–28.
95. Ling W., Amass L., Shoptaw S., Annon J. A., Babcock D., Brigham G. *et al.* A multi-center randomized trial of buprenorphine–naloxone and clonidine for opioid detoxification: findings from the National Institute on Drug Abuse Clinical Trials Network. *Addiction* 2005; **100**: 1090–100.
96. Ball S. A., Van Horn D., Crits-Christoph P., Woody G. E., Farentinos C., Martino S. *et al.* Site matters: multisite randomized trial of motivational enhancement therapy in community drug abuse clinics. *J Consult Clin Psychol* 2007; **75**: 556–67.
97. Petry N. M., Peirce J. M., Stitzer M. L., Blaine J., Roll J. M., Cohen A. *et al.* Effect of prize-based incentives on outcomes in stimulant abusers in outpatient psychosocial treatment programs: a National Drug Abuse Treatment Clinical Trials Network study. *Arch Gen Psychiatry* 2005; **62**: 1148–56.
98. Riggs P. D. New findings from a randomized controlled trial: osmotic-release methylphenidate (OROS-MPH) for ADHD in adolescents with substance use disorders. American Academy of Child and Adolescent Psychiatry (AACAP), 26–31 October. New York, NY. 2010. Available from: http://ctndisseminationlibrary.org/display/560.htm (accessed 17 June 2011; archived by Webcite at http://www.webcitation.org/5zXDgL2Fj)
99. Nunes E. V. Web-delivery of evidence-based, psychosocial treatment for substance use disorders. In: National Institute on Drug Abuse Center for Clinical Trials Network. Bethesda, MD, 2010. Available from: http://clinicaltrials.gov/ct2/show/NCT01104805 (accessed 17 June 2011; archived by Webcite at http://www.webcitation.org/5zXDU5ZqM)
100. Tiffany S. T., Friedman L., Greenfield S. F., Hasin D. S., Jackson R. Beyond drug use: a systematic consideration of other outcomes in evaluations of treatments for substance use disorders. *Addiction*; in press; 2011.

APPENDIX I

Clinically meaningful substance abuse treatment outcome measure for effectiveness trials
15–16 December 2009
Participant list

George Bigelow PhD
Johns Hopkins University School of Medicine
James Bjork PhD
National Institute on Drug Abuse
Michael Bogenschutz MD
University of New Mexico
Gregory Brigham PhD
Maryhaven, Inc.
Rita Cardim
Johns Hopkins Health System
Kathleen Carroll PhD
Yale University School of Medicine
Ling Chen PhD
US Food and Drug Administration
Allan Cohen MA, MFT
Bay Area Addiction Research and Treatment, Inc.
Dennis Daley PhD, LSW
University of Pittsburgh Medical Center
Marta De Santis PhD
National Institute on Drug Abuse
Traci Donnelly MS
Phoenix House of New York
Dennis Donovan PhD
University of Washington
Sarah Q. Duffy PhD
National Institute on Drug Abuse
Jeffrey Leimberger PhD
Duke Clinical Research Institute
Michael Levy PhD
CAB Health and Recovery Services
Esther Lewis MPA
Indiana Access to Recovery
Robert Lindblad MD
The EMMES Corporation
Walter Ling MD
University of California, Los Angeles
G. Alan Marlatt PhD
University of Washington
Dennis McCarty PhD
Oregon Health and Science University
Patrick B. McEneaney MBA
Phoenix House of New England and Florida
David Metzger PhD
University of Pennsylvania
Mary Ellen Michel PhD
National Institute on Drug Abuse
Karen Nielsen MBA, MPA
Public Sector Healthcare
Neal Oden PhD
The EMMES Corporation
Harold Perl PhD
National Institute on Drug Abuse
Kenzie Preston PhD
National Institute on Drug Abuse
Cendrine Robinson (PhD student)
Uniformed Services University of Health Sciences

David Epstein PhD
National Institute on Drug Abuse
Daniel J. Feaster PhD
University of Miami
Patrick Flynn PhD
Texas Christian University
Lawrence Friedman MD
National Institute on Drug Abuse
John Gardin PhD
ADAPT, Inc.
Udi Ghitza PhD
National Institute on Drug Abuse
Shelly Greenfield MD, MPH
Harvard University Medical School
John Hamilton
Regional Network of Programs, Inc.
Deborah Hasin PhD
Columbia University
Marilyn Huestis PhD
National Institute on Drug Abuse
John Hughes MD
University of Vermont
Petra Jacobs MD
National Institute on Drug Abuse
Ron Jackson MSW
Evergreen Treatment Services
Michael Klein PhD
US Food and Drug Administration
Carmen Rosa MS
National Institute on Drug Abuse
Jeffrey Selzer MD
Medical Society of the State of New York
Eugene Somoza MD, PhD
University of Cincinnati
Maxine Stitzer PhD
Johns Hopkins School of Medicine
Michele Straus MS, RPh
National Institute on Drug Abuse
Jose Szapocznik PhD
University of Miami
Betty Tai PhD
National Institute on Drug Abuse
Stephen Tiffany PhD
University at Buffalo, The State University Of New York
Madhukar Trivedi MD
University of Texas Southwestern Medical Center
Paul Wakim PhD
National Institute on Drug Abuse
Elizabeth Wells PhD
University of Washington
Roger Weiss MD
Harvard University Medical School
Celia Winchell MD
US Food and Drug Administration
Joan Zweben PhD
University of California, San Francisco