

# Forum

## EBAM\* For Practitioners

(\*Evidence-Based Addiction Medicine)

Evaluating & Using Research Evidence in Clinical Practice

Stewart B. Leavitt, PhD

Sponsored by an educational grant from Mallinckrodt, Inc.

### New Outlooks on Addiction Research

**Medical research is an imperfect science.** Research in substance dependency (addiction) is no exception and, in fact, has its own limitations. Understanding those imperfections is essential for becoming a more critical reader of addiction research literature and a more discriminating consumer of scientific evidence.

Just as juries need evidence from reliable witnesses or forensic investigations to arrive at impartial and fair verdicts, addiction treatment providers need credible information to answer health-care questions and make clinical decisions. Yet, the amount of information in the addiction field has rapidly increased, bringing with it challenges of navigating efficiently through the mushrooming number of articles and identifying evidence that is of valid and reliable quality.[1]

Added to this, regulatory oversight of addiction treatment programs requires adopting scientifically-validated practices for improving patient care and outcomes. Increasingly, programs also will need to gather and interpret their own research data for reporting results of their efforts as part of the ongoing accreditation process.[2]

Toward those ends, this booklet describes some fundamental principles for evaluating and using research evidence in clinical practice. It provides knowledge to determine if a research article is relevant to clinical information needs, and if the results are likely to be valid for a particular purpose.

Many of the concepts may be unfamiliar to readers and will require careful study. This booklet is best “digested” slowly and then used as a reference when evaluating research in addiction medicine.

### Contents

<b>New Outlooks</b> .....1	Cohort Studies .....4	Importance of Power .....9
Caveat Lector .....2	RCTs .....5	“Bottom-line” Clinical Effects .....9
The Role of EBAM .....2	Reviews & Meta-analyses .....5	Estimates of Effect .....10
Problems of Proof .....2	<b>Assessing Validity of Research</b> .....5	Confidence Intervals .....11
Reasoning with Research .....2	<b>Methodology &amp; Outcomes</b> .....5	“Forest Plots” .....12
<b>Addiction Research Approaches</b> .....3	<b>Detecting Bias</b> .....6	“Survival” Analyses .....12
<b>Hypothesis Testing</b> .....3	Publication Bias .....6	Correlation .....13
The Null Hypothesis .....3	Patient Selection .....6	<b>Why Good Research Goes Bad</b> .....13
Statistical Significance.....3	Confounding Factors .....7	Design, Execution, Reporting .....13
Clinical Significance .....4	Randomization .....7	Post Hoc Analyses .....14
<b>Types of Studies</b> .....4	Blinding (Masking).....7	Fallacies of Anecdotes .....14
Levels of Scientific Evidence .....4	Placebo Effects .....7	Mass Media Distortions .....14
Perspectives Articles .....4	Run-in Periods .....7	<b>Putting Research Into Practice</b> .....15
Case Reports .....4	Compliance & Followup .....8	Everyday Relevance .....15
Cross-Sectional Studies .....4	ITT vs Per-Protocol .....8	Healthy Skepticism .....15
Case-Control Studies .....4	<b>Clinical Relevance of Statistics</b> .....9	References .....16

## EBAM helps answer 3 key questions:

- Where did you hear about that treatment?
- How do you know the information is valid?
- What do you propose doing and what results do you expect?

### Caveat Lector (Reader Beware)

*It cannot be assumed that everything appearing in print is worthwhile or valid.* Authorities in medical publishing have conceded that many wrong, or at least unreliable, therapeutic answers are being generated due to biased studies, representing small numbers of patients, and relying on inappropriate analyses.[3] Numerous investigations of reputable medical journals, spanning many years, have found a surprising number of faults:

- On average, approximately half or more – ranging from 25% to 90% [3-5] – of all articles in the journals examined contained errors varying from omissions of crucial information to significant design flaws affecting validity.
- Abstracts accompanying journal articles often receive the greatest attention. Yet, a review of articles chosen randomly from 6 major medical journals found that from 18% to 68% of the abstracts in examined journals contained data that were inconsistent with or absent from the main body of the articles.[6]
- In one investigation, 80% (40/50) of the systematic reviews and meta-analyses randomly selected were judged to have serious and extensive flaws.[7]

While the above claims themselves might be subjected to critical review, they do point toward the need for very cautious examinations of published medical literature. *Caveat lector* (reader beware) is sound advice, and nowhere is this more important than in the addiction treatment field.

### The Role of EBAM

During the early-to-mid 1990s there was an intensive movement worldwide to adopt principles of “evidence-based medicine” in all healthcare disciplines. Such efforts were directed to the needs of busy clinicians and staff, enabling them to critically interpret research rather than accepting at face value what was presented. The philosophical origins of evidence-based medicine date back to mid 19th century Paris and earlier, and may be defined as, *“the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients.”*[8,9]

Applied to addiction treatment, Evidence-Based Addiction Medicine (EBAM) involves combining clinical expertise with the best available external evidence on a topic of concern gathered from various sources. EBAM approaches empower addiction

treatment providers to clearly differentiate between clinical practices based on sound evidence versus those founded more on traditional practice, long-standing prejudices, or physiological rationales that might be outdated.[10]

Furthermore, in this era of managed care and increasing pressures of accountability, evidence-based practices can help staff respond convincingly to questions such as:

1. Where did you hear about that treatment?
2. How do you know the information is valid?
3. What do you propose doing and what results do you expect?

### Problems of Proof

*Accepting or rejecting research evidence depends on tolerance for uncertainty.* Due to the imperfect nature of research, there is always reasonable doubt that the observed outcomes might have been due to chance or random events, at least to some degree. Researchers know this and specify in advance how much uncertainty – or play of chance – they are willing to tolerate in presenting their evidence as either favorable (positive) or unfavorable (negative) regarding the intervention(s) studied.[10]

Scientific research, by its nature, *does not “prove” anything.* There are always limitations of some sort in study design and execution resulting in a degree of uncertainty as to the validity of the findings. Yet, there is increasing interest in providing a higher level of patient care, which requires valid research evidence for making clinical decisions.[11]

### Reasoning with Research

*Medical research reporting often is biased.* This may seem like a strong generalization; however, it merely reflects the fact that there is a specific goal behind any research endeavor to begin with and a particular point of view expressed in the presentation of data. Research articles are a form of persuasive communication, no matter how scientifically objective they might appear through the skillful use of language.

However, bias must be distinguished from prejudice, which would include presenting a purposely unsupported, distorted, or slanted (one-sided) interpretation of facts or data to manipulate readers’ perceptions and opinions. Genuinely prejudiced communications are fortunately uncommon in the medical literature and are considered unethical.

At the very least, addiction treatment providers need to bring a healthy skepticism to bear and apply more critical reviewing and evaluative skills in their reading and interpretation of research literature. Being aware of the potential for biases, flawed study designs or analyses, and inappropriate reporting methods can help in avoiding untrustworthy data and selecting publications that provide the best evidence for particular informational needs.

## Addiction Research Approaches

*Research begins with a question or hypothesis and proceeds to the design and execution of a study in search of an answer.*

Many research approaches have been devised over the years, and each has weaknesses, limitations, or biases that can affect validity. From a clinical *treatment* perspective, which is the primary concern here, addiction research seeks improvements in specific aspects of patient care via certain **interventions**, such as medications, drug dosages, or behavioral therapies.

Usually, a group of patients exposed to an *Experimental* intervention is compared with patients in a *Control* group receiving a placebo or comparison intervention. Performance on specific **outcome measures** (eg, retention in treatment, illicit-drug abstinence) is used as evidence of effectiveness.[9]

No matter how large the study, each group contains only a relatively tiny sampling of patients representative of a much larger patient population. Statistical analyses use data collected from the samples to estimate what the effects might be in that larger population, and whether the performance outcomes of the Experimental versus Control groups demonstrate either beneficial, harmful, or neutral effects of treatment.[9]

There are many rules for proper scientific method in designing, conducting, and reporting research – which are never perfectly followed. Understanding those rules and the limits of acceptable deviation is important for judging the value of individual research studies.

## Importance of Hypothesis Testing

***What is the purpose of the study? Are the results significant?***

The purpose of the research investigation should be very clearly stated by the authors as a **hypothesis**, which is essentially a prediction of certain results that can be measured, tested, and either supported or refuted.[12] In treatment studies, researchers examining if different interventions affect some outcome variable(s) in the same or different ways might state the hypothesis as, “The purpose of this study was to determine the benefits of Experimental treatment X in decreasing illicit-opioid use compared with standard therapy with Treatment Y.”

### The Null Hypothesis

Research designs and analyses are based on an assumption that any differences found between the interventions are due to random effects or chance. The role of chance here recognizes that all things in nature are possible and may occur with or without outside intervention. For example, some patients do recover from illness whether or not they receive medical treatment.[13]

**Research designs are based on an assumption that any differences found between the interventions are due to random effects or chance.**

The assumption of *no effect due specifically to the intervention* is known as the **null hypothesis**. [14] The goal of most treatment studies is to demonstrate that significant and valid differences between the examined groups *do exist and are of sufficient size that the null hypothesis can be rejected*. [15]

### Statistical Significance

A level of probability, called a *p*-value, that is considered as *statistically significant* is selected in advance as a basis for rejecting the null hypothesis. Traditionally, this is a probability value equal to or less than 0.05, expressed as  $p \leq .05$ . [14,15]

**Example:** When there is an observed difference between interventions, a  $p < .05$  suggests there is less than a 5% probability — or less than 1 chance in 20 — that the result was due merely to chance or some random effect of nature, rather than the interventions themselves. On this basis, the null hypothesis of no effect would be rejected.

To compute the *p*-value, various test statistics are used. Authors should report the specific test used to determine the *p*-value, with an explanation of what the test measures. [13,15]

A low *p*-value, signifying a statistically significant difference between study groups, is often viewed as the most important sign of a “good” study. However, the *p*-value merely describes the degree of uncertainty or error in asserting that a difference exists when there actually is no difference. That is, how often the researcher might be wrong by saying that the Experimental treatment made a difference *when it really did not*: 1 time out of 20 in the case of  $p = .05$ ; or, 1 time out of 100 as with  $p = .01$ . [13,16]

**Example:** A hypothetical paper concludes, “Treatment X significantly reduced illicit-opioid use compared with standard therapy Y;  $t = 4.5$ ,  $p = .009$ .” The researcher is suggesting that according to the data analysis – using a *t*-test – there was less than 1 chance in 100 ( $<.01$ ) that such results were due to random error or mere coincidence. Therefore, the null hypothesis was rejected, proposing that Treatment X did make a statistically significant difference.

With all aspects of any study, there is the possibility of unknown, chance factors producing false or misleading outcomes, no matter how small the likelihood. The lower the probability of such errors — ie, the lower the *p*-value — the *less uncertainty* in the results and the more confidently a treatment or intervention might be accepted for clinical practice.

## Clinical Significance

**Statistical significance does not automatically denote clinical significance.** Improvements produced by an Experimental intervention may be large enough to achieve statistical significance, but too small in absolute terms to provide significant clinical benefits for patients. This is especially the case if there are offsetting factors, such as increased costs, less convenience, or greater side effects with the new treatment.[13]

**Example:** A treatment might increase average time to opioid relapse from 10 days to 20 days, a statistically large 100% improvement. However, the new treatment might entail every day rather three-times weekly psychotherapy sessions, at greatly added cost and inconvenience, to gain only a relatively small clinical advantage (10 days added abstinence).

## Types of Studies

### Levels of Scientific Evidence

Various types of research studies assessing therapeutic effects may be ranked according to a “hierarchy of evidence.” This is based on the relative strength of each study for providing results that are likely to be free of bias and valid. Rankings from weakest at the bottom to strongest at the top are reflected in **Table 1**. [9,11,17-20]

The ranking does not question the essential ability of each research approach to be valid and of value for a particular purpose. However, it does recognize that certain forms of evidence may be given greater emphasis for guiding clinical decision-making. Following, from lowest to highest ranking, is a discussion of each type of research approach that addresses treatment effectiveness in addiction medicine.

### Perspectives Articles

“Perspectives” is a coined term to represent overviews, reports, commentary, and interviews. These are the most prevalent types of communications in the addiction field and are often cited as evidence. They are at the bottom of the evidence hierarchy because they summarize or comment on research that was done by others, rather than generate original data from clinical experimentation. They are highly subject to bias, such as favoring one viewpoint over another.

Still, perspectives articles can be invaluable sources of information by consolidating existing evidence and offering interpretations to aid understanding and further inquiry by the reader. Such articles are of greatest evidentiary value when they fully cite sources of information and, in some cases, offer opposing viewpoints or are peer reviewed.

**Table 1: Hierarchy of Evidence**

<b>Systematic Reviews &amp; Meta-Analyses of RCTs</b>
<b>Randomized Controlled Trials (RCTs)</b>
<b>Cohort Studies</b> (Follow-up, Incidence, Longitudinal, Prospective Study)
<b>Case-Control Studies</b> (Case-referent, Case-comparison, Retrospective Study)
<b>Cross-Sectional Surveys</b> (Prevalence Study)
<b>Case Reports</b> (Case History, Case Series, Anecdote)
<b>Perspectives</b> (Overviews, Reports, Commentary, and Interviews)

However, even when there appears to be a balanced presentation of relevant data it must be assumed that the author has a particular viewpoint (perspective) that is not necessarily all encompassing.[21,22] In fact, space limitations in publications usually prohibit fully exploring all sides of a topic.

### Case Reports

Also called case histories, case series, or anecdotes, these draw upon personal observations or medical records reviews to report unusual or unexpected events in conjunction with a medication or therapy. There are many biases associated with such reports, including errors in observation or data interpretation, inadequate documentation, and unsupported conclusions. (Also see below, “Fallacies of Anecdotes as Evidence.”)

### Cross-Sectional Surveys

These investigations – also called, prevalence or epidemiological studies – examine the relationship between medical conditions and other variables of interest as they exist in a defined population at a particular time. Researchers examine interventions or exposures (what occurred) and outcome conditions (what happened as a result). This type of study can establish *associations* but not causality. There may be problems with recall bias (not remembering exactly what occurred), and extraneous factors (confounders) can be unequally distributed among subjects.

### Case-Control Studies

Also called case-referent, case-comparison, or retrospective studies, these identify patients with the outcome(s) of interest (cases) and Control patients without the same outcome(s). The researchers then look back in time to compare how many subjects in each group had the same interventions or exposures of interest. This is a relatively fast and inexpensive method and may be the only feasible alternative for examining long-term treatment effects or other outcomes with long lag times between interventions and outcomes. A problem is that this method is highly subject to recall bias or inconsistent records in determining what had occurred in the past.

### Cohort Studies

Cohort studies – also called, followup, incidence, longitudinal, or prospective studies – are the most common form of clinical trials in addiction medicine, but are subject to various forms of bias. A single group may be involved, but usually two or more groups of patients (cohorts) are enrolled that either receive the treatment of

interest (Experimental group) or do not (Controls). The groups are followed forward in time to observe outcomes of interest.

The groups should be evenly matched, with eligibility criteria and outcome assessments standardized. Problems include a lack of randomization, and difficulty in identifying Control patients similar to those treated or lack of a suitable Control group entirely. Treatment effects also may be linked to unknown or uncontrolled factors (confounders).

In the addiction research literature, these trials are sometimes called “observational” or “naturalistic” studies. This reflects the fact that subjects are drawn from an existing patient population (sometimes called a “convenience,” “unselected,” or “fortuitous” sample) and allocated to groups based on their existing condition and/or treatment, rather than being recruited specifically for the study and randomly assigned to Experimental treatment or Control groups.

**Historical Note:** The very first clinical trial, in 1747, was a cohort study. A ship’s doctor, seeking a treatment for scurvy, took 12 seamen so afflicted and treated them 2 at a time with either daily doses of cider, elixir vitriol, vinegar, sea water, nutmeg, or oranges and lemons. These were compared with afflicted shipmates receiving none of the treatments. He observed a rapid and beneficial effect of the citrus fruit therapy.[23]

### Randomized Controlled Clinical Trials (RCTs)

RCTs are considered by many as the “gold standard” when addressing questions of medication or therapeutic efficacy.[8] In this design, patients are recruited, carefully selected, and then randomly assigned to Experimental and Control groups, which are followed for the outcomes of interest. The groups are equally matched demographically (age, sex, etc.) and any extraneous factors (confounders) are assumed to be equally distributed across groups.

Unfortunately, this type of study can be the most costly in terms of time and money. Furthermore, in addiction medicine, there may be ethical problems with Control conditions –such as denial of treatment or inadequate treatment resulting in adverse outcomes – and there may be volunteer bias in terms of the characteristics of patients who are willing to be randomly assigned to Experimental or Control procedures for treating their addiction problems.

### Systematic Reviews & Meta-analyses

Systematic reviews gather all available evidence of the highest quality available to address clearly focused clinical questions.[1] Clinical practice guidelines often result from the systematic review process, which can be quite involved.

Explicitly defined methods for gathering research evidence help limit bias in identifying and selecting appropriate studies, and validity criteria for the inclusion of evidence from each study accepted should be clearly stated. Conclusions tend to be reliable and accurate, depending on the quality of available evi-

## Questions of validity consider whether reported research outcomes represent the most accurate directions and size of the intervention effects.

dence. Most important, a systematic review facilitates the relatively rapid assimilation of large amounts of research by readers. [21,24]

However, critics have expressed concerns about the validity of combining studies that were done on different patient populations, in different settings, at different times, and sometimes for different reasons.[21] Another limitation is the search procedure used to identify studies for inclusion. Commonly used electronic databases, such as MEDLINE and EMBASE, are convenient but usually do not include all studies that may be relevant and important.[18]

Meta-analyses take systematic reviews a step further by combining statistical evidence from multiple investigations and using mathematical techniques to analyze the results. Hence, these are research projects in which *the unit of analysis is the individual study rather than an individual patient*. This approach allows for achieving greater precision and clinical applicability of results than is possible with any individual study or systematic review (see also, “Meta-Analysis ‘Forest Plots’” below). [9,18,24,25]

### Assessing Validity of Research

Questions of validity consider whether reported research outcomes represent the most accurate direction and size of intervention effects. Basically, can the research be trusted?[26]

Validity may be conceptualized along two broad dimensions: [27]

- **Internal Validity** is the degree to which a research result is likely to be correct and free of bias. It refers to observed effects that are applicable to the *subjects in a particular study* (as opposed to external validity).
- **External validity** – also called generalizability, relevance, or transferability – is the extent to which the results of an investigation might be expected in typical addiction treatment settings and/or apply to populations beyond those included in the study.

Validity is determined in large part by examining a study’s methodology, outcomes, and sources of potential bias.

### Methodology & Outcomes

Study **methodology** — how it was planned (the protocol) and executed — should be described in great detail by the authors.[28] A good hypothetical question to ask is: “If someone wanted to repeat this research study, is there sufficient information telling exactly what to do?”

Replication is the only way the *reliability* of results can be confirmed and the ultimate validity of the findings established. If the study under review never has been repeated in any fashion, the reliability of the results and their relevance for clinical decision-making are less certain.

Even when studies are replicated they may be different in so many ways as to make comparisons difficult or impossible.

**Outcomes** of an investigation are determined by measurements or observations of **endpoints**. These should be specified in advance as part of the study protocol to avoid bias in analyzing the data. Essentially, endpoints are the “payoff” – the results responding to the hypotheses of the study – and the type of endpoints and their measurement determine data accuracy and contribute to study validity. There are two types:[29]

1. **Primary endpoints** most directly, objectively, and definitively portray the target condition or result of interest. For example, retention in treatment can be directly observed and measured in days, weeks, or months.
2. **Surrogate endpoints** are *indirect* measures, serving as “markers” of an outcome of interest, and are common in addiction research. For example, periodic urinalyses can be surrogate markers of either illicit-drug use or abstinence, but they do not necessarily indicate length of abstinence between tests or quantity of any drug consumed. Furthermore, it is important that surrogate endpoints do not reflect some confounding variable (eg, a person testing positive for opioids due to poppy seed or cough syrup ingestion).

Many endpoints pose concerns regarding their accuracy and reliability, such as those assessing: a) symptomatic effects (eg, nausea, fatigue); b) psychological affect (eg, drug craving, anxiety); c) functional status (eg, ability to work or go to school); or, d) social outcomes (eg, family relationships).[28] There should be a complete description in the study report of how such outcome measures were validated, and how changes in the measures accurately reflected status changes in the patients.

## Detecting Bias

Bias in research studies has been defined as anything that influences conclusions about the groups under investigation and potentially distorts comparisons.[28] Since all research is imperfect, the question is not *if* a particular study reflects bias, but *how much*, and whether the biases are sufficient to negate the internal and external validity of the results.[11] Bias takes many forms, some more obvious than others (**Table 2**).

### Publication Bias

To begin, there are potential biases influencing which studies even appear in print. Investigations with significantly positive results, favoring the Experimental treatment, are more likely to be

**Table 2: Areas of Potential Bias**

• Publication Biases	• Placebo Effects
• Patient Selection	• Run-in Periods
• Confounding Factors	• Subject Compliance
• Randomization	• Followup Duration
• Blinding (Masking)	• ITT vs Per-Protocol

submitted for publication than those with negative or equivocal outcomes (*publication bias*). This can make it appear that certain treatments are more effective than might be the case.[4,26]

It also is important to consider that, even in the best of circumstances, it can take years from the time of data gathering until a study appears in print (*time-lag bias*). The latest revelations appearing in today’s journals may be completely overruled by studies already waiting in the publication pipeline.

### Patient Selection

When two or more groups of subjects are compared, an important goal is for them to be as similar as possible, *except* for any specific differences under examination. However, even when study-group composition is equivalent, or if only a single group is examined, the chosen subjects may exhibit obvious selection biases that could affect external validity (eg, a study may include only males or younger persons).[27,28]

There also is the question of who is excluded from a study. For example, a trial may be restricted to patients with only mild forms of a disease, those who respond in certain ways to the treatment in question, or those who are compliant with particular treatment regimens.[28,30] Such approaches may lead to more efficient study designs, and more dramatic results, but they limit the validity of the conclusions.[27]

*Example:* There has been a great deal of controversy surrounding the selection of subjects for trials of antidepressant medications. It has been noted that, while some exclusion criteria are clearly necessary, others – such as, rejecting persons with substance-use disorders in 84% of trials – are used primarily to maximize medication-versus-placebo differences. Thus, the generalizability of the research to everyday clinical practice has been questioned.[31]

Authors should provide comparative descriptions explaining the baseline (entry level) characteristics of groups at the beginning of the study. These show how the groups were similar or dissimilar in terms of age, sex, race, and other key variables that might have influenced outcomes, such as preexisting psychiatric conditions. This also helps readers determine how closely the study subjects match the patients that are of concern in a particular clinical setting (external validity).[16]

Additionally, other than the intervention under investigation, Experimental and Control groups should have been treated or managed equally; otherwise, the results could be weakened or slanted.[26] This can be difficult in addiction research using behavioral interventions, whereby psychosocial therapies delivered by more than one psychologist or counselor, or at multiple research sites, may not be precisely identical.

## Confounding Factors

Many researchers use subject exclusion/inclusion criteria to reduce the presence of extraneous factors – called “confounding” variables – that may bias outcomes in some way. Such factors might prevent the outcome of interest from occurring or cause it to occur when it otherwise might not.

Although it is possible to control for confounders that are *known and measured* using certain statistical manipulations, it makes data interpretation more complex and potentially less precise. If patients with known potential confounders are excluded from the study, it can lead to sampling bias, since those chosen for investigation might not be representative of a typical clinic population.[12]

## Randomization

Randomization is the only way to control for confounders that are *not known or not measured*. [12,26] However, many research designs in addiction medicine do not use randomization, and the likely existence of potential confounders and associated bias always needs to be carefully considered.

The essential principle of randomization is that any research subject has the same and equal chance of assignment to any study condition.[4] Randomization procedures must be objective, such as using tables of random numbers or randomization computer software, and these should be described by the authors.[28]

Non-randomized studies need to be examined cautiously. On average, such research designs tend to overestimate the effects of healthcare interventions; although, the extent and even the direction of this bias is often impossible to predict.[18,26]

## Blinding (Masking)

Patients, clinicians, and other study personnel who are aware of just who is, and who is not, receiving an intervention or treatment of interest are likely to form opinions about its efficacy. Such opinions, whether optimistic or pessimistic, may systematically distort other aspects of treatment as well as the reporting of outcomes.

The most effective way of avoiding such bias is by double-blinding (also called, “double-masking”), in which neither patients nor study personnel know who is in the Control or Experimental groups. When separate study evaluators are involved, it is called “triple-blinding” if they are blinded.[26]

In addiction research, blinding can be difficult or impossible to achieve. Subjects can often correctly guess which drug or drug dosages they are receiving, based on effects or side effects.[32] With behavioral interventions, group assignment is often apparent to investigators and subjects; such as, motivational therapy with or without incentives. The authors should acknowledge and address such limitations in their reports.

**One the most insidious forms of bias occurs when a favorable response to a drug therapy is attributable to the mere *expectation of some benefit.***

## Placebo Effects

One of the most insidious forms of bias occurs when a favorable response to drug therapy — regardless of whether it is active medication or an inert placebo — is attributable to the mere *expectation of some benefit*. [4] This is called the “placebo effect.”

Conversely, a “nocebo phenomenon” also has been described. Subjects receiving medications or placebos may report *adverse effects* simply due to the anticipation of sickness or other negative consequences.[33]

The only way to control for placebo and nocebo influences is via effective double-blind research designs.[4] At that, the inactive placebo agent must be completely identical to the active medication in appearance, and other procedures that assist in masking should be described by the authors (such as, the addition of agents to the placebo that would mimic expected side effects of the active drug).

The use of placebos, in general, has raised ethical concerns. Some authorities have argued that placebo administration is not appropriate if effective treatment for a condition exists and active medications can be used as a Control. The other side of the debate contends that, when assessing efficacy of an agent, it must be acknowledged that some persons have favorable outcomes without *any* intervention and only an inactive placebo condition would detect that effect.[34]

This controversy might be of particular importance in addiction research, if certain patients are assigned to a placebo Control group (or essentially no treatment), even though prior research has demonstrated the failure of such approaches. The result is that many of these subjects drop out or continue aberrant behaviors during the course of a study – eg, heroin injecting – which can have harmful consequences.

## Run-in Periods

Some studies in addiction have used “run-in periods” involving a time before the “official” trial begins when no active treatment under investigation is given. This period sometimes serves a role in screening out ineligible or potentially noncompliant subjects and/or ensuring that participants are in a stable condition.

Data from this stage of a trial are only occasionally of value and this methodology can bias results. One researcher observed,

“Compared with results that might have been observed without the run-in period, the reported results may overestimate the benefits and underestimate the risks of treatment.”[30] In addiction research, a run-in period may result in high rates of pretrial dropouts or ineligible subjects, and the remaining participants may not be representative of typical patients.

**Example:** In almost all clinical trials of naltrexone for alcoholism, subjects were required to be alcohol-abstinent for a period ranging from a few days to several weeks prior to starting the medication. This may have negatively biased certain outcomes, such as time to first drink or relapse.[35] A more recent trial, omitting the run-in period, demonstrated improvements in relapse rates and study retention with naltrexone, and without added adverse reactions.[36]

### Compliance & Followup

Participant adherence to study protocol (ie, the plan for conducting a study specified in the methodology section of a report) and trial completion are essential ingredients for valid outcomes and conclusions. Yet, these can be difficult to achieve.

For various reasons, some subjects will disregard instructions and/or drop out during the course of a trial. Some authorities have suggested that having more than 10% of participants lost to followup may be cause for concern, and greater than 20% may invalidate study results.[37]

The problem is that reasons for subject noncompliance or dropout are often unknown. Adverse events (side effects, relapse, death) can be a cause; or, conversely, subjects may be doing so well that they stop taking medication and never return for followup assessments.

Noncompliance with treatment protocols and dropouts are especially problematic in addiction research, which often involves an unstable population that is not known for adhering to instructions. However, enrolling only patients with the most potential for full compliance and participation raises questions about selection bias and external validity.

**Example:** Participant discontinuations of up to 80% have been observed in some methadone dosing trials,[32,38] notably among Control subjects receiving inadequate medication. However, in many addiction treatment studies it is common to have 30%-50% of enrolled subjects drop out, which greatly diminishes the strength of results and may bias outcomes.

Another important concern is the duration of followup.[28] A study must continue long enough for the effect of the intervention to be reflected in the outcome variables. For practical and economic reasons, many addiction treatment studies may be too brief to account for outcomes that could take many months or even years to be fully realized.

Whatever the length of the study, for purposes of determining validity, every patient considered for and entering an experimental investigation should be accounted for at its conclusion.[28]

Guidelines have recommended that authors provide a flow diagram of participants' progress through all phases of a study, for each group, giving reasons for any exceptions.[39, see **Figure 1**] Although this originally was advised for RCTs, better journals are requiring it for most clinical study designs.

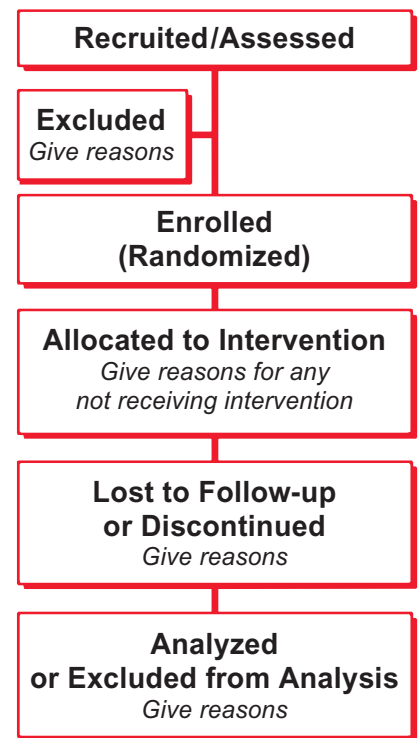
### ITT vs Per-Protocol Analyses

The most rigorous approach to overcoming bias in assessing outcomes is called an *intention-to-treat (ITT) analysis*. This statistically takes into account all data on *all* patients originally allocated to study groups. That is, all patients who enrolled in a study with the intention of being treated are analyzed as if they received full treatment; even if they dropped out early, did not take all of their medication, or deviated from the protocol in other ways.[9,12,40,41]

It is assumed that this approach reflects most accurately how patients act in everyday clinical practice. In randomized trials, dropouts or departures from protocol for reasons *other than the treatment itself* would be, on average, equally distributed across the groups. Therefore, any differences in outcomes would likely be due only to effects of the intervention or treatment.[26] However, the higher the rates of noncompliance or discontinuation, the greater the likelihood that ITT analyses will produce distorted conclusions about treatment efficacy.[41]

Another strategy is a “per-protocol” analysis of data. For this, researchers take into account *only* those patients who complete all, or a specified proportion, of the study and are compliant with the study protocol to at least a certain degree. Those who dropped out early, did not take adequate medication or attend sufficient therapy sessions, or otherwise departed from the protocol in significant ways are excluded from the analysis.

Sometimes also called an “endpoint,” “on-treatment,” “efficacy,” or “as treated” analysis, this approach can bias results in favor of the Experimental treatment, but it makes sense when the groups are not randomized to begin with or many subjects are lost to followup. Per-protocol analyses also might be justified in addiction research when outcome success hinges specifically on retention in treatment and compliance with the protocol regimen.



**Figure 1:** Flow diagram of subject progress through a trial. Adapted from [39].

**Example:** Studies of naltrexone in treating alcoholism [35] and in preventing opioid relapse [42] have consistently found, using per-protocol analyses, that treatment is effective in patients who are compliant with continuing to take the medication. Whereas, ITT analyses, which included subjects who were noncompliant in taking naltrexone or left treatment early, demonstrated less remarkable outcome results.

Authors should carefully explain their strategies for data gathering and analysis, and ideally present both ITT and per-protocol analyses in their reports. Readers can then more completely judge the validity of results and conclusions.

## Clinical Relevance of Statistics

Medical research depends on mathematics for the interpretation of outcome data. However, expertise in statistics is not necessary for assessing the validity and clinical relevance of addiction studies.

## Importance of Power

**How many subjects should be enrolled in a research study to achieve valid results?**

When researchers design a study, statistical methods allow them to determine how many subjects will be needed to have a moderate, high, or very high chance of detecting significant differences between groups. This is called a “**power analysis**.”

Unfortunately, few authors report how they determined sample size – the power – and this omission is particularly prevalent in addiction research literature. If a power analysis is not specifically mentioned, it cannot be assumed that one was done.

Power analyses determine the probability of accurately rejecting the null hypothesis (that any differences are due to chance). In other words, **power determines the chances of detecting a true difference between groups, when one exists**. [4, 14, 43, 44]

Several factors in study design determine power, such as: 1. the size of the treatment effect that is considered clinically significant; 2. the amount of uncertainty that is acceptable (statistical significance level); and, especially, 3. the group sizes. [4, 43]

It is common for researchers to calculate required group sizes based on a power of at least 80%. This provides for an 80% likelihood of reporting a statistically significant difference between groups when one actually exists. [4, 28] Considerable numbers of subjects may be required to achieve even that minimum level of power, and studies with smaller sample sizes may falsely conclude that the Experimental and Control groups do not differ, when too few patients were evaluated to validly make such a claim. [4, 14, 43]

**Example:** In a trial comparing two doses of methadone, the authors stated: “Power analyses based on effects detected in an earlier clinical trial of methadone treatment indicated that 96 patients in each group would be needed to detect a medium effect size (0.20) with

**Small-scale studies might be useful in generating hypotheses for further study; however, they are unreliable for guiding clinical practices.**

80% power in an intent-to-treat analysis.” [45] Therefore, they appropriately randomly divided 192 patients into two groups.

Small-scale studies might be useful in generating hypotheses for further study; however, they are unreliable for guiding clinical practices. Consequently, there has been some concern expressed about the ethical propriety of enrolling patients in underpowered trials that do not provide valid answers to clinical questions. [44] According to the authors, patients in these circumstances are being exposed to the hazards and burdens of experimental procedures, and time and money are expended, for limited ends. Although this proposition is debatable, it seems worthy of consideration; particularly, in view of the many smaller-scale studies in addiction medicine.

## “Bottom-line” Clinical Effects

Practitioners need answers to basic questions about the chances of success or failure when considering new interventions, such as:

- How does the new treatment compare with other therapies?
- How many patients will achieve better outcomes with the new treatment?

For the non-statistician healthcare professional, there are several, easy-to-perform calculations that provide answers. These may be called “bottom-line” clinical effects, because they help put research into everyday practice. [16]

**To begin, consider the following:** The statements below compare two treatments that were tested for reducing illicit-opioid use. Which seems most effective?

- A. This treatment produced a 17% reduction in illicit-opioid-positive urine screens.
- B. Compared with the other treatment, illicit-opioid use was reduced by 31% with this treatment.
- C. Illicit-opioid use in patients receiving this treatment was only 69% of that in the other treatment group.
- D. For every 6 patients administered this treatment one additional patient achieved an illicit-opioid-free outcome.

Actually, all 4 statements pertain to the exact same set of results reported in a study by Ling and colleagues, [32] which favored maintenance treatment with 80 mg/d methadone (Experimental group) compared with 8 mg/d buprenorphine (Control

group). This is an example of how the same data may be expressed as different **estimates of effect** for clinical decision-making purposes,[8,15,46] and these are essential ingredients of evidence-based medicine.

### Estimates of Effect

These are quite useful for healthcare providers and becoming commonplace in the general medical literature. However, the measures often are not included in addiction research reports.

Estimates of effect can be easily derived from data that are usually provided in published papers.[8,9,15,46] As will be seen below, descriptions of these estimates incorporate the term “risk” to signify the occurrence of an outcome event of interest – its event rate.

The event rate (ER) – in a Control group (CER) and an Experimental group (EER) – represents the proportion of subjects in whom the event was observed, expressed as a fraction or percentage ranging from 0 (0%) to 1.0 (100%). For example, the CER and EER may designate the percentage of patients in each group achieving illicit-opioid abstinence during a trial. The estimate of effect measure behind each of the 4 statements above (A-D) and its formula is explained below and summarized in **Table 3**.

#### A. Absolute Risk Reduction (ARR)

The absolute risk reduction (ARR) responds to a basic question: How much better did one group do than the other? Sometimes called the “risk difference,” it is the difference in the event rate (outcome risk) between the *Control group (CER)* and *Experimental group (EER)*.

$$ARR = CER - EER$$

An ARR of 0 would indicate *no difference* between comparison groups; whereas, the highest possible value of +1.0 denotes a 100% reduction of events in the Experimental group. A negative value indicates a reverse effect (a risk *increase*, with more events occurring in the Experimental than the Control group). However, the clinical meaning of this must be interpreted within the context of the study, because more events in the Experimental condition may be the preferable outcome to suggest treatment success (eg, greater retention in treatment).

**Example:** Using data from the Ling et al. study,[32] mentioned above, the mean percentage (average event rate) of urinalyses positive for illicit opioids in the Experimental group receiving 80 mg/d of methadone was 38% (or, M80 = .38). In comparison, the illicit-opioid-positive event rate was 55% in Control group subjects administered 8 mg/d buprenorphine (BUP = .55).

Thus, **CER (BUP) = .55; EER (M80) = .38**.

The  $ARR = CER (BUP) - EER (M80) = .55 - .38 = .17$ . In other words, treatment with 80 mg/d methadone reduced the risk of illicit-opioid-positive urinalyses during the trial by an average of 17%.

#### B. Relative Risk Reduction (RRR)

A next question might be, how did the *reduction* in outcome risk as a result of the Experimental treatment (the ARR, above) compare with the outcome in the Control group? The answer is provided by the “relative risk reduction”:

$$RRR = (CER - EER) / CER \text{ or } ARR / CER$$

An RRR of 0 indicates *no effect* of the Experimental treatment relative to the Control condition. As with the ARR, an RRR other than zero can be a positive or negative value depending on the context.

**Example:** Continuing the Ling et al. example from above,  $RRR = ARR / CER = .17 / .55 = .31$ . Thus, relative to treatment with 8 mg/d buprenorphine, 80 mg/d methadone reduced average illicit-opioid use by 31%.

#### C. Risk Ratio (RR)

As a further question, what was the advantage of the Experimental treatment in comparison with the Control condition? The “risk ratio” provides an answer, since it is the ratio of the event rate (outcome risk) in the Experimental group to the event rate in the Control group:

$$RR = EER / CER$$

Also known as the “relative risk,” an RR of 1.0 indicates that outcomes in the comparison groups are the same. Values may be less than or greater than 1.0 indicating whether the Experimental treatment reduced or increased the outcome risk relative to that in the Control group, respectively. Again, the corresponding meaning of this value must be interpreted within the context of the study.

**Example:**  $RR = EER / CER = .38 / .55 = 0.69$  in the Ling et al. study.[32] So, the risk of illicit-opioid use in M80 group patients was only 69% of the risk in BUP-treated subjects.

#### D. Numbers Needed to Treat (NNT)

How many patients need to receive the Experimental treatment to either: a) prevent one additional undesired outcome (eg, illicit drug use or treatment dropout); or, b) achieve one additional desired outcome? An answer is provided by the “numbers needed to treat,” calculated as:

$$NNT = 1 / ARR \text{ or } 1 / (CER - EER)$$

Measure	Formula	Description
<b>ARR</b> Absolute Risk Reduction	<b>= CER - EER</b>	Difference between outcome event rate in Control group vs Experimental group.
<b>RRR</b> Relative Risk Reduction	<b>= (CER - EER) / CER</b> or ARR / CER	Percent reduction of event rate in the Experimental group vs Control group.
<b>RR</b> Risk Ratio or Relative Risk	<b>= EER / CER</b>	Ratio of outcome event risk in Experimental group vs Control group.
<b>NNT</b> Number Needed to Treat	<b>= 1 / ARR</b> or 1 / (CER - EER)	The number of patients treated to prevent or gain one additional outcome event.

The NNT can be very helpful for deciding on the clinical advantage of an intervention. For example, a research trial demonstrating an NNT of 4 for a particular treatment would indicate that only 4 patients must be administered the treatment to prevent an undesired outcome or gain a desired effect beyond what would be expected without it.

On the other hand, if 50 patients must be treated to achieve the effect, the new treatment's cost/benefit advantages over the comparison (Control) therapy described in the research might be questionable. Whether an NNT represents gaining a desired effect or preventing a negative outcome, and the value of this, depends on the context of the study. For example, treating 50 patients to save an additional life seems reasonable; treating the same number to gain only an extra month of illicit-drug abstinence in one patient might be of less certain value.

**Example:** In the Ling et al. study,  $NNT = 1/ARR = 1/.17 = 6$  (rounded). Thus, for every 6 patients treated with 80 mg/d methadone, one additional patient would remain illicit-opioid free, beyond what might be achieved with 8 mg/d buprenorphine.

In addiction medicine, the NNT may sometimes understate a new treatment's potential value for many patients. As one authority has noted, "The higher the probability that a patient will experience an adverse outcome if we don't treat, the more likely the patient will benefit from treatment, and the fewer such patients we need to treat to prevent one event." [15]

The NNT also may be influenced by the length of followup in a particular study. During a longer study more patients receiving placebo or no treatment might be expected to experience negative outcomes, which, in turn, would increase the ARR and reduce the NNT for demonstrating benefits of the Experimental treatment. [47]

## Confidence Intervals (CIs)

A basic tenet of research is that no matter how many times the same experiment is repeated, slightly different outcome event rates will emerge each time.[16] So, then, which is the "true" result?

Mean (average) values or other measurements representing outcome event rates are actually approximations of "true" effects – the ones most expected to be accurate, but never exactly known. Each measurement is called a **point estimate**, "to remind us that although the true value lies somewhere in its neighborhood, it is unlikely to be precisely correct." [15] The larger "neighborhood" in which the "true value" is likely to reside is portrayed by a **confidence interval**. It is expressed as a range with a given degree of expected certainty or "confidence"; most often 95%.

A 95% CI represents a range that includes the "true" point estimate of interest 95% of the time. Seldom will the true result be at the extremes of the range, and it will be *outside* the range only 5% of the time — 2.5% of the time it will be above the range and 2.5% below.[30] Researchers can calculate CIs for almost any statistical test or measure of effect, such as, mean, ARR, RRR, RR, NNT, etc. (See examples in box, **Figures 2 & 3**.)

Confidence intervals are related to *p*-values, but CIs provide more meaningful information for clinical decision-making purposes. Some authorities have even suggested that CIs could replace *p*-values, since hypothesis testing using only *p*-values applies cut-off points for statistical significance that are arbitrary and do not portray a broader picture of the most accurate or "true" effect.[13] As one author states, "The confidence interval around the result in a clinical trial indicates the limits within which the 'real' difference between the treatments is likely to lie, and hence the strength of the inference that can be drawn from the result." [16]

## Understanding Confidence Intervals

If CIs are graphically plotted it is easier to assess differences between groups.

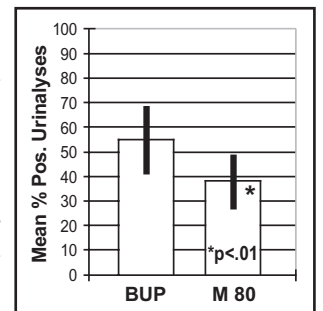
**Figure 2** demonstrates several variations of CI plots with means (large dots) for two groups: **A**. If there is an overlap of the CIs, also including the mean of at least one group, there is no statistical difference between groups; **B**. If there is no CI overlap, then a statistically significant

difference exists; **C**. If the CIs overlap, but the means are outside the overlapping portions, a statistical difference may or may not exist and a hypothesis test will indicate the *p*-value. Examining the CIs helps determine effect size and if the Experimental treatment might be of clinical benefit, even if the differences are statistically nonsignificant. CI comparisons also can involve more than two groups.

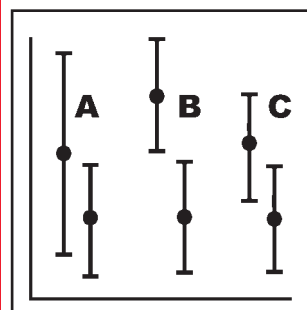
In the Ling et al. study, discussed above,[32] the values for percent positive illicit-opioid urinalyses were (mean; 95% CI): M80 (38%; 28-48%); BUP (55%; 42-68%). See **Figure 3**.

The point estimate (mean) for M80 was statistically significantly lower

**Figure 3:** Bars represent point estimate means; vertical lines denote 95% CIs derived from Ling et al.[32]



than for BUP,  $p < .01$ . However, the CIs suggest that there is a possible overlap, or a chance that the BUP treatment may be equivalent to M80. This is visualized more clearly in Figure 3. However, taking the high and low extremes of the CIs, there also is a possible 40% (68% - 28%) mean difference (absolute risk reduction) between M80 and BUP treatments.



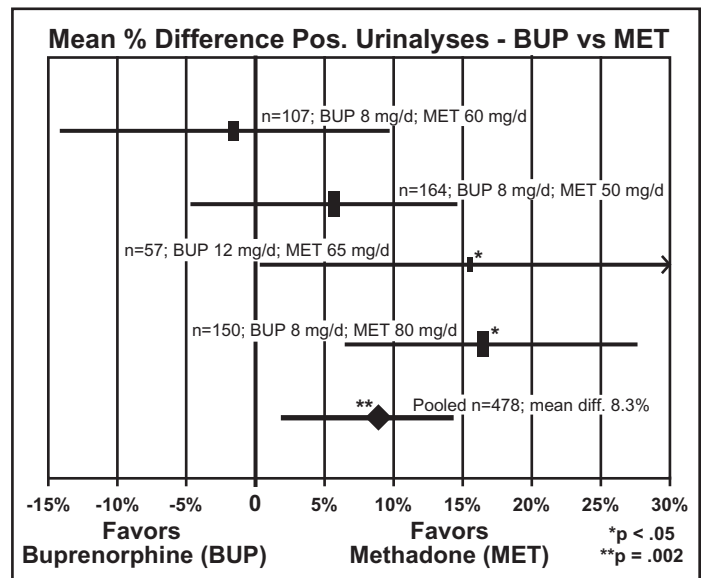
**Figure 2:** Visualizing differences between groups using confidence intervals. [48]

## Understanding “Forest Plots”

**Figure 4** represents a “forest plot” of illicit-opioid-positive urinalysis data from a meta-analysis by Barnett and colleagues of 4 studies comparing the efficacy of buprenorphine with methadone.[50] A solid vertical line at 0% corresponds to no statistical differences between groups and is the “line of no effect.”

The horizontal line for each study represents the 95% confidence interval (CIs) for the mean difference – essentially the absolute risk reduction (ARR) – with a block representing the point estimate weighted according to study size. If the CI line crosses the vertical line of no effect it means there was no statistically significant difference between treatments. Since, the mean difference is calculated by subtracting methadone (MET) percent positive urinalyses from those of buprenorphine, negative numbers favor buprenorphine. (Note: The 4th line from the top represents data from the Ling et al. article [32] described above.)

The bottom line with the diamond-shaped mark represents pooled (mathematically combined) data from the 4 studies. It lies entirely to the right of the line of no effect and has a relatively narrow CI, representing a highly statistically significant benefit of methadone (mean difference [ARR] = 8.3%;  $p = .002$ ). At the upper extent of the pooled CI, the advantage approaches 15%.



**Figure 4:** Forest plot of data from Barnett et al.[50]

In this plot, it becomes visually apparent that, as methadone doses increased above 60 mg/d, there was a stronger advantage over either 8 or 12 mg/d buprenorphine.

The poor retention of participants and small enrollments in some addiction research trials can be problematic. As the number of subjects enrolled in a study decreases, and/or as more subjects drop out during the course of a trial, the results can become less valid; that is, the study loses statistical power, the CI ranges become wider, and the point estimates are less likely to be accurate. Hence, it becomes more difficult to confidently apply such research in clinical decision-making.[15,16,26]

## Meta-Analysis “Forest Plots”

CIs also may be used to graphically portray relationships between multiple studies of the same treatment or intervention. For this, the point estimates and confidence intervals for each trial are displayed in what is known as a “forest plot.”[49]

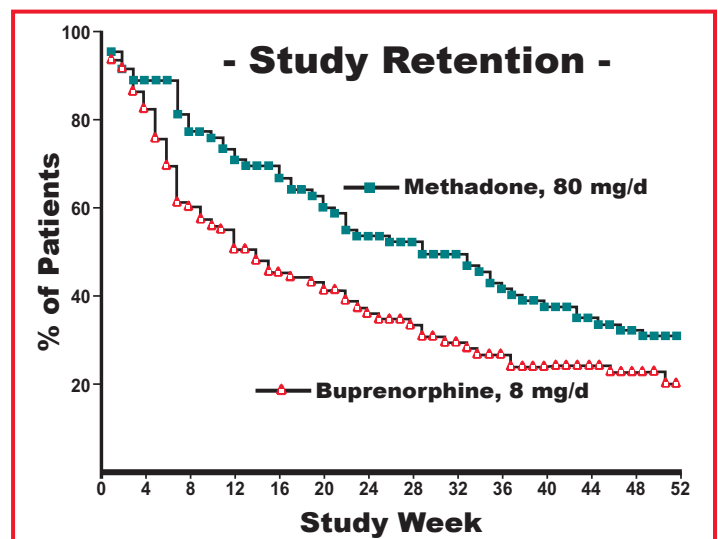
This has been especially useful in systematic reviews or meta-analyses for comparing outcomes of the different trials evaluated. There are variations of these very useful diagrams depending on the data available – such as means, mean differences (ARRs), risk ratios (RR), and other parameters – but forest plots typically follow the pattern illustrated in **Figure 4** (see box).[4,18,21,24,49]

Forest plots were allegedly so-named because they graphically resemble a “forest of lines,” especially when many individual studies are represented.[49] Such plots allow quickly visualizing individual study results, as well as how their combined or pooled results lead to clinically useful conclusions; however, it is essential that authors also provide the numerical data supporting the graphical interpretations.

## Time-to-Event “Survival” Analyses

How do outcome event rates in patients receiving different treatments change and compare over time?

In many addiction treatment studies, outcomes of interest include the time to certain events. Often, researchers are interested in how long participants remain in treatment and refrain from illicit-drug use. It is conventional to describe the cumulative events over time as “survival data,” from which survival curves can be plotted as shown in **Figure 5**. [51] The overall differences between group curves can be statistically tested and authors should provide  $p$ -values for the tests of significance.



**Figure 5:** Percentage of patients in each group still in treatment at the end of each study week. From Ling et al.[32]

The statistical test used should assess differences between the entire survival curves. At certain times during the course of a study the groups may be somewhat equivalent, but there may be significant differences overall. So, looking only at specific timepoints can be misleading by over- or underestimating treatment effects, compared with the value of a full course of treatment.[4,43]

**Example:** Retention in treatment was an important outcome variable in the Ling et al. study,[32] discussed above. As **Figure 5** depicts, during the entire course of the trial, the M80 group (top line) had much better overall retention than the BUP group ( $p = .009$ ). However, at *specific* timepoints of 26 and 52 weeks taken in isolation, the differences between groups failed to achieve statistical significance ( $p = .16$ ), even though there were marked differences in the percentages of patients retained in each group at those times.

### Correlation

How might a change in one variable (eg, drug dose) affect another (eg, retention in treatment)? Many research designs describe relationships between variables using correlation analyses.

The degree of relationship is usually depicted by a calculated correlation coefficient – or,  $r$  value – which can be positive or negative, ranging from +1.0 to -1.0, with 0 indicating no correlation. If the coefficient is positive, it means that one variable increases along with the other; a negative  $r$  indicates that one variable decreases as the other increases.[14] A guide for interpreting the meaning of  $r$ -values is presented in **Table 4**. [14,48]

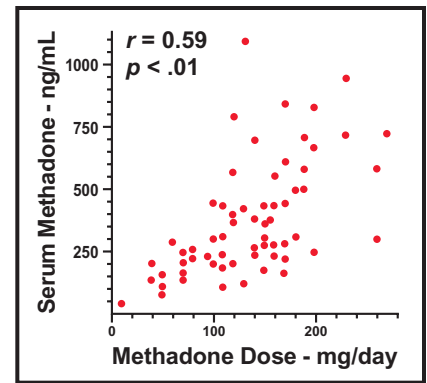
**Table 4: Interpreting Correlation Coefficients**

$r$ Value	Correlation/Relationship
< .20	Minor/Slight relationship
.20 - .40	Low/Small relationship
.40 - .70	Moderate/Substantial relationship
.70 - .90	High/Strong relationship
> .90	Very high/Greatly dependable relationship

The authors should include a discussion of what the relationship means and its extent, and it can be essential that data points are visually depicted in a “scatterplot” as an aid to interpretation (**Figure 6**). Also, the correlation should be statistically tested, and a significant  $p$ -value would indicate that the null hypothesis – ie, any relationship being due merely to chance – may be rejected.[16]

**Example:** **Figure 6** is a scatterplot depicting a substantial positive relationship between methadone dose and trough (low point) methadone concentration in blood serum ( $r = 0.59$ ). This also is statistically significant ( $p < .01$ ). However, at each level of methadone dose there are patients with widely differing serum concentrations, so it would be improper to state that a particular dose “causes” a specific methadone serum concentration; other factors might be important in many patients.[52]

**Figure 6:** Methadone serum level plotted against daily dose for 69 illicit-opioid abstinent patients.[52]



Correlational data is highly subject to bias, since multiple factors can influence the relationships in most cases and  $r$ -values alone do not explain all of the reasons; that is, the causes. However, research reports describing such relationships may misinterpret or misrepresent the data to imply causality and this bias is sometimes subtle, as when authors “suggest” that there may be a direct influence of one variable on another.

## Why Good Research Goes Bad

### Problems of Design, Execution, Reporting

The role of researcher-author can be difficult. Many challenges must be overcome to initiate a project and see it through to publication. Some of these are listed in **Table 5**.

**Table 3: Challenges Faced by Researcher Authors**

<b>Insufficient Funding</b>	Good research can be costly and researchers may opt to do “something” even when funds are not available to “do it right.”
<b>Inadequate Training</b>	Researchers who are primarily clinicians may not have adequate training in research methodology, analysis, and/or reporting.
<b>Lack of Tech Support</b>	Many researchers do not have access to data processing and statistical support staff.
<b>“Publish or Perish”</b>	Researchers may be under pressure to publish, influencing them to generate more data than necessary and to “dredge” available data for positive results or multiple articles.
<b>Hidden Agendas</b>	There may be “political” influences promoting the exploration of certain hypotheses or demonstrating particular results that are beyond the researcher-author’s control.
<b>Journal Space Limitations</b>	Most journals limit the length of articles, causing authors to carefully choose what is included or deleted. Subtle censorship often fosters inadequate reporting of analyses.
<b>“Too Many Cooks”</b>	Multiple authors and/or peer reviewers complicate the writing process and presentation of results.

## “The plural of anecdote is not evidence.” – Leshner

The research design-to-publication process can take years and the final product may be disappointing to both the authors and intended audience. Guidelines from international organizations have been developed for improving the reporting quality of medical manuscripts;<sup>[39,53]</sup> however, many authors are unfamiliar with these and few journals in the addiction treatment field rigorously follow the recommendations.

Perhaps, the greatest frustration for all concerned comes from requirements and limitations imposed by journal editors. Rarely, if ever, is an author allowed sufficient space to fully explore the subject at hand. So, decisions must be made about deleting certain data, abbreviating the statistical presentations, and limiting the amount of explanatory text. If there are multiple authors, the writing process becomes much more tortuous. Some flawed papers are reminiscent of the old saw about a camel being a horse that was designed by a committee.

Most medical journals are peer reviewed, which is a quality-control measure but can interject bias. Reviewers may have their own viewpoints or favorite styles of analysis that are imposed on the author. The review process is usually blinded; neither the author nor readers are informed of peer-reviewer identities, so their qualifications cannot be judged. If debates arise, and the author does not fully comply with requested changes, the article is likely to be rejected.

### Dangers of *Post Hoc* Analyses

Researchers are often tempted to examine their data after the study ends (*post hoc*) searching for “interesting results” that were not hypothesized. This is called *retrospective subgroup analysis* or “*data dredging*.”<sup>[4,9,15,46]</sup>

For example, outcomes in subgroups of subjects from the overall study population may be analyzed looking for significant differences. These groups may include only males, persons of a certain age, or those with specific preexisting conditions, etc.

These analyses often show that the treatment had a different or better effect in the subgroup. In some cases, such patients may be more typical of those in a particular clinic setting and the information can be helpful. However, if these analyses were not planned in advance their validity might be doubtful, since the initial statistical power of the study is dissipated and, in the case of randomized trials, the subgroups represent derandomized samples.<sup>[15]</sup>

As a further concern, there is a greater likelihood, purely by chance, of finding statistically significant differences when multiple tests are performed on the same data. For example, if the data is reorganized and reanalyzed 20 times, at least one sig-

nificant result at the  $p = .05$  level would likely be found merely due to random effects, and the authors should acknowledge this possibility.<sup>[4,16,43]</sup>

### Fallacies of Anecdotes as Evidence

As noted earlier (*Table 1*), case reports or anecdotes are weak evidence – toward the bottom of the hierarchy. Such cases may be of interest, provided they are well-documented with complete descriptions of relevant facts, but are of very limited clinical value until confirmed by more extensive and higher quality investigations.

An evidence-based medicine approach would suggest that case reports have associated *null hypotheses*. That is, the observed events should be assumed due to random effects or unknown causes unless significantly demonstrated otherwise.

**Example:** There have been case reports that motivated possibly excessive concerns and/or far-reaching actions in the addiction treatment field. This may have happened with LAAM, wherein 10 patients exhibiting cardiac arrhythmias while on LAAM were a primary basis for discontinuing the drug in the European Union; although there also were confounding factors that might have influenced the heart problems.<sup>[54]</sup>

Small-scale investigations, enrolling few participants and sometimes called “pilot studies,” might be viewed merely as a case series or collection of anecdotes. And, as Alan Leshner, PhD, former Director of NIDA, frequently stressed, “The plural of anecdote is not evidence.”<sup>[55]</sup>

### Mass Media Distortions

Finally, persons engaged in reporting research results in the mass media are obligated to consider the likely public reaction. Misleading information is potentially harmful; yet, few guidelines have been provided in this regard.<sup>[56]</sup>

Medical journalists usually lack an in depth understanding of clinical research methodology and of common errors in such investigations. Meanwhile, news editors seek sensational health stories that will seize the public’s attention.<sup>[57]</sup>

One investigation found that research studies reporting negative outcomes are more likely to appear in print or broadcast news. Furthermore, results from randomized controlled trials are underreported by the press in favor of less rigorous observational studies.<sup>[57]</sup>

Journals, research organizations, and researchers themselves often issue press releases on newsworthy investigations. However, these usually do not sufficiently emphasize study limitations, and data are often presented in formats that exaggerate the perceived importance of the findings. In many cases statistical data are omitted entirely in favor of headline-making conclusions.<sup>[58]</sup>

Frequently in press releases, and subsequent news reports, there is a biased leap of interpretation from laboratory data to

potential effects in humans. This is especially the case with research abstracts presented at scientific meetings. These sometimes receive prominent attention in the mass media, even though such reports are preliminary, with uncertain validity, and often are never accepted for formal publication by peer-reviewed journals.[59]

While medical journalists are perhaps not as educated or vigilant as they might be in assessing provided information, more rigorous standards also are needed for the quality of reports released to the press by researchers or their representatives. Research results are too often misrepresented to the public as scientifically sound evidence when that is not the case,[59] and critical readers need to be watchful for what is essentially *medical propaganda*.

## Putting Research Into Practice

### Everyday Relevance

Clinical research reflects “probable possibilities” – that is, the potential for interventions to have certain outcome effects within specified limitations – but not necessarily the “realities” of everyday practice. To be usefully applied in particular addiction treatment settings, research must satisfy essential questions about relevance:[15]

- Overall, does the study present high quality, valid evidence?
- Are the questions (hypotheses) addressed by the study relevant to your needs?
- Are the study subjects similar to your own patients?
- Is the research methodology free of bias and clearly explained?
- Are the results understandable and statistically significant?
- Do the conclusions make sense from patient-care perspectives (clinically significant)?

Above all else, clinical research outcomes should satisfy the last question by helping to define best practices, with important benefits for patients outweighing any disadvantages. There also needs to be a clear connection between the intervention and the stated or implied benefit.[15] For example, retention in an addiction treatment program, in itself, is of little benefit to the patient. However, if it results in fewer drug relapses and improved quality of life the advantages are obvious.

Furthermore, even if investigators report favorable effects of an intervention on one clinically important outcome, care must be taken that there are no deleterious effects on other outcomes.[15] Retention in treatment may enhance drug abstinence; however, if clinic attendance requirements are such that the person cannot hold down a job, go to school, or care for a family, important life goals may not be achieved.

## Developing a Healthy Skepticism

***Skepticism becomes a virtue when critically assessing addiction research with the objective of making evidence-based clinical decisions.***

Generally, if some aspect of a study report seems sloppy, incomplete, puzzling, or questionable, there may be deep flaws in the research.[27] Ideally, several studies will have been conducted in response to a particular clinical question of interest, and these can be compared and contrasted to help resolve skepticism. This is where systematic reviews and meta-analyses, summarizing a body of research on a topic, can be especially helpful.

It is expected that educated readers, having absorbed the information in this booklet, will demand a stronger level of research evidence, a higher quality of analyses, and better reporting. For starters, research reports in the addiction treatment field should reflect essential concepts presented above, such as: power analyses, evidence-based estimates of effect, confidence intervals for all outcomes, numbers needed to treat (NNT), and helpful visual presentations of data (eg, tables, plots, graphs). There are many resources available via the Internet where readers can learn more about the intricacies of these concepts (see **Table 6**).

Readers can help promote better addiction research by supporting and subscribing to only those publications offering the highest quality articles presenting valid research evidence. Similarly, research-funding organizations and agencies should be pressured to adopt higher standards of expected quality in all phases of addiction research, from design to publication. Finally, biased or slanted research reporting in the mass media or other information sources can be challenged by writing letters of complaint.

There always may be a degree of uncertainty about what will work best in particular patients, for that is the nature of scientific research and medical practice. However, applying principles of evidence-based addiction medicine (EBAM) will help pave a path toward more rational and sensible clinical decision making, leading to improved patient care and more enlightened healthcare policies.

### Table 6: Resources on the Internet

- <http://www.nettingtheevidence.org.uk> – master site based in the UK for finding a wide variety of evidence-based medicine resources via the Internet.
- <http://cebm.net> – Centre for Evidence-Based Medicine (England) provides access to various resources and instructional text.
- <http://www.ebmny.org/thecentr2.html> – site sponsored by the New York Academy of Medicine links to various resources for evidence-based medicine.
- <http://www.med.ualberta.ca/ebm/main.htm> – “Evidence Based Medicine Tool Kit” at Univ. of Alberta, Canada.
- <http://www.goldenhour.co.il> – gateway to evidence-based medicine resources developed by Israeli scientists.
- <http://www.statsoft.com/textbook/glosfra.html> – extensive glossary of statistical terms.
- <http://www.ruf.rice.edu/~lane/rvls.html> – online statistics book provides many examples and simulations using actual data.
- <http://www.theresearchassistant.com/research/link.asp> – online statistics calculators and resource links.

Access to all checked December 2002

## References

1. Systems to Rate the Strength of Scientific Evidence. Fact Sheet. Rockville, MD: Agency for Healthcare Research and Quality; 2002. AHRQ Publication No. 02-P0022.
2. Jackson RT. Treatment practice and research issues in improving opioid treatment outcomes. *Science & Practice Perspectives (NIDA)*. 2002;1(1):22-28.
3. Altman DG. Poor-quality medical research: what can journals do? *JAMA*. 2002;287(21):2765-2767.
4. Glantz, SA. *Primer of Bio-Statistics*. 4th ed. New York, NY: McGraw-Hill, Health Prof. Div.; 1997.
5. McGuigan SM. The use of statistics in the *British Journal of Psychiatry*. *Br J Psychiatry*. 1995;167:683-688.
6. Pitkin RM, Branagan MA, Burmeister LF. Accuracy of data in abstracts of published research articles. *JAMA*. 1999;281(12):1110-1111.
7. Jadad AR, Moher M, Browman GP, et al. Systematic reviews and meta-analyses on the treatment of asthma: critical evaluation. *BMJ*. 2000;320:537-540.
8. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice & Teach EBM*. New York, NY: Churchill Livingstone; 1997.
9. Guyatt G, Rennie D (eds). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago, IL: AMA Press; 2002.
10. Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature: I. How to get started. *JAMA*. 1993;270(17):2093-2095.
11. Sheldon TA, Guyatt GH, Haines A. Getting research findings into practice: when to act on the evidence. *BMJ*. 1998;317:139-142.
12. Iverson C (chair). *American Medical Association Manual of Style*. 9th ed. Baltimore: Williams & Wilkins; 1998.
13. Sterne JAC, Smith GD. Sifting the evidence - what's wrong with significance tests? *BMJ*. 2001;322:226-231.
14. Williams F. *Reasoning with Statistics*. New York: Holt, Rinehart and Winston; 1968.
15. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature: II. How to use an article about therapy or prevention; B. What were the results and will they help me in caring for my patients? *JAMA*. 1994;271(1):59-63.
16. Greenhalgh T. How to read a paper: Statistics for the non-statistician, II - "Significant" relations and their pitfalls. *BMJ*. 1997;315(7105).
17. Greenhalgh T. How to read a paper: Getting your bearings (deciding what the paper is about). *BMJ*. 1997;315(7102).
18. Mulrow CD, Oxman AD (eds). *Cochrane Collaboration Handbook*. In: *The Cochrane Library [database on disk and CDROM]*. Oxford, UK: The Cochrane Collaboration; updated 1997.
19. Centre for Evidence-Based Medicine, NHS Research and Development. *EBM Toolbox*. Available at: <http://cebmrj2.ox.ac.uk/>.
20. Levels of evidence and grades of recommendations. Oxford, UK: Centre for Evidence-Based Medicine; May 2001. Available at: <http://minerva.minervation.com/cebmr/docs.levels.html>.
21. Greenhalgh T. How to read a paper: Papers that summarize other papers (systematic reviews and meta-analyses). *BMJ*. 1997;315(7109).
22. Rodwin MC. The politics of evidence-based medicine. *J Health Politics, Policy and Law*. 2001;26(2). Available at: <http://www.ahrq.gov/clinic/jhpl/>.
23. Hurst JW. Who organized the first clinical trial? *Medscape Cardiology*. September 19, 2002. Available at: <http://www.medscape.com>. Accessed 9/25/02.
24. Egger M, Smith GD, Phillips AN. Meta-analysis: Principles and procedures. *BMJ*. 1997;315(7121).
25. Oxman AD, Cook DJ, Guyatt. Users' guides to the medical literature: VI. How to use an overview. *JAMA*. 1994;272(17):1367-1371.
26. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature: II. How to use an article about therapy or prevention; A. Are the results of the study valid? *JAMA*. 1993;270(21):2598-2601.
27. Juni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *BMJ*. 2001;323:42-46.
28. Greenhalgh T. How to read a paper: Assessing the methodological quality of published papers. *BMJ*. 1997;315(7103).
29. Greenhalgh T. How to read a paper: Papers that report drug trials. *BMJ*. 1997;315(7106).
30. Pablos-Mendez A, Barr RG, Shea S. Run-in periods in randomized trials: Implications for the application of results in clinical practice. *JAMA*. 1998;279(3):222-224.
31. Posternak MA, Zimmerman M, Keitner GI, Miller IW. A reevaluation of the exclusion criteria used in antidepressant efficacy trials. *Am J Psychiatry*. 2002;159(2):191-200.
32. Ling W, Wesson DR, Charuvastra C, Klett CJ. A controlled trial comparing buprenorphine and methadone maintenance in opioid dependence. *Arch Gen Psychiatry*. 1996;53:401-407.
33. Barsky AJ, Saintfort R, Rogers MP, Borus JF. Nonspecific medication side effects and the nocebo phenomenon. *JAMA*. 2002;287:622-627.
34. Kupfer DJ, Frank E. Placebo in clinical trials for depression [editorial]. *JAMA*. 2002;287(14).
35. Leavitt SB. Evidence for the efficacy of naltrexone in the treatment of alcohol dependence (alcoholism). *Addiction Treatment Forum - Clinical Update*. March 2002. Available at: <http://www.atforum.com>.
36. Heinala P, Alho H, Kiianmaa K, Lonnqvist J, Kuoppasalmi K, Sinclair JD. 2001. Targeted use of naltrexone without prior detoxification in the treatment of alcohol dependence: a factorial double-blind, placebo-controlled trial. *J Clin Psychopharmacol*. 21(3):287-292.
37. How to read clinical journals: III. To learn the clinical course and prognosis of disease. *Can Med Assoc J*. 1981;124:869-872.
38. Johnson RE, Chutuaep MA, Strain EC, Walsh SL, Stitzer ML, Bigelow. A comparison of levomethadyl acetate, buprenorphine, and methadone for opioid dependence. *New Engl J Med*. 2000;343(18):1290-1297.
39. Moher D, Schulz KF, Altman D, for CONSORT Group. The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*. 2001;285(15):1987-1991.
40. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomized controlled trials. *BMJ*. 1999;319:670-674.
41. Fogg L, Gross D. Threats to validity in randomized clinical trials. *Res in Nurs Health*. 2000;23:79-87.
42. Leavitt SB. Naltrexone in the prevention of opioid relapse. *Addiction Treatment Forum*. August 2002. Available at: <http://www.atforum.com>.
43. Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. *Biostatistics in Clinical Medicine*. 3rd ed. New York, NY: McGraw-Hill, Health Prof. Div.; 1994.
44. Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA*. 2002;288(3):358-362.
45. Strain EC, Bigelow GE, Liebson IA, Stitzer ML. Moderate- vs high-dose methadone in the treatment of opioid dependence. *JAMA*. 1999;281(11):1000-1005.
46. Greenhalgh T. How to read a paper: Statistics for the non-statistician. *BMJ*. 1997;315(7104).
47. Chatellier G, Zapletal E, Lemaitre D, Menard J, Degoulet P. The number needed to treat: A clinically useful nomogram in its proper context. *BMJ*. 1996;312:426-429.
48. Fink A. *How to Analyze Survey Data*. Thousand Oaks, CA: Sage Publications; 1995.
49. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. *BMJ*. 2001;322:1479-1480.
50. Barnett PG, Rodgers JH, Bloch DA. A meta-analysis comparing buprenorphine to methadone for treatment of opioid dependence. *Addiction*. 2001;96:683-690.
51. Altman DG, Bland MJ. Time to event (survival) data. *BMJ*. 1998;317:468-469.
52. Okruhlica L, Devinsk F, Valentova J, Klempova D. Does therapeutic threshold of methadone concentration in plasma exist? *Heroin Add & Rel Clin Probl*. 2002;4(1):29-36.
53. International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts submitted to Biomedical Journals. *Ann Intern Med*. 1997;126:36-47.
54. European Agency for the Evaluation of Medicinal Products. EMEA public statement on the recommendation to suspend the marketing authorization for Orlaam (levacetylmethadol) in the European Union. 2001. Publication EMEA/8776/01.
55. Vastag B. Talking with Alan I. Leshner, PhD, National Institute on Drug Abuse director. *JAMA*. 2001;285(9):1141-1143.
56. *Guidelines on Science and Health Communication*. London: Social Issues Research Centre; 2001(November). ISBN 0 85403 570 2.
57. Bartlett C, Sterne J, Egger M. What is newsworthy? Longitudinal study of the reporting of medical research in two British newspapers. *BMJ*. 2002;325:81-84.
58. Woloshin S, LM Schwartz. Press releases: translating research into news. *JAMA*. 2002;287(21):2856-2858.
59. Schwartz LM, Woloshin S, Baczek L. Media coverage of scientific meetings: too much, too soon? *JAMA*. 2002;287(21):2859-2863.

### ADDICTION TREATMENT

# Forum

is published by: Clinco Communications, Inc.  
P.O. Box 685  
Mundelein, IL 60060

©2003 Clinco Communications, Inc.

A.T. Forum is made possible by an educational grant from Mallinckrodt Inc., a manufacturer of methadone & naltrexone.

March 2003